Lourdes Agapito
Michael M. Bronstein
Carsten Rother (Eds.)

LNCS 8926

Computer Vision – ECCV 2014 Workshops

Zurich, Switzerland, September 6–7 and 12, 2014
Proceedings, Part II

Springer

Lecture Notes in Computer Science 8926

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

More information about this series at http://www.springer.com/series/7412

Lourdes Agapito · Michael M. Bronstein
Carsten Rother (Eds.)

Computer Vision – ECCV 2014 Workshops

Zurich, Switzerland, September 6–7 and 12, 2014
Proceedings, Part II

Springer

Editors
Lourdes Agapito
University College London
London
UK

Michael M. Bronstein
University of Lugano
Lugano
Switzerland

Carsten Rother
Technische Universität Dresden
Dresden
Germany

Videos to this book can be accessed at
http://www.springerimages.com/videos/978-3-319-16180-8

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-16180-8 ISBN 978-3-319-16181-5 (eBook)
DOI 10.1007/978-3-319-16181-5

Library of Congress Control Number: 2015933663

LNCS Sublibrary: SL6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Foreword

Welcome to Zurich !

As you know, the European Conference on Computer Vision is one of the top conferences on computer vision. It was first held in 1990 in Antibes (France) with subsequent conferences in Santa Margherita Ligure (Italy) in 1992, Stockholm (Sweden) in 1994, Cambridge (UK) in 1996, Freiburg (Germany) in 1998, Dublin (Ireland) in 2000, Copenhagen (Denmark) in 2002, Prague (Czech Republic) in 2004, Graz (Austria) in 2006, Marseille (France) in 2008, Heraklion (Greece) in 2010, and Firenze (Italy) in 2012. Many people have worked hard to turn the 2014 edition into as great a success. We hope you will find this a mission accomplished.

The Chairs have decided to adhere to the classical single-track scheme. In terms of the time ordering, we have decided to largely follow the Firenze example (typically starting with poster sessions, followed by oral sessions), which offers a lot of flexibility to network and is more forgiving for the not-so-early-birds and hardcore gourmets.

A large conference like ECCV requires the help of many. They made sure you again get a full program including the main conference, tutorials, workshops, exhibits, demos, proceedings, video streaming/archive, and web descriptions. We want to cordially thank all those volunteers! Please have a look at the conference website to see their names (http://eccv2014.org/people/). We also thank our generous sponsors. You will see their logos around at several occasions during the week, and also prominently on the ECCV 2014 website (http://eccv2014.org/). Their support has been vital to keep prices low and to enrich the program. And it is good to see such level of industrial interest in what our community is doing!

Please do not forget to take advantage of your free travel pass. It allows you to crisscross our splendid city with its fabulous public transportation.

We hope you will enjoy ECCV 2014 to the full.

Also, willkommen in Zürich!

September 2014

Marc Pollefeys
Luc Van Gool

Preface

Welcome to the Workshop proceedings of the 13th European Conference on Computer Vision, held during September 6–12, 2014 in Zurich, Switzerland. We are delighted that the main ECCV 2014 was accompanied by 28 workshops.

We received 38 workshop proposals on diverse computer vision topics. The evaluation process was not easy because of the high quality of the submissions, and the final 28 selected workshops complemented the main conference program. Nearly all of the workshops were running for a full day, with the exception of two half-day workshops and one two-day workshop. In the end, the addressed workshop topics constituted a good mix between novel current trends and traditional issues, without forgetting to address the fundamentals of the computational vision area.

We would like to thank all the Workshop Organizers for their hard work and for making the workshop sessions a great success. We hope that participants enjoyed the workshops, together with the associated papers included in these volumes.

Kind regards / mit freundlichen Grüßen,

November 2014

Michael M. Bronstein
Lourdes Agapito
Carsten Rother

Organization

General Chairs

Luc Van Gool	ETH Zurich, Switzerland
Marc Pollefeys	ETH Zurich, Switzerland

Program Chairs

Tinne Tuytelaars	Katholieke Universiteit Leuven, Belgium
Bernt Schiele	MPI Informatics, Saarbrücken, Germany
Tomas Pajdla	Czech Technical University Prague, Czech Republic
David Fleet	University of Toronto, Canada

Local Arrangement Chairs

Konrad Schindler	ETH Zurich, Switzerland
Vittorio Ferrari	University of Edinburgh, UK

Workshop Chairs

Lourdes Agapito	University College London, UK
Carsten Rother	Technische Universität Dresden, Germany
Michael M. Bronstein	University of Lugano, Switzerland

Tutorial Chairs

Bastian Leibe	RWTH Aachen, Germany
Paolo Favaro	University of Bern, Switzerland
Christoph H. Lampert	IST, Austria

Poster Chair

Helmut Grabner	ETH Zurich, Switzerland

Publication Chairs

Mario Fritz	MPI Informatics, Saarbrücken, Germany
Michael Stark	MPI Informatics, Saarbrücken, Germany

Demo Chairs

Davide Scaramuzza	University of Zurich, Switzerland
Jan-Michael Frahm	University of North Carolina at Chapel Hill, USA

Exhibition Chair

Tamar Tolcachier	University of Zurich, Switzerland

Industrial Liason Chairs

Alexander Sorkine-Hornung	Disney Research Zurich, Switzerland
Fatih Porikli	ANU, Australia

Student Grant Chair

Seon Joo Kim	Yonsei University, Korea

Air Shelters Accommodation Chair

Maros Blaha	ETH Zurich, Switzerland

Website Chairs

Lorenz Meier	ETH Zurich, Switzerland
Bastien Jacquet	ETH Zurich, Switzerland

Internet Chair

Thorsten Steenbock	ETH Zurich, Switzerland

Student Volunteer Chairs

Andrea Cohen	ETH Zurich, Switzerland
Ralf Dragon	ETH Zurich, Switzerland
Laura Leal-Taixé	ETH Zurich, Switzerland

Finance Chair

Amael Delaunoy	ETH Zurich, Switzerland

Conference Coordinator

Susanne H. Keller	ETH Zurich, Switzerland

Workshop Organizers

W01 - Where Computer Vision Meets Art (VISART)

Gustavo Carneiro	The University of Adelaide, Australia
Alessio Del Bue	Italian Institute of Technology, Italy
Joao Paulo Costeira	Instituto Superior Tecnico, Lisbon, Portugal

W02 - Computer Vision in Vehicle Technology with Special Session on Micro Aerial Vehicles

David Geronimo	KTH, Sweden
Friedrich Fraundorfer	Technische Universität München
Davide Scaramuzza	University of Zurich, Switzerland

W03 - Spontaneous Facial Behavior Analysis

Guoying Zhao	University of Oulu, Finland
Stefanos Zafeiriou	Imperial College London, UK
Matti Pietikäinen	University of Oulu, Finland
Maja Pantic	Imperial College London, UK

W04 - Consumer Depth Cameras for Computer Vision

Andrea Fossati	ETH Zurich, Switzerland
Jürgen Gall	University of Bonn, Germany
Miles Hansard	Queen Mary University London, UK

W05 - ChaLearn Looking at People: Pose Recovery, Action/Interaction, Gesture Recognition

Sergio Escalera	Computer Vision Center, UAB and University of Barcelona, Catalonia, Spain
Jordi González	Universitat Autònoma de Barcelona and Computer Vision Center, Catalonia, Spain
Xavier Baró	Universitat Oberta de Catalunya and Computer Vision Center, Catalonia, Spain
Isabelle Guyon	Clopinet, Berkeley, California, USA
Jamie Shotton	Microsoft Research Cambridge, UK

W06 - Video Event Categorization, Tagging, and Retrieval toward Big Data

Thomas S. Huang	University of Illinois at Urbana-Champaign, USA
Tieniu Tan	Chinese Academy of Sciences, China
Yun Raymond Fu	Northeastern University, Boston, USA
Ling Shao	University of Sheffield, UK
Jianguo Zhang	University of Dundee, UK
Liang Wang	Chinese Academy of Sciences, China

W07 - Computer Vision with Local Binary Patterns Variants

Abdenour Hadid	University of Oulu, Finland
Stan Z. Li	Chinese Academy of Sciences, China
Jean-Luc Dugelay	Eurecom, France

W08 - Reconstruction Meets Recognition Challenge (RMRC)

Nathan Silberman	New York University, USA
Raquel Urtasun	University of Toronto, Canada
Andreas Geiger	MPI Intelligent Systems, Germany
Derek Hoiem	University of Illinois at Urbana-Champaign, USA
Sanja Fidler	University of Toronto, Canada
Antonio Torralba	Massachusetts Institute of Technology, USA
Rob Fergus	New York University, USA
Philip Lenz	Karlsruher Institut für Technologie, Germany
Jianxiong Xiao	Princeton, USA

W09 - Visual Object Tracking Challenge

Roman Pflugfelder	Austrian Institute of Technology, Austria
Matej Kristan	University of Ljubljana, Slovenia
Ales Leonardis	University of Birmingham, UK
Jiri Matas	Czech Technical University in Prague, Czech Republic

W10 - Computer Vision + ONTology Applied Cross-disciplinary Technologies (CONTACT)

Marco Cristani	University of Verona, Italy
Robert Ferrario	ISTC-CNR, Trento, Italy
Jason Corso	SUNY Buffalo, USA

W11 - Visual Perception of Affordances and Functional Visual Primitives for Scene Analysis

Karthik Mahesh Varadarajan	Technical University of Vienna, Austria
Alireza Fathi	Stanford University, USA
Jürgen Gall	University of Bon, Germany
Markus Vincze	Technical University of Vienna, Austria

W12 - Graphical Models in Computer Vision

Michael Yang	Leibniz University Hannover, Germany
Qinfeng (Javen) Shi	University of Adelaide, Australia
Sebastian Nowozin	Microsoft Research Cambridge, UK

W13 - Human-Machine Communication for Visual Recognition and Search

Adriana Kovashka	University of Texas at Austin, USA
Kristen Grauman	University of Texas at Austin, USA
Devi Parikh	Virginia Tech, USA

W14 - Light Fields for Computer Vision

Jingyi Yu	University of Delaware, USA
Bastian Goldluecke	Heidelberg University, Germany
Rick Szeliski	Microsoft Research, USA

W15 - Computer Vision for Road Scene Understanding and Autonomous Driving

Bart Nabbe	Toyota, USA
Raquel Urtasun	University of Toronto, Canada
Matthieu Salzman	NICTA, Australia
Lars Petersson	NICTA, Australia
Jose Alvarez	NICTA, Australia
Fatih Porikli	NICTA, Australia
Gary Overett	NICTA, Australia
Nick Barnes	NICTA, Australia

W16 - Soft Biometrics

Abdenour Hadid	University of Oulu, Finland
Paulo Lobato Correia	University of Lisbon, Portugal
Thomas Moeslund	Aalborg University, Denmark

W17 - THUMOS Challenge: Action Recognition with a Large Number of Classes

Jingen Liu	SRI International, USA
Yu-Gang Jiang	Fudan University, China
Amir Roshan Zamir	UCF, USA
George Toderici	Google, USA
Ivan Laptev	Inria, France
Mubarak Shah	UCF, USA
Rahul Sukthankar	Google Research, USA

W18 - Transferring and Adapting Source Knowledge (TASK) in Computer Vision (CV)

Antonio M. Lopez	Computer Vision Center and Universitat Autónoma de Barcelona, Spain
Kate Saenko	University of Massachusetts Lowell, USA
Francesco Orabona	Toyota Technological Institute Chicago, USA
José Antonio Rodríguez	Xerox Research EuroFrance
David Vázquez	Computer Vision Cente, Spain
Sebastian Ramos	Computer Vision Center and Universitat Autónoma de Barcelona, Spain
Jiaolong Xu	Computer Vision Center and Universitat Autónoma de Barcelona, Spain

W19 - Visual Surveillance and Re-identification

Shaogang Gong	Queen Mary University of London, UK
Steve Maybank	Birkbeck College, University of London, UK
James Orwell	Kingston University, UK
Marco Cristani	University of Verona, Italy
Kaiqi Huang	National Laboratory of Pattern Recognition, China
Shuicheng Yan	National University of Singapore, Singapore

W20 - Color and Photometry in Computer Vision

Theo Gevers	University of Amsterdam, The Netherlands
Arjan Gijsenij	Akzo Nobel, The Netherlands
Todd Zickler	Harvard University, USA
Jose M. Alvarez	NICTA, Australia

W21 - Storytelling with Images and Videos

Gunhee Kim	Disney Research, USA
Leonid Sigal	Disney, USA
Kristen Grauman	University of Texas at Austin, USA
Tamara Berg	University of North Carolina at Chapel Hill, USA

W22 - Assistive Computer Vision and Robotics

Giovanni Maria Farinella	University of Catania, Italy
Marco Leo	CNR- Institute of Optics, Italy
Gerard Medioni	USC, USA
Mohan Triverdi	UCSD, USA

W23 - Computer Vision Problems in Plant Phenotyping

Hanno Scharr	Forschungszentrum Jülich, Germany
Sotirios Tsaftaris	IMT Lucca, Italy

W24 - Human Behavior Understanding

Albert Ali Salah	Boğaziçi University, Turkey
Louis-Philippe Morency	University of Southern California, USA
Rita Cucchiara	University of Modena and Reggio Emilia, Italy

W25 - ImageNet Large-Scale Visual Recognition Challenge (ILSVRC2014)

Olga Russakovsky	Stanford University, USA
Jon Krause	Stanford University, USA
Jia Deng	University of Michigan, USA
Alex Berg	University of North Carolina at Chapel Hill, USA
Fei-Fei Li	Stanford University, USA

W26 - Non-Rigid Shape Analysis and Deformable Image Alignment

Alex Bronstein	Tel-Aviv University, Israel
Umberto Castellani	University of Verona, Italy
Maks Ovsjanikov	Ecole Polytechnique, France

W27 - Video Segmentation

Fabio Galasso	MPI Informatics Saarbrücken, Germany
Thomas Brox	University of Freiburg, Germany
Fuxin Li	Georgia Institute of Technology, Germany
James M. Rehg	Georgia Institute of Technology, USA
Bernt Schiele	MPI Informatics Saarbrücken, Germany

W28 - Parts and Attributes

Rogerio S. Feris	IBM, USA
Christoph H. Lampert	IST, Austria
Devi Parikh	Virginia Tech, USA

Contents – Part II

W10 - Computer Vision + ONTology Applied Cross-Disciplinary Technologies

W11 - Visual Perception of Affordances and Functional Visual Primitives for Scene Analysis

W06 - Video Event Categorization, Tagging and Retrieval toward Big Data

Grading Tai Chi Performance in Competition with RGBD Sensors

Hui Zhang(✉), Haipeng Guo, Chaoyun Liang, Ximin Yan, Jun Liu, and Jie Weng

Department of Computer Science, United International College, 28, Jinfeng Road, Tangjiawan, Zhuhai, Guangdong, China
{amyzhang,hpguog}@uic.edu.hk
{f030300021,f030300052,f030300030,f030300047}@mail.uic.edu.hk

Abstract. In order to grade objectively, referees of Tai Chi practices always have to be very concentrated on every posture of the performer. This makes the referees easy to be fatigue and thus grade with occasional mistakes. In this paper, we propose using Kinect sensors to grade automatically. Firstly, we record the joint movement of the performer skeleton. Then we adopt the joint differences both temporally and spatially to model the joint dynamics and configuration. We apply Principal Component Analysis (PCA) to the joint differences in order to reduce redundancy and noise. We then employ non-parametric Nave-Bayes-Nearest-Neighbor (NBNN) as a classifier to recognize the multiple categories of Tai Chi forms. To give grade of each form, we study the grading criteria and convert them into decision on angles or distances between vectors. Experiments on several Tai Chi forms show the feasibility of our method.

Keywords: Tai Chi · RGBD sensor · Kinect

1 Introduction

Tai Chi, as shortened to Tai Chi Chuan, is a traditional Chinese martial art, which is practiced for both its defense training and its health benefits. Because of its soft and continuously flowing movements, Tai Chi is able to cultivate both peoples mind and physical body into a balance system [11]. Tai Chi has become popular internationally and many Tai Chi schools have been opened around the world. Trainees can follow the coach in order to learn different forms in Tai Chi. Meanwhile, there are a lot of Tai Chi national or international competitions for the performers to improve their skills, such as London Competition for Traditional Tai Chi Chuan or Tai Chi Competition in New York, etc.

In a national Tai Chi competition, there are generally eight referees sitting in six position around the playground (see figure 1 for details). The five referees on the edge of the playground will first manually record the scores from their own view points and show them to the three chief referees after the performer finishing

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 3–13, 2015.
DOI: 10.1007/978-3-319-16181-5_1

his performance. The chief referees will finally give out the final score according to all of the scores collected. Such grading s largely based on manual works. There are also electronic systems for Tai Chi grading utilized in national competitions. Referees press keys on a joystick to deduce a score when he found that the performer makes a mistake. This system, along with the manual procedures, requires the referees concentrating on observing every posture of the performers movement in order to give a justice grade. The referees are easy to get tired and thus subject errors are inevitable during grading. Therefore, an automatic and objective method is urgently needed to solve these problems.

Fig. 1. The position of the referees

To facilitate the manual works, the first task is to work on recognizing different forms of Tai Chi performance. For automatic human action recognition, traditional methods may work on video sequences captured by a single camera. In this case, a video is a sequence of 2D RGB frames in time series. In [1,3,8–10], the spatio-temporal volume-based method have been proposed to compute and the similarity between two action volumes are compared to recognize the action. Another trend of methods is based on motion trajectory for recognizing human activities [13,14]. Human actions were interpreted by the movement of a set of body joints. In [18], naive Bayes mutual information maximization (NBMIM) is introduced as a discriminative pattern matching criterion for action classification.

However, it is not easy to extract and track skeleton joints from 2D video sequences quickly and accurately until the launch of Microsoft Kinect sensors. The Kinect sensor is able to capture RGB sequences as well as depth maps of human action in real time. With its associated SDK or OpenNI, we could model human actions by the motion of a set of key joints [6] with reasonable accuracy. There are applications or research with Kinect supporting martial art practices, such as the Kinect Sports game, the posture classification of Muay

Thai [7], etc. Human action and activity recognition with Kinect become popular research topics recently [5,12,16]. In order to have a fast, simple yet powerful recognition,[15] proposes an actionlet ensemble model to characterize the human actions, which represents the interaction of a subset of human joints. Zanfir et. al. [19] introduce a non-parametric Moving Pose (MP) descriptor considering both pose information as well as differential quantities (speed and acceleration) of the human body joints.

Inspired by [17], this paper first record the joint movement of the performers skeleton. Then we adopt the joint differences both temporally and spatially to model the joint dynamics and configuration. We apply Principal Component Analysis (PCA) to the joint differences by reducing redundancy and noise. We then employ non-parametric Nave-Bayes-Nearest-Neighbor (NBNN) as a classifier to recognize the multiple categories of Tai Chi actions.

After the system has recognized the action of the performer, the next task is to mark the quality of the performers action. We convert the text description of the criteria into the grading decisions on angles or distances between vectors. Experiments on several sample Tai Chi Chuan actions show the feasibility of our method. Note that we have used only one Kinect sensor for grading. We plan to use six Kinect later similar to the position configuration of the referees in figure 1 so that the grading results will be comparable to those by the referee.

This paper is organized as follows. Section 2 introduces the feature extraction and dimension reduction. Section 3 provides our classifier for action recognition. Section 4 studies the grading rules and converts them into programmable decisions. Then Section 5 summarized the implementation steps. The experimental results are shown in section 6. Finally, section 7 gives the conclusions.

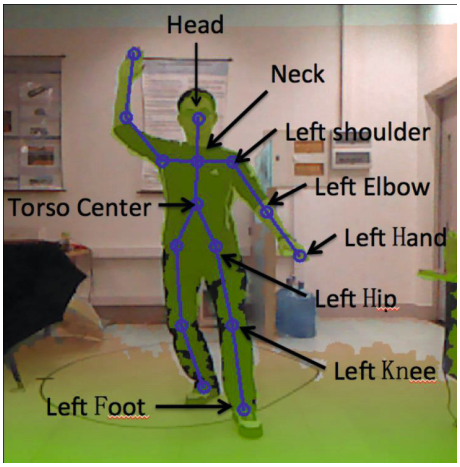


Fig. 2. The joints on the skeleton in OpenNI.

2 Feature Extraction

The human skeleton captured by Kinect sensor could have n joints and their respective 3D positions are $\mathbf{X}_k$ $(k = 1, ..., n)$. In OpenNI, $n = 15$. The joint 3D positions $\mathbf{X}_k = \{x_k, y_k, d_k\}$ are indicated by head, neck, torso center, left shoulder, left elbow, left hand, left hip, left knee, left foot, etc. as shown in figure 2. These joints are defined by the Kinect skeletal tracking system. The joints have hierarchy that the torso center joint as the root and extends to the head, feet and hands. Note that the three coordinate of a joint $\mathbf{X}_k = \{x_k, y_k, d_k\}$ are of inconsistent coordinates, e.g. $\{x_k, y_k\}$ are in screen coordinates and d_k is in world coordinate. Therefore the data normalization has to be first applied to $\mathbf{X}_k$ to avoid bias attributes in greater numeric ranges dominating those in smaller numeric ranges.

An action $\mathbf{A}_i$ could be represented by a sequence of frames f_{ij} $(j = 1, ..., N_i)$, where f_{ij} is a vector containing n coordinates of skeleton joints,

$$\mathbf{A}_i = \{f_{i1}, f_{i2}, ..., f_{iN_i}\}, \tag{1}$$

$$f_{ij} = \begin{pmatrix} \mathbf{X}_{head} \\ \mathbf{X}_{neck} \\ \mathbf{X}_{leftshoulder} \\ \mathbf{X}_{leftelbow} \\ \cdots \\ \mathbf{X}_{rightfoot} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \\ \mathbf{X}_4 \\ \cdots \\ \mathbf{X}_n \end{pmatrix}. \tag{2}$$

To characterize the action features, we first set the initial frame to approximate the neutral posture. Then we form the preliminary feature representation for each frame by the combination of three feature channels as $f_c = [f_{cc}, f_{cp}, f_{ci}]$ (see figure 3 in detail).

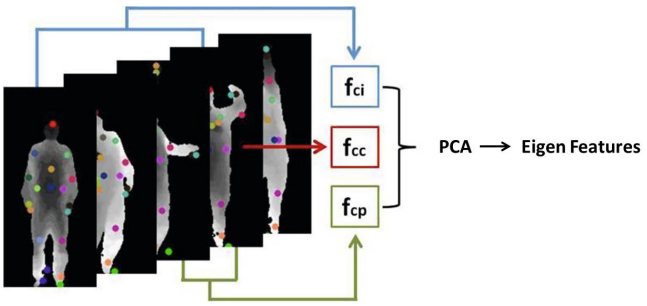


Fig. 3. The framework of representing Eigen features. In each frame, we obtain three feature sets, f_{cc}, f_{cp} and f_{ci} to capture the information of offset, posture, and motion. The normalization and PCA are then applied to obtain Eigen features descriptor for each frame.

Here f_{cc} is the pair-wise joints differences within the current frame, i.e.,

$$f_{cc} = \{\mathbf{X}_s^c - \mathbf{X}_t^c | s, t = 1, 2, ..., n; s \neq t\}, \tag{3}$$

which is used to characterize the joints' static posture information of current frame-c. f_{cp} is the pair-wise joints differences between the current frame-c and its preceding frame-p, i.e.,

$$f_{cp} = \{\mathbf{X}_s^c - \mathbf{X}_t^p | s, t = 1, 2, ..., n\}. \tag{4}$$

f_{cp} is used to capture the dynamic property of current frame-c. Finally, to represent the overall dynamics of the current frame-c with respect to the initial frame-i, the pair-wise joints differences f_{ci} are computed between frame-c and frame-i, i.e.,

$$f_{ci} = \{\mathbf{X}_s^c - \mathbf{X}_t^i | s, t = 1, 2, ..., n\}. \tag{5}$$

By making use of PCA, we could then reduce redundancy and noise in f_c. As a result, we obtain the Eigen features $\mathbf{E}_j$ representation for each frame f_{ij}. Most energy could be covered in the first few leading eigenvectors and 95% redundant data could be removed.

3 Action Recognition with NBNN Classifier

The Naive-Bayes-Nearest-Neighbor (NBNN) [2] is used here as the classifier for Tai Chi action recognition. The Nearest-Neighbor (NN) has several advantages over most learning-based classifiers. First, it doesn't require the time-consuming learning process. Second, the Nearest-Neighbor naturally deals with a large number of classes. Third, it avoids the over fitting problem. Instead of using NBNN-based image classification [3], we use NBNN-based video classification for Tai Chi action recognition. We directly compute Video-to-Class distance rather than Video-to-Video distance. Therefore the action recognition is performed by

$$C^* = arg\ min \sum_{j=1}^{N_i} ||\mathbf{E}_j - NN_c(\mathbf{E}_j)||^2, \tag{6}$$

where $NN_c(\mathbf{E}_j)$ is the nearest neighbor of $\mathbf{E}_j$ in class-C.

4 Converting Grading Criteria to Angles or Distances between Vectors

From the methods of previous sections, each Tai Chi form could be recognized correctly. Now the next task is to convert the Tai Chi grading criteria of each action into programmable rules. We first need to study the details of the Tai Chi grading criteria [4].

Let's look at Tai Chi Chuan 24 forms. We analyze each form and its grading criteria and found that some of the criteria are related with angles between two

bones. Here is an example that a straight arm is forbidden in Tai Chi Chuan competitions. As the competition rules, the arms should always in bending (see figure 4 for detail).

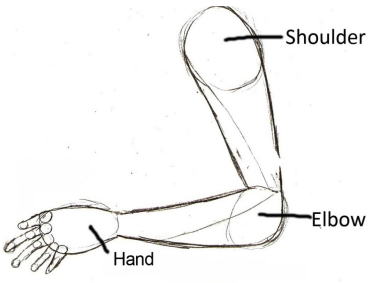


Fig. 4. Only a bend instead of a straight arm is allowed in Tai Chi Chuan competitions

Therefore, we first get the joint points of the shoulder, the elbow and the hand $\mathbf{X}_{shoulder}$, $\mathbf{X}_{elbow}$, $\mathbf{X}_{hand}$. Then we can get the upper arm bone as the vector $\mathbf{B}_{ua} = \mathbf{X}_{shoulder} - \mathbf{X}_{elbow}$ and the lower arm bone as the vector $\mathbf{B}_{la} = \mathbf{X}_{elbow} - \mathbf{X}_{hand}$. Then the angle θ between $\mathbf{B}_{ua}$ and $\mathbf{B}_{la}$ can be calculated as

$$\theta = arc\cos\left(\frac{\mathbf{B}_{ua} \cdot \mathbf{B}_{la}}{|\mathbf{B}_{ua}||\mathbf{B}_{la}|}\right). \tag{7}$$

Therefore according to the creteria, if the angle θ is close to 180^o, corresponding marks will be deducted.

Other criteria may relate to the distance between two joint points or the distance between a joint point and the ground plane. For example, if the performer performs the lunge motion (see figure 5 for detail), he is not allowed to drag his step on the ground when he moves his left foot. So we need to calculate the distance between the left foot and the ground plane. With the depth of the points on ground captured by Kinect, we can easily calculate the ground plane $\mathbf{P}_g : ax+by+cz+d = 0$. The normal of the ground plane can be directly obtained as $\mathbf{N}_g = (a, b, c)$. Thus the distance $\mathbf{D}_{xp}$ between the joint point $\mathbf{X} = (x, y, z)$ and the plane $\mathbf{P}_g$ is

$$\mathbf{D}_{xp} = \frac{|ax + by + cz + d|}{\sqrt{a^2 + b^2 + c^2}}. \tag{8}$$

Here that $\mathbf{D}_{xp}$ is the least distance that performers left foot should raise from the ground. Note that different people has different height, so that the distance would be varied. The body size should be scaled to a reference size first before we measure $\mathbf{D}_{xp}$.

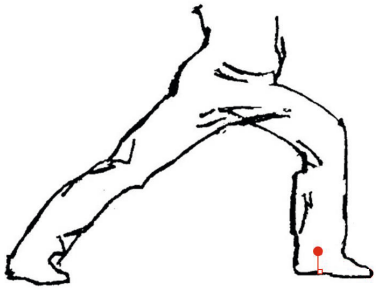


Fig. 5. In a lunge motion arm, the performer is not allowed to drag his step

5 Implementation

To implement this Tai Chi grading system, first we have to prepare a database for storing the joint positions of standard expert's actions captured by a single Kinect sensor. For each action, we normalize the data and form the feature matrix. Then we apply PCA to the feature data to reduce the data dimension.

During testing, when the system detect a new video input from Kinect, the referee has to indicate the start and end frame for different actions of the performer. Then for each action, the joint positions are stored and then normalized. We now use them to form the feature matrix. PCA will be applied to the performer feature data to reduce the data dimension. Thereafter, we can decide which category the performer belongs to by using NBNN classifier.

Within each action category, corresponding grading criteria are applied to postures such that the postures are graded objectively. Finally, the overall grade for the performer is provided automatically by the system.

The detailed procedures can be described in the following algorithm.

6 Experimental Results

Since we use OpenNI for developing our Kinect system, there are 15 joints in each frame. After normalization, we will have a huge feature dimension. f_{cc}, f_{cp} and f_{ci} contains 105, 255 and 255 pair-wise comparisons, respectively. Since each comparison generates three values $(\Delta x, \Delta y, \Delta d)$, this results in a dimension of $3 \times (105 + 255 + 255) = 1845$. Then PCA is applied to reduce redundancy and noise to obtain the Eigen features representation for each frame. From our experiments, the 95% energy is covered in the first 13 leading eigenvectors.

We take Tai Chi Chuan 24 forms as the example. Table 1 lists the details of the 24 forms and totally there are 33 postures.

Here we use the commencing position and its corresponding grading criteria for example. We will describe how to qualify criteria with conditional decisions. In figure 6, it is clear to see the postures for the commencing position.

Algorithm 1. Tai Chi Grading Procedure with a Kinect Sensor.

1: prepare a database for storing the joint positions of standard expert's actions captured by a Kinect sensor;
2: normalize the data stored;
3: for each action, form the feature matrix f_c with equation (3), (4), (5) (see section 2);
4: apply PCA to the feature data to reduce the data dimension;
5: during testing, the referee indicates the start and end frame for each action of the performer;
6: then for each action, the joint positions are stored and also normalized;
7: form the feature matrix;
8: apply PCA to the performer feature data to reduce the data dimension;
9: decide which category the performer belongs to by using NBNN classifier with equation (6);
10: within each action category, apply corresponding grading criteria to each postures by making use of equation (7) and (8);
11: the overall grade for each action is summed automatically in order to get the total mark.

Table 1. Tai Chi Chuan 24 forms

1. Commencing position	2. Part the wild horses mane to both sides (3)
3. White crane spreads its wings	4. Brush knee and twist hip on both sides (3)
5. Hand strums the lute	6. Repulse the monkey both sides (4)
7. Grasp the birds tail, left side	8. Grasp the birds tail, right side
9. Single whip	10. Wave hands like clouds (3)
11. Single whip	12. High pat on horse
13. Kick with the right heel	14. Strike opponents temple with fists
15. Turn body and kick left heel	16. Squatting and standing on one leg left side
17. Squatting and standing on one leg right side	18. A fair maiden threads the shuttle both sides
19. Pluck needle from the sea bottom	20. Open fan through the back
21. Turn body wrench, parry, punch	22. Apparent close-up
23. Cross-hands	24. Closing form

(a) (b) (c) (d) (e)

Fig. 6. Commencing position.

The following shows the grading rules and how to translate it into conditional decisions.

– Open two feet (see figure 6 (b) for detail). If the feet do not have the same width with that of the shoulders, 0.1 point will be deducted. To convert the criteria into qualified rules, we first connect the two foot joints and also connect the two shoulder joints. If the length of the two line segments has apparent difference or they are not perpendicular to the normal of the ground plane, we will deduct 0.1 point.
– Slowly raise the two arms forward horizontally (see figure 6 (c) for detail). If the hand or elbow joints are higher than the shoulders, 0.1 point will be deducted. We calculate the distance from the hand / elbow joints to the ground plane and the distance from the shoulder joints to the ground plane. If the former ones are larger than the later, we will deduct 0.1 point.
– Move the arms up (see figure 6 (d)) and then down (see figure 6 (e) for detail). If one of the elbow joints is above the hand joints, 0.1 point will be deducted. We calculate the distance from the elbow joints to the ground plane and the distance from the hand joints to the ground plane. If the former is larger than or equal to the latter, we will deduct 0.1 point.

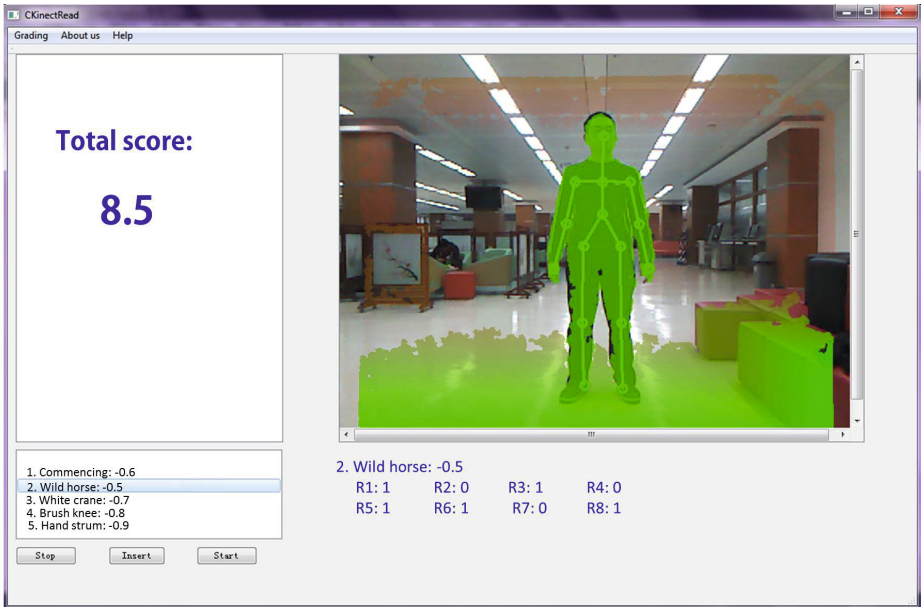


Fig. 7. Tai Chi Grading Interface

Here we only show three rules for grading the commencing position. There are in fact 8 rules in our implementation for grading each form in Tai Chi 24 forms. Through the study and on-the-spot investigation of Tai Chi, Tai Chi grading

criteria are converted into the quantified rules by applying different algorithms introduced in section 4.

Figure 7 illustrates the user interface of our system. The skeleton joints are shown together with the input video. The deducted grade and which rule is broken are illustrated in the bottom. And the total grade is given in the top left panel.

7 Conclusions

In this paper, we introduced a Tai Chi Chuan grading system with the Microsoft Kinect sensor. We first capture the joint movement of the performers skeleton. Then we record the joint differences both temporally and spatially to model the joint dynamics and configuration. Principal Component Analysis is then allied to the joint differences in order to reduce redundancy and noise. Then non-parametric Nave-Bayes-Nearest-Neighbor (NBNN) is employed as a classifier to recognize the multiple categories of Tai Chi forms. To grade the quality of each posture, we convert the competition grading criteria into decision on angles or distances between vectors. Experiments on several sample Tai Chi Chuan forms show the feasibility of our method.

Due to the slow and smooth motion of Tai Chi Quan, our method works well in the good indoor environment. In the future, we need to extend our work so that the method could be used to grade Tai Chi performance in real playground environment. Furthermore, separate forms are evaluated but not the motion coherence which is very important in Tai Chi performance. We would next focus on the motion coherence. Another future work is to use multiple Kinect sensors to capture skeleton joints and provide grading. Six Kinect sensors are required as their positions can be located as those of the referees in figure 1. The individual grading will be collected and a statistical result is expected to give the final grade. This could also solve the self-occlusion problem caused by the performer rotation.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Project no. 61005038) and an internal funding from United International College (Project no. R201312).

References

1. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. IEEE Trans. on Pattern Analysis and Machine Intelligence **23**(3), 257–267 (2001)
2. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
3. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features

4. Federation, I.W.: International wushu competition rules. International Wushu Federation (2005)
5. Han, J., Shao, L., X, D., Shotton, J.: Enhanced computer vision with microsoft kinect sensor: A review. IEEE Transactions on Cybernetics, **43**(5) 1317–1333
6. Johansson, G.: Visual perception of biological motion and a model for its analysis. Journal of Attention Perception and Psychophysics **14**(2), 201–211 (1973)
7. Kaewplee, K., Khamsemanan, N., Nattee, C.: Muay thai posture classification using skeletal data from kinect and k-nearest neighbors. In: Proceedings of the International Conference on Information and Communication Technology for Embedded Systems (ICICTES 2014) (2014)
8. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d gradients. In: Proceedings of British Machine Vision Conference (2008)
9. Laptev, I.: On space-time interest points. International Journal of Computer Vision, **64**(2)
10. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
11. Lee, M.S., Ernst, E.: Systematic reviews of tai chi: An overview. British Journal of Sports Medicine **46**(10), 713–718 (2011)
12. Liu, L., Shao, L.: Learning discriminative representations from rgb-d video data. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)
13. Parameswaran, V., Chellappa, R.: View invariance for human action recognition. Journal of Attention Perception and Psychophysics **66**(1), 83–101 (2001)
14. Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, pp. 2004–2011 (2009)
15. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3d human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36**(5), 914–927 (2014)
16. Wu, D., Shao, L.: Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA (2014)
17. Yang, X., Tian, Y.: Eigenjoints-based action recognition using nave-bayes-nearest-neighbor. In: Proc. Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 14–19 (2012)
18. Yuan, J., Liu, Z., Wu, Y.: Discriminative video pattern search for efficient action detection. IEEE Trans. Pattern Anal. Mach. Intell. **33**(9), 1728–1743 (2011)
19. Zanfir, M., Leordeanu, M., Sminchisescu, C.: The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection

Human Action Recognition by Random Features and Hand-Crafted Features: A Comparative Study

Haocheng Shen[1(✉)], Jianguo Zhang[1], and Hui Zhang[2]

[1] School of Computing, University of Dundee, Dundee, UK
hyshen@dundee.ac.uk, jgzhang@computing.dundee.ac.uk
[2] Department of Computer Science & Technology, United International College, Zhuhai, China
amyzhang@uic.edu.hk

Abstract. One popular approach for human action recognition is to extract features from videos as representations, subsequently followed by a classification procedure of the representations. In this paper, we investigate and compare hand-crafted and random feature representation for human action recognition on YouTube dataset. The former is built on 3D HoG/HoF and SIFT descriptors while the latter bases on random projection. Three encoding methods: Bag of Feature(BoF), Sparse Coding(SC) and VLAD are adopted. Spatial temporal pyramid and a two-layer SVM classifier are employed for classification. Our experiments demonstrate that: 1) Sparse Coding is confirmed to outperform Bag of Feature; 2) Using a model of hybrid features incorporating frame-static can significantly improve the overall recognition accuracy; 3) The frame-static features works surprisingly better than motion features only; 4) Compared with the success of hand-crafted feature representation, the random feature representation does not perform well in this dataset.

Keywords: Action recognition · Hand-crafted feature · Random representation

1 Introduction

Recognizing human action is a significant branch of computer vision and attracting increasing attentions due to its widely applications like crime monitoring and human-computer interaction. Generally, the recognition task can be simply viewed as a combination of two subtasks: extract features as representations from video frame sequence, and subsequent classification of the video representations. Among the two subtasks, one key point is to built such a feature representation which contains the main structure of an action and robust to background cluttering, illumination and scale changes etc. Substantial approaches of exploring the feature representation have been proposed and proven successful in action recognition, such as 3D HoG [8], HoG/HoF [10], extended SURF [19]. These feature

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 14–28, 2015.
DOI: 10.1007/978-3-319-16181-5_2

representations are all hand-crafted and need to be computed by a specific mathematical manner. Recently, random feature representation has been popular in texture recognition [15], face recognition [20], and medical image analysis [13]. However, little work has been reported on applying the random feature representation into video based action recognition. Therefore, in this paper, we evaluate and compare these two different feature representations for action recognition task. We have three main contributions: (1) a comparative study of different combinations of existing schemes for video action recognition based on hand-crafted feature representation and report the best combination whose performance is competitive to one of the state-of-art techniques on the same dataset; (2) Investigate the popular random feature representation to see whether it is a feasible approach for video based human action recognition; (3) Investigate the role of frame-static features and motion features for action recognition on the popular YouTube dataset.

The rest of this paper is organized as follow: Section 2 reviews relevant literature of approaches for action recognition; Section 3 describes each component of designed algorithm in details; Section 4 indicates the implementations and experiment results; Conclusions and future work are given in Section 5.

2 Related Work

The approach for action representations can be generally divided into two categories: global representations and local representations. For the former, the human body is first located in the image. Then the person referred as interest of region (ROI) would be encode as a whole. Local representation is a more popular approach which describes the observation as a collection of local descriptors or patches. Dollár et al.[5] extract a cuboid by 3D Gabor filter and then concatenated the gradients for each pixel in the cuboid to form the descriptors. Laptev et al.[10] introduced the HoG/HoF descriptors which compute histograms of both spatial gradient and optic flow accumulated in neighbourhood regions around the interest points. Klaser et al.[8] extend HoG to 3D and build the 3D HoG descriptor. It is based on histograms of 3D gradient orientations which is uniformly quantized by regular polyhedrons in an integral video representation. Based on the image SURF descriptor [1], Willems et al.[19] present the extended SURF descriptors for videos. The 3D patches is uniformly divided into small girds first then each cell is represented by a vector of weighted sums of responses of the Haar-wavelets along the three axes. Liu et al.[14] firstly extract static feature in a video frame as the complementary of motion feature to build representation of the video, which outperforms using motion feature only. Le et al.[12] combined with deep learning techniques to use unsupervised feature learning as a way to learn features directly from unlabelled video data. Wang et al.[18] extract the dense trajectories and motion boundary descriptors from the video as the representation. As the motion boundary descriptors can reduce the affects of camera motions effectively, it makes a huge progress in the realistic videos based action recognition and can be treated as the state-of-the-art.

3 Method

Motived by [14] and consider the large variation in realistic videos, we strongly believe that static feature like a static pose in a single frame also contains important action contextual information which can provide strong cues and thus be served as a complementary of motion feature for action recognition. Motivated by this observation, we investigate the role of motion and static feature for action recognition and build a hybrid model upon them for both hand-crafted and random feature representation. The flowchart of our work is shown in Figure 1. We will follow this flowchart to describe our algorithm in details.

Fig. 1. The flowchart of video based recognition

3.1 Spatial-Temporal Interest Points Detection

Spatial-temporal interest points are the locations in space and time domain where a significant variation occurs in the local neighborhood. We apply the extension of Gabor filter proposed by Dollar et al. [5] to extract the 3D interest cuboids, which capture the most important characteristics of the movement occurring in the video. The response function has the form:

$$R = (I * g * h_{ev})^2 + (I * g * h_{od})^2 \tag{1}$$

where $g(x, y; \sigma)$ is the 2D Gaussian smoothing kernel for spatial dimensions, and h_{ev} and h_{od} are a quadrature pair of 1D Gabor filters applied temporally. They are defined as:

$$h_{ev}(t; \tau, \omega) = -cos(2\pi t\omega)e^{-t^2/\tau^2} \tag{2}$$

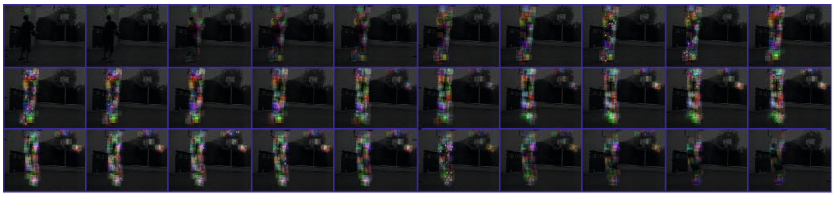
$$h_{od}(t; \tau, \omega) = -sin(2\pi t\omega)e^{-t^2/\tau^2} \tag{3}$$

where $\omega = 4/\tau$. The interest points are located in the local maxima corresponding to the response function. The parameter σ and τ correspond to the spatial and temporal scale of the detected cuboid. We set the size of the cuboid to $19 \times 19 \times 11$ pixels. Some examples of interest cuboids detected by the 3D Gabor detector on video frame sequence are shown in Figure 2.

3.2 Hand-Crafted Feature Representation

The visual content of a video segment can be represented by a set of descriptors computed at every interest point position within its near cuboid region. It is obvious that the oriented gradient can capture spatial information while optic flow is able to catch the movement information. Therefore, we adopt the 3D

(a)



(b)

Fig. 2. Example of 3D interest cuboids detection. (a) original frames; (b) cuboids detected by 3D Gabor filter (best viewed in colour).

HoG/HoF descriptor similar to Laptev et al. [10], which computes histograms of both oriented gradient and optic flow accumulated in spatial-temporal interest cuboids.

Specifically, the 3D interest cuboid is firstly smoothed and divided into $3 \times 3 \times 2$ grid of cells; for each cell, 4-bin histograms of gradient (HoG) and 5-bin histograms of optic flow (HoF) are calculated based on the oriented direction. Then the normalized histograms from each small grid are concatenated to form the local descriptor. We employ PCA to reduce the dimensionality to 200 experimentally.

Using motion feature only may not be distinct enough, especially for the unrestricted videos like YouTube action dataset. Intuitively, The static feature can be viewed as a very strong complementary. To extract static feature, we sample temporally at every 15 frames from the frame sequence of the video. For each frame, we build SIFT descriptors [16] upon dense sampling grid. Additionally, multi-scale static feature is achieved by changing the size of the static image by multiplying $1/\sqrt{2}$.

3.3 Random Feature Representation

We employ random projection to build random feature representation. The key idea of random projection originated from the Johnson-Lindenstrauss lemma [4]: if points in a high dimension are projected onto a randomly selected subspace of suitable dimension, then the distance between points are approximately preserved. In practice, the original d-dimensional data is projected to a k-dimensional ($k << d$)

subspace using a random matrix $k \times d$ matrix $\mathbf{R}$ whose columns have unit lengths. It can be represented by:

$$\mathbf{X}^{RP}_{k\times N} = \mathbf{R}_{k\times d}\mathbf{X}_{d\times N} \tag{4}$$

As before, we apply the random projection on both motion feature and static feature to form the random feature representation. Specifically, for each extracted cuboid, we first normalize the intensity of each pixel within the cuboid and then uniformly divide the cuboid into $2 \times 2 \times 2$ grids. Assume the size of each grid is $w \times h \times t$ pixels so for each grid we identify gray-scale vector $\mathbf{v} \in \mathbb{R}^d (d = wht)$ by stacking the intensities; then use random projection to reduce dimensionality and form the random feature descriptor. The random matrix $\mathbf{R}$ is defined as the Gaussian measurement matrix whose elements are independent, zero-mean, unit-variance Gaussian random variables. Finally the projected vectors for each sub-cuboid are concatenated to form the local descriptor of the whole cuboid.

Similarly, for the static feature extraction, we use dense sampling as before on each sampled video frame sequence. For each dense point, the patch whose size is the same as that of SIFT descriptors is extracted and the gray-scale vector is formed by stacking the intensities. Then random projection is employed to generate the local static random descriptors.

3.4 Descriptors Encoding

As the number of local descriptors extracted by the above methods varies from each video, distinguishing these descriptors from different classes of action directly is not straightforward. A popular approach is to firstly learn a codebook containing a fixed number of visual words based on the training descriptors set, then encode the descriptors with the codebook.

A simple but effective method to learn the codebook is K-means clustering algorithm. The main idea is to minimize the sum of squared Euclidean distances between points $\mathbf{x}_j$ and their nearest cluster $\mathbf{v}_k$:

$$\arg\min_{\mathbf{V}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{V}_i} \|\mathbf{x}_j - \mathbf{v}_k\|^2 \tag{5}$$

where $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_k]^\top$ are the target codebook with K cluster centers. We propose 2-level K-means clustering to generate the codebook: for each class of action, apply K-means for the first level clustering, then based on the first level results, the K-means clustering is applied again to create the final codebook. The size of codebook is set to 256.

We mainly evaluate two popular encoding methods: Bag of Feature (BoF) and Sparse Coding (SC) [21] for both feature representations. Moreover, we extra evaluate Vector of Locally Aggregated Descriptors (VLAD) [7] for random feature representation.

Bag of Feature. Let $\mathbf{X}$ be a set of descriptors in a D-dimensional feature space, $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_M]^\top \in \mathbb{R}^{M \times D}$. The Bag of Feature quantization problem can be re-formulated into a matrix factorization problem:

$$\begin{aligned} \min_{\mathbf{U}} \quad & \sum_{m=1}^{M} \|\mathbf{x}_m - \mathbf{u}_m \mathbf{V}\|^2 \\ \text{subject to} \quad & Card(\mathbf{u_m}) = 1, |\mathbf{u_m}| = 1, \mathbf{u}_m \geqslant 0, \forall m \end{aligned} \tag{6}$$

where $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_M]^\top$ is the cluster membership indicators and $\mathbf{V}$ is the pre-calculated codebook. The cardinality constraint $Card(\mathbf{u_m}) = 1$ means that only one element of $\mathbf{u}_m$ is nonzero, and $|\mathbf{u_m}|$ indicates that the summation of the absolute value of each element in $\mathbf{u}_m$. After obtaining the encoded descriptor set $\mathbf{U}$, the video can be represented by frequencies of each visual word. Since the number of visual words is fixed for all descriptors sets, a video with arbitrary number of descriptors is then converted into a single histogram vector whose length equals to the number of visual words. This provides extreme convenience for the future classification processing.

Sparse Coding. The constraint for BoF model $Card(\mathbf{u_m}) = 1$ is too restrictive to reconstruction $\mathbf{X}$ with low error. We can relax the constraint by making $\mathbf{u_m}$ to have a small number of nonzero element. Meanwhile, the number of nonzero element is enforced to be minimum. Then the BoF is turned into another problem known as Sparse Coding:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \sum_{m=1}^{M} \|\mathbf{x}_m - \mathbf{u}_m \mathbf{V}\|^2 + \lambda |\mathbf{u_m}| \\ \text{subject to} \quad & \|\mathbf{v_k}\| \leqslant 1, \forall k = 1, 2, ..., K \end{aligned} \tag{7}$$

Similar to BoF, in the training stage a set of training descriptors are used to solve Equation 7 with respect to $\mathbf{U}$ and $\mathbf{V}$. The conventional way for such a optimization problem is to iteratively optimize either over $\mathbf{U}$ or $\mathbf{V}$ while fixing the other. We set the initial codebook $\mathbf{V}$ of Sparse Coding as the result generated by K-means algorithm described above instead of using a random initialization. This processing can make the objective function more optimized when the number of iteration is fixed.

Each column of $\mathbf{U}$ corresponds to the coefficients of all the local descriptors to one specific visual word in the codebook $\mathbf{V}$, we adopt the max pooling function for SC, which has been well established by biophysical evidence and empirically justified by many image categorization algorithms. It can be represented by:

$$z_j = \max\left\{|u_{1,j}|, |u_{2,j}|, ..., |u_{M,j}|\right\} \tag{8}$$

where z_j is the j-th element of $\mathbf{z}$, $u_{i,j}$ is the matrix element at i-th row and j-th column of $\mathbf{U}$.

Vector of Locally Aggregated Descriptors. Besides BoF and SC, we evaluate another encoding method: vector of locally aggregated descriptors (VLAD) [7] for random feature representation. The idea of the VLAD is to accumulate the difference between each visual word $\mathbf{v}_i$ and the descriptor $\mathbf{x}_i$ which is assigned to that visual word. Note that VLAD can be viewed as a non-probabilistic version of the Fisher Vector [17]. Therefore, if the local descriptor is d-dimensional, the dimension D of VLAD would be $D = k \times d$. A component $u_{i,j}$ of VLAD can be obtained by summing over all the local random feature descriptors:

$$u_{i,j} = \sum_{\mathbf{x} \text{ belong to } \mathbf{v_i}} x_j - v_{i,j} \tag{9}$$

where the indices $i = 1...k$ and $j = 1...d$ index the visual word and the local descriptor component respectively.

3.5 Spatial-Temporal Pyramid

All the encoding methods described above only capture the statistical characteristic of the descriptors set. None of spatial and temporal layout of geometrical features has been taken into consideration. Spatial Pyramid Matching (SPM) proposed by [11] overcomes this limitation in still image classification. It works by partitioning the image into increasingly fine sub-regions and computes histograms of local descriptors over the resulting sub-regions. The final feature vector is formed by concatenating histograms of each sub-region with the corresponding weight of each level of pyramid. The spatial pyramid is a simple and computationally efficient complement of an orderless BoF image representation. It has shown significantly improved performance over the standard BoF model as it describes the observations as a collection of local representations, which are somewhat invariant to changes in scale, illumination and partial occlusions.

We extend this approach to 3D by adding subregions with respect to temporal domain. The spatial-temporal pyramid is built by uniformly dividing the frame sequence of video into $2 \times 2 \times 2$ grids for the first level and $3 \times 3 \times 3$ grids for the second level. The descriptor set of each subregion is a set of descriptors whose corresponding interest points are located within such a subregion. Then the local characteristics in totally 36 subregions are calculated by BoF or SC or VLAD with corresponding local descriptors set. Finally, for BoF or VLAD, we concatenate the weighted histograms of each subregion of video to form a feature vector of the video; while for SC, the corresponding coefficients to the local descriptor sets in each subregion are concatenated, then the max pooling function is applied to form the representation of the video.

3.6 Support Vector Machine

The size of feature vectors of videos generated by Spatial-Temporal Pyramid (STP) approach would be very large. For example, a feature vector of a video constructed by 3-level uniformly distributed pyramid and 256 visual words would

have 9216 attributes. If these feature vectors are directly classified by SVM classifiers, it would be very computationally expensive on the training stage, especially for large dataset involving thousands of videos.

We build a two-layer SVM classifiers system for classification processing based on [22]. The structure of the two-layer SVM classifiers system is shown in Figure 3. In the first layer, the vectors produced by the same pyramid level in different videos are classified separately using χ^2 kernel. The decision values outputted by the first layer for each video against the corresponding class label can be viewed as an abstract descriptors of the particular pyramid level of videos. Then the decision values from each pyramid level are concatenated and classified again by RBF kernel.

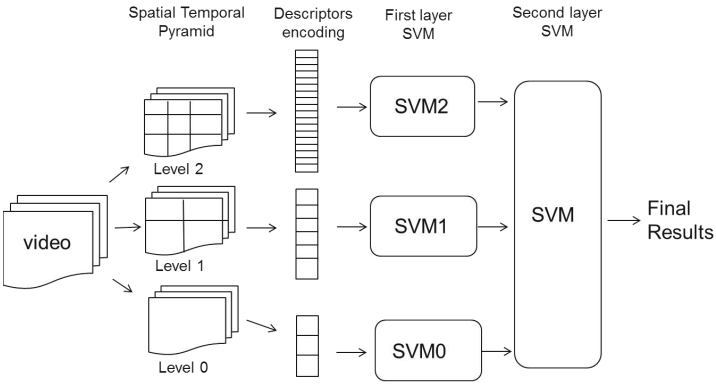


Fig. 3. The two layer SVM classifier structure

The two-layer SVM classifiers have the following attractive properties: 1) The decision values represent the descriptors set in a more concise way and are more robust to the effect of noise. 2) it can combine different types of feature effectively. In our case, it is better to match the kernel in different spatial temporal pyramid. 3) the second layer SVM assigns weights based on action classes for each pyramid level instead of assigning it to the visual words of different levels directly. 4) this would enable parallelized computing to make the overall process more efficient. All these properties leads to better results than the standard SPM method with traditional one-layer SVM classifier.

4 Experiments

4.1 Dataset

The video dataset we used is the YouTube action dataset from [14]. The videos in this dataset are mostly collected from YouTube and captured under uncontrolled condition so they contain significant camera motion, background clutter,

illumination changes, viewpoint changes and objects scale changes. All these properties of this video dataset make it closer to the realistic video data, but also push precise recognition more highly challenging.

YouTube action dataset contains 11 action categories: basketball *shooting*, cycling, diving, golf *swing*, horse-back *riding*, soccer *juggling*, swinging, tennis *swinging*, trampoline *jumping*, volleyball *spiking* and *walking* with a dog. In order to remove the unfair effect of the same background in recognition, the videos in each kind of action are split into 25 groups, where each group has different actors, backgrounds, viewpoints. Our experiments setup is the same as that proposed in [14]. There are totally 1168 videos for use and leave-one-out group cross validation is used. All the colorful videos are convert into gray-level before further processing.

4.2 Hand-Crafted Feature Representation

Firstly, we evaluated BoF and SC encoding combined with spatial temporal pyramid based on the motion feature only. The results are shown in Figure 4. As expected, it can be observed that SC achieves higher accuracies in most classes of action as well as the overall accuracy 64.98% than that 60.10% of BoF. We explained this improvement as that SC can achieve a much lower reconstruction error due to the less restrictive constraint, although it is more computationally expensive.

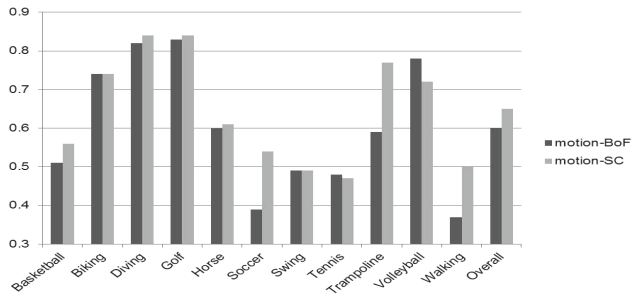


Fig. 4. The classification accuracies generated by BoF and SC based on the motion feature.

The number of static local descriptors can be tens of thousands. Because of the high memory requirement, the static feature is built by only BoF due to its low computational complexity. The overall accuracy based on static feature built on original frames is 65.33%, while the accuracy based on static feature built on multi-scale frames is 66.52%. There is no significant improvement between multi-scale and original static feature. Therefore, we discard multi-scale static feature for decreasing the computational complexity and use the original static feature only for the rest experiments.

We also evaluated the motion model, static model and hybrid model. The motion model is based only on motion feature encoded by SC while the static model is based only on static feature encoded by BoF. The hybrid model is to combine the motion feature and static feature. The results are shown in Figure 5. Intuitively, motion feature and static feature are complementary for action recognition. And this has been proven by our experiment that the accuracy of hybrid model is higher than eithor motion or static model in every class of action recognition as well as the overall accuracy, which is 75.51%, 64.98%, 65.33% for hybrid, motion and static model respectively. The hybrid model has the better performance over 10% than both motion and static model, which is impressive. Hence, it can be concluded that the hybrid model can achieve the best results, and not only motion feature but also static feature plays a significant role in action recognition. It can be also observed that the static model works surprisingly better than motion model. We explain this improvement by the fact that the dense feature contain more useful information than the interest points based feature. The confusion table for classification using hybrid model is shown in Figure 6.

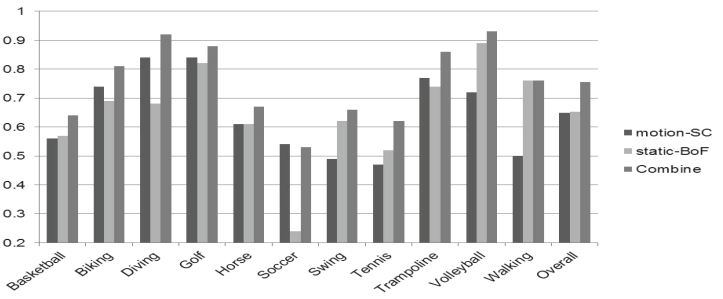


Fig. 5. The classification accuracies generated by motion, static and hybrid model

Lastly, we compared our method based on hand-crafted feature representation with the state-of-art on the same dataset (see Table 1). It can be clearly seen that our method is competitive with the state-of-art. Specifically, our framework is quite similar with Liu et al. [14] but our overall accuracy (75.51%) is higher than theirs (71.2%). Note that the highest accuracy (85.4%) proposed by Wang et al. [18] is much higher (over 10%) than all other methods because they adopted the dense trajectories feature on building motion feature, which is very computational intensive and memory consuming.

4.3 Random Feature Representation

The parameter settings for building random feature representation is the same as building the hand-crafted feature representation described above. We also evaluated the parameters of random feature representation by grid search. Firstly,

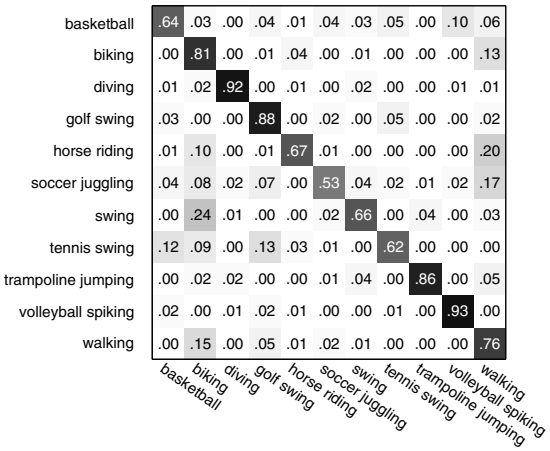


Fig. 6. The confusion table for classification using hybrid model

Table 1. The comparison with the state-of-art

Liu et al. (2009) [14]	**71.2%**
Ikizler-Cinbis and Sclaroff (2010) [6]	75.21%
Brendel and Todorovic (2010) [3]	77.8%
Le et al. (2011) [12]	75.8%
Bhattacharya et al. (2011) [2]	76.5%
Wang et al. (2013) [18]	85.4%
Our method	**75.51%**

we searched for the appropriate projected dimension n. The sub-feature vector is projected into 25, 50, 100, 200 dimensions so that the dimensionality of the final local descriptor would be 200, 400, 800 and 1600 respectively. Note that for the descriptors with 1600 dimensionality, we sampled 400 descriptors from each video to generate the codebook due to the high memory requirement. Another parameter we try to optimize is the size of the cuboid, the size employed in building hand-crafted feature representation ($19 \times 19 \times 11$ pixels) is taken as the benchmark.

The results based on the diverse projected dimensions are shown in Figure 7. It can be seen that the accuracies over diverse dimensions of descriptors are all fluctuated around 50% and there is no significant difference between each dimensionality. In addition, no obvious tendency of improvement or decreasing over the diverse dimensions can be observed. Therefore, we conclude that the projected dimension is not an important factor that affects the final classification accuracy. The projected dimension is then fixed to 200 as same as hand-crafted descriptors due to its lower computational complexity and for the sake of comparisons.

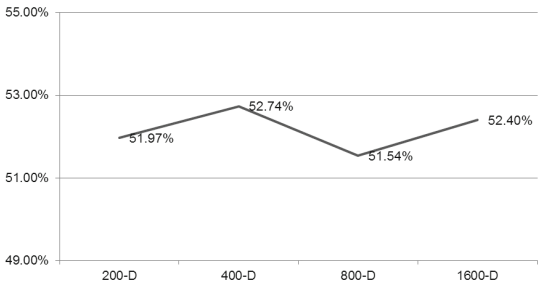


Fig. 7. The accuracies based on the descriptors projected to 200, 400, 800 and 1600 dimensions by random projection on the extracted cuboid with the size of $19 \times 19 \times 11$ pixels

To investigate the effect of size of cuboid, we conducted a set of experiments based on 3 different spatial size and 2 different temporal size. The results of total 6 experiments with different cuboid sizes are shown in Table 2. Again, all the results fluctuated between 45% and 50% and there is no significant improvement among them. The best result we got is 51.97% with the $19 \times 19 \times 11$ pixels cuboid size. Therefore, changing the size of the extracted cuboid would not improve the performance.

In addition, the result applying VLAD encoding method is 55.65% based on the 200 dimension random descriptors and 128 visual words, which is similar to that of using SC (55.31%). As expected, the result of VLAD is better than BoF (51.97%) and the computational time is much less than SC but at the cost of consuming memory.

Table 2. The results based on different sizes of the cuboid with a fixed projected 200 dimension

$11 \times 11 \times 11$	$11 \times 11 \times 23$
50.26%	45.89%
$19 \times 19 \times 11$	$19 \times 19 \times 23$
51.97%	48.03%
$39 \times 39 \times 11$	$39 \times 39 \times 23$
44.09%	46.40%

4.4 Comparisons

We also evaluated the random descriptors on the static model and the hybrid model. Again, the hybrid model can achieve about 8% improvement on overall accuracy over the motion and static model. The best results we obtained for random feature representation and the corresponding results generated by the hand-crafted feature are list in Table 3.

Table 3. The comparison between the random feature and the hand-crafted feature representation

Method	Random Feature	Hand-crafted Feature
motion - BoF	51.97%	**60.10%**
motion - SC	55.31%	**64.98%**
static - BoF	51.54%	**65.33%**
Combined	62.50%	**75.51%**

From the Table 3, we can see that there is a big difference between results from the two proposed feature representations. For each evaluation of encoding method, the performance of hand-crafted feature is over 10% higher than that of random feature, which cannot be ignored. As the framework, encoding and classifiers parameter settings are totally the same for evaluating both feature representation, we can conclude that the random feature representation does not perform well in this YouTube action dataset although it is simple to be implemented and successful in other recognition domains. Recall that random projection is a power tool in dimensionality reduction and should be beneficial in the cases where the distances of the original high dimensional data are meaningful. Therefore, we explained this failure of random feature representation for possibly one reason that the original distance or similarities information contained by the extracted cuboids are themselves suspect so that the random feature descriptors are not distinct enough to be classified.

5 Conclusions

In this paper, we investigate and compare two different feature representations for video based human action recognition: hand-crafted and random feature representation. The former is built by 3D HoG/HoF descriptors for motion feature and SIFT descriptors for static feature while the latter is based on random projection. Three popular approaches of encoding descriptors: BoF, SC and VLAD are applied. Additionally, spatial temporal pyramid and a two-layer SVM classifier are employed for classification processing.

For the motion feature of both representations, we evaluated both BoF and SC encoding methods. The results confirms that SC outperforms BoF as indicated in object recognition community. Based on the performance of the motion, static and hybrid model, we found that using hybrid features of motion and static can significantly improve the overall recognition accuracy which only uses motion features. Therefore, as complementary of the motion feature, the static feature plays an essential role in action recognition on this dataset and suprisingly it even works better than motion feature only. Compared with the success of the popular hand-crafted feature representation such as 3D HoG/HoF, SIFT descriptors for action recognition, the proposed random feature representation

based on random projection does not perform well in this dataset. This is probably due to the suspect of original information contained by the extracted cuboids as well as the random error.

The overall accuracies over YouTube action dataset based on random features is far behind the state-of-art performance. For the future work, the random feature based approach would be experimented on other datasets, such as HMDB51 dataset [9] and Hollywood movie dataset [10].

Acknowledgments. This work is partially supported by the National Natural Science Foundation of China (Project no. 61005038) and an internal funding from United International College (Project no. R201312).

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
2. Bhattacharya, S., Sukthankar, R., Jin, R., Shah, M.: A probabilistic representation for efficient large scale visual recognition tasks. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2593–2600. IEEE (2011)
3. Brendel, W., Todorovic, S.: Activities as time series of human postures (2010)
4. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of johnson and lindenstrauss. Random Structures & Algorithms **22**(1), 60–65 (2003)
5. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. IEEE (2005)
6. Ikizler-Cinbis, N., Sclaroff, S.: Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 494–507. Springer, Heidelberg (2010)
7. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3304–3311. IEEE (2010)
8. Klaser, A., Marszalek, M.: A spatio-temporal descriptor based on 3d-gradients. In: IEEE Conference on British Machine Vision Conference, BMVC 2009 (2009)
9. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision (ICCV) (2011)
10. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, pp. 1–8. IEEE (2008)
11. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
12. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3361–3368. IEEE (2011)

13. Li, W., Zhang, J., McKenna, S.J., Coats, M., Carey, F.A.: Classification of colorectal polyp regions in optical projection tomography. ISBI (2013)
14. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 1996–2003. IEEE (2009)
15. Liu, L., Fieguth, P.: Texture classification from random features. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(3), 574–586 (2012)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
17. Perronnin, F., Sánchez, J., Mensink, T.: Improving the Fisher Kernel for Large-Scale Image Classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 143–156. Springer, Heidelberg (2010)
18. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision, 1–20 (2013)
19. Willems, G., Tuytelaars, T., Van Gool, L.: An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 650–663. Springer, Heidelberg (2008)
20. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(2), 210–227 (2009)
21. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2009, pp. 1794–1801. IEEE (2009)
22. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision **73**(2), 213–238 (2007)

Modeling Supporting Regions for Close Human Interaction Recognition

Yu Kong[1(✉)] and Yun Fu[1,2]

[1] Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA
[2] College of Computer and Information Science, Northeastern University, Boston, MA, USA
{yukong,yunfu}@ece.neu.edu

Abstract. This paper addresses the problem of recognizing human interactions with close physical contact from videos. Different from conventional human interaction recognition, recognizing close interactions faces the problems of ambiguities in feature-to-person assignments and frequent occlusions. Therefore, it is infeasible to accurately extract the interacting people, and the recognition performance of an interaction model is degraded. We propose a patch-aware model to overcome the two problems in close interaction recognition. Our model learns discriminative supporting regions for each interacting individual. The learned supporting regions accurately extract individuals at patch level, and explicitly indicate feature assignments. In addition, our model encodes a set of body part configurations for one interaction class, which provide rich representations for frequent occlusions. Our approach is evaluated on the UT-Interaction dataset and the BIT-Interaction dataset, and achieves promising results.

1 Introduction

Automatic understanding human actions in videos is important to several real-world applications, for example, video retrieval, video annotation, and visual surveillance. These videos often contain close interactions between multiple people with physical contact (e.g., "hug" and "fight"). This raises two major challenges in understanding this type of interaction videos: the body part occlusion and the ambiguity in feature assignments (features such as interest points are difficult to be uniquely assigned to a particular person in close interactions).

Unfortunately, the aforementioned problems are not addressed in existing interaction recognition methods [1,11,12,24]. Methods in [1,11] use trackers/detectors to roughly extract people, and assume interactions do not contain close physical contact (e.g., "walk" and "talk"). Their performance are limited in close interactions since the feature of one single person may contain noises from background or the other interacting people. Feature assignment problem is avoided in [12,24] by treating the interaction people as a group. However, they do not utilize the intrinsic rich context of the interaction. Interest points have

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 29–44, 2015.
DOI: 10.1007/978-3-319-16181-5_3

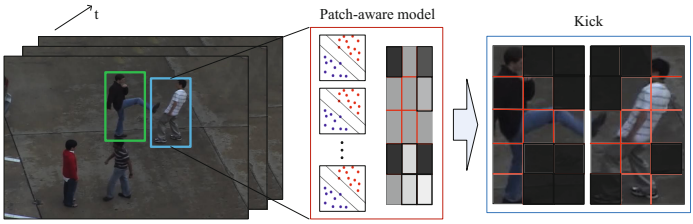


Fig. 1. Example of the inference results of our patch-aware model. Our model recognizes human interaction and discriminatively learns the supporting regions for each interacting people.

shown that they can be mainly associated with foreground moving human bodies in conventional single-person action recognition methods [13,21]. However, since multiple people present in interactions, it is difficult to accurately assign interest points to a single person, especially in close interactions. Therefore, action representations of people are extremely noisy and consequently degrade the recognition performance.

In this paper, we propose a novel patch-aware model for solving the aforementioned problems in close human interaction recognition from videos (Figure 1). Our model learns discriminative supporting regions for each interacting person, which accurately separate the target person from background. The learned supporting regions also indicate the feature-to-person assignments, which consequently help better represent individual actions. In addition, each interaction class associates with a variety of supporting region configurations, thereby providing rich and robust representations for different occlusion cases.

We propose a rich representation for close interaction recognition. Specifically, we introduce a set of binary latent variables for 3D patches indicating which subject the patch is associated with (background, person 1 or person 2), and encourage consistency of the latent variables across all the training data. The appearance and structural information of patches is jointly captured in our model, which captures the motion and pose variations of interacting people. To address the challenge of an exponentially large label space, we use a structured output framework, employing a latent SVM [6]. During training, the model learns which patterns belong to the foreground and background, allowing for better labeling of body parts and identification of individual people. Results show that the learned supporting patches significantly facilitate the recognition task.

Our work differs from [1,2,11,14] in that they can only deal with interactions that do not contain close physical contact (e.g. "queueing" and "talking") while our method specifically aims at recognizing close interactions. Different from [17,19,24] which treat the interacting people as a group, our model provides fine-grained supporting regions for each interacting person, which allows us to recognize individual action. Although methods in [18,22] can roughly extract

each interacting person using a tracker or detector, they do not model 3D patches and background, and cannot accurately separate people. Our method, in contrast, captures different importance of 3D patches in interaction classes and thus can accurately separate people.

2 Related Work

Multi-person activity recognition has been receiving much attention in computer vision community. Methods in [2,11] studied the collective activity recognition problem using crowd context. People in a collective activity have no close physical contact with each other and perform similar action, e.g. "crossing the road", "talking", or "waiting". Specifically, Choi *et al.*[2] utilized human pose, velocity and spatiotemporal distribution of individuals to represent the crowd context information. They further developed a system that can simultaneously track multiple people and recognize their interactions [1]. Lan *et al.*[11] represented crowd context by action co-occurrence of interacting people. Odashima *et al.*[14] proposed the Contextual Spatial Pyramid to detect the action of multiple people.

Human interactions, e.g. "hug", "push", and "hi-five", usually involve frequent close physical contact. Perez *et al.*[15] investigated interaction recognition between two people in realistic scenarios. They adopted a human detector to extract individual in videos. However, the ambiguities in feature-to-person assignments during close physical contact remains a problem. Ryoo and Aggarwal [18] utilized body part tracker to extract each individual in videos and then applied context-free grammar to describe spatial and temporal relationships between people. To avoid the extraction of individual people, approaches in [12,19,24] treat interacting people as a group and recognize their interactions based on group motion patterns.

Human-object and object-object interaction have also been investigated in recent work. Gupta *et al.*[8] incorporated rich context derived from object class, object reaction, and manipulation motion into Bayesian models for recognizing human-object interaction from videos and static images. Mutual context of objects and human poses was explored by Yao and Fei-Fei [23]. Their work showed that using mutual context, solving human pose estimation and object detection problems simultaneously can greatly benefit each other. A dynamically multi-linked Hidden Markov Model was proposed by Gong and Xiang [7] for recognizing group actions involving multiple objects. Desai *et al.*[3] encoded geometric configurations of objects and human pose in contextual models for recognizing human-object interactions (e.g. tennis-serve and tennis-forehand).

3 Interaction Representation

Our approach takes advantage of 3D spatiotemporal local features to jointly recognize interaction and segment people in the interaction. Given a video, a visual tracker is applied to extract interacting people from each other, and also differentiate them from the background at a patch-level. In each bounding box,

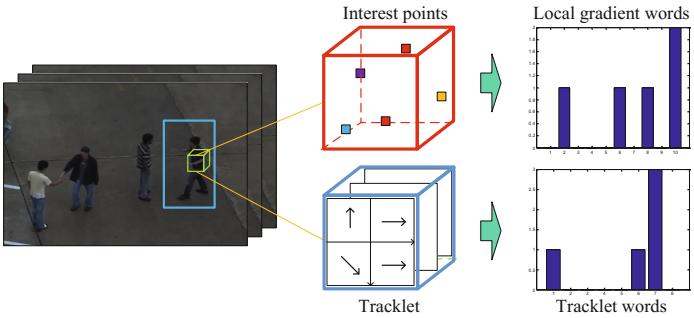


Fig. 2. Illustration of feature representation. We extract both interest points and tracklet from 3D patches.

spatiotemporal interest points [5] and tracklet [16] are computed within each 3D patch, and described using the bag-of-word model [5,12,13] (Figure 2). Spatiotemporal patches are obtained by decomposing a video of size $R \times C \times T$ into a set of non-overlapping spatiotemporal 3D patches, each of which is of size $r \times c \times t$. Similar to action representation based on histograms of video words [5,13,17], we describe each patch by the histogram of video words within the patch.

Noted that the detected interest points and tracklet are mainly associated with salient regions in human body; few of them are associated with background. This results in an inexpressive representation for background. Our aim in this paper is to extract each interacting people from the interactions and thus the background must be described.In this paper, we augment *virtual video words* (VVWs) to describe background.

The idea of VVWs is to build a discriminative feature for background so that background and foreground can be well differentiated. Consider the features of patches as data points in a high-dimensional space. Then patch features associated with foreground are distributed subjecting to an unknown probability. We would like to define some virtual data points for background and make them as far as possible from those foreground data points in order to make these two-class data points well separated. Since we use linear kernel in the model, the best choice for virtual data points is the one that can be linearly separated from foreground data points). In our work, we use origin point for virtual data points, i.e. all the bins in the histogram of a 3D patch which have no video words in it are set to 0.

4 Patch-Aware Model

Given the representation of an interaction video, our goal is to determine the interaction class (e.g. "push") as well as infer supporting regions for each interacting person. These 3D regions in this work can be associated with background or one of the interacting people.

Suppose we are given N training samples $\{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$, where $\mathbf{x} \in \mathbb{R}^D$ denotes the video feature and $y \in \mathcal{Y}$ is the interaction class. Our purpose is to learn a discriminative function $f_{\mathbf{w}} : \mathbf{x} \rightarrow y$, which infers the interaction class for an unknown interaction video. To model the supporting regions for each interacting person, we introduce a set of auxiliary binary latent variables $\{h_j\}_{j=1}^{M} \in \mathcal{H}$ ($h_j \in \{0, 1\}$), each of which associates with one patch. $h_j = 0$ denotes that the j-th patch is associated with the background and $h_j = 1$ means it is with foreground. Note that intra-class variability leads to different patch configurations in certain interaction classes. For instance, in "handshake", some people would like to pat the other people while shaking hands with the people but some do not like that. We solve this problem by treating regions as latent variables and inferring the most probable states of latent variables in training. An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is employed to encode the configurations of these patches. A vertex $h_j \in \mathcal{V}$ ($j = 1, \cdots, M$) corresponds to the j-th patch and an edge $(h_j, h_k) \in \mathcal{E}$ corresponds to the dependency between the two patches.

We define the discriminative function as

$$f(\mathbf{x}; \mathbf{w}) = \arg\max_{y} \left[\max_{\mathbf{h}} F(\mathbf{x}, \mathbf{h}, y; \mathbf{w}) \right] \tag{1}$$

where $\mathbf{h}$ is vector of all latent variables. The scoring function $F(\mathbf{x}, \mathbf{h}, y; \mathbf{w})$ is used to measure the compatibility between between the video data $\mathbf{x}$, the interaction class y and the latent patch labels $\mathbf{h}$.

We model the scoring function $F(\cdot)$ as a linear function $F(\mathbf{x}, \mathbf{h}, y; \mathbf{w}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{h}, y) \rangle$ with $\mathbf{w}$ being model parameter and $\Phi(\mathbf{x}, \mathbf{h}, y)$ being a feature vector. Specifically, the scoring function $F(\cdot)$ is defined as the summation of four components:

$$\begin{aligned} F(\mathbf{x}, \mathbf{h}, y; \mathbf{w}) = & \sum_{j \in \mathcal{V}} \alpha^{\mathrm{T}} \psi(\mathbf{x}_j, h_j, y) + \sum_{j \in \mathcal{V}} \beta^{\mathrm{T}} \theta(\mathbf{x}_j, h_j) \\ & + \sum_{j \in \mathcal{V}} \gamma_j^{\mathrm{T}} \eta(h_j, y) + \lambda^{\mathrm{T}} \pi(\mathbf{x}, y), \end{aligned} \tag{2}$$

where $\mathbf{w} = \{\alpha, \beta, \gamma, \lambda\}$ is model parameter, $\mathbf{x}_j$ is the feature extracted from the j-th patch.

Class-specific Patch Model. $\alpha^{\mathrm{T}} \psi(\mathbf{x}_j, h_j, y)$ models the agreement between the observed patch feature $\mathbf{x}_j$, the patch label h_j and the interaction class y. The definition of the feature vector $\psi(\mathbf{x}_j, h_j, y)$ is given by

$$\psi(\mathbf{x}_j, h_j, y) = \mathbf{1}(y = a) \cdot \mathbf{1}(h_j = b) \cdot f(\mathbf{x}_j), \tag{3}$$

where $f(\mathbf{x}_j)$ denotes the local feature of the j-th patch and $\mathbf{1}(\cdot)$ is an indicator function. In our work, $f(\mathbf{x}_j)$ encodes both appearance information and structural information of the j-th patch: $f(\mathbf{x}_j) = [f_a(\mathbf{x}_j), f_s(\mathbf{x}_j)]$. The appearance information $f_a(\mathbf{x}_j)$ is the distribution of words in the patch, and the structural information $f_s(\mathbf{x}_j)$ is the location of the patch. To compute the structural feature $f_s(\mathbf{x}_j)$, we discretize the bounding box into M patches and the spatial location

feature of a patch $\mathbf{x}_j$ can be represented as a vector of all zeros with a single 1 for the bin occupied by $\mathbf{x}_j$. We apply a template α of size $(D+M) \times H \times Y$ on the feature function $\psi(\mathbf{x}_j, h_j, y)$ to weigh the different importance of elements in the feature function, where Y is the number of interaction classes, and H is the number of patches labels. Each entry in α_{yhcm} can be interpreted as, for patch of state h, how much the proposed model prefers to see a discriminative word in the m-th bin when the codeword is c and the interaction label is y. The class-specific patch model $\alpha^{\mathrm{T}}\psi(\mathbf{x}_j, h_j, y)$ can be regarded as a linear classifier and scores the feature vector $\psi(\mathbf{x}_j, h_j, y)$.

The model encodes class-specific discriminative patch information which is of great importance in recognition. Note that the patch label h is unobserved during training and the feature function defined above models the implicit relationship between an interaction class and supporting regions. During training, the model automatically "aware" the supporting regions for an interaction class by maximizing the score $F(\mathbf{x}, \mathbf{h}, y; \mathbf{w})$.

Global Patch Model. $\beta^{\mathrm{T}}\theta(\mathbf{x}_j, h_j)$ measures the compatibility between the observed patch feature $\mathbf{x}_j$ and the patch label h_j. We define the feature function $\theta(\mathbf{x}_j, h_j)$ as

$$\theta(\mathbf{x}_j, h_j) = \mathbf{1}(h_j = b) \cdot f(\mathbf{x}_j), \tag{4}$$

where $f(\mathbf{x}_j)$ is the local feature of the j-th patch used in the class-specific patch model. This model encodes shared patch information across interaction classes. It is a standard linear classifier trained to infer the label (0 or 1) of the j-th patch given patch feature $\mathbf{x}_j$. The parameter β is a template, which can be considered as the parameter of a binary linear SVM trained with data $\{\mathbf{x}_j, h_j\}_{j=1}^{M}$.

Essentially, the global patch model encodes the shared patch information across interaction classes. For example, since we use a tracker to obtain a bounding box of an interacting person, this person tends to appear in the middle of the box and thus the patches in the middle of the box are likely to be labeled as foreground. This information is shared across all interaction classes and can be elegantly encoded by our global patch model.

Class-specific Structure Model. $\gamma_j^{\mathrm{T}}\eta(h_j, y)$ encodes the structural information of patches in one interaction class. Intuitively, human poses are different in various interaction classes. Although this information are unobserved in training samples, we treat them as latent variables so that they can be automatically discovered during model training. The class-specific structure model is given by

$$\eta(h_j, y) = \mathbf{1}(h_i = b) \cdot \mathbf{1}(y = a). \tag{5}$$

Clearly, the label of a patch is related to its location. Therefore, we use a set of untied weights $\{\gamma\}_{j=1}^{M}$ for the j-th patch, each of which is of size $H \times Y$, where M is the number of patches. The class-specific structure model expresses the prior that, without observing any feature, given an interaction class a, which state of the j-th patch is likely to be.

The class-specific structure model expresses the idea that, without observing any low-level feature, given an interaction class a, which state of the j-th patch is likely to be. The model shows its preference by scoring the feature vector

$\eta(h_j, y)$ using a weight vector γ_j. Since the feature vector is a $0-1$ vector, if an entry in $\gamma_j(b, a)$ is positive, the model encourages labeling the j-th patch as b when current interaction class is a.

Global Interaction Model. $\lambda^{\mathrm{T}}\pi(\mathbf{x}, y)$ is used to differentiate different interaction classes. We define this feature vector as

$$\pi(\mathbf{x_0}, y) = \mathbf{1}(y = a) \cdot \mathbf{x_0}, \tag{6}$$

where $\mathbf{x}_0 \in \mathbb{R}^d$ is a feature vector extracted from the whole action video. Here we use the bag-of-words representation for the whole video. This potential function is essentially a standard linear model for interaction recognition if other components are not considered. If other potential functions in Eq.(2) are ignored, and only the global interaction potential function is considered, the parameter λ can be learned by a standard multi-class linear SVM.

Discussion. The proposed patch-aware model is specifically designed for interaction recognition with close physical contact. Compared with exiting interaction recognition methods [1,2,11,14,17–19,22,24], our model accounts for motion at a fine-grain patch level using the three components, the class-specific patch component, the global patch component, and the class-specific structure component. These three components model the appearance and structural information of local 3D patches and allow us to accurately separate interacting people at patch-level. To our best knowledge, our work is the first one that provides supporting patches for close interaction recognition, which can be used to separate interacting people.

5 Model Learning and Testing

Learning. The latent SVM formulation is employed to train our model given the training examples $\mathcal{D} = \{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$:

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_n(\xi_n + \sigma_n) \tag{7}$$

$$\text{s.t.} \max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{(n)}, \mathbf{h}_{y^{(n)}}, y^{(n)}) - \max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}^{(n)}, \mathbf{h}, y) \tag{8}$$

$$\geqslant \Delta(y, y^{(n)}) - \xi_n, \forall n, \forall y,$$

$$\mu(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y) \leqslant \sigma_n, \forall n, \forall y, \tag{9}$$

where $\mathbf{w}$ denotes model parameter, ξ and σ are slack variable that allow for soft margin, and C is the soft-margin parameter. $\Delta(y, y^{(n)})$ represents the 0-1 loss function. $\mu(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y)$ in Constraint (9) enforces the similarity over latent regions for training videos. Our assumption is that, for videos in the same category, they are likely to have the same latent variable values. We define $\mu(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y)$ as

$$\mu(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}_y, y) = \frac{1}{M} d(\mathbf{h}_{y^{(n)}}, \mathbf{h}_y) \cdot \mathbf{1}(y = y^{(n)}), \tag{10}$$

where $d(\cdot,\cdot)$ computes the Hamming distance between the two vectors. The optimization problem (7-9) can be solved using the latent SVM framework [6].

Computing Subgradient. The above optimization problem can be efficiently solved by the non-convex cutting plane algorithm [4]. The key idea of this algorithm is that, it iteratively approximates the objective function by increasingly adding new cutting planes to the quadratic approximation. The two major steps of the algorithm are to compute the empirical loss $R(\mathbf{w}) = \sum_n (\xi_n + \sigma_n)$ and the subgradient $\frac{\partial R}{\partial \mathbf{w}}$.

The computation of a subgradient is relatively straight-forward, assuming the inference over $\mathbf{h}$ can be done. Denote the empirical loss $R(\mathbf{w})$ as $R(\mathbf{w}) = \sum_n R^n(\mathbf{w})$, then the subgradient can be computed by

$$\frac{\partial R}{\partial \mathbf{w}} = \Phi(\mathbf{x}^{(n)}, \mathbf{h}^*, y^*) - \Phi(\mathbf{x}^{(n)}, \mathbf{h}', y^{(n)}), \tag{11}$$

where $(\mathbf{h}^*, y^*)$ and $\mathbf{h}'$ are computed by

$$(\mathbf{h}^*, y^*) = \arg\max_{y,\mathbf{h}} \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}^{(n)}, \mathbf{h}, y) + \Delta(y^{(n)}, y), \tag{12}$$

$$\mathbf{h}' = \arg\max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}^{(n)}, \mathbf{h}, y^{(n)}) - \mu(\mathbf{h}_{y^{(n)}}, y^{(n)}, \mathbf{h}, y). \tag{13}$$

Testing. Given an unknown interaction video, we assume that the interaction region in the video is known. Our aim is to infer the optimal interaction label y^* and the optimal configurations of 3D patches $\mathbf{h}^*$:

$$\max_{y} \max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}, \mathbf{h}, y). \tag{14}$$

To solve the above optimization problem, we enumerate all possible interaction classes $y \in \{\mathcal{Y}\}$ and solve the following optimization problem:

$$\mathbf{h}_y^* = \arg\max_{\mathbf{h}} \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}, \mathbf{h}, y), \forall y \in \mathcal{Y}. \tag{15}$$

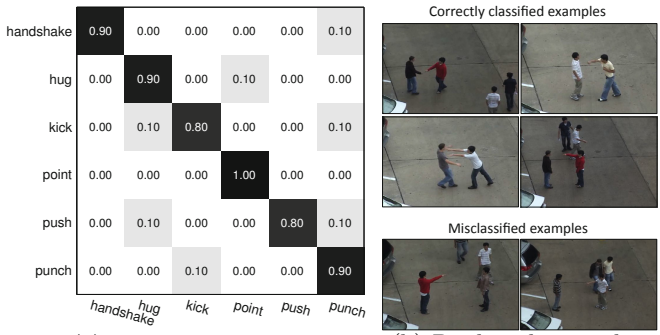
Here, the latent variables $\mathbf{h}$ are connected by a lattice. In this work, we adopt loopy belief propagation to solve the above optimization problem.

Given the latent variable vector $\mathbf{h}_y^*$, we then compute the score $f_{\mathbf{w}}(\mathbf{x}, \mathbf{h}_y^*, y) = \mathbf{w}^{\mathrm{T}} \Phi(\mathbf{x}, \mathbf{h}_y^*, y)$ for all interaction classes $y \in \mathcal{Y}$ and pick up the optimal interaction class y^* which maximizes the score $F(\mathbf{x}, \mathbf{h}_y^*, y; \mathbf{w})$.

6 Experiments

6.1 Datasets

We test our method on the UT-Interaction dataset [20] and the BIT-Interaction dataset [9]. UT dataset consists of 6 classes of human interactions: handshake, hug, kick, point, punch and push. The UT dataset was recorded for the human activity recognition contest (SDHA 2010) [20], and it has been used by several state-of-the-art action recognition methods [17,19,24]. BIT dataset consists of 8 classes of human interactions: bow, boxing, handshake, high-five, hug, kick, pat, and push. Each class contains 50 videos, with a total of 400 videos.



(a) Confusion matrix. (b) Predicted examples.

Fig. 3. Confusion matrix and classification examples of our method on UT dataset

6.2 Experiment Settings

We extract 300 interest points [5] from a video on both datasets. Gradient descriptors are utilized to characterize the motion around interest points. Principal component analysis algorithm is applied to reduce the dimensionality of descriptors to 100 and build a visual word vocabulary of size 1000. We use a visual tracker to obtain a bounding box for each interacting people. Then a 3D volume computed by stacking bounding boxes along temporal axis is split into non-overlapping spatiotemporal cuboids of size $15 \times 15 \times 15$. We use the histogram of the video words in a 3D patch as the patch feature.

We adopt the leave-one-out training strategy on the UT dataset. The split training strategy is applied on BIT dataset to train our model. 272 videos are randomly chosen for training our patch-aware model and the remaining videos are used for testing.

6.3 Experimental Results

Results on UT-Interaction dataset. On UT dataset, we first evaluate the recognition accuracy of our method and report supporting region results. Then we compare with state-of-the-art methods [10,11,13,17,24].

Recognition Accuracy. We test our method on UT dataset and show the confusion matrix in Figure 3. Our method achieves 88.33% recognition accuracy. Confusions are mainly due to visually similar movements in two classes (e.g. "push" and "punch") and the influence of moving objects in the background. Classification examples are illustrated in Figure 3.

Eq.(5) defines a class-specific structure model for all classes. It would be interesting to investigate the performance of a shared pose prior. We replace the class-specific structure prior in Eq.(5) with a shared one which is defined as $\eta(h_j, y) = 1(h_i = b)$. Results are shown in Table 1. The accuracy difference between the two priors is 5%. This is mainly due to that motion variations in

Fig. 4. The learned supporting regions on the UT dataset

individual actions are significant. The model with class-specific prior is able to learn pose under different classes, and benefits the recognition task.

Table 1. Accuracies of different pose prior on UT dataset

Pose prior	shared	class-specific
Accuracy	83.33%	88.33%

Supporting Regions. The learned supporting regions on the UT dataset are shown in Fig. 4. Our model can accurately discover supporting regions of interacting people. This is achieved by finding the most discriminative regions (e.g. hand and leg) that support an interaction class. Note that some videos in the UT dataset have background motion, e.g.,"point", which introduces noise in the video. However, our model uses the structure prior component in Eq. (5) and the consistency Constraint (9) to enforce a strong structure prior information on the patches, and thus can determine which patches are unlikely to be associated with foreground. This leads to accurate patch labeling results. Some of the patch labels are incorrect mainly due to intra-class variations. People in an interaction class may behave differently according to their personal habits. This increases the difficulty of learning class-specific pose prior.

Comparison Results. We evaluate the value of components in the proposed model, including the global interaction model, the structure prior model, and the patch models. We remove these from our patch-aware model respectively, and obtain three different methods: the no-GI method that removes global interaction potential $\lambda^{\mathrm{T}}\pi(\mathbf{x}, y)$, the no-SP method that removes the structure prior potential $\gamma_j^{\mathrm{T}}\eta(h_j, y)$, and the no-CGP method which removes both class-specific and global patch model $\alpha^{\mathrm{T}}\psi(\mathbf{x}_j, h_j, y)$ and $\beta^{\mathrm{T}}\theta(\mathbf{x}_j, h_j)$ from the full model.

We compare our full model with previous methods [11,13,17,24], the no-GI method, no-SP method and no-CGP method, and adopt a bag-of-words representation with a linear SVM classifier as the baseline. Results in Table 2 show that our method outperforms all the comparison methods. It should be noted that our method learns supporting regions, which can be used to separate people while the methods in [11,13,17,24] cannot achieve this goal.

Results in Table 2 show that our method outperforms [11,13,17,24]. The baseline bag-of-words method simply uses low-level features for recognition. By

Table 2. Recognition accuracy (%) of methods on the UT dataset.

Methods	Function	handshake	hug	kick	point	punch	push	Overall
bag-of-words	only Rec.	70	70	80	90	70	70	75
no-GI method	Rec. and Seg.	20	30	40	30	10	20	25
no-SP method	Rec. and Seg.	70	80	70	70	80	80	75
no-CGP method	Rec. and Seg.	80	90	70	90	80	80	81.67
Liu *et al.*[13]	only Rec.	60	70	100	80	60	70	73.33
Lan *et al.*[11]	only Rec.	70	80	80	80	90	70	78.33
Yu *et al.*[24]	only Rec.	100	65	75	100	85	75	83.33
Ryoo & Aggarwal [17]	only Rec.	80	90	90	80	90	80	85
Our method	Rec. and Seg.	90	90	80	100	80	90	**88.33**

comparison, our method treats cuboid variables as mid-level features and utilize them to describe local motion information. With rich representation of interaction, our method achieves superior performance. Our method outperforms the method proposed in [17]. Their method uses structural information between interest points to aid recognition. In this work, we adopt a different scheme to encode structure information of interest points. The information is encoded by the location of spatiotemporal cuboids which contains the interest points. Besides, the learned supporting regions in our model can also be used to separate people in interactions while their method cannot. Lan *et al.*[11] utilized action context to recognize interactions. We argue that action context may not able to capture complex action co-occurrence since individual motion could be totally different in an interaction class. Thus modeling the action context may not capture significant motion variations in individual actions. We infer an interaction based on the mid-level patch features. The mid-level features we build can provide detailed regional motion information of interactions and thus improve recognition results. Compared with [24], our method learns supporting regions to separate people while [24] treats interacting people as a group and do not consider separation.

Evaluation on BIT-Interaction Dataset. We conduct two groups of experiments on BIT dataset. First, we test the recognition performance of our method, and show the results on supporting regions and the structure prior. We then test the effectiveness of each component in our patch-aware model.

Recognition Results. In the first experiment, we test the proposed method on BIT dataset. The confusion matrix is shown in Figure 5(a). Our method achieves 85.38% accuracy in classifying human interactions. Results show that the method can differentiate interactions in various challenging situations, e.g. partially occlusion and background clutter (Figure 5(b)). This is mainly due to the modeling of the supporting regions. In such challenging scenarios, the supporting regions of each interacting people can be accurately inferred by the patch-aware model according to their appearance and structural information. If

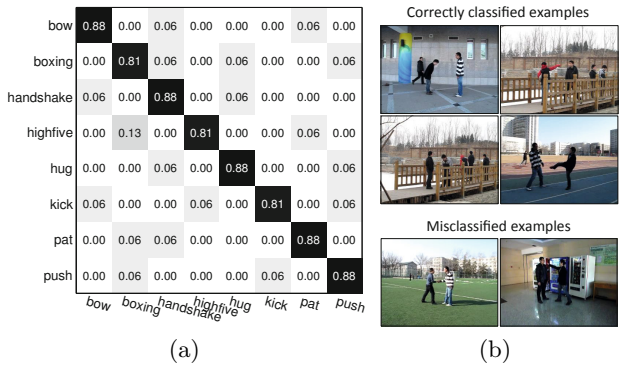


Fig. 5. (a) Confusion matrix of our method and (b) classification examples on the BIT dataset. Our method achieves 85.38% accuracy on the BIT dataset.

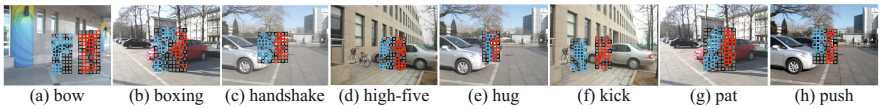


Fig. 6. The learned supporting regions on the BIT dataset

the region belongs to the interacting people, the model would assign high weights to the feature extracted from that region and thus more trust the region. If not, the feature extracted from the region will receive low weight and thus play trivial role in interaction recognition.

Most of the misclassifications are due to the visual similarity, e.g. "boxing" and "push", "pat" and "boxing". In addition, some temporal segments in the classes of "boxing" and "push" are shared with other classes. For example, in "boxing", some early segments are visually similar to some "hug" segments. Both of them are "stretching out hand". Since we adopt voting strategy for classification, these misclassified segments would result in the misclassification of the video. Moreover, some of misclassifications are due to significant occlusion in which the extracted interest points are no discriminative enough for differentiating interactions.

Supporting Regions. Results in Figure 6 show that our model can accurately find supporting regions for interacting people in close interactions. For example, in "hug" interaction (Figure 6(e)), the supporting regions of two close people can be accurately labeled. Our model essentially conducts a refinement in the bounding box. It utilizes both appearance and structure information of patches, and learns latent pose prior for each interaction class. The optimal patch label configuration (supporting regions) that maximizes the score of an interaction class are automatically discovered in the learning procedure. The learned

supporting regions overcome the problem of ambiguity in feature assignments and thus facilitate the recognition task. Some of the labels are incorrect. This is mainly due to intra-class variations. People in an interaction class may behave differently according their personal habits. This increases the difficulty of learning class-specific pose priors. We do not fully use temporal information in our model since the inference on a loopy graph is inefficient.

Structure Prior. We encode the shared structure prior potential function into our patch-aware model (refer to as SS model) and compare it with the proposed model defined in Eq.(2) (called full model). Results in Table 3 indicate that the full model outperforms the SS model. The reason can be explained from the view of parameters. For j-th patch, a shared prior for the patch is associated with parameter γ_j where γ_j is a vector of length H. This shared model is too simple to capture pose variations among all the classes. By comparison, a class-specific prior for the cuboid is associated with parameter γ_j where γ_j is a vector of length $Y \times H$. With a more complex structure prior, the full model can easily capture large pose variations and separate background and foreground for each interaction class. Thus the recognition performance is improved.

Table 3. Accuracies of different pose prior on BIT dataset

Pose prior	shared	class-specific
Accuracy	80.47%	85.38%

Comparison Results. In this experiment, we evaluate the value of components in the proposed method, including the global interaction potential, the structure prior potential, and the potential encoding appearance and structure information of observations. We remove these from our patch-aware model respectively, and obtain three different methods: the no-GI method that removes global interaction potential $\lambda^{\mathrm{T}}\pi(\mathbf{x}, y)$, the no-SP method that removes the structure prior potential $\gamma_j^{\mathrm{T}}\eta(h_j, y)$, and the no-CGP method which removes the appearance and structure information of observations $\alpha^{\mathrm{T}}\psi(\mathbf{x}_j, h_j, y)$ and $\beta^{\mathrm{T}}\theta(\mathbf{x}_j, h_j)$ from the full model. Our patch-aware model is compared with these three methods as well as the baseline bag-of-words representation with a linear SVM classifier.

Table 4 indicates that our method outperforms all the baseline methods. The performance gain achieved by our method over the baseline bag-of-words method is significant since our model is able to automatically infer the supporting regions and treat them as mid-level features. As expected, our method significantly outperforms the no-GI method, which emphasizes the importance of global interaction potential in interaction recognition. The global interaction potential function can be considered as a standard linear model for interaction recognition without considering other components. Without this potential, the

Table 4. Recognition accuracy (%) on the BIT dataset. R. and S. are short for recognition and segmentation, respectively.

Methods	Func.	bow	boxing	handshake	high-five	hug	kick	pat	push	Overall
bag-of-words	only R.	81.25	75	50	75	81.25	68.75	62.5	68.75	70.31
no-GI model	R. & S.	20.31	20.31	25	18.75	37.5	18.75	31.25	18.75	23.83
no-SP model	R. & S.	75	68.75	68.75	75	68.75	87.5	81.25	68.75	74.22
no-CGP model	R. & S.	62.5	56.25	62.5	87.5	81.25	87.5	87.5	68.75	75
Lan *et al.*[11]	only R.	81.25	75	81.25	87.5	87.5	81.25	81.25	81.25	82.03
Ours	R. & S.	87.5	81.25	87.5	81.25	87.5	81.25	87.5	87.5	**85.38**

model mainly focuses on the cuboid features which would be not discriminative enough. The results of the proposed method are higher than the no-SP method, which indicates the effectiveness of the structure prior in recognition. Without the structure prior, the no-SP method is unable to capture mid-level features in cuboids. The information the no-SP model can capture is simply the noisy low-level features rather than meaningful regional information. Since the segmentation and recognition tasks are smoothly connected in our work, the lack of semantic understanding of cuboids would influence the recognition results. As a result, the recognition accuracy of the no-SP method is decreased. The full model outperforms the no-CGP method. The appearance and structure information in the full model serves as local features and complements the global interaction information. The local features are able to describe local motion of interaction and provide detailed information. With appearance and structure information, our method can recognize more challenging interaction videos and thus achieves higher results.

7 Conclusion

We have proposed a novel model for jointly recognizing human interaction and segmenting people in the interaction. Our model is built upon the latent structural support vector machine in which the patches are treated as latent variables. The consistency of latent variables are encouraged across all the training data. The learned patch labels indicate the supporting regions for interacting people, and thus solve the problems of feature assignment and occlusion. Experiments show that our method achieves promising recognition results and can segment people at patch level during an interaction, even in a close interaction.

Acknowledgments. This research is supported in part by the NSF CNS award 1314484, ONR award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218.

References

1. Choi, W., Savarese, S.: A unified framework for multi-target tracking and collective activity recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 215–230. Springer, Heidelberg (2012)
2. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: CVPR (2011)
3. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for static human-object interactions. In: CVPR Workshop on Structued Models in Computer Vision (2010)
4. Do, T.M.T., Artieres, T.: Large margin training for hidden markov models with partially observed states. In: ICML (2009)
5. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: VS-PETS (2005)
6. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
7. Gong, S., Xiang, T.: Recognition of group activities using dynamic probabilistic networks. In: ICCV, vol. 2, pp. 742–749 (2003)
8. Gupta, A., Kembhavi, A., Davis, L.: Observing human-object interactions: Using spatial and functional compatibility for recognition. PAMI **31**(10), 1775–1789 (2009)
9. Kong, Y., Jia, Y., Fu, Y.: Learning human interaction by interactive phrases. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 300–313. Springer, Heidelberg (2012)
10. Kong, Y., Jia, Y., Fu, Y.: Interactive phrases: semantic descriptions for human interaction recognition. PAMI (2014)
11. Lan, T., Wang, Y., Yang, W., Robinovitch, S.N., Mori, G.: Discriminative latent models for recognizing contextual group activities. PAMI **34**(8), 1549–1562 (2012)
12. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR (2008)
13. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: CVPR (2011)
14. Odashima, S., Shimosaka, M., Kaneko, T., Fukui, R., Sato, T.: Collective activity localization with contextual spatial pyramid. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part III. LNCS, vol. 7585, pp. 243–252. Springer, Heidelberg (2012)
15. Patron-Perez, A., Marszalek, M., Reid, I., Zissermann, A.: Structured learning of human interaction in tv shows. PAMI **34**(12), 2441–2453 (2012)
16. Raptis, M., Soatto, S.: Tracklet descriptors for action modeling and video analysis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 577–590. Springer, Heidelberg (2010)
17. Ryoo, M.S.: Human activity prediction: early recognition of ongoing activities from streaming videos. In: ICCV (2011)
18. Ryoo, M., Aggarwal, J.: Recognition of composite human activities through context-free grammar based representation. In: CVPR, vol. 2, pp. 1709–1718 (2006)
19. Ryoo, M., Aggarwal, J.: Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In: ICCV, pp. 1593–1600 (2009)

20. Ryoo, M., Aggarwal, J.: UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA) (2010)
21. Shi, Q., Cheng, L., Wang, L., Smola, A.: Human action segmentation and recognition using discriminative semi-markov models. IJCV **93**, 22–32 (2011)
22. Vahdat, A., Gao, B., Ranjbar, M., Mori, G.: A discriminative key pose sequence model for recognizing human interactions. In: ICCV Workshops, pp. 1729–1736 (2011)
23. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR, pp. 17–24 (2010)
24. Yu, T.H., Kim, T.K., Cipolla, R.: Real-time action recognition by spatiotemporal semantic and structural forests. In: BMVC (2010)

W07 - Computer Vision with Local Binary Patterns Variants

Fast Features Invariant to Rotation and Scale of Texture

Milan Sulc(✉) and Jiri Matas

Center for Machine Perception, Department of Cybernetics,
Faculty of Electrical Engineering, Czech Technical University in Prague,
Prague, Czech Republic
sulcmila@cmp.felk.cvut.cz

Abstract. A family of novel texture representations called Ffirst, the Fast Features Invariant to Rotation and Scale of Texture, is introduced. New rotation invariants are proposed, extending the LBP-HF features, improving the recognition accuracy. Using the full set of LBP features, as opposed to uniform only, leads to further improvement. Linear Support Vector Machines with an approximate χ^2-kernel map are used for fast and precise classification.

Experimental results show that Ffirst exceeds the best reported results in texture classification on three difficult texture datasets KTH-TIPS2a, KTH-TIPS2b and ALOT, achieving 88 %, 76 % and 96 % accuracy respectively. The recognition rates are above 99 % on standard texture datasets KTH-TIPS, Brodatz32, UIUCTex, UMD, CUReT.

Keywords: Texture · Classification · LBP · LBP-HF · Histogram · SVM · Feature maps · Ffirst

1 Introduction

Texture description and recognition techniques have been the subject to many studies for their wide range of applications. The early work focused on the problem of terrain analysis [12,36] and material inspection [37]. Later applications of texture analysis include face recognition [1], facial expressions [30,42] and object recognition [39]. The relation between scene identification and texture recognition is discussed by Renninger and Malik [28]. Texture analysis is a standard problem with several surveys available, e.g. [6,19,24,40]. Many texture description methods are based on the Local Binary Patterns [10,11,17,20–23,41], which is a computationally simple and powerful approach.

We introduce a family of novel texture representations called Ffirst - the Fast Features Invariant to Rotation and Scale of Texture. It is based on LBP-HF-S-M, the rotation invariant features obtained from sign- and magnitude-LBP histograms using Fourier transform proposed by Zhao et al. [41]. We enrich the LBP-HF-S-M representation by proposing additional rotational invariants and by the use of non-uniform patterns.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 47–62, 2015.
DOI: 10.1007/978-3-319-16181-5_4

The scale invariance of Ffirst is obtained by the technique recently applied in the context of bark recognition [32].

We show that the novelties improve performance in texture recognition experiments with a feature-mapped linear SVM classifier approximating the χ^2 kernel.

The rest of this paper is organized as follows: The state-of-the-art approaches to texture recognition are briefly reviewed in Section 2. The new family of texture representations called Ffirst is introduced and described in Section 3. Section 4 is dedicated to the proposed extensions of LBP-HF and Ffirst. Section 5 presents our experiments on standard texture datasets. Section 6 concludes the paper.

2 State of the Art

Several recent approaches to texture recognition report fine results on the standard datasets, often using complex description methods. Sifre and Mallat [31] used a cascade of invariants computed using scattering transforms to construct an affine invariant texture representation. A sparse representation based Earth Mover's Distance (SR-EMD) presented by Li et al. [15] achieves good results in both image retrieval and texture recognition. Quan et al. [27] propose a texture feature constructed by concatenating the lacunarity-related parameters estimated from the multi-scale local binary patterns. Local Higher-Order Statistics (LHS) proposed by Sharma et al. [30] describe higher-order differential statistics of local non-binarized pixel patterns. The method by Cimpoi et al. [7] uses Improved Fisher Vectors (IFV) for texture description. This work also shows further improvement when combined with describable texture attributes learned on the Describable Textures Dataset (DTD).

3 The Ffirst Method

In order to describe texture independently of the pattern size and orientation in the image, a description invariant to rotation and scale is needed. For practical applications we also demand computational efficiency.

In this section we introduce a new texture description called Ffirst (Fast Features Invariant to Rotation and Scale of Texture), which combines several state-of-the-art approaches to satisfy the given requirements. This method builds on and improves a texture descriptor for bark recognition introduced in [32].

3.1 Completed Local Binary Pattern and Histogram Fourier Features

The Ffirst description is based on the Local Binary Patterns (LBP) [20,22]. The common LBP operator (further denoted as sign-LBP) computes the signs of differences between pixels in the 3×3 neighbourhood and the center pixel.

LBP have been generalized [21] to arbitrary number of neighbours P on a circle of radius R, using an image function $f(x, y)$ and neighbourhood point coordinates (x_p, y_p):

$$\mathrm{LBP}_{P,R}(x,y) = \sum_{p=0}^{P-1} s(f(x,y) - f(x_p,y_p))2^p, \quad s(z) = \begin{cases} 1 : z \leq 0 \\ 0 : \quad \text{else} \end{cases}. \tag{1}$$

To achieve rotation invariance[1], Ffirst uses the so called LBP Histogram Fourier Features (LBP-HF) introduced by Ahonen et al. [2], which describe the histogram of uniform patterns using coefficients of the discrete Fourier transform. Uniform LBP are patterns with at most 2 spatial transitions (bitwise 0-1 changes). Unlike the simple rotation invariants using $\mathrm{LBP}^{\mathrm{ri}}$[21,25], which assign all uniform patterns with the same number of 1s into one bin,

$$\mathrm{LBP}^{ri}_{P,R} = \min\left\{\mathrm{ROR}\left(\mathrm{LBP}_{P,R}, i\right) \mid i = 0, 1, .., P-1\right\}, \tag{2}$$

the LBP-HF features preserve the information about relative rotation of the patterns.

Denoting a uniform pattern $U_p^{n,r}$, where n is the "orbit" number corresponding to the number of "1" bits and r denotes the rotation of the pattern, the DFT for given n is expressed as:

$$H(n,u) = \sum_{r=0}^{P-1} h_I(U_p^{n,r}) e^{-i2\pi ur/P}, \tag{3}$$

where the histogram value $h_I(U_p^{n,r})$ denotes the number of occurrences of a given uniform pattern in the image.

The LBP-HF features are equal to the absolute value of the DFT magnitudes (which are not influenced by the phase shift caused by rotation):

$$\text{LBP-HF}(n,u) = |H(n,u)| = \sqrt{H(n,u)\overline{H(n,u)}}. \tag{4}$$

Since h_I are real, $H(n,u) = H(n,P-u)$ for $u = (1, .., P-1)$, and therefore only $\lfloor \frac{P}{2} \rfloor + 1$ of the DFT magnitudes are used for each set of uniform patterns with n "1" bits for $0 < n < P$. Three other bins are added to the resulting representation, namely two for the "1-uniform" patterns (with all bins of the same value) and one for all non-uniform patterns.

The LBP histogram Fourier features can be generalized to any set of uniform patterns. In Ffirst, the LBP-HF-S-M description introduced by Zhao et al. [41] is used, where the histogram Fourier features of both sign- and magnitude-LBP are calculated to build the descriptor. The combination of both sign- and magnitude-LBP called Completed Local Binary Patterns (CLBP) was introduced by Guo and Zhang [10]. The magnitude-LBP checks if the magnitude of the difference of

[1] LBP-HF (as well as LBP^{ri}) are rotation invariant only in the sense of a circular bit-wise shift, e.g. rotation by multiples 22.5° for $\mathrm{LBP}_{16,R}$.

the neighbouring pixel (x_p, y_p) against the central pixel (x, y) exceeds a threshold t_p:

$$\text{LBP-M}_{P,R}(x,y) = \sum_{p=0}^{P-1} s(|f(x,y) - f(x_p,y_p)| - t_p)2^p. \tag{5}$$

We adopted the common practice of choosing the threshold value (for neighbours at p-th bit) as the mean value of all m absolute differences in the whole image:

$$t_p = \sum_{i=1}^{m} \frac{|f(x_i,y_i) - f(x_{ip},y_{ip})|}{m}. \tag{6}$$

The LBP-HF-S-M histogram is created by concatenating histograms of LBP-HF-S and LBP-HF-M (computed from uniform sign-LBP and magnitude-LBP).

3.2 Multi-scale Description and Scale Invariance

A scale space is built by computing LBP-HF-S-M from circular neighbourhoods with exponentially growing radius R. Gaussian filtering is used[2] to overcome noise.

Unlike the MS-LBP approach of Mäenpää and Pietikäinen [17], where the radii of the LBP operators are chosen so that the effective areas of different scales touch each other, Ffirst uses a finer scaling with a $\sqrt{2}$ step between scales radii R_i, i.e. $R_i = R_{i-1}\sqrt{2}$.

This radius change is equivalent to decreasing the image area to one half. The finer sampling uses more evenly spaced information compared to [17], as illustrated in Figures 1a, 1b. The first LBP radius used is $R_1 = 1$, as the LBP with low radii capture important high frequency texture characteristics.

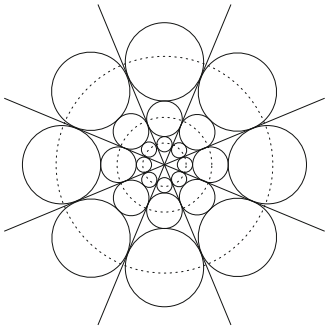
Similarly to [17], the filters are designed so that most of their mass lies within an effective area of radius r_i. We select the effective area diameter, such that the effective areas at the same scale touch each other: $r_i = R_i \sin \frac{\pi}{P}$.

LBP-HF-S-M histograms from c adjacent scales are concatenated into a single descriptor. Invariance to scale changes is increased by creating n_{conc} multi-scale descriptors for one image. See Algorithm 1 for the overview of the texture description method.

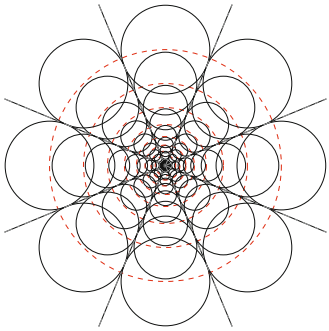
3.3 Support Vector Machine and Feature Maps

In most applications, a Support Vector Machine (SVM) classifier with a suitable non-linear kernel provides higher recognition accuracy at the price of significantly higher time complexity and higher storage demands (dependent on the number of support vectors). An approach for efficient use of additive kernels via explicit feature maps is described by Vedaldi and Zisserman [35] and can be

[2] The Gaussian filtering is used for a scale i only if $\sigma_i > 0.6$, as filtering with lower σ_i leads to significant loss of information.



(a) Scale space of Mäenpää and Pietikäinen [17]



(b) Scale space from [32] used in Ffirst

Fig. 1. The effective areas of filtered pixel samples in a multi-resolution $\mathrm{LBP}_{8,R}$ operator

Algorithm 1. The Ffirst description method overview

```
R_1 := 1
for all scales i = 1...(n_conc + c − 1) do
  σ_i := R_i sin π/P / 1.96
  if σ_i > 0.6 then
    apply Gaussian filter (with std. dev. σ_i) on the original image
  end if
  extract LBP_{P,R_i}-S and LBP_{P,R_i}-M and build the LBP-HF-S-M descriptor
  for j = 1...n_conc do
    if i ≥ j and i < j + c then
      attach the LBP-HF-S-M to the j-th multi-scale descriptor
    end if
  end for
  R_{i+1} := R_i √2
end for
```

combined with a linear SVM classifier. Using linear SVMs on feature-mapped data improves the recognition accuracy, while preserving linear SVM advantages like fast evaluation and low storage (independent on the number of support vectors), which are both very practical in real time applications. In Ffirst we use the explicit feature map approximation of the χ^2 kernel.

The "One versus All" classification scheme is used for multi-class classification, implementing the Platt's probabilistic output [16,26] to ensure SVM results comparability among classes. The maximal posterior probability estimate over all scales is used to determine the resulting class.

In our experiments we use a Stochastic Dual Coordinate Ascent [29] linear SVM solver implemented in the VLFeat library [34].

4 Adding Rotational Invariants

The LBP-HF features used in the proposed Ffirst description are built from the DFT magnitudes of differently rotated uniform patterns, as described in Section 3.1. We propose 3 more variants for the description, which will appear in our experiments in Section 5.

The variant denoted as Ffirst^{+} creates additional rotational invariants, LBP-HF^{+} features, computed from the first harmonics for each orbit:

$$\text{LBP-HF}^{+}(n) = \sqrt{H(n,1)\overline{H(n+1,1)}} \tag{7}$$

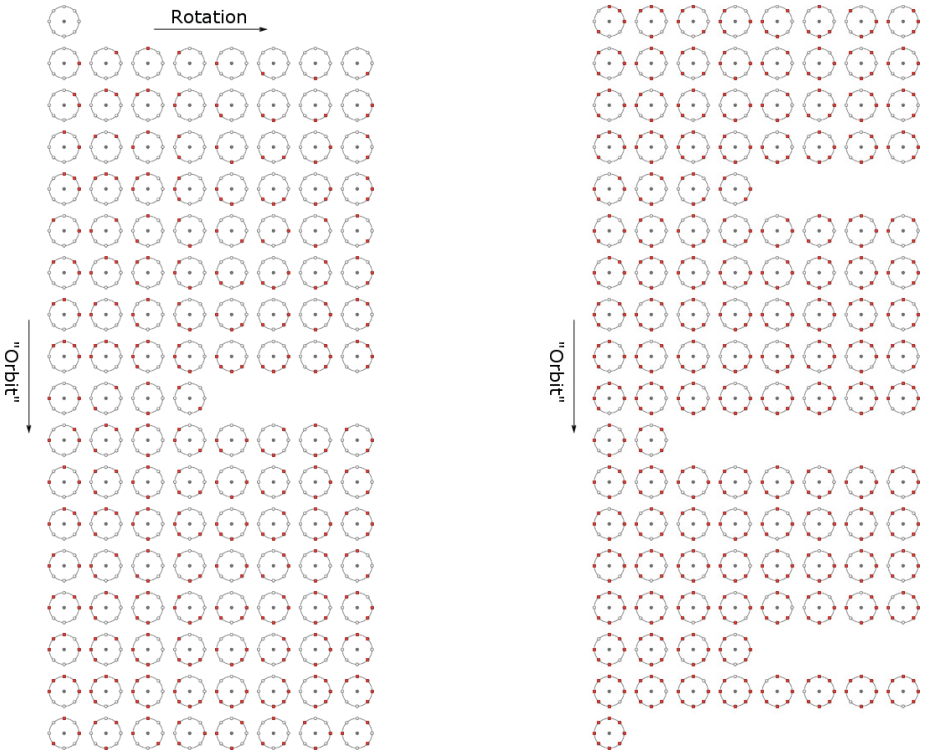


Fig. 2. Ordering the full set of Local Binary Patterns for the Histogram Fourier features

Another variant, Ffirst$^{\forall}$, uses all LBP instead of only the subset of uniform patterns. Note that in this case, some orbits have a lower number of patterns, as some non-uniform patterns have less possible rotations, as illustrated in Figure 2.

The last variant, denoted as Ffirst$^{\forall+}$, uses the full set of patterns for LBP-HF features, adding also the additional LBP-HF^{+} features

5 Experiments

5.1 Datasets

The proposed Ffirst method for texture classification was tested using the standard evaluation protocols on the following texture datasets:

The KTH-TIPS texture database [9,13] contains images of 10 materials. There are 81 images (200x200 px) of each material with different combination of pose, illumination and scale.

The standard evaluation protocol on the KTH-TIPS dataset uses 40 training images per material.

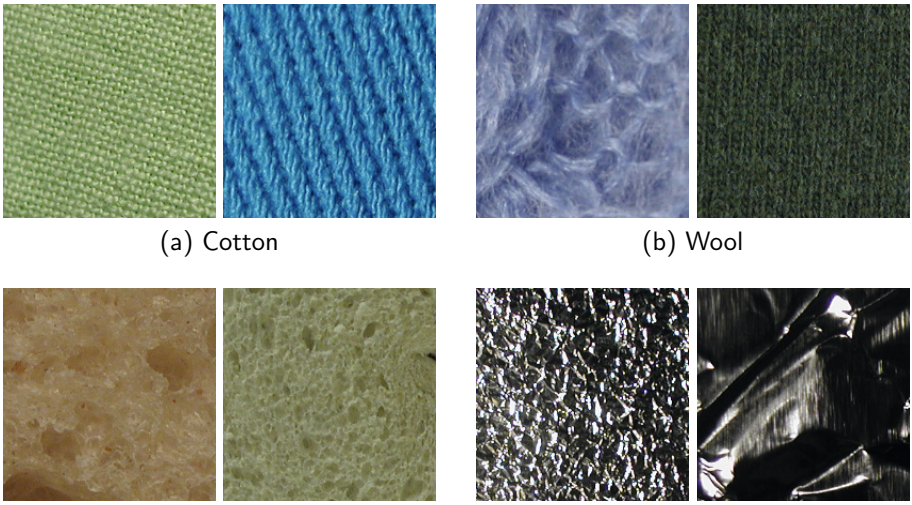


Fig. 3. Examples of 4 texture classes from the KTH-TIPS2 database

The KTH-TIPS2 database was published [5,18] shortly after KTH-TIPS. It builds on the KTH-TIPS database, but provides multiple sets of images - denoted as "samples" - per material class (examples in Figure 3).

There are 4 "samples" for each of the 11 materials in the KTH-TIPS2 database, containing 108 images per "sample" (again with different combination of pose, illumination and scale). However, in the first version of this dataset, for 4 of those 44 "samples" only 72 images were used. This first version is usually denoted as KTH-TIPSa, and the standard evaluation method uses 3 "samples" from each class for training and 1 for testing. The "complete" version of this database, KTH-TIPSb, is usually trained only on 1 "samples" per class and tested on the remaining 3 "samples".

The Brodatz32 dataset [33] was published in 1998 and it contains low resolution (64x64 px) grey-scale images of 32 textures from the photographs published by Phil Brodatz [3] in 1966, with artificially added rotation (90°) and scale change (a 64x64 px scaled block obtained from 45x45 pixels in the middle). There are 64 images for each texture class in total.

The standard protocol for this dataset simply divides the data into two halves (i.e. 32 images per class in the training set and 32 in the test set).

Even though the original images are copyrighted and the legality of their usage in academic publications is unclear[3], Brodatz textures are one of the most popular and broadly used sets in texture analysis.

The UIUCTex database, sometimes referred to as the Ponce Group Texture Database, was published by Lazebnik et al. [14] in 2005 and features 25 different texture classes, 40 samples each. All images are in VGA resolution (640x480 px) and in grey-scale.

The surfaces included in the database are of various nature (wood, marble, gravel, fur, carpet, brick, ..) and were acquired with significant viewpoint, scale and illumination changes and additional sources of variability, including, but not limited to, non-rigid material deformations (fur, fabric, and water) and viewpoint-dependent appearance variations (glass). Examples of images from different classes are in Figure 4.

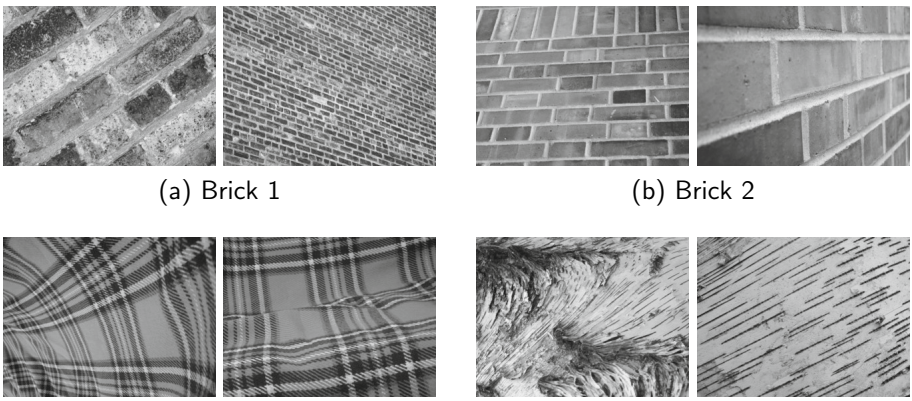


Fig. 4. Examples of 4 texture classes from the UIUCTex database

The results on this dataset are usually evaluated using 20 or 10 training images per class. In our experiments, the former case with a larger training set is performed.

[3] http://graphics.stanford.edu/projects/texture/faq/brodatz.html

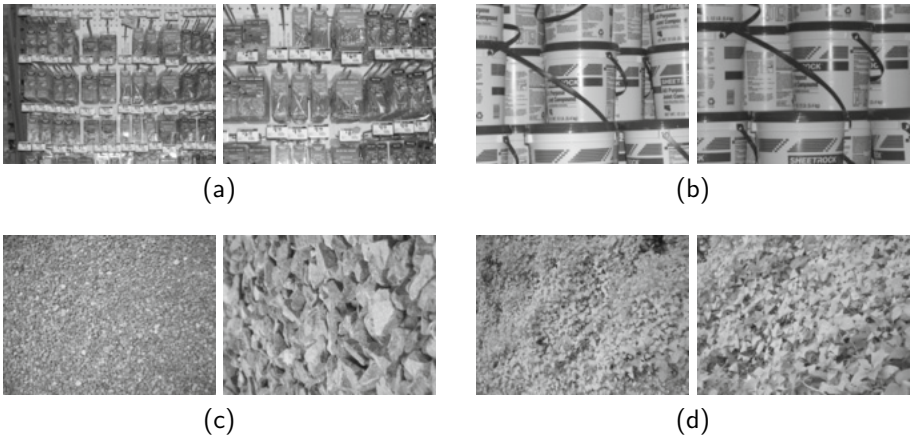


Fig. 5. Examples of 4 texture classes from the UMD database

The UMD dataset [38] consists of 1000 uncalibrated, unregistered grey-scale images of size 1280x960 px, 40 images for each of 25 different textures. The UMD database contains non-traditional textures like images of fruits, shelves of bottles and buckets, various plants, or floor textures.

The standard evaluation protocol for UMD is dividing the data into two halves (i.e. 20 images per class in the training set and 20 in the test set).

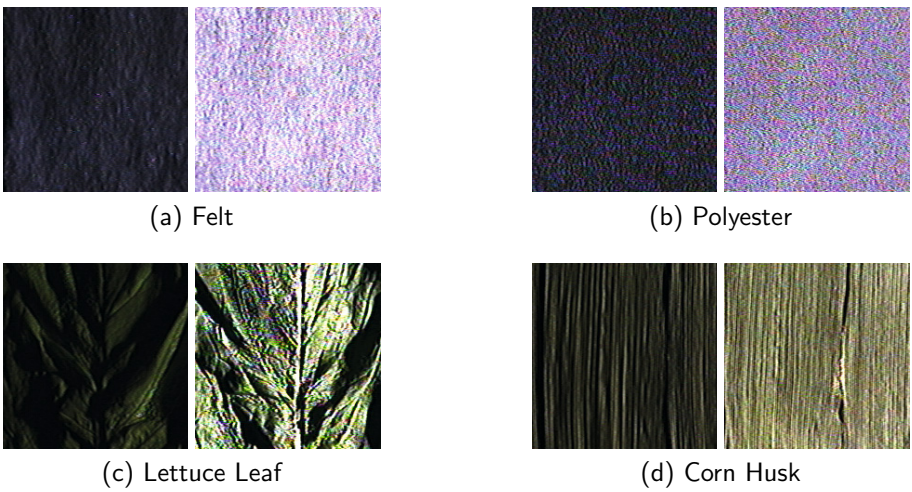


Fig. 6. Examples of 4 texture classes from the CUReT database

The CUReT image database [8] contains textures from 61 classes, each observed with 205 different combinations of viewing and illumination directions. In the commonly used version, denoted as the cropped CUReT database[4], only 92 images are chosen, for which a sufficiently large region of texture is visible across all materials. A central 200x200 px region is cropped from each of these images, discarding the remaining background. There are thus 61x92=5612 images in the cropped database.

Though CUReT also contains a BRDF (bidirectional reflectance distribution function) database, for purposes of standard texture recognition methods, only the image database is used. We use 46 training images per class, which is a standard evaluation protocol for the CUReT database.

Fig. 7. Examples of 4 texture classes from the ALOT database

The Amsterdam Library of Textures [4], denoted as ALOT, contains 250 texture classes. Each class contains 100 images obtained with different combinations of viewing and illumination directions and illumination color.

To compare our results on the ALOT dataset to the state-of-the-art [27] we use 20 training images and 80 test images per class.

5.2 Parameter setting

In all following experiments, we use the same setting of our method: $n_{\text{conc}} = 3$ multi-scale descriptors per image are used, each of them consisting of $c = 6$ scales described using LBP-HF-S-M. A higher number of concatenated scales offers only minimal improvement in accuracy, while increasing the processing time. The final histogram is kernelized using the approximate χ^2 feature map, although using the intersection kernel would provide similar results. In the application,

[4] http://www.robots.ox.ac.uk/~vgg/research/texclass/setup.html

the data are only trained once and the training precision is more important than the training time. Thus we demand high accuracy, setting SVM parameters to: regularization parameter $\lambda = 10^{-7}$, tolerance for the stopping criterion $\epsilon = 10^{-7}$, maximum number of iterations: 10^8. We use the unified setting in order to show the generality of the Ffirst description, although setting the parameters individually for a given dataset might further increase the accuracy.

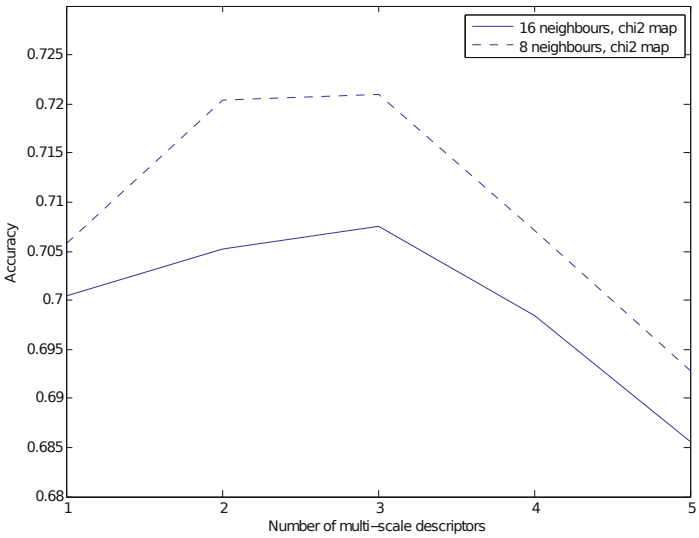


Fig. 8. Dependence of the KTH-TIPS2b recognition rate on the number of multiscale descriptors in Ffirst, denoted c

Figures 8 and 9 illustrate the effect of different parameter settings on the recognition accuracy for the KTH-TIPS2b texture database.

To reduce the effect of random training and test data choice, the presented results are averaged from 10 experiments.

5.3 Classification Results

The experimental results in texture classification are compared to the state-of-the-art in Tables 1, 2, containing the results on the KTH-TIPS datasets and on other standart texture datasets respectively.

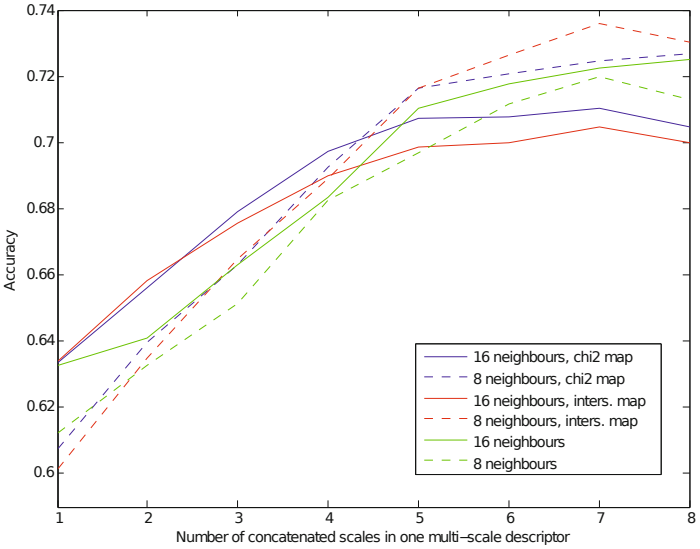


Fig. 9. Feature mapping and concatenating features from multiple scales in Ffirst, KTH-TIPS2b

Table 1. Evaluation of Ffirst on other standard datasets, compared to the state-of-the-art methods

	Brodatz32	*UIUCTex*	*UMD*	*CUReT*	*ALOT*
Num. of classes	32	25	25	61	250
Ffirst	99.2±0.3	98.6±0.6	99.3±0.3	98.5±0.2	92.9±0.3
Ffirst^{+}	99.3±0.3	98.7±0.7	99.3±0.3	98.6±0.3	93.4±0.3
Ffirst$^{\forall}$	99.6±0.2	99.0±0.5	99.3±0.3	99.1±0.2	95.0±0.3
Ffirst$^{\forall+}$	**99.7±0.2**	99.3±0.4	99.3±0.3	99.2±0.2	**95.9±0.5**
IFV_{SIFT} [7]	–	97.0±0.9	99.2±0.4	99.6±0.3	–
IFV_{SIFT} [7] + DeCAF [5]	–	99.0±0.5	99.5±0.3	**99.8±0.2**	–
Scattering [31]	–	**99.4±0.4**	**99.7±0.3**	–	–
LHS [30]	99.5±0.2	–	–	–	–
SR-EMD-M [15]	–	–	**99.9**	99.5	–
PLS [27]	–	96.6	98.99	–	93.4
MS-LBP-HF-KlSVM [32]	96.2±0.6	96.4±0.6	–	–	–

[5] Results from http://www.robots.ox.ac.uk/~vgg/data/dtd/

Table 2. Evaluation of Ffirst on the KTH-TIPS datasets, compared to the state-of-the-art methods

	KTH-TIPS2a	*KTH-TIPS2b*	*KTH-TIPS*
Num. of classes	11	11	10
Ffirst	86.2±5.5	72.1±5.1	98.9±0.7
Ffirst^{+}	86.4±5.0	72.7±5.2	98.9±0.8
Ffirst$^{\forall}$	88.0±6.5	75.8±4.1	99.1±0.5
Ffirst$^{\forall+}$	**88.2±6.7**	**76.0±4.1**	99.1±0.5
IFV_{SIFT} [7]	82.5±5.2	69.3±1.0	**99.7±0.1**
IFV_{SIFT} [7] + DeCAF [6]	84.4±1.8	**76.0±2.9**	99.8±0.2
IFV_{SIFT}[7] + DeCAF +DTD_{RBF} [6] [7]	–	77.4±2.2	–
Scattering [31]	–	–	99.4±0.4
LHS [30]	73.0±4.7	–	–
SR-EMD-M [15]	–	–	**99.8**
PLS [27]	–	–	98.4

[6] Results from http://www.robots.ox.ac.uk/~vgg/data/dtd/
[7] The method requires an additional training set (the DTD dataset)

5.4 Suitability for Real-Time Applications

Table 3 shows a comparison of our image processing times to the state-of-the-art texture recognition method by Cimpoi et al. [7] based on IVF_{SIFT}. Both the implementation of Ffirst and IVF_{SIFT}[8] used MATLAB scripts with a C code in the VLFeat [34] framework (after adding a new CLBP implementation for our method). The processing times were measured on a standard laptop (1.3 GHz Intel Core i5, 4 GB 1600 MHz DDR3) without parallelization.

The average description time for a low resolution (200x200px) image for Ffirst is at most 0.05 s, while for higher resolutions the processing time will grow proportionally to the image resolution, as the number of local operations will increase with the number of pixels.

6 Conclusions

We proposed a family of novel texture representations called Ffirst, the Fast Features Invariant to Rotation and Scale of Texture, using several state-of-the-art approaches. The first variant, Ffirst^{+}, uses newly proposed rotational invariants, another, denoted as Ffirst$^{\forall}$, allows to build the features from the full set of LBP, including non-uniform patterns.

[8] Using the code kindly provided by the authors of [7]

Table 3. Average image description time for one image, compared to IFV_{SIFT}

	KTH-TIPS2b	*KTH-TIPS*	*CUReT*
Image resolution	200x200 px	200x200 px	200x200 px
Ffirst	0.029 s / im.	0.028 s / im.	0.029 s / im.
Ffirst^{+}	0.032 s / im.	0.032 s / im.	0.032 s / im.
$\text{Ffirst}^{\forall}$	0.035 s / im.	0.035 s / im.	0.036 s / im.
$\text{Ffirst}^{\forall+}$	0.048 s / im.	0.049 s / im.	0.049 s / im.
IFV_{SIFT} [7]	0.089 s / im.	0.088 s / im.	0.090 s / im.

The $\text{Ffirst}^{\forall+}$ method, using both proposed improvements, achieves the best results, exceeding the best reported results in texture classification on three difficult texture datasets, KTH-TIPS2a, KTH-TIPS2b and ALOT, achieving 88%, 76% and 96% accuracy respectively. The recognition rates were above 99% on standard texture datasets KTH-TIPS, Brodatz32, UIUCTex, UMD, CUReT.

The Ffirst description and the evaluation based on linear Support Vector Machines are fast, making the proposed method suitable for real time applications.

Acknowledgments. Jiri Matas was supported by Czech Science Foundation Project GACR P103/12/G084, Milan Sulc by Czech Technical University project SGS13/142/OHK3/2T/13.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
2. Ahonen, T., Matas, J., He, C., Pietikäinen, M.: Rotation Invariant Image Description with Local Binary Pattern Histogram Fourier Features. In: Salberg, A.-B., Hardeberg, J.Y., Jenssen, R. (eds.) SCIA 2009. LNCS, vol. 5575, pp. 61–70. Springer, Heidelberg (2009)
3. Brodatz, P.: Textures: a photographic album for artists and designers, vol. 66. Dover, New York (1966)
4. Burghouts, G.J., Geusebroek, J.M.: Material-specific adaptation of color invariant features. Pattern Recognition Letters **30**(3), 306–313 (2009)
5. Caputo, B., Hayman, E., Mallikarjuna, P.: Class-specific material categorisation. In: Tenth IEEE International Conference on Computer Vision ICCV 2005, vol. 2, pp. 1597–1604. IEEE (2005)
6. Chen, C.h., Pau, L.F., Wang, P.S.p.: Handbook of pattern recognition and computer vision. World Scientific (2010)
7. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. arXiv preprint arxiv:1311.3618 (2013)
8. Dana, K.J., Van Ginneken, B., Nayar, S.K., Koenderink, J.J.: Reflectance and texture of real-world surfaces. ACM Transactions on Graphics (TOG) **18**(1), 1–34 (1999)

9. Fritz, M., Hayman, E., Caputo, B., Eklundh, J.O.: The kth-tips database (2004)
10. Guo, Z., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Transactions on Image Processing **19**(6), 1657–1663 (2010)
11. Guo, Z., Zhang, L., Zhang, D.: Rotation invariant texture classification using lbp variance (lbpv) with global matching. Pattern Recognition **43**(3), 706–719 (2010)
12. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics **6**, 610–621 (1973)
13. Hayman, E., Caputo, B., Fritz, M., Eklundh, J.-O.: On the Significance of Real-World Conditions for Material Classification. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3024, pp. 253–266. Springer, Heidelberg (2004)
14. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. PAMI **27**(8), 1265–1278 (2005)
15. Li, P., Wang, Q., Zhang, L.: A novel earth movers distance methodology for image matching with gaussian mixture models. In: Proc. of IEEE International Conference on Computer Vision (ICCV) (2013)
16. Lin, H.T., Lin, C.J., Weng, R.C.: A note on platts probabilistic outputs for support vector machines. Machine Learning **68**(3) (2007)
17. Mäenpää, T., Pietikäinen, M.: Multi-scale Binary Patterns for Texture Analysis. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 885–892. Springer, Heidelberg (2003)
18. Mallikarjuna, P., Fritz, M., Targhi, A., Hayman, E., Caputo, B., Eklundh, J.: The kth-tips and kth-tips2 databases (2006). http://www.nada.kth.se/cvap/databases/kth-tips
19. Mirmehdi, M., Xie, X., Suri, J.: Handbook of texture analysis. Imperial College Press (2009)
20. Ojala, T., Pietikainen, M., Harwood, D.: Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In: Proc. IAPR 1994, vol. 1, pp. 582–585 (1994)
21. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24**(7), 971–987 (2002)
22. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996)
23. Ojala, T., Valkealahti, K., Oja, E., Pietikäinen, M.: Texture discrimination with multidimensional distributions of signed gray-level differences. Pattern Recognition **34**(3), 727–739 (2001)
24. Pietikäinen, M.: Texture recognition. Computer Vision: A Reference Guide, 789–793 (2014)
25. Pietikäinen, M., Ojala, T., Xu, Z.: Rotation-invariant texture classification using feature distributions. Pattern Recognition **33**(1), 43–52 (2000)
26. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, **10**(3) (1999)
27. Quan, Y., Xu, Y., Sun, Y., Luo, Y.: Lacunarity analysis on image patterns for texture classification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2014)
28. Renninger, L.W., Malik, J.: When is scene identification just texture recognition? Vision Research **44**(19), 2301–2311 (2004)
29. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. arXiv preprint arxiv:1209.1873 (2012)

30. Sharma, G., ul Hussain, S., Jurie, F.: Local Higher-Order Statistics (LHS) for Texture Categorization and Facial Analysis. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 1–12. Springer, Heidelberg (2012)
31. Sifre, L., Mallat, S.: Rotation, scaling and deformation invariant scattering for texture discrimination. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1233–1240. IEEE (2013)
32. Sulc, M., Matas, J.: Kernel-mapped histograms of multi-scale lbps for tree bark recognition. In: 2013 28th International Conference of Image and Vision Computing New Zealand (IVCNZ), pp. 82–87 (2013)
33. Valkealahti, K., Oja, E.: Reduced multidimensional co-occurrence histograms in texture classification. PAMI **20**(1), 90–94 (1998)
34. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/
35. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. PAMI, **34**(3) (2011)
36. Weszka, J.S., Dyer, C.R., Rosenfeld, A.: A comparative study of texture measures for terrain classification. IEEE Transactions on Systems, Man and Cybernetics **4**, 269–285 (1976)
37. Weszka, J.S., Rosenfeld, A.: An application of texture analysis to materials inspection. Pattern Recognition **8**(4), 195–200 (1976)
38. Xu, Y., Ji, H., Fermüller, C.: Viewpoint invariant texture description using fractal analysis. International Journal of Computer Vision **83**(1), 85–100 (2009)
39. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. IJCV **73**(2), 213–238 (2007)
40. Zhang, J., Tan, T.: Brief review of invariant texture analysis methods. Pattern Recognition **35**(3), 735–747 (2002)
41. Zhao, G., Ahonen, T., Matas, J., Pietikainen, M.: Rotation-invariant image and video description with local binary pattern features. IEEE Transactions on Image Processing **21**(4), 1465–1477 (2012)
42. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(6), 915–928 (2007)

Local Binary Patterns to Evaluate Trabecular Bone Structure from Micro-CT Data: Application to Studies of Human Osteoarthritis

Jérôme Thevenot[1,2,3(✉)], Jie Chen[3], Mikko Finnilä[1,2], Miika Nieminen[2], Petri Lehenkari[2], Simo Saarakkala[1,2], and Matti Pietikäinen[3]

[1] Department of Medical Technology, University of Oulu, Oulu, Finland
jerome.thevenot@oulu.fi
[2] MRC, Oulu University Hospital and University of Oulu, Oulu, Finland
[3] Department of Computer Science and Engineering, University of Oulu, Oulu, Finland

Abstract. Osteoarthritis (OA) causes progressive degeneration of articular cartilage and pathological changes in subchondral bone. These changes can be assessed volumetrically using micro-computed tomography (μCT) imaging. The local descriptor, i.e. local binary pattern (LBP), is a new alternative solution to perform analysis of local bone structures from μCT scans. In this study, different trabecular bone samples were prepared from patients diagnosed with OA and treated with total knee arthroplasty. The LBP descriptor was applied to correlate the distribution of local patterns with the severity of the disease. The results obtained suggest the appearance and disappearance of specific oriented patterns with OA, as an adaptation of the bone to the decrease of cartilage thickness. The experimental results suggest that the LBP descriptor can be used to assess the changes in the trabecular bone due to OA.

Keywords: Bone structural analysis · Micro-CT · Osteoarthritis · Multiscale LBP

1 Introduction

The local binary pattern (LBP) descriptor [24][27] has been widely used for object recognition, image segmentation, texture analysis, face analysis, *et al.* in computer vision field. However, most of the LBP studies are based on texture and material classification [24][29], and the possibilities offered by the LBP descriptor for structural analysis of objects in the medical field are still poorly known. In general, the low sensitivity of the LBP for monotonic greyscale variations [23] makes it ideal for medical image processing. Furthermore, the LBP descriptor offers the possibility to assess the distribution of local patterns within a region/volume of interest both in 2D and 3D. Despite these facts very few LBP studies have been performed for 3D data obtained by conventional medical 3D imaging techniques, such as computed tomography (CT) or magnetic resonance imaging (MRI).

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 63–79, 2015.
DOI: 10.1007/978-3-319-16181-5_5

The nature of bone tissue, and especially its inner architecture, makes it a perfect candidate for structural analysis. As the bone adapts itself to environmental factors, the changes in its structure not only give information on its strength, but also provide symptomatic indications of diseases [34]. As such, CT images has been used to assess the structural bone alterations in osteoporosis [32] or in osteoarthritis (OA) [6]. OA is a disease of the whole joint primarily causing degeneration of the articular cartilage and remodeling of subchondral bone. With current clinical diagnostic techniques, the disease is often diagnosed at the end stage when the joint replacement surgery is the only effective treatment available. In OA the subchondral bone is specifically affected by sclerosis and undergoes structural changes such as the formation of osteophytes and bone cysts [4]. It has been even suggested that the structural bone changes in OA might occur before the changes at the articular cartilage [5]. However, despite some evidence that the changes in subchondral bone could contribute to the development of OA [11][12][22][18], clinical diagnostic methods mostly focus on the cartilage alone by measuring its erosion and degeneration. Following this clinical trend, a recent histopathological method to grade the severity of OA at the tissue level, *i.e.* OARSI grading system, is assessing the depth and extent of the lesions in the articular cartilage [28]. The high reliability and repeatability of the OARSI grading system [8][26] suggest this method to be a consistent reference for comparative studies assessing OA at different stages of the disease.

Micro-computed tomography (μCT) is an imaging technique similar to clinical CT, but at micro-scale level instead of macro-scale level. A significant advantage of the LBP method if applied to μCT scans is its ability to take into account the impact of partial volume effect. The partial volume effect is always a downside of the conventional analysis of bone structure, as a result of the binarization of the data for volumetric reconstruction. Eventually, μCT allows to analyze the inner structure of bone (see Fig. 1 up) with high resolution, enabling to study the size, organization and connectivity of individual bone fibers. The analysis of μCT data allows to assess the micro-level changes in bone structure related to the adaptation of the wear of overlying articular cartilage. In several animal studies characteristic bone microstructural changes related to OA have been reported [20][33][9][15][16]. Similar trends in human studies have been observed in microscopic studies [19][2]. However, μCT-derived bone structural parameters have yet to be compared with the severity of OA in humans.

There is an evident lack of studies involving LBP method to bone analysis, and in each case it has been based on plain radiographic imaging [14][35][13], showing already the potential of this tool to assess OA [13]. The most interesting aspect of applying the LBP method on bone microarchitecture is to obtain local distribution of patterns on high amount of data for each sample.

The aim of this study was to establish a new protocol to assess the local changes in bone structure using LBP descriptor and to correlate them with the severity of OA assessed by OARSI grading as a ground truth. The procedure has been divided in two parts: the selection of relevant pixels by using multiscale LBP to assess the continuity of the patterns, and the grouping of the patterns based

on both their orientation and their amount of markers. The methodological concepts and results introduced here have been performed in 2D slice-by-slice as a preliminary study before the development of real volumetric analysis.

2 LBP Method to Perform Bone Structural Analysis from Micro CT data

2.1 Background

In this section, we will present the background of the bone structural analysis from μCT data. After that, we describe the existing methods and their limitations. The bone includes two compartments: the trabecular bone and the cortical bone surrounding it (see Fig. 5). While the cortical compartment is more compact and mainly corresponds the bending resistance of the bone [30], the trabecular compartment is metabolically more active and affected by remodeling [34]. From the structural point of view, the adaptation of the internal compartment to environmental loading conditions makes the trabecular bone a crucial region of interest for analysis of structural changes.

Micro-computed tomography is an imaging method enabling to obtain volumetric data of microstructures [3]. Briefly, X-ray transmission measurements are used to create cross-sections of a physical object, similarly to clinical CT but at higher resolutions (<100 microns). While applied to bone analysis, it allows to visualize the trabecular structures and obtain information on the density of the bone. The average grey level value within a region of interest is also correlated to bone strength. The grey level value of each pixels for each scan is highly dependent on the mineralization of the structures: poorly mineralized structure (*i.e.* fibers being resorbed or created) will have a lower grey level value than highly mineralized structures. Another phenomenon affecting the grey level value of a pixel is the partial volume effect [31]: this phenomenon occurs for features not being totally within the slice thickness of the considered image or smaller than the pixel size, resulting in a lowered grey level value of the pixels affected. This peculiar event is expected to be more likely at the edges of the fibers, areas the most affected by remodeling. Another artefact in CT analysis is the beam hardening [21], causing the edges of an object to appear brighter than the center: this artefact is caused by the attenuation of the X-rays. All of these artefacts deteriorate the quality of CT analysis.

Traditionally, in bone structural analysis several parameters of the trabecular bone is evaluated to indicate the bone inner architecture and its quality [3]. The conventional method consists of a binarization of the image stack by a pre-defined threshold value believed to represent the minimum gray scale value of the bone. Subsequently, volumetric representation of the structure can be reconstructed and parameters calculated from the 3D model. The most relevant parameters are the bone volume fraction representing the amount of bone reconstructed over the volume, the trabecular thickness and spacing giving indication on morphometry of trabecular bone structure, and the structure model index

(SMI) categorizing the trabecular structure as plate- or rod-like.

A main limitation of the conventional method to analyze bone structure is the loss of the original pixel values. Only the information related to the location of the bone and its organization remains. Furthermore, the selection of a binarization threshold is often performed subjectively, increasing inter-repeatability inaccuracies. Finally, image pixels/voxels representing poorly mineralized structure, or pixels affected by partial volume effect, are typically excluded from the analysis to avoid overestimation of the bone structures (as shown in Fig. 1).

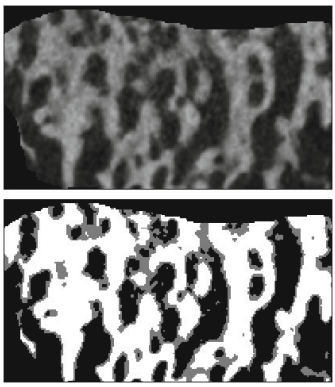


Fig. 1. up: μCT scan of a trabecular bone; **down:** white region are the bone and grey region are the pixels affected by partial volume effect / poorly mineralized. These gray regions are typically excluded from a conventional analysis of bone structure

2.2 Basics of LBP

The basics of LBP [24][27] is shown in Fig. 2. The neighborhood of a center pixel is checked for evaluating the occurrences of equal/higher grey level values than in the center pixel. A specific local pattern is then determined based on the locations of these occurrences. This pattern depicts the local structure surrounding the studied pixel such as edges, contours and flat regions.

The local structure of each pixels within a region of interest can be mathematically assessed by the following function:

$$LBP = \sum_{k=0}^{n-1} s(g_k - g_c)2^k, \quad s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0, \end{cases} \tag{1}$$

where n is the amount of neighbors evaluated, g_k the grey level value of the k-th neighbor, and g_c the pixel value of the central (studied) pixel. Depending on the radius and amount of neighbors considered the grey level value g_k might require to be estimated by interpolation.

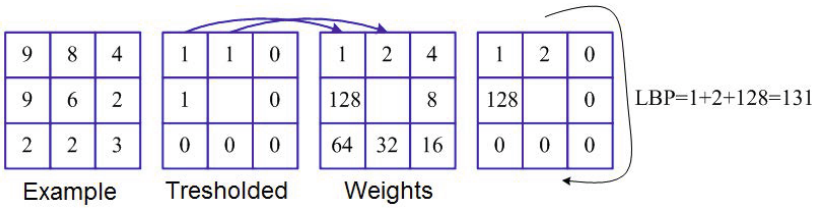


Fig. 2. Basic LBP. In the example, the center pixel (value 6) is used as a threshold. The tresholded values are then multiplied by their corresponding pixels in the weights matrix and summed to obtain the LBP value of the center pixel

Eventually, the number of different patterns that can be assessed by the LBP method is related to the amount n of neighbors evaluated for each center pixel; the number of possible patterns being 2^n. While in texture analysis the full histograms of patterns might be required in the methods, in structural analysis some grouping of patterns are required to avoid redundant information of similar patterns corresponding to identical local structures assessed at different locations.

2.3 Selection of Relevant Pixels Using Multiscale LBP

Another difference between texture analysis and bone structural analysis using the LBP is the selection of the evaluated pixels. In texture analysis, every pixels within a region of interest are usually considered in the calculations. For bone structural analysis, some pre-selection has to be performed to evaluate only relevant information. As mentioned previously, in traditional analysis of bone structure from μCT scans, the images are binarized by a threshold representing the minimum grey level value of the bone. This step allows the users to separate the relevant information (the bone) from the irrelevant empty spaces. Based on this principle, structural analysis using LBP method should be applied solely to pixels located nearby or within bone structures.

The selection of relevant pixels to compute LBP features has been divided in two steps, as shown in Algorithm 1. The first step is to filter out the empty space while the second step is to filter out the isolated noise by a connection test. In our case, the Otsu method [25] was chosen to extract the foreground information representing the bone from the background representing the empty spots. However, to verify that relevant pixels with lower grey level value are not considered as background, the averaged minimum value returned by Otsu method for all the slices is lowered by an arbitrary percentage δ of its value ($\delta = 95\%$ in our case), as shown in Step 1 of Algorithm 1. Here S_0 is the region filtered out the empty space by Otsu method.

An important factor in the bone structural analysis using the LBP method is to assess not only the pixels within the bone, as in traditional analysis, but also pixels that are in the neighborhood of the bone structures. This factor is crucial since the modelling and remodeling of the bone can be assessed in the surrounding of the structures (Step 2 of Algorithm 1 and Fig. 3).

Algorithm 1. The selection of relevant pixels

Input Images I obtained by μCT scan
Output The selected of relevant pixels S

Step 1 Filter out the empty space by Otsu method
1.1 $S_0 = \Phi$
1.2 Compute threshold by Otsu method g_{otsu}
1.3 For $g_c \in I$
If $g_c > \delta g_{otsu}$
S_0=S_0+$\{g_c\}$
End
End

Step 2 Filter out the isolate noise by connection test $g_{k,1}$
2.1 $S_1 = S_2 = \Phi$
Let $g_{k,1}$ be any of *k-th* neighbor at radius 1 from center pixel $\{g_c\}$; $g_{m,2}$ be *m-th* neighbor at radius 2 from g_c. *k*=0,1,...7 and *m*=0,1,...15
2.2 For $g_c \in S_0$
If $\exists\ g_{k,1} > \delta g_{otsu}$ and *k*=0,1,...7
S_1=S_1+$\{g_c\}$
End
End
2.3 Let *R* be the max distance used for connectivity test
For $g_c \in I - S_0$
If $\exists\ g_{k,1} > \delta g_{otsu}$ and $\exists\ g_{m,2} \geq \delta g_{otsu}$,
and $\parallel g_{k,1} - g_{m,2} \parallel \leq R$, *m*=0,1,...15
$S_2 = S_2$+$\{g_c\}$
End
End
2.4 $S = S_1 + S_2$

Specifically, the structures of the bone are continuous. Based on this principle, we propose using multiscale LBP method to select only bone structures and get rid of irrelevant artefacts. The selection of relevant pixels can be performed by two steps: the first one assessing the pixels within the structures, and the second assessing the pixels next to the edges of the structures. Both of the steps can be performed at multiple scales depending on the resolution of the images. In our case, we use two scales, *i.e.* it is considered that only areas with at least two consecutive pixels higher than a specific threshold are actual bone.

As shown in Step 2.4 of Algorithm 1, S is the resulting pixels, which are used for the following structure analysis using LBP. Increasing the amount of neighbors for a higher radius improves the accuracy of the method as more similar patterns can be considered. An option is to increase the amount of neighbors by 8 for each increase of radius, as suggested earlier [23].

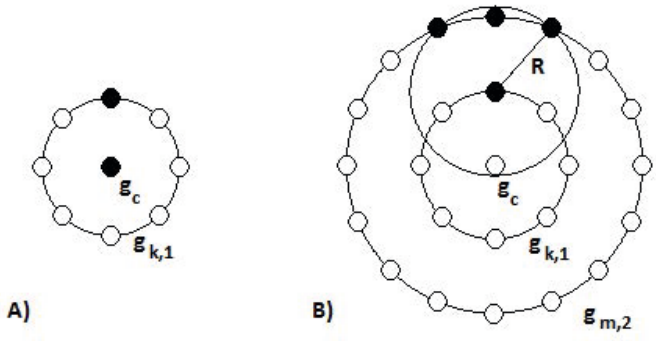


Fig. 3. Examples of center pixels considered in the analysis. Black markers have equal/higher grey level value than the threshold. **A**) Both the center pixel and one of its neighbor k have higher grey level value than the threshold. **B)** The empty center pixel is neighbor with an edge at radius 1. The continuation of the bone is validated at radius 2 within the length R (three neighbors m checked in this case)

2.4 Grouping of Patterns Using Principal Component Analysis

After applying the LBP method to an image, a histogram is generated with a size corresponding to the amount of different patterns assessed. However, in bone structural analysis a large amount of assessed patterns are redundant, as they represent the same information obtained from different locations. This suggests that grouping of the patterns is required to obtain reduced histograms with truly relevant information. Therefore, we propose to group the patterns both by their main orientation, but also by taking into account the amount of markers they consist. The term marker describes a neighbor with an equal/higher grey level value than both the central pixel g_c and the threshold. The justification for this grouping can be explained as follows:

- The main orientation of the patterns provides information on the orientation of the structures of recognized bone fiber. For example, healthy bone fibers are expected to have structures organized mainly along the daily loading orientation. Structures with different orientations could suggest an adaptation of the inner architecture of the bone fiber due to extra factors.
- The amount of markers in each pattern gives information on the nature of the local structures of bone fiber. For example, if 8 neighbors are considered in the analysis a pattern with 2-3 consecutive markers will suggest a straight structure, while a pattern with more consecutive markers will suggest a corner or a spot.

Principal component analysis (PCA) is performed for each possible pattern to obtain its main orientation. A score of the PCA is assessed to exclude the patterns without a consistent orientation. For a specific pattern, the principal components of each markers were assessed and averaged by axis. Then, the score

corresponded to the value of the axis with highest weight, this axis being the most affected by the location of the markers. A high value suggests a distribution sparse along this axis and eventually along the main orientation of the pattern. To resume, this score evaluates how well the markers of a pattern are fitted towards the line representing the principal component, if the score is higher than a given threshold value, then the orientation of the pattern is not clear.

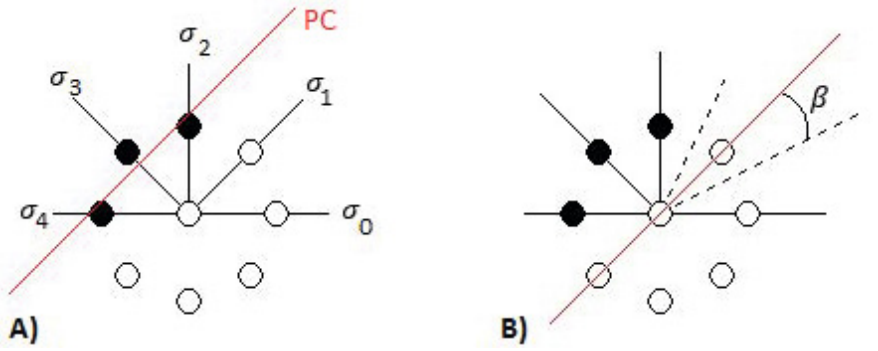


Fig. 4. Affectation of a pattern in a group. In this example, 5 different angles (σ_0 to σ_4) are used to group the patterns, as well as 8 neighbors. **A)** The principal component (PC) orientation of the pattern is defined from the 3 markers. **B)** The line representing the PC is translated to the center pixel and the angle σ_p is established within the tolerance β. For this specific case, the pattern belongs to the group (σ_1,M=3)

An example of the grouping using PCA can be seen in the Fig. 4. A group (σ_p,M) correspond to the sum of patterns with an orientation σ_p and with M markers. It is defined as follows:

$$group(\sigma_p, M) = \sum_{j=1}^{2^n} pattern_j, \quad if \begin{cases} |angle(pattern_j)| \leq \beta \\ score(pattern_j) < score_{tresh} \\ markers(pattern_j) = M, \end{cases} \tag{2}$$

with *angle(pattern$_j$)* being the orientation angle of the j pattern, β the angle deviation tolerated and $score_{tresh}$ the maximum threshold value of the score from the PCA defining a clear orientation. Eventually, the amount of different groups will be defined by the amount of markers (M) and also by the number of angles (σ_p) (note that an angle σ_p being equivalent to an angle $\pi + \sigma_p$). The angle β represents the tolerance for the patterns orientation towards the angle σ_p. The angles σ_p and β are defined as such:

$$\sigma_p = \frac{\pi p}{N}, \quad with\ 0 \leq p \leq N, \tag{3}$$

$$and \quad \beta = \frac{\pi}{2N}, \tag{4}$$

with p the considered angle division and N being the amount of different angles used for grouping the patterns.

3 Experiment: Assessment of Osteoarthritis

3.1 Sample Preparation, Imaging and Histopathology

In our experiments, 24 osteochondral samples [10] were earlier prepared from 14 patients with OA (age 76 ± 9 years: 2 males and 12 females), treated with total knee arthroplasty at Oulu University Hospital. Sample collection and their use were approved by the Ethical Committee of the Northern Ostrobothnia Hospital District, Oulu, Finland (*Diary 187/2013, Ethical Committee statement 78/2013*). Samples were prepared from tibial plateus which are always extracted during routine total knee endoprosthesis surgery. Tibial plateus were first visually classified into three categories in terms of degeneration of the articular cartilage: 1) most inviolable (or intact) cartilage, 2) moderate cartilage degeneration and wear, and 3) partly or fully exposed subchondral bone. Samples were stored in phosphate-buffered saline (PBS) for μCT imaging. While albeit samples had various cartilage thickness, they were selected from comparable anatomical area.

Osteochondral samples were scanned with μCT device at isotropic voxel size of 27.8 μm (Skyscan 1172, Bruker microCT, Kontich, Belgium). The scanned trabecular bone was located below the subchondral plate of the proximal tibia. While the bones were oriented along the proximal-distal axis during the scanning, no information regarding the medio-lateral or antero-posterior axes were available. This limitation is shown in Fig. 5.

After the μCT imaging, samples were formalin-fixed and decalcified in EDTA. Paraffin-embedded blocks were sectioned to 5 μm and stained with Safranin O. Histological sections were graded from three slices by three independent evaluators according to the standardized OARSI grading system [28]. The average from three evaluators was used as a final OARSI grade in further analysis.

3.2 Data Selection and Structural Assessment

Since only the proximal-distal axis was known, all analysis presented here were performed in both partly sagittal / coronal planes and the results obtained were averaged. Since the specific remodeling of the trabecular bone is unknown for patients at different stages of OA, it can be hypothesized that the internal architecture is affected differently along the antero-posterior axis and the medio-lateral one. Thus, considering both the perpendicular planes along the proximal-distal axis can reduce the error implied by the unknown rotation of the sample before the scanning.

The selection of the pixels implemented within the structural analysis using LBP was performed as suggested in Fig. 6. First, OTSU method within the trabecular bone was applied for each slice of both planes. Then, multiscale LBP was applied at 2 levels for the selection of relevant pixels:

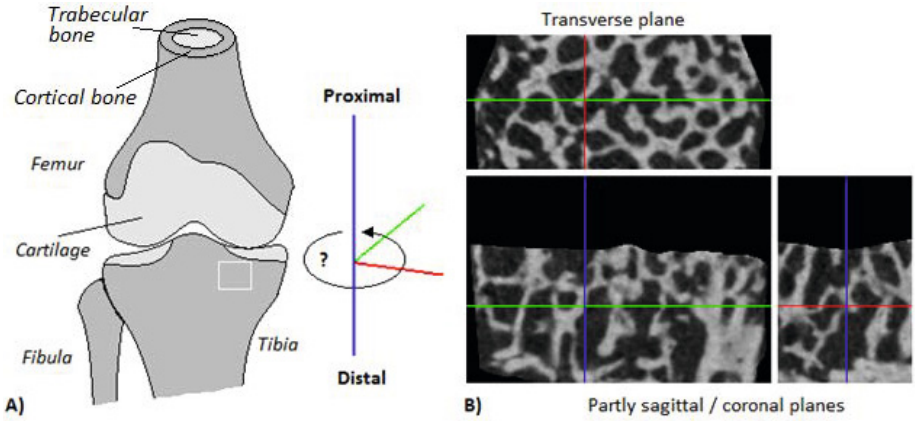


Fig. 5. A) Location of one trabecular sample (indicated by white square); only the proximal-distal axis is known. **B)** MicroCT scans of one trabecular sample along the three perpendicular planes. The lack of anatomical references suggests that the data on the lower pictures are neither fully in the sagittal nor coronal planes

- Radius 0 (the center pixel itself) and radius 1 for a center pixel with higher/ equal value than g_{Otsu}.
- Radius 1 and radius 2 for a center pixel within an empty space. Similarly than in Fig. 3, 16 neighbors at radius 2 were considered. The center pixel was included in the analysis if at least one marker at radius 1 and one marker within the 3 closest neighbors at radius 2 existed.

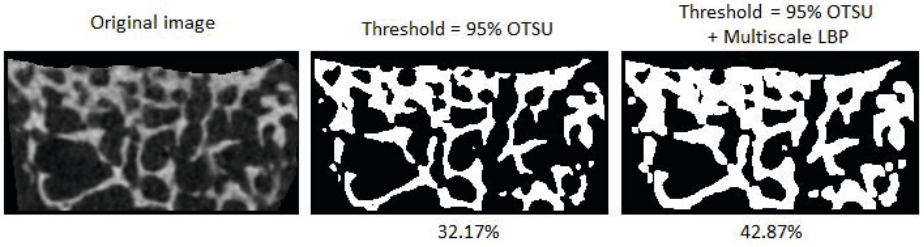


Fig. 6. Comparison between different methods to select the relevant pixels within an original μCT scan. The percentage represents the amount of selected pixels within the segmented area of the original image. Using Otsu thresholding alone gives an initial estimation for the bone pixels. Combination of Otsu thresholding and multiscale LBP selects both bone areas and the relevant surrounding pixels, and it is considered to better select relevant pixels in the analysis

Once the relevant pixels were selected, LBP analysis at radius 1 and with 8 neighbors was performed for the stack of scans. Grouping of the patterns using

PCA was performed in order to evaluate 3 angles: 0, 45 and 90 ($\pm$ β=22.5). A $score_{tresh}$ of 0.75 was used in the analysis to select the oriented patterns. This value was experimentally chosen to allow one blank neighbor (non-marker) within a set of consecutive markers of a pattern. The range of the possible angles was limited to 0-90 since the analysis was performed in 2 planes perpendicular to each other. Grouping the angles 0/180 and 45/135 by symmetry was then required to keep relevance of the results. A representation of grouping is presented in Fig. 7.

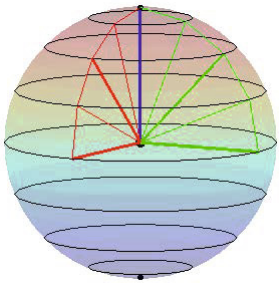


Fig. 7. Grouping of angles using PCA. Thick lines represents main angles and thin lines the deviation β

3.3 Correlations between Patterns Distribution and the Severity of the Disease

The OARSI grades of the samples were distributed from 0.89 to 6.25 (mean 3.51 $\pm$ 1.69). The intra-observer and inter-observer repeatability (CV_{rms}) of the OARSI grading were 8.78% and 11.84%, respectively. The intra-observer and inter-observer reliability (ICC) in the OARSI grading were 0.96% and 0.95%, respectively. According to previous literature [8][26], both ICC values represent an excellent reproducibility. Following this validation, the mean OARSI grade for each sample is considered as the ground truth for the stage of OA in the resulting LBP analysis.

For each sample, once the LBP method is applied to the stack of scans, the full histograms are converted to obtain the occurrences of each specific pattern as a percentage of all the patterns recognized within a volume of interest. Different parameters are assessed from the full histograms and then correlated with the OARSI grades:

- *The percentage of studied pixels*: ratio of relevant pixels used in the analysis by the total number of pixels available in the segmented trabecular bone.
- *The mean amount of markers*: corresponds to the mean value of markers for all the local patterns of the sample pooled together.

– *The amount of different patterns*: for 8 neighbors, the maximum amount of different patterns is 256. This parameter provides a count of all the local patterns recognized for a sample.
– *The entropy of local patterns*: describes the randomness of local patterns in the volume of interest. The entropy of local patterns was calculated as follows:

$$E = -\sum_i P_i log_2(P_i) \tag{5}$$

where P_i contains the count of a specific local pattern i occurring in the stack of scans. If an image contains only one local pattern, the entropy of the patterns within the image is zero.

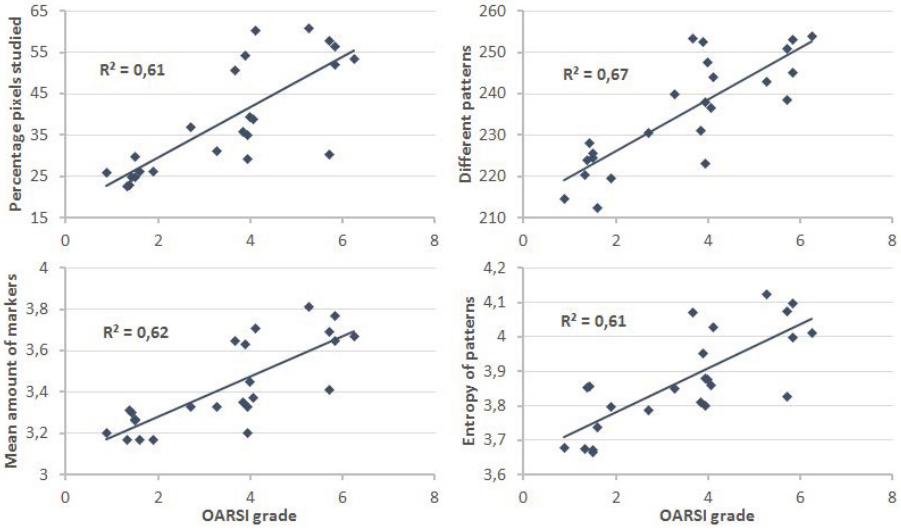


Fig. 8. Linear regression analysis between OARSI grades and trabecular bone parameters (N=24) derived from LBP analysis

Based on Fig. 8, the percentage of pixels studied is positively correlated to the severity of the disease, as suggested previously in literature [17][1]. Similarly to the results obtained in the radiographical study of Hirvasniemi *et al.* [13], the entropy of local patterns was proportional to the increase of OA level. An increase of the entropy of local patterns with OA corresponds to a higher variation in different patterns, supported by the increase of amount of different patterns, which could be explained by the appearance of bone sclerosis at higher OARSI grades [4][28].

The mean distribution of local patterns after the reduction of the histograms by grouping the patterns using PCA is shown in Fig. 9. Based on the results

obtained from the Pearson correlation analysis between each group and the OARSI grades, it can be seen that the occurrence of patterns with lower amount of markers (≤ 4) tends to disappear, while the occurrence of patterns with more markers (>4) increases. This result is also supported in Fig. 8 by the relation between the mean amount of markers and the OARSI grade. As an explanation, while the severity of OA increases, the trabecular bone tends to create more connections to improve its strength.

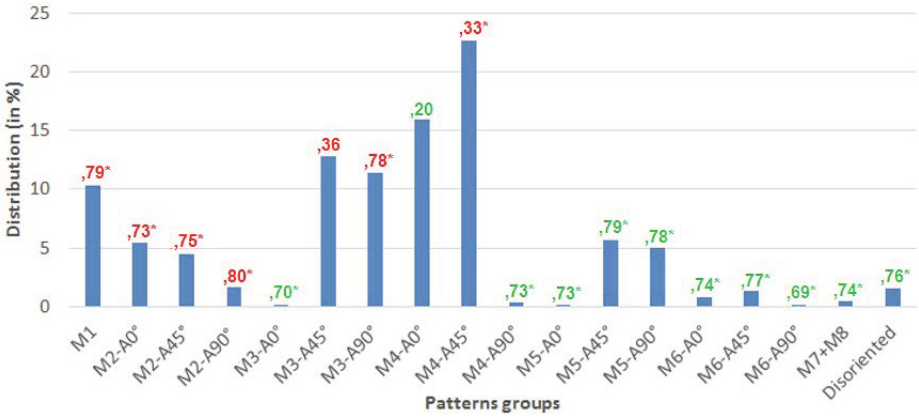


Fig. 9. Mean distribution of the pattern groups for all samples (n=24) pooled together. A group is defined by the amount of markers M and the main orientation A of its patterns. The group *Disoriented* corresponds to all the patterns without clear orientation, with markers M between 2 and 6. Correlation coefficients between OARSI grades and groups are red for a negative correlation and green for a positive one. $*p<0.001$

One interesting observation concerning the orientation of the patterns can be seen for groups with 3 markers. These groups are highly relevant since they are mainly representing straight edges of the fibers (as shown in Fig. 10).

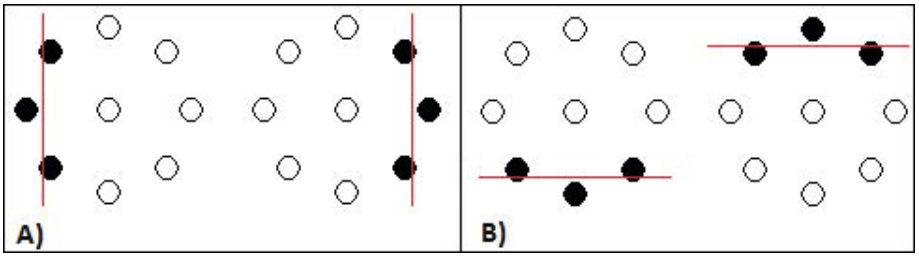


Fig. 10. Examples of patterns from two different groups with 3 markers. **A)** *M3-A90.* **B)** *M3-A0*

As expected, because the bone fibers are mostly oriented along the daily loading, the group *M3-A0* is smaller than the group *M3-A90*. However, for an increase in the severity of the disease, the group *M3-A90* decreases while the group *M3-A0* increases, suggesting the apparition of horizontal patterns connecting trabecular fibers between each other like arches. This hypothesis is furthermore supported by the increases of groups with more markers, suggesting the creation of corners. The non-significant decrease of the group *M3-45* can be explained as it represents the transition between horizontal and vertical patterns.

The hypothesized creation of bridges (Fig. 11) between the fibers suggests a reduction of the degree of anisotropy of the trabecular bone proportionally with the disease, as previously suggested [6]. This result is furthermore supported by the decrease of trabecular separation and structure model index, suggesting the trend of the trabecular structures to change from a rode-like type towards a plate-like shape [36][7].

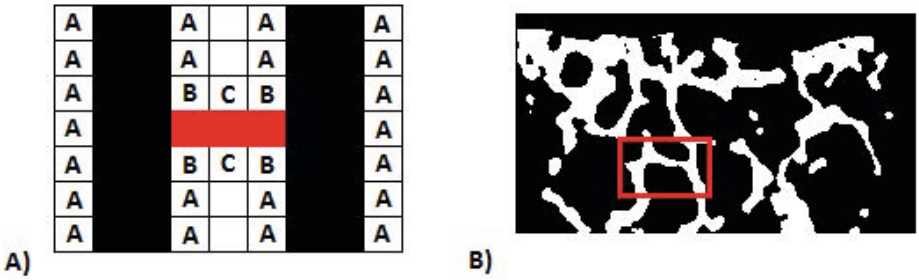


Fig. 11. A) Representation of the creation of a bridge in red between fibers in black. *A:M3-A90*, *B:M5-A45* and *C:M3-A0*. **B)** Example from a μCT slice with a bridge shown in the red rectangle

4 Conclusion

This study proposes a novel application of the local binary pattern method to perform bone structural analysis from μCT data. The experiment performed here suggests that this method can be used to assess the changes in the trabecular bone due to OA. While traditional bone structural analysis is affected by phenomena, such as partial volume effect or beam hardening, the LBP method is less subject to these issues due to its nature being based on the comparison of neighborhood intensity related to studied pixels, instead on the direct analysis of grey level values. The results obtained here are complementary to the traditional structural parameters and suggest that the assessment of other visual features could enhance the understanding of bone remodelling in OA. Further development of the present method should be performed, such as applying 3D LBP with real volumetric neighborhood analysis, using local ternary patterns method to improve the robustness, or using classifiers to estimate the severity of the disease.

Acknowledgments. This study was supported by the strategic funding of the University of Oulu, Infotech Oulu and the Academy of Finland.

References

1. Bennell, K., Creaby, M., Wrigley, T., Hunter, D.: Tibial subchondral trabecular volumetric bone density in medial knee joint osteoarthritis using peripheral quantitative computed tomography technology. Arthritis Rheum. **58**(9), 2776–2785 (2008)
2. Bobinac, D., Spanjol, J., Zoricic, S., Maric, I.: Changes in articular cartilage and subchondral bone histomorphometry in osteoarthritic knee joints in humans. Bone **32**(3), 284–290 (2003)
3. Bouxsein, M., Boyd, S., Christiansen, B., Guldberg, R., Jepsen, K., Müller, R.: Guidelines for assessment of bone microstructure in rodents using micro-computed tomography. J. Bone Miner Res. **25**(7), 1468–1486 (2010)
4. Buckwalter, J., Mankin, H.: Articular cartilage: degeneration and osteoarthritis and repair and regeneration and and transplantation. Instr. Course Lect. **47**, 487–504 (2012)
5. Burr, D., Gallant, M.: Bone remodelling in osteoarthritis. Nat. Rev. Rheumatol. **8**(11), 665–673 (2002)
6. Chappard, C., Peyrin, F., Bonnassie, A., Lemineur, G., Brunet-Imbault, B., Lespessailles, E., Benhamou, C.: Subchondral bone micro-architectural alterations in osteoarthritis: a synchrotron micro-computed tomography study. Osteoarthritis Cartilage **14**(3), 215–223 (2006)
7. Chiba, K., Ito, M., Osaki, M., Uetani, M., Shindo, H.: In vivo structural analysis of subchondral trabecular bone in osteoarthritis of the hip using multi-detector row ct. Osteoarthritis Cartilage **19**(2), 180–185 (2011)
8. Custers, R., Creemers, L., Verbout, A., VanRijen, M., Dhert, W., Saris, D.: Reliability and reproducibility and variability of the traditional histologic histochemical grading system vs the new oarsi osteoarthritis cartilage histopathology assessment system. Osteoarthritis Cartilage **15**(11), 1241–1248 (2007)
9. Ding, M., Danielsen, C., Hvid, I.: Effects of hyaluronan on three-dimensional microarchitecture of subchondral bone tissues in guinea pig primary osteoarthrosis. Bone **36**(3), 489–501 (2005)
10. Finnilä, M., Aho, O.M., Tiitu, V., Thevenot, J., Rautiainen, J., Nieminen, M., Valkealahti, M., Lehenkari, P., Saarakkala, S.: Correlation of subchondral bone morphometry and oarsi grade in osteoarthritic human knee samples. Osteoarthritis Cartilage **22**(S), 350–351 (2014)
11. Goldring, M., Goldring, S.: Articular cartilage and subchondral bone in the pathogenesis of osteoarthritis. Ann. N.Y. Acad. Sci. **1192**, 230–237 (2010)
12. Goldring, S., Goldring, M.: Bone and cartilage in osteoarthritis: is what's best for one good or bad for the other? Arthritis Res. Ther. **12**(5), 143 (2010)
13. Hirvasniemi, J., Thevenot, J., Immonen, V., Liikavainio, T., Pulkkinen, P., Jämsä, T., Arokoski, J., Saarakkala, S.: Quantification of differences in bone texture from plain radiographs in knees with and without osteoarthritis. Osteoarthritis Cartilage (2014). doi:10.1016/j.joca.2014.06.021
14. Houam, L., Hafiane, A., Boukrouche, A., Lespessailles, E., Jennane, R.: One dimensional local binary pattern for bone texture characterization. Pattern Anal. Appl. **17**(1), 1–15 (2012)

15. Intema, F., Hazewinkel, H., Gouwens, D., Bijlsma, J., Weinans, H., Lafeber, F., Mastbergen, S.: In early oa and thinning of the subchondral plate is directly related to cartilage damage: results from a canine aclt-meniscectomy model. Osteoarthritis Cartilage **18**(5), 691–698 (2010)
16. Intema, F., Sniekers, Y., Weinans, H., Vianen, M., Yocum, S., Zuurmond, A., DeGroot, J., Lafeber, F., Mastbergen, S.: Similarities and discrepancies in subchondral bone structure in two differently induced canine models of osteoarthritis. J. Bone Miner Res. **25**(7), 1650–1657 (2010)
17. Kamibayashi, L., Wyss, U., Cooke, T., Zee, B.: Trabecular microstructure in the medial condyle of the proximal tibia of patients with knee osteoarthritis. Bone **17**(1), 27–35 (1995)
18. Li, G., Yin, J., Gao, J., Cheng, T., Pavlos, N., Zhang, C., Zheng, M.: Subchondral bone in osteoarthritis: insight into risk factors and microstructural changes. Arthritis Res. Ther. **15**(6), 223 (2013)
19. Matsui, H., Shimizu, M., Tsuji, H.: Cartilage and subchondral bone interaction in osteoarthrosis of human knee joint: a histological and histomorphometric study. Microsc. Res. Tech. **37**(4), 333–342 (1997)
20. Mohan, G., Perilli, E., Kuliwaba, J., Humphries, J., Parkinson, I., Fazzalari, N.: Application of in vivo micro-computed tomography in the temporal characterisation of subchondral bone architecture in a rat model of low-dose monosodium iodoacetate-induced osteoarthritis. Arthritis Res. Ther. **13**(6), 210 (2011)
21. Nakashima, Y., Nakano, T.: Optimizing contrast agents with respect to reducing beam hardening in nonmedical x-ray computed tomography experiments. J. Xray Sci. Technol. **22**(1), 91–103 (2014)
22. Neda, S.C., Roman-Blas, J., Largo, R., Herrero-Beaumont, G.: Subchondral bone as a key target for osteoarthritis treatment. Biochem. Pharmacol. **83**(2), 315–323 (2012)
23. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on feature distributions. In: British Machine Vision Conference, vol. 25, pp. 51–59 (1996)
24. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (2002)
25. Otsu, N.: A treshold selection method from gray-level histograms. IEEE Transactions of Systems and Man and Cybernetics **9**(1) (1979)
26. Pearson, R., Kurien, T., Shu, K., Scammell, B.: Histopathology grading systems for characterisation of human knee osteoarthritis : reproducibility and variability and reliability and correlation and validity. Osteoarthritis Cartilage **19**(3), 324–331 (2011)
27. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer vision using local binary patterns. Springer (2011)
28. Pritzker, K., Gay, S., Jimenez, S., Ostergaard, K., Pelletier, J., Revell, P., Salter, D., Berg, W.V.: Osteoarthritis cartilage histopathology: grading and staging. Osteoarthritis Cartilage **14**(1), 13–29 (2006)
29. Qi, X., YuQiao, Y., Li, C., Guo, J.: Multi-scale joint encoding of local binary patterns for texture and material classification. In: British Machine Vision Conference (2013)
30. Seeman, E., Delmas, P.: Bone quality-the material and structural basis of bone strength and fragility. N. Engl. J. Med. **354**(21), 2250–2261 (2010)
31. Souza, A., Udupa, J., Saha, P.: Volume rendering in the presence of partial volume effects. IEEE Trans. Med. Imaging **24**(2), 223–235 (2005)

32. Thevenot, J., Hirvasniemi, J., Finnilä, M., Pulkkinen, P., Kuhn, V., Link, T., Eckstein, F., Jämsä, T., Saarakkala, S.: Trabecular homogeneity index derived from plain radiograph to evaluate bone quality. J. Bone Miner Res. **28**(12), 2584–2591 (2013)
33. Wang, T., Wen, C., Yan, C., Lu, W., Chiu, K.: Spatial and temporal changes of subchondral bone proceed to microscopic articular cartilage degeneration in guinea pigs with spontaneous osteoarthritis. Osteoarthritis Cartilage **21**(4), 574–581 (2013)
34. Wolff, J.: Das gesetz der transformation der knochen. The Law of Bone Remodelling (1892)
35. Woloszynski, T., Podsiadlo, P., Stachowiak, G., Kurzynski, M.: A signature dissimilarity measure for trabecular bone texture in knee radiographs. Med. Phys. **37**(5), 2030–2042 (2010)
36. Zhang, Z., Li, Z., Jiang, L., Jiang, S., Dai, L.: Micro-ct and mechanical evaluation of subchondral trabecular bone structure between postmenopausal women with osteoarthritis and osteoporosis. Osteoporos. Int. **21**(8), 1383–1390 (2010)

Impact of Topology-Related Attributes from Local Binary Patterns on Texture Classification

Thanh Phuong Nguyen[1(✉)], Antoine Manzanera[1], and Walter G. Kropatsch[2]

[1] ENSTA-ParisTech, 828 Bd des Maréchaux, 91762 Palaiseau Cedex, France
{thanh-phuong.nguyen,antoine.manzanera}@ensta-paristech.fr
[2] PRIP Group, 9/186-3 Favoritenstrasse, 1040 Wien, Austria
krw@prip.tuwien.ac.at

Abstract. A general texture description model is proposed, using topology related attributes calculated from Local Binary Patterns (LBP). The proposed framework extends and generalises existing LBP-based descriptors like LBP-rotation invariant uniform patterns (LBP^{riu2}), and Local Binary Count (LBC). Like them, it allows contrast and rotation invariant image description using more compact descriptors than classic LBP. However, its expressiveness, and then its discrimination capability, is higher, since it includes additional information, including the number of connected components. The impact of the different attributes on texture classification performance is assessed through a systematic comparative evaluation, performed on three texture datasets. The results validate the interest of the proposed approach, by showing that some combinations of attributes outperform state-of-the-art LBP-based texture descriptors.

Keywords: Local binary pattern · Local descriptor · Texture classification

1 Introduction

Texture recognition is a very active research topic in computer vision and pattern recognition. One of the most popular approaches for texture classification is based on feature distribution using Local Binary Pattern (LBP), introduced in [1]. Since the generalised work of Ojala et al. [2], LBP is widely considered as an efficient descriptor for capturing local properties of images. The decisive advantages of LBPs are their low computational cost and their invariance to monotonic changes of illumination. These good properties allow to successfully apply LBPs not only to texture recognition, but also to many other areas of computer vision.

In the wake of LBP's success, many authors have introduced variants of LBP descriptors [3] to improve the performance of classic LBP, or to better suit it to a

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 80–93, 2015.
DOI: 10.1007/978-3-319-16181-5_6

specific problem. Many different aspects have been considered. For preprocessing step, Gabor filters [4] have been used for capturing more global information. Different neighbourhoods, such as elliptical neighbourhood [5], three or four-patch approaches [6] have been employed to exploit anisotropic information. To address the issue of LBP instability on near constant image areas, the Local Ternary Patterns [7] use three values $\{-1, 0, 1\}$ in the encoding step. Multi-scale or multi-structure approaches [8,9] are considered to represent information at larger scales. Liao [10] chooses the most frequent patterns to improve the recognition accuracy. Guo et al. [11] use a complementary component related to magnitude to improve the texture classification.

In this paper, we propose a generic approach to improve the discrimination power of LBP by considering different geometrical and topological attributes extracted from LBPs. The proposed framework extends and generalises several existing LBP variants, and is also compatible (and then can be combined) with most of the other variants.

The remaining of the paper is organised as follows. The next section presents related works. Section 3 introduces the proposed framework, based on a family of rotation invariant attributes extracted from LBP. Section 4 is an evaluation of the descriptors derived from our models, compound with state-of-the-art descriptors, for the texture classification task applied on three classic datasets.

2 LBP and Its Rotation Invariant Forms

Local Binary Patterns [2] were introduced by Ojala et al. as a contrast invariant, binary version of the texture unit to represent its spatial structure. The binary pattern is formed by comparing a pixel value with its surrounding neighbours. The LBP encoding can be defined as follows:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) \cdot 2^p, s(x) = \begin{cases} 1, x \geq 0 \\ 0, \text{otherwise} \end{cases}$$

where g_c represents the gray value of the centre pixel and g_p $(0 \leq p < P)$ denotes the gray value of the neighbour pixel on a circle of radius R, and P is the total number of neighbours. The sample values can be calculated by interpolation.

The concept of circular neighbourhood allows to introduce the notions of uniform LBP, and also of rotation invariant LBP. A LBP is said uniform if the number of bit-transitions (0-1 and 1-0) in a circular scan of the pattern is at most 2. In texture description based on uniform LBP (denoted LBP^{u2}), non uniform LBPs are considered irrelevant, and then discarded or put in a single class. The rotation invariant LBP is defined as follows: $\mathrm{LBP}^{ri}_{P,R} = \min_{0 \leq i < P} \{ROR(\mathrm{LBP}_{P,R}, i)\}$, where $ROR(x, i)$ corresponds to the right circular bit-wise shift of i bits on P-bit number x. Very good texture classification results have been reported [2] using rotation invariant uniform patterns (LBP^{riu2}).

Zhao et al. [12] introduced Local Binary Count as a variant of LBP. It ignores the local binary structure of LBP by only counting the number of "1" in the pattern. Although they dramatically simplify the geometric structure, LBC features have been used with success for texture classification.

3 Core Texture Model

3.1 Topology Related LBP Attributes

The local descriptors used by our texture model embed and generalise several rotation invariant descriptors, including uniform patterns and Local Binary Count. They are based on a family of numerical attributes that are calculated on the original LBP. Consider the support of $\mathrm{LBP}_{P,R}$ as a set of P points on a circle, where 2 consecutive points are said adjacent (see Fig. 1). Topological information can then be extracted from the LBP using the connected components (circular runs) of 1s in the pattern. We will consider the following attributes:

- Number of connected components of 1s (#)
- Length of the largest run of 1s (M)
- Length of the smallest run of 1s (m)
- Total number of 1s (Σ)

All these attributes are rotation invariant. # is a topological measure, whose importance in the characterisation of shape is attested by a number of works in digital topology, in particular in the detection of critical points in thinning algorithms [13]. The uniform patterns correspond to # = 1 or 0. M and m can be seen as extensions of the uniform pattern values to non uniform patterns. Σ is equivalent to the Local Binary Count. Figure 1 illustrates a non-uniform binary pattern (10111010) of 8 bits; with # = 3, M = 3, m = 1 and Σ = 5.

These attributes are not independent; all configurations of values are not possible and must respect the following constraints:

1. $\mathrm{m} \leq \mathrm{M} \leq \Sigma$
2. $0 \leq \# \leq \lfloor P/2 \rfloor$
3. if $\# = 0$, $\mathrm{m} = \mathrm{M} = \Sigma = 0$
4. if $\# = 1$, $1 \leq \mathrm{m} = \mathrm{M} = \Sigma \leq P$
5. if $\# > 1$, $1 \leq \mathrm{M} \leq P$ - 2# + 1
6. if $\# > 1$, $1 \leq \mathrm{m} \leq \lfloor P/\# \rfloor$ - 1
7. if $\# > 1$, $\# \leq \Sigma \leq P$ - #

These properties imply that for a combination of two or more attributes, the number of different configurations is relatively small compared to 2^P (see also Table 1).

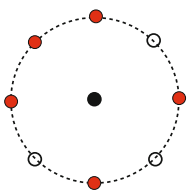


Fig. 1. A non-uniform pattern where 1 (resp. 0) is represented by red filled circle (resp. black circle)

3.2 Texture Modelling

The purpose of this work is to evaluate the contribution of the different attributes in texture description. Every version of the descriptor used in the experiments is then related to a vector of s attributes $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_s)$.

Basically, we describe a texture by computing, for each pixel, the LBP and its s attributes, and then by calculating, for the whole image, the joint histogram of the s attributes. The number of different attribute vectors depends on P, and on the chosen subset of attributes. In general, it is much smaller than 2^P, the number of different LBPs.

In practice, to reduce the size of the histogram and the computation time of the descriptor, we associate a unique label to every value of the attribute vector, and pre-compute the label of all the LBP values in a label table Λ.

To do this, for every subset of s attributes, we create an s-dimensional array T initialized with zeros, and a scalar counter c initialized to zero. Then we enumerate all the LBP values n from 0 to $2^P - 1$, and calculate the vector attribute $\mathcal{A}(n)$. If $T(\mathcal{A}(n))$ is equal to zero, we increment c, and set $T(\mathcal{A}(n)) = c$. In all cases, we set the label table $\Lambda(n) = T(\mathcal{A}(n))$. The final value of the counter is denoted $N_{\mathcal{A}}$, the number of distinct vectors of attributes.

Finally we represent a texture by a histogram of labels:

$$H(l) = |\{\mathbf{p}; \Lambda\left(\mathrm{LBP}_{P,R}(\mathbf{p})\right) = l\}|$$

In the experiments, we shall denote the texture descriptor based on the subset of attributes $\mathcal{A}$ as $\mathrm{LBP}^{\mathcal{A}}_{P,R}$, following the conventional notations $u2$ or $riu2$ in LBP based models. Figure 2 shows a texture image with its corresponding label images and label histograms for the different configurations of $\mathrm{LBP}^{\mathcal{A}}_{1,8}$. In addition, Figure 3 shows images and histograms of labels corresponding to $\mathrm{LBP}^{\#Mm}_{1,8}$ for different images, from the same texture class (first row), and from different classes (second row). The visual (di)similarity of histograms depending on the class is apparent on the figure.

To assess the interest of differentiating uniform patterns or not, a mixed texture representation ($\mathrm{LBP}^{riu2+\mathcal{A}}_{P,R}$) is also evaluated in our work, by taking into account the above encoding only on non-uniform patterns, and using $riu2$ encoding for uniform patterns:

$$\mathrm{LBP}^{riu2+\mathcal{A}}_{P,R}(\mathbf{p}) = \begin{cases} \mathrm{LBP}^{riu2}_{P,R}(\mathbf{p}), \text{ if } \mathrm{LBP}_{P,R}(\mathbf{p}) \text{ is uniform} \\ p + 1 + \Lambda\left(\mathrm{LBP}_{P,R}(\mathbf{p})\right), \text{ otherwise} \end{cases}$$

3.3 Relation with Previous Works

As mentioned before, $\mathrm{LBP}^{\mathcal{A}}_{P,R}$ are related with other rotation invariant patterns: $\mathrm{LBP}^{riu2}_{P,R}$ [2] and LBC [12]. We now discuss further those relations.

- $\text{LBP}^{\Sigma}_{P,R}$ *is exactly* LBC. It means that if $\Sigma \in \mathcal{A}$, $\text{LBP}^{\mathcal{A}}_{P,R}$ is a generalisation of LBC.
- *When* $\text{card}(\mathcal{A}) \geq 2$ *and* $(\# \in \mathcal{A}$ *or* $\Sigma \in \mathcal{A})$, $\text{LBP}^{\mathcal{A}}_{P,R}$ *is a superset of* $\text{LBP}^{riu2}_{P,R}$ *patterns.* In that case indeed, *riu*2 patterns are distinguished, either by the value of # and anyone among $\{\text{M}, \text{m}, \Sigma\}$, or by one of the identity $\text{M} = \Sigma$ or $\text{m} = \Sigma$.[1] Therefore, for such combination of attributes $\mathcal{A}$, $LBP^{\mathcal{A}}_{P,R}$ inherits the distinctive properties of $LBP^{riu2}_{P,R}$, while containing more information. In this sense, $LBP^{\mathcal{A}}_{P,R}$ generalises $\text{LBP}^{riu2}_{P,R}$.
- As a consequence, with the same conditions on $\mathcal{A}$, *the performance of* $\text{LBP}^{\mathcal{A}}_{P,R}$ *and* $\text{LBP}^{riu2+\mathcal{A}}_{P,R}$ *are the same.*
- *When* $\text{card}(\mathcal{A}) = 1$ *or* $\mathcal{A} = \{\text{M}, \text{m}\}$, $\mathcal{A}$ and *riu*2 are complementary, and $\text{LBP}^{riu2+\mathcal{A}}_{P,R}$ *can be better than* $\text{LBP}^{\mathcal{A}}_{P,R}$ *or* $\text{LBP}^{riu2}_{P,R}$ *alone.*

Table 1 displays the number of labels (and then of histogram bins) for the different configurations of attributes. Note that the numbers for $\text{LBP}^{\mathcal{A}}_{P,R}$ and $\text{LBP}^{riu2+\mathcal{A}}_{P,R}$ are different only if $\text{card}(\mathcal{A}) = 1$ or $\mathcal{A} = \{\text{M}, \text{m}\}$. In addition, Table 2 shows the number of labels for several existing LBP-based methods.

Table 1. Number of different labels, i.e. number of histogram bins of the texture descriptor in the different configurations

Schema	#	M	m	Σ	M#	m#	MΣ	Mm	#Σ	mΣ	Mm#	#MΣ	MmΣ	#mΣ	#MmΣ
$\text{LBP}^{\mathcal{A}}_{8,1}$	5	9	9	9	18	14	21	15	18	18	22	23	23	22	23
$\text{LBP}^{\mathcal{A}}_{16,2}$	9	17	17	17	66	36	92	59	66	66	125	180	159	125	212
$\text{LBP}^{\mathcal{A}}_{24,3}$	13	25	25	25	146	62	225	135	146	146	353	680	557	353	989
$\text{LBP}^{riu2+\mathcal{A}}_{8,1}$	12	14	14	14	18	14	21	18	18	18	22	23	23	22	23
$\text{LBP}^{riu2+\mathcal{A}}_{16,2}$	24	30	30	30	66	36	92	66	66	66	125	180	159	125	212
$\text{LBP}^{riu2+\mathcal{A}}_{24,3}$	36	46	46	46	146	62	225	146	146	146	353	680	557	353	989

Table 2. Number of different labels in several encodings

Method	(P,R)=(8,1)	(P,R)=(16,2)	(P,R)=(24,3)
$\text{LBP}^{riu2}_{P,R}$	10	18	26
$\text{LBP}^{u2}_{P,R}$	59	243	555
$\text{CLBP}^{riu2}_{P,R}$	200	648	1352

There is a strong link between $\text{LBP}^{riu2+\mathcal{A}}_{P,R}$ with previous works aiming at exploiting information from non-uniform patterns to improve the texture descriptors. In particular $\text{LBP}^{riu2+\{\Sigma\}}_{P,R}$ and $\text{LBP}^{riu2+\{\#\}}_{P,R}$ are close to [14]. In this work, the authors extended the notion of uniform pattern (as the Σ attribute does), and the other patterns were encoded by the number of 0-1 transitions, which corresponds to the # attribute.

[1] Note that $\{\text{M}, \text{m}\}$ alone do not allow to distinguish uniform patterns, since the identity $\text{M} = \text{m}$ can occur with several connected components.

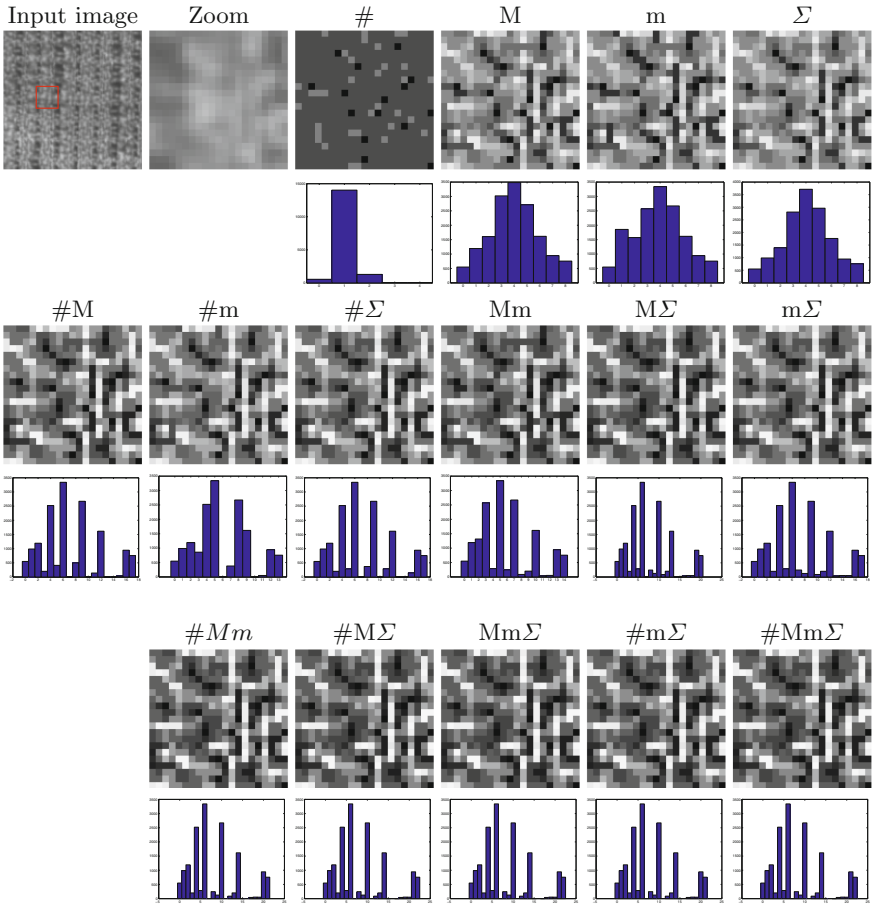


Fig. 2. A texture image and its label images and label histograms for the different configurations of attributes, with $(P, R) = (8, 1)$. For the best visualization, the label images are zoomed from a part corresponding to the red square of the texture image.

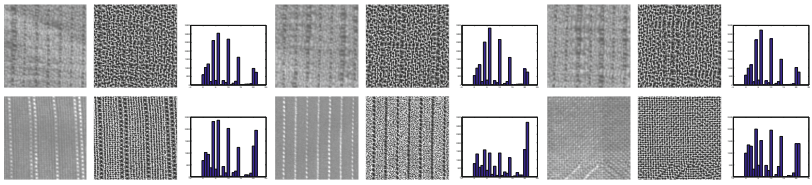


Fig. 3. Texture images and their label images and histograms for $\mathrm{LBP}^{\#Mm}_{8,1}$. The first row contains images of the same class, the second row images of different classes.

3.4 Completed Texture Descriptor

Guo et al [11] proposed a state-of-the-art variant of LBP by coding the local differences as two complementary components: signs ($s_p = s(g_p - g_c)$) and magnitudes ($m_p = |g_p - g_c|$). They proposed to use two binary patterns, called CLBP-Sign (CLBP_S) and CLBP-Magnitude (CLBP_M). The first pattern is identical to the LBP. The second one which measures the local variance of magnitude is defined as follows:

$$\text{CLBP\_M}_{P,R} = \sum_{p=0}^{P-1} s(m_p - \tilde{m}).2^p,$$

where $\tilde{m}$ is the mean value of m_p for the whole image. In addition, Guo et al. observed that the local value itself carries important information. Therefore, they defined the operator CLBP-Center (CLBP_C) as follows:

$$\text{CLBP\_C} = s(g_c - \tilde{g}),$$

where $\tilde{g}$ is the mean gray level for the whole image. Because these operators are complementary, their combination leads to a significant improvement, and then CLBP is now considered a reference method in texture classification.

Inspired from this work, we also evaluated our descriptors by complementing the difference sign information (CLBP_S) by the difference magnitude (CLBP_M) and gray level (CLBP_C). For CLBP_S and CLBP_M, instead of using $riu2$ mapping, we apply our proposed encoding to obtain $\text{CLBP}^{\mathcal{A}}_{P,R}$ and $\text{CLBP\_M}^{\mathcal{A}}_{P,R}$. Finally, the feature vector of the whole image is constructed by considering the joint histograms of $\text{CLBP\_S}^{\mathcal{A}}_{P,R}$, $\text{CLBP\_M}^{\mathcal{A}}_{P,R}$ and CLBP_C. Then, if $\text{LBP}^{\mathcal{A}}_{P,R}$ has n different labels, $\text{CLBP}^{\mathcal{A}}_{P,R}$ has $2n^2$ labels (see also Table 1).

3.5 Texture Classification

Because the contribution of this work is focused on texture descriptors, and the competing LBP based methods all used χ^2 distance as similarity metrics [2], and nearest neighbour as classification criterion, we used the same classification method for fair comparison purposes. If H_1 and H_2 are two attribute label histograms, the χ^2-dissimilarity between the two textures is:

$$\chi^2(H_1, H_2) = \sum_{i=1}^{N} \frac{(H_1(i) - H_2(i))^2}{H_1(i) + H_2(i)},$$

with $N = N_{\mathcal{A}}$ or $N = N_{riu2+\mathcal{A}}$ the number of labels.

4 Experiments

4.1 Datasets

The effectiveness of the proposed method is assessed by a series of experiments on three large and representative databases: Outex [15], CUReT [16] and UIUC [17].

The Outex database (examples are shown in Figure 6) contains images captured from a wide variety of real materials. We consider the two commonly used test suites, Outex_TC_00010 (TC10) and Outex_TC_00012 (TC12), containing 24 classes of textures. Each image may be seen under nine different rotation angles between 0 and 90°. For TC10, The texture images at angle 0° are chosen for training the classifier, all the remaining images are used for testing. For TC12, aside from the different viewing angles, the images can have three types of illumination: "inca", used for learning, and "t184" or "horizon", used for testing (test sets are denoted TC12t and TC12h respectively).

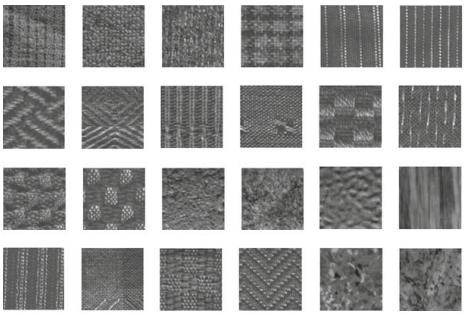
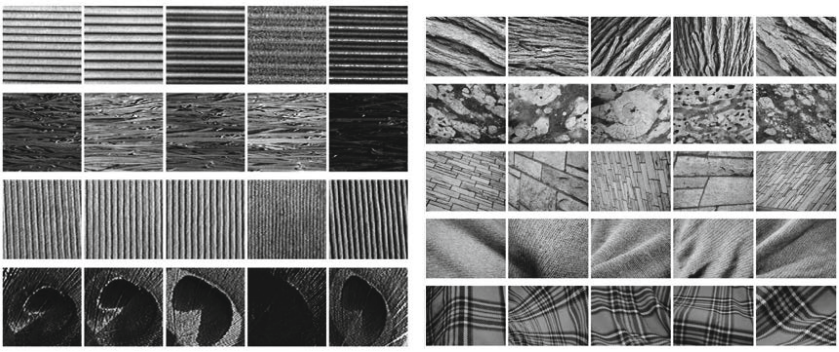


Fig. 4. Texture images with the illumination condition "inca" and zero degree rotation angle from the 24 classes of textures on the Outex database

The CUReT database contains 61 texture classes (see Figure 5.a), each having 205 images acquired at different viewpoints and illumination orientations. We follow the experimental protocol proposed in [18,19], using 4 different learning sets made of 6, 12, 23 and 46 images (first line of Table 7).



(a) CUReT dataset (b) UIUC dataset

Fig. 5. Examples of texture images

The UIUC texture database includes 25 classes with 40 images in each class. The resolution of each image is 640×480. The database contains materials imaged under significant viewpoint variations (examples are shown in Figure 5.b). Following [17], to eliminate the dependence of the results on the particular training images used, four different learning sets of 5, 10, 15 and 20 images are used while the remaining images per class are used as test set.

In the upcoming result sections, the performance measure will be given in percentage of correct classification. As the typical size of the test sets is around 5 000, the percentage values are rounded to the first decimal. Furthermore, our methods is practically deterministic (ignoring the slight influence of interpolation in the computation of the LBP). Finally, the test protocol of the Outex dataset is also deterministic, and the typical observed standard deviation in the cross validation schemes of Curet and UIUC is less than 0.1%.

4.2 Computational Cost

We consider in this section the computational cost of our descriptors with respect to other LBP-based operators. Experiments on Outex TC10 test suite containing 4320 images of 128 × 128 pixels were performed on a machine with 2.0 GHz CPU, 4Go RAM and Linux 3.2.0-23 kernel. Table 3 presents the computation time (in seconds) of different descriptors in the configuration $(P, R) = (2, 16)$ and reports the total time (in seconds) for classifying the 3840 test images against the 480 reference images.

Table 3. Complexity of our different descriptors with respect to LBP^{riu2} (FET: Feature Extraction Time, MT: Matching Time)

Method	$riu2$	#	M	m	Σ	M#	m#	MΣ	Mm	#Σ	mΣ	Mm#	#MΣ	MmΣ	#mΣ	#MmΣ
FET	78.1	79.2	78.4	78.3	80.2	80.9	80.8	79.2	78.7	79.5	78.9	80.3	80.6	83.3	83.3	82.2
MT	1.2	0.9	1.1	1.1	1.2	4.7	2.1	4.7	2.7	4.5	3.2	6.9	11.4	10.1	6.5	13.1

As can be seen from Table 3, the descriptor construction time does not vary much from one method to the other, while the classification time is proportional to the length of the feature vector.

4.3 1st Experiment: $\mathrm{LBP}^{\mathcal{A}}_{P,R}$ and $\mathrm{LBP}^{riu2+\mathcal{A}}_{P,R}$

Table 4 compares our descriptors ($\mathrm{LBP}^{\mathcal{A}}_{P,R}$) with the classic $\mathrm{LBP}^{riu2}_{P,R}$ on Outex dataset, for different (P, R) configurations. Those results can be interpreted as follows:

- The four attributes have distinct properties. Considered alone, their performance is comparable to $\mathrm{LBP}^{riu2}_{P,R}$, except for #, whose expressiveness is too weak if taken alone.
- Jointly considering 2 attributes, the results are (except in one case) better than $\mathrm{LBP}^{riu2}_{P,R}$, with an average improvement which can reach 6%.

Table 4. Comparison between LBP^{riu2} and the basic $\text{LBP}^{\mathcal{A}}$ on Outex dataset

Method	(P,R)=(8,1)				(P,R)=(16,2)				(P,R)=(24,3)			
	TC10	TC12 t	TC12 h	Mean	TC10	TC12 t	TC12 h	Mean	TC10	TC12 t	TC12 h	Mean
LBP^{riu2} [2]	84.8	65.5	63.7	71.3	89.4	82.3	75.2	82.3	95.1	85.0	80.8	87.0
$\text{LBP}^{\#}$	52.4	37.4	32.0	40.6	66.5	51.3	47.1	55.0	77.0	67.5	58.2	67.6
Gain	-	-	-	-	-	-	-	-	-	-	-	-
LBP^{M}	83.6	67.1	64.4	71.7	87.6	82.2	78.9	82.9	95.9	88.1	86.4	90.1
Gain	-	1.6	0.7	0.4	-	-	3.7	0.6	0.8	3.1	5.6	3.1
LBP^{m}	84.8	64.5	62.2	69.9	91.0	82.9	77.0	83.7	96.5	86.2	80.4	87.7
Gain	0.0	-	-	-	1.6	0.6	1.8	1.4	1.4	1.2	-	0.7
LBP^{Σ}	82.9	65.0	63.2	70.4	88.7	82.6	77.4	82.9	91.3	83.8	82.7	86.0
Gain	-	-	-	-	-	0.3	2.2	0.6	-	-	1.9	-
$\text{LBP}^{M\#}$	86.5	69.9	66.2	74.2	93.7	85.3	82.1	87.1	96.8	88.7	84.3	89.9
Gain	1.7	4.4	2.5	2.9	4.3	3.0	6.9	4.8	1.7	3.7	3.5	2.9
$\text{LBP}^{m\#}$	85.7	67.4	66.4	73.1	93.1	85.9	81.4	86.8	97.5	89.3	85.0	90.6
Gain	0.9	1.9	2.7	1.8	3.7	3.6	6.2	4.5	2.4	4.3	4.2	3.6
$\text{LBP}^{M\Sigma}$	85.8	69.7	66.6	74.0	92.5	85.9	82.3	86.9	96.9	89.9	86.0	91.0
Gain	1.0	4.2	2.9	2.7	3.1	3.6	7.1	4.6	1.8	4.9	5.2	4.0
LBP^{Mm}	85.6	66.8	63.6	72.0	92.5	85.4	82.4	86.8	**98.1**	**92.2**	**87.2**	**92.5**
Gain	0.8	1.3	-	0.7	3.1	3.1	7.2	4.5	**3.0**	**7.2**	**6.4**	**5.5**
$\text{LBP}^{\#\Sigma}$	**87.1**	69.8	67.8	74.9	93.4	84.6	79.7	85.9	96.5	87.5	83.6	89.2
Gain	**2.3**	4.3	4.1	3.6	4	2.3	4.5	3.6	1.4	2.5	2.8	2.2
$\text{LBP}^{m\Sigma}$	86.0	70.1	66.8	74.3	92.9	85.8	83.4	87.4	97.8	91.4	86.8	92.0
Gain	1.2	4.6	3.1	3.0	3.5	3.6	8.2	5.1	2.7	6.4	6.0	5.0
$\text{LBP}^{Mm\#}$	85.8	70.5	**68.2**	74.8	**94.3**	86.8	84.2	88.4	97.2	90.9	86.7	91.6
Gain	1.0	5.0	**4.5**	3.5	**4.9**	3.5	9.0	6.1	2.1	5.9	5.9	4.6
$\text{LBP}^{\#M\Sigma}$	86.0	70.6	67.9	74.8	93.7	87.0	84.3	88.3	97.2	90.4	86.2	91.3
Gain	1.2	5.1	4.2	3.5	4.3	4.7	9.1	6.0	2.1	5.4	5.4	4.3
$\text{LBP}^{Mm\Sigma}$	86.0	70.6	67.9	74.8	94.1	87.3	84.1	88.5	97.0	90.3	86.4	91.2
Gain	1.2	5.1	4.2	3.5	4.7	5.0	8.9	6.2	1.9	5.3	5.6	4.2
$\text{LBP}^{\#m\Sigma}$	86.1	**70.8**	67.8	**74.9**	94.1	87.0	84.1	88.4	97.4	91.0	86.7	91.7
Gain	1.3	**5.3**	4.1	**3.6**	4.7	4.7	8.9	6.1	2.3	6.0	5.9	4.7
$\text{LBP}^{\#Mm\Sigma}$	86.0	70.6	67.9	74.8	94.1	**87.6**	**84.5**	**88.7**	97.1	90.2	86.4	91.2
Gain	1.2	5.1	4.2	3.5	4.7	**5.3**	**9.3**	**6.4**	2.0	5.2	5.6	4.2

– Using 3 or 4 attributes further improves the results, except when $P = 24$. This can be explained by the fact that in this case, the number of labels is too high, which makes the histogram too sparse for the χ^2 distance.

In addition, Table 5 presents a comparison between $\text{LBP}^{riu2+\mathcal{A}}_{P,R}$ and $\text{LBP}^{riu2}_{P,R}$, when $\mathcal{A}$ is mono attribute or $\{Mm\}$. It can be seen that the performance of $\text{LBP}^{riu2+\mathcal{A}}_{P,R}$ is, in most cases, better than $\text{LBP}^{\mathcal{A}}_{P,R}$ or $\text{LBP}^{riu2}_{P,R}$ alone.

4.4 2nd Experiment: $\text{CLBP}^{\mathcal{A}}_{P,R}$

Because, when the number of labels become too large ($P = 24$), the use of several attributes become really inefficient due to the very high dimension of feature vectors, in this experiment we consider only a combination of at most two attributes.

Table 6, 7, 8 present the results obtained by our methods $\text{CLBP}^{\mathcal{A}}_{P,R}$ on the three datasets (Outex, CUReT and UIUC) in comparison with other LBP-based methods. From these tables, we can make the following remarks

Table 5. Comparison between LBP^{riu2}, $\text{LBP}^{\mathcal{A}}$ and the mixed $\text{LBP}^{riu2+\mathcal{A}}$ on Outex dataset when $\mathcal{A}$ is a mono attribute or $\{Mm\}$

Method	(P,R)=(8,1)				(P,R)=(16,2)				(P,R)=(24,3)			
	TC10	TC12 t	TC12 h	Mean	TC10	TC12 t	TC12 h	Mean	TC10	TC12 t	TC12 h	Mean
LBP^{riu2} [2]	84.8	65.5	63.7	71.3	89.4	82.3	75.2	82.3	95.1	85.0	80.8	87.0
$\text{LBP}^{\#}$	52.4	37.4	32.0	40.6	66.5	51.3	47.1	55.0	77.0	67.5	58.2	67.6
$\text{LBP}^{riu2+\#}$	85.3	66.6	65.7	72.5	91.5	83.5	78.2	84.4	96.1	86.3	81.6	88.0
Gain_{riu2}	0.5	1.1	2.0	1.3	2.1	1.2	3.0	2.1	1.0	1.3	0.8	1.0
$\text{Gain}_{\#}$	32.9	29.2	33.7	31.9	25.0	32.2	31.1	29.47	19.17	18.80	23.37	20.44
LBP^{M}	83.6	67.1	64.4	71.7	87.6	82.2	78.9	82.9	95.9	88.1	86.4	90.1
LBP^{riu2+M}	**86.3**	68.5	64.6	73.2	91.0	84.5	79.9	85.1	96.7	88.1	84.2	89.8
Gain_{riu2}	1.5	3.0	0.9	1.9	1.6	2.2	4.7	2.8	1.6	3.1	3.4	2.8
Gain_{M}	2.7	1.4	0.2	1.5	3.4	2.3	1.0	2.2	0.8	-	-	-
LBP^{m}	84.8	64.5	62.2	69.9	91.0	82.9	77.0	83.7	96.5	86.2	80.4	87.7
LBP^{riu2+m}	85.1	66.7	65.3	72.4	92.2	85.1	80.3	85.9	97.5	89.1	84.6	90.0
Gain_{riu2}	0.3	1.2	1.6	1.1	2.8	2.8	5.1	3.6	2.4	4.1	3.8	3.0
Gain_{m}	0.3	2.2	3.1	2.5	1.2	2.2	3.3	2.2	1.0	2.9	4.2	2.3
LBP^{Σ}	82.9	65.0	63.2	70.4	88.7	82.6	77.4	82.9	91.3	83.8	82.7	86.0
$\text{LBP}^{riu2+\Sigma}$	86.1	68.1	65.9	**73.4**	90.1	83.7	77.1	83.6	96.0	86.3	82.8	88.4
Gain_{riu2}	1.3	2.6	2.2	2.1	0.7	1.4	1.9	1.3	0.9	1.3	2.0	1.4
Gain_{Σ}	3.2	3.1	2.7	3.0	1.4	1.1	-	0.7	4.7	2.5	0.1	2.4
LBP^{Mm}	85.6	66.8	63.6	72.0	92.5	85.4	82.4	86.8	98.1	**92.2**	**87.2**	**92.5**
$\text{LBP}^{riu2+Mm}$	85.9	**69.8**	**66.8**	74.2	**93.0**	**85.5**	**82.8**	**87.1**	**98.2**	91.8	87.1	92.4
Gain_{riu2}	1.1	4.3	1.1	2.9	3.6	3.2	7.6	4.8	3.1	6.8	6.3	5.4
Gain_{Mm}	0.3	3.0	3.2	2.2	0.5	0.1	0.4	0.3	0.1	-	-	-

Table 6. Comparison between $\text{CLBP}^{\mathcal{A}}_{P,R}$ and other LBP-based methods on Outex dataset

Method	(P,R)=(8,1)				(P,R)=(16,2)				(P,R)=(24,3)			
	TC10	TC12 t	TC12 h	Mean	TC10	TC12 t	TC12 h	Mean	TC10	TC12 t	TC12 h	Mean
LBP^{riu2} [2]	84.8	65.5	63.7	71.3	89.4	82.3	75.2	82.3	95.1	85.0	80.8	87.0
LTP [7]	94.1	75.9	74.0	81.3	97.0	90.2	86.9	91.3	98.2	93.6	89.4	93.8
DLBP [10]					97.7	92.1	88.7	92.8	98.1	91.6	87.4	92.4
DLBP + NGF [10]					99.1	93.2	90.4	94.2	98.2	91.6	87.4	92.4
CLBP_S_M/C [11]	94.5	81.9	82.5	86.3	98.0	91.0	91.1	93.4	98.3	94.0	92.4	94.9
CLBP_S/M [11]	94.7	82.7	83.1	86.8	97.9	90.5	91.1	93.2	99.3	93.6	93.3	95.4
CLBP_S/M/C [11]	96.6	90.3	92.3	93.0	98.7	93.5	93.9	95.4	98.9	95.3	94.5	96.3
Our proposed descriptors												
$\text{CLBP}^{\#}$	85.4	71.7	70.7	75.9	90.6	83.4	80.9	85.0	89.0	81.3	79.3	83.2
CLBP^{M}	96.5	90.9	93.0	93.4	98.4	95.5	96.2	96.7	99.1	97.2	96.8	97.7
CLBP^{m}	96.7	90.5	91.6	92.9	99.0	95.6	94.9	90.5	99.5	96.7	96.0	97.4
CLBP^{Σ}	**97.2**	89.8	92.9	93.3	98.5	93.3	94.1	95.3	98.8	94.0	95.4	96.1
$\text{CLBP}^{M\#}$	96.2	90.6	93.5	93.5	98.9	95.5	95.8	96.8	99.5	**97.0**	**96.3**	**97.6**
$\text{CLBP}^{m\#}$	96.3	90.4	92.2	93.0	98.9	95.3	95.1	96.4	99.3	96.4	96.0	97.2
$\text{CLBP}^{M\Sigma}$	96.3	**91.1**	93.4	93.6	98.9	95.3	**96.5**	**96.9**	99.4	94.3	93.4	95.7
CLBP^{Mm}	96.8	90.9	93.2	**93.6**	98.9	**95.7**	95.2	96.6	99.3	95.5	93.8	96.2
$\text{CLBP}^{\#\Sigma}$	96.7	90.7	**93.6**	93.6	**99.0**	95.4	96.1	96.8	**99.6**	96.2	95.8	97.2
$\text{CLBP}^{m\Sigma}$	96.5	90.6	92.3	93.1	**99.0**	95.23	96.1	96.8	99.1	94.6	93.0	95.5

- Except when $\mathcal{A} = \{\#\}$, $\text{CLBP}^{\mathcal{A}}_{P,R}$ outperforms the previous methods in all configurations.
- Between mono attributes, M is the best configuration. It means that CLBP^{M} outperforms CLBC [12] that is exactly CLBP^{Σ}.

Table 7. Experimentation of $CLBP^{\mathcal{A}}_{P,R}$ on CUReT dataset[2].

Method	(P,R)=(8,1)				(P,R)=(16,3)				(P,R)=(24,5)			
	N=46	N=23	N=12	N=6	N=46	N=23	N=12	N=6	N=46	N=23	N=12	N=6
LTP [7]	85.13	79.25	72.25	63.09	92.66	87.30	80.22	70.50	91.81	85.78	77.88	67.77
$LBP^{riu2}/VAR_{P,R}$ [21]	93.87	88.76	81.59	71.03	94.20	89.12	81.64	71.81	91.87	85.58	77.13	66.04
CLBP S/M/C [11]	95.6	91.3	84.9	74.8	95.9	92.1	86.1	77.0	94.7	90.3	83.8	74.5
CLBP S/M[11]	93.5	88.7	81.9	72.3	94.4	90.4	84.2	75.4	93.6	89.1	82.5	73.3
Our proposed descriptors												
$CLBP^{\#}$	81.8	72.2	61.8	52.0	82.3	72.8	61.6	50.2	77.7	67.6	56.5	45.2
$CLBP^{M}$	95.7	91.5	84.1	74.0	96.1	92.5	85.7	76.3	96.4	92.3	85.9	77.6
$CLBP^{m}$	95.8	91.3	83.8	73.7	96.8	92.5	85.8	77.1	95.4	91.6	85.4	77.8
$CLBP^{\Sigma}$	94.8	90.1	82.7	72.1	94.7	89.8	82.3	72.0	93.9	87.5	80.6	69.0
$CLBP^{m\#}$	95.9	91.7	84.4	74.5	97.0	**93.2**	86.7	**78.2**	**96.1**	**92.6**	**85.9**	**78.6**
$CLBP^{M\Sigma}$	**96.3**	**92.1**	84.8	74.8	96.0	92.0	85.7	76.5	x	x	x	x
$CLBP^{Mm}$	96.2	91.9	84.7	74.6	96.7	93.1	86.7	78.1	94.5	90.45	83.8	77.0
$CLBP^{\#\Sigma}$	96.2	91.8	84.6	74.8	96.5	92.7	86.2	77.1	93.7	89.0	81.5	73.8
$CLBP^{m\Sigma}$	**96.3**	91.9	**84.9**	74.4	96.7	92.5	86.7	77.1	92.27	88.1	80.6	73.7
$CLBP^{M\#}$	**96.3**	**92.1**	84.7	**75.1**	**96.8**	**93.2**	**87.3**	**78.2**	95.01	91.2	84.24	77.1

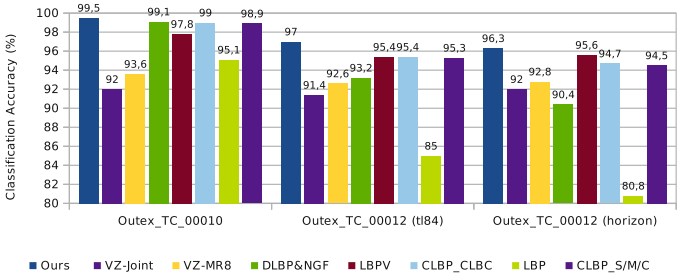


Fig. 6. Comparing the best results of $CLBP^{M\#}$ with the best results of recent methods on Outex dataset

– The combination of two attributes still improves the performance of our descriptors. The improvement in relation with CLBP_S/M/C varies from 0.5% to 4.5% depending on test configurations.

In addition, Figure 6 presents the best results of one configuration ($M\#$) in comparison with the best results of recent methods on Outex dataset: LBP^{riu2} [2], LTP [7], DLBP + NGF [10], VZ-MR8 [20], VZ-Joint [20], CLBP_S_M/C [11] and CLBP_CLBC [12]. As can be seen from it, because the topology-related attributes bring more information than the typical mapping $riu2$, our best results in complementary scheme improve significantly with respect to CLBP.

[2] $CLBP^{M\Sigma}_{3,24}$ is not tested on this dataset.

Table 8. Experimentation of $\text{CLBP}^{\mathcal{A}}_{P,R}$ on UIUC dataset

Method	(P,R)=(8,1)				(P,R)=(16,2)				(P,R)=(24,3)			
	20	15	10	5	20	15	10	5	20	15	10	5
LBP^{riu2}	54.6	52.9	47.1	39.7	61.3	56.4	51.2	42.7	64.0	60.0	54.2	44.6
CLBP_S/M [19]	81.8	78.5	74.8	64.8	87.9	85.1	80.6	71.6	89.2	87.4	81.9	72.5
CLBP_S/M/C [19]	87.6	85.7	82.6	75.0	91.0	89.4	86.3	78.6	91.2	89.2	85.9	78.0
CRLBP($\alpha = 1$) [22]	86.9	85.7	82.2	73.9	92.9	91.8	88.1	82.0	93.3	92.0	89.5	81.9
Number of training images N =	20				15				10			5
Xu et al. [23]	93.8				91.3				89.7			83.3
Our proposed descriptors												
$\text{CLBP}^{\#}$	75.0	70.8	67.0	59.5	70.8	65.7	60.3	49.9	66.4	60.9	55.4	44.3
CLBP^{M}	88.0	85.8	82.9	75.2	92.1	90.7	87.9	81.1	93.1	92.3	88.7	81.9
CLBP^{m}	87.3	84.5	81.6	73.7	91.3	89.6	86.2	78.3	92.1	90.3	86.5	78.1
CLBP^{Σ}	88.1	85.6	82.8	75.2	90.8	89.4	86.7	79.9	91.2	89.9	86.9	79.4
$\text{CLBP}^{M\#}$	88.1	86.2	83.2	76.0	92.5	90.9	88.4	80.8	93.8	92.0	89.2	81.6
$\text{CLBP}^{m\#}$	87.8	85.7	82.5	75.4	92.4	90.6	88.0	80.3	93.5	91.6	88.5	80.6
$\text{CLBP}^{M\Sigma}$	88.2	**86.4**	83.6	**76.3**	93.0	91.7	**89.2**	**82.2**	94.2	**92.7**	90.0	82.7
CLBP^{Mm}	88.2	86.2	83.4	76.0	92.9	91.6	89.1	81.8	**94.4**	92.8	**90.1**	82.6
$\text{CLBP}^{\#\Sigma}$	**88.4**	86.4	**83.6**	76.3	92.3	90.6	88.2	80.9	93.1	91.4	88.6	80.7
$\text{CLBP}^{m\Sigma}$	88.2	86.4	83.4	76.3	**93.0**	**91.4**	88.8	81.9	94.3	92.6	89.9	**82.8**

5 Conclusions

We have proposed a versatile and efficient variant of LBP for texture description. The proposed framework extends existing rotation invariant LBP based coding, including *riu*2 and LBC, while enhancing their expressiveness and improving their discrimination capability. The classification results on three recent texture datasets prove the relevance of our framework. Used in combination with the complemented LBP coding, it even outperforms the state-of-the art LBP based descriptors. In the future, we plan to address the problem of high dimensionality when using more attributes in complemented LBPs.

References

1. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996)
2. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. PAMI **24**, 971–987 (2002)
3. Matti, P., Abdenour, H., Guoying, Z., Timo, A.: Computer Vision Using Local Binary Patterns (2011)
4. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition. In: ICCV pp. 786–791 (2005)
5. Liao, S., Chung, A.C.S.: Face recognition by using elongated local binary patterns with average maximum distance gradient magnitude. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 672–679. Springer, Heidelberg (2007)

6. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: Real-Life Images Workshop, ECCV (2008)
7. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Processing **19**, 1635–1650 (2010)
8. Mäenpää, T., Pietikäinen, M.: Multi-scale binary patterns for texture analysis. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 885–892. Springer, Heidelberg (2003)
9. Liao, S.C., Zhu, X.X., Lei, Z., Zhang, L., Li, S.Z.: Learning Multi-scale Block Local Binary Patterns for Face Recognition. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 828–837. Springer, Heidelberg (2007)
10. Liao, S., Law, M.W.K., Chung, A.C.S.: Dominant local binary patterns for texture classification. IEEE Trans. Image Processing **18**, 1107–1118 (2009)
11. Guo, Z., Zhang, Z., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Processing **19**(6), 1657–1663 (2010)
12. Zhao, Y., Huang, D.S., Jia, W.: Completed local binary count for rotation invariant texture classification. IEEE Trans. Image Processing **21**, 4492–4497 (2012)
13. Yokoi, S., Toriwaki, J.I., Fukumura, T.: An analysis of topological properties of digitized binary pictures using local features. CGIP **4**, 63–73 (1975)
14. Fathi, A., Naghsh-Nilchi, A.R.: Noise tolerant local binary pattern operator for efficient texture analysis. Pattern Recognition Letters **33**, 1093–1100 (2012)
15. Ojala, T., Menp, T., Pietikinen, M., Viertola, J., Kyllnen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: ICPR. 701–706 (2002)
16. Dana, K.J., van Ginneken, B., Nayar, S.K., Koenderink, J.J.: Reflectance and texture of real-world surfaces. ACM Trans. Graph. **18**, 1–34 (1999)
17. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE Trans. PAMI **27**, 1265–1278 (2005)
18. Varma, M., Zisserman, A.: A statistical approach to material classification using image patch exemplars. IEEE Trans. PAMI **31**, 2032–2047 (2009)
19. Guo, Z., Zhang, L., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Trans. Image Processing **19**, 1657–1663 (2010)
20. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision **62**, 61–81 (2005)
21. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. PAMI **24**, 971–987 (2002)
22. Zhao, Y., Jia, W., Hu, R.X., Min, H.: Completed robust local binary pattern for texture classification. Neurocomputing **106**, 68–76 (2013)
23. Xu, Y., Ji, H., Fermüller, C.: Viewpoint invariant texture description using fractal analysis. International Journal of Computer Vision **83**, 85–100 (2009)

Gait-Based Person Identification Using Motion Interchange Patterns

Gil Freidlin(✉), Noga Levy, and Lior Wolf

Tel-Aviv University, Tel Aviv, Israel
gilfreid@post.tau.ac.il

Abstract. Understanding human motion in unconstrained 2D videos has been a central theme in Computer Vision research, and over the years many attempts have been made to design effective representations of video content. In this paper, we apply to gait recognition the Motion Interchange Patterns (MIP) framework, a 3D extension of the LBP descriptors to videos that was successfully employed in action recognition. This effective framework encodes motion by capturing local changes in motion directions. Our scheme does not rely on silhouettes commonly used in gait recognition, and benefits from the capability of MIP encoding to model real world videos. We empirically demonstrate the effectiveness of this modeling of human motion on several challenging gait recognition datasets.

Keywords: MIP · LBP · Gait recognition · CASIA · TUMGAID

1 Introduction

Human gait is a valuable biometric characteristic describing the coordinated, cyclic movements of a walking person. Gait analysis is available where other biometrics cannot be measured, as gait can be recognized from a distance, does not require cooperation or even awareness of the subject, and works well on low resolution videos as recorded by standard surveillance cameras. The main challenge of gait recognition is the inherent large variability due to physical factors such as injuries or fatigue, carrying a load or wearing motion restrictive clothes.

Over the years many attempts have been made to design effective representations of video content. These range from high-level shape representations, to methods which consider low-level appearance and motion cues. In the task of Action recognition, the video representation aims to distinguish among human actions regardless of their performer. Interestingly, motion representations developed for action recognition and applied for gait recognition [5,9,13,15,19,32] demonstrate good perception within the same action (walking).

In this work, we adopt the Motion Interchange Patterns (MIP) [21] representation that was developed for action recognition applications. MIP encodes motion directly from video frames, and does not require preprocessing such as

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 94–110, 2015.
DOI: 10.1007/978-3-319-16181-5_7

extracting the silhouette from the background or finding the cycles of the motion as other methods do. This rich local representation of human motion produces a discriminative signature of human cyclic gait motion. We suggest adaptations of the original MIP scheme to gait based identification.

2 Gait Recognition

Gait recognition approaches can be roughly divided into model-based and model-free categories. The model-based family of methods use knowledge about the body shape for the gait analysis. Model matching is performed in each frame in order to measure the physical gait parameters such as trajectories, limb length and angular speed.

Model-free techniques capture gait characteristics by analyzing the feature distribution over the space and time extent of the motion. These techniques often rely on extracting the human silhouette in every frame under the assumption that the interesting information about gait pattern lies in the body shape and contour. Popular methods such as the GEI [11] variants estimate the gait period and average the silhouettes over the gait cycle. Motion features are then computed either directly on the silhouette characteristics or by modeling the silhouette sequence using, for example, optical flow [25] or dynamic texture descriptors [23].

The human silhouette represents human body motions in a compact and efficient way but requires background subtraction, a challenging task for realistic backgrounds. Identification performance is sensitive to the silhouettes quality (as demonstrated in [4]), hence silhouette-based methods are not well adjusted to unconstrained environment. Additionally, relying merely on silhouettes might miss out details containing significant motion information.

In a recent line of work, descriptors extracted directly from video frames, that were originally developed for action recognition, are applied to gait recognition. A few examples are LBP descriptors [16], HOG variants [5,13,15], and dense trajectories [9].

3 Action Recognition Descriptors

A central family of action recognition approaches uses low-level representation schemes of the information in a video. These approaches can be further categorized as local descriptors [26], optical flow based methods [1] and dynamic-texture representations [36].

Local descriptors [22,28,34] capture the locality of the human motion in time and space. As a first stage, pixels that are potentially significant to understand the scenario are detected and the region around them is represented by a local descriptor. To represent the entire video, these descriptors are processed and combined using, for example, a bag-of-words representation [27]. A major drawback of this approach is the sensitivity to the number of interest points detected. When a small number of interest points is detected, there is insufficient information for recognition. Videos with too much motion (e.g., background motion

such as waves or leaves in the wind) may provide a lot of information irrelevant for recognition.

The optical flow between successive frames [1,31], sub-volumes of the video [18], or surrounding the central motion [7,8] is highly valuable for Action Recognition. A drawback of optical flow methods is committing too soon to a particular motion estimate at each pixel. When these estimates are mistaken, they affect subsequent processing by providing incorrect information.

Dynamic-texture representations extend existing techniques for recognizing textures in 2D images to time-varying "dynamic textures" [12,20]. One such technique is Local Binary Patterns (LBP) [29], that extracts texture using local comparisons between a pixel and the pixels surrounding it, and encodes these relations as a short binary string. The frequencies of these binary strings are combined to represent the entire image region.

The Local Trinary Patterns (LTP) descriptor of [36] is an LBP extension to videos. An LTP code of a pixel is a trinary string that is computed by considering the relations among patches centered around the pixel in consecutive frames. A video is partitioned into a regular grid of non-overlapping cells and the histograms of the LTP codes in each cell are then concatenated to represent the entire video.

In this work, we adopt a dynamic-texture based representation, the Motion Interchange Patterns (MIP) [21], a recent video representation that was developed and evaluated on action recognition applications. This representation reflects the range of possible changes in motion and their likelihoods of occurring at each pixel in the video. Static edges are indicated by identifiable combinations of the MIP values, and may be ignored by subsequent processing. MIP codes also allow effective camera motion compensation, required in unconstrained videos.

4 Motion Interchange Patterns

Given an input video, the MIP encoding [21] assigns eight trinary strings consisting of eight digits each, to every pixel in every frame. A single digit compares the compatibility of one motion in a specific direction from the previous frame to the current frame, and one motion in another direction from the current frame to the next one. Figure 1 illustrates the motion structure extracted from comparing different patches.

The code of a given pixel p in the current frame, denoted $S(p)$, is constructed by considering eight possible 3×3 patches around p in both preceding and successive frames. Each digit in $S(p)$ refers to a pair of patches, one from the preceding frame and another from the following frame, out of 64 such pairs.

The sum of squared differences (SSD) patch-comparison operator is used to set the matching bit. Denote by SSD1 (SSD2) the sum of squared differences between the patch in the previous (next) frame and the patch in the current frame, as depicted in Figure 2. Each trit, $S_{i,j}(p)$, is computed as follows, for some threshold parameter θ:

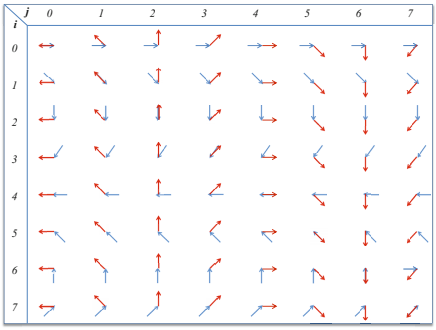


Fig. 1. Representation of motion comparisons of patches from three successive frames. For a given pixel and frame, blue arrows show the motion from a patch in the preceding frame and red arrows show the motion to a patch in the succeeding frame.

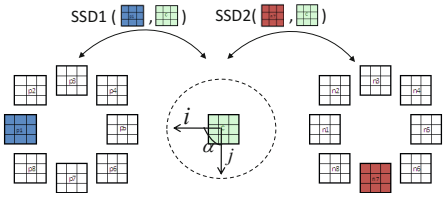


Fig. 2. Each trinary digit in the MIP encoding represents a comparison of two SSD scores, both referring to the same central patch (in green). SSD1 is computed between the central patch and a patch in the previous frame (in blue), and SSD2 is computed between the central patch and a patch in the next frame (in red).

$$S_{i,j}(p) = \begin{cases} 1 & if \quad SSD1 - \theta > SSD2 \\ 0 & if \ |SSD2 - SSD1| \leq \theta \\ -1 & if \quad SSD1 < SSD2 - \theta \end{cases} \tag{1}$$

A value of -1 indicates that the former motion is more likely and 1 indicates that the latter is more likely. The 0 value indicates that both are compatible in approximately the same degree or that there is no motion is this location. MIP compares all eight motions to the eight subsequent motions, obtaining a comprehensive characterization of the change in motion at each video pixel.

MIP Global Descriptor. Denote by i and j the patch locations taken from the previous and following frames respectively, and let α be the angle between direction i and direction j out of the eight possible angle values. There are eight (i, j) pairs for each α, and the concatenation of their $S_{i,j}(p)$ values creates a trinary string. Each 8-trit string is separated into two binary strings, a positive string indicating the ones and a negative string indicating the minus ones, and translated into an integer in the range 0-255. Each pixel obtains 16 integer values, two values per α, that represent the complete motion interchange pattern for that pixel.

For each angle α, two histograms of size 256 are pooled (for the values taken from the positive and negative binary strings ,separately) for each 16×16 cell placed inside the image and concatenated, thus creating 512-dimensional MIP features. A dictionary containing 5000 code words is constructed using k-means on a random subset of MIP features (50000 in our experiments), taken from the encoded gallery set videos. Then, each local string is assigned to the closest word in the dictionary. Denote by u^α the histogram of the dictionary code words in the entire movie, normalized to the sum of one and containing the square root of each element. The global descriptor of a video clip is a concatenation of the eight u^α histograms of all channels.

5 MIP-Based Gait Recognition

Our baseline method employs MIP encoding on videos to find a motion signature of a walking person. We compute the MIP encoding for each video, and then use the local features to create a global descriptor for the whole video as described in section 4.

The MIP encoding is well adapted to gait recognition. The MIP descriptor is a normalized histogram of a bag-of-words of the patterns, hence contains pattern frequencies and does not require finding the gait cycles explicitly (We assume that each video contains at least one gait cycle). Moreover, significant motion patterns tend to repeat in each cycle while noise is random, and are therefore better represented in the histogram.

Another advantage is that MIP does not require silhouette extraction but rather works directly on the video frames. When MIP encoding is applied to moving silhouettes, the boundaries of the body motion are well encoded but other relevant details in the raw video are lost (e.g. the hand swing when passing over the body).

Designed for the action recognition task, MIP implicitly decodes all moving objects in the scene. Therefore, in a video clip containing a single walking person, MIP implicitly decodes the moving person without prior knowledge of the body location, while other methods require external human detection [15] or bounding box assignment. However, when the scene contains other consistently moving objects, their motion is encoded as well, hence narrowing down the area of interest might be needed.

We suggest two modifications of MIP adjusted for gait recognition - confounding details removal and temporal MIP.

Confounding Details Removal. MIP is an appearance-based method, hence, along with the action of interest, it encodes other details that can be misleading in the background or outfit. The standard MIP partly overcomes confusing information by downscaling the input images into a fixed size (100×134 in our experiments) before applying MIP. However, the degraded image quality affects the expressiveness of pose description that might be valuable for analyzing the motion, for example in the elbows region. Hence, after downscaling we

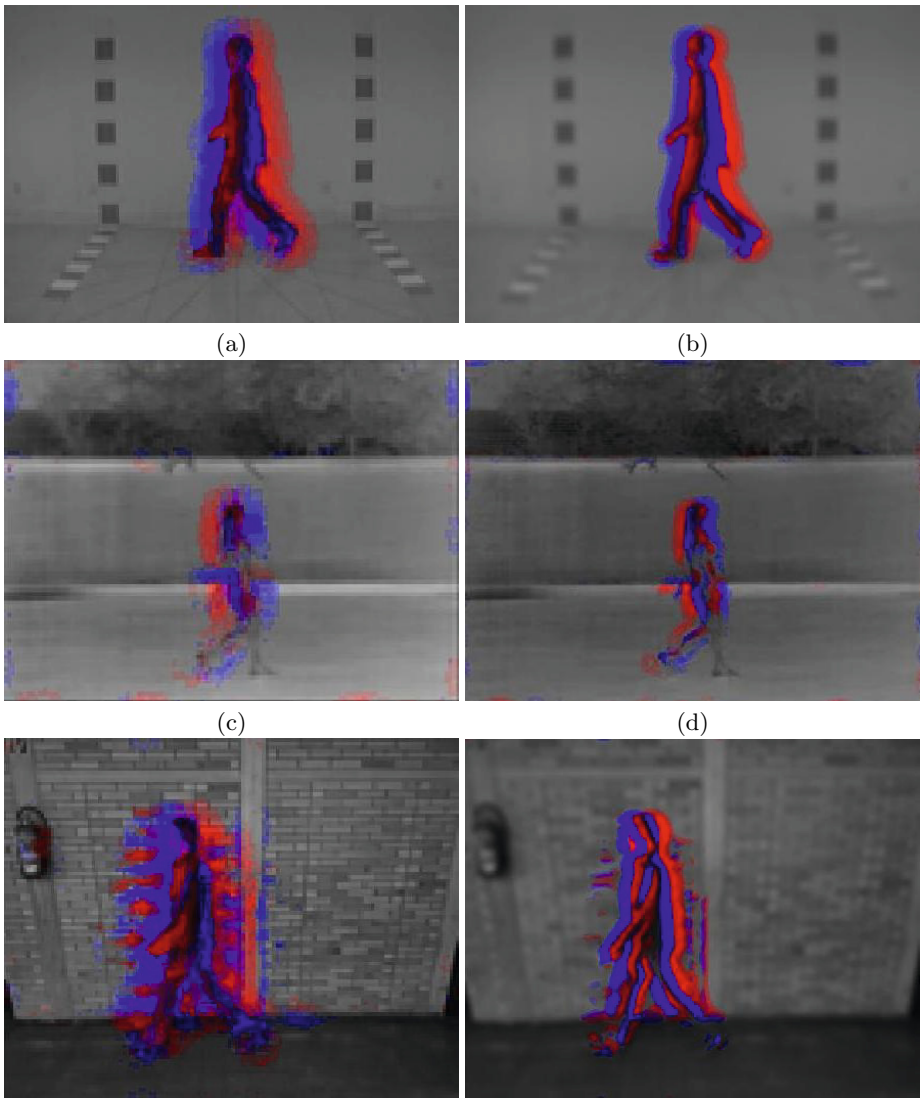


Fig. 3. MIP encoding. The first row contains images from CASIA-B, the second row contains images from CASIA-C, and the bottom row contains images from TUMGAID. In each row, the left image shows the standard MIP encoding and the right image shows MIP with confounding details removal. The encoding after details removal is sharpened and represent the moving human body in greater accuracy. The coded motions are illustrated by color coding pixels by their 8-trit strings content, for a specific α between the compared directions. Blue - motion from the previous frame to the current frame, red - motion from the current frame to the next frame. In image (e), the bricks shape within the shade is encoded, contributing misleading motion patterns. In image (h), details removal is applied and the shade is not encoded as a part of the moving object.

upscale the frames to their original size by interpolation and compute MIP on the original size frames. We acquire a precise MIP encoding of moving body parts represented by significantly more features compared to MIP on the downscaled images, without being distracted by misleading details. This form of low-pass filtering is more suitable compared to conventional direct smoothing on the original image, as it tends to remove textures while keeping depth boundaries without distorting the moving shape. By removing confounding patterns, the weight of the motion patterns relevant for gait identification is increased, thus improving the representation of the motion in the learned dictionaries.

As shown in Figure 3, the resulting encoding follows the moving body parts accurately.

Temporal MIP. The local motion pattern used in the standard MIP compares local motion in a three-sequential-frame scope, symmetrical in both preceding and successive directions. The temporal MIP suggested here enlarges the temporal scope by considering temporal a-symmetric scopes of motion.

The MIP encoding described in section 4 is computed for a given frame t on frames $t-1$, t and $t+1$. The temporal MIP further encodes MIP on frames $t-2$, t and $t+1$ and on frames $t-1$, t and $t+2$, and illustrated in Figure 4. A normalized histogram is constructed separately for every α in each of these encodings. Finally, the global descriptor is a concatenation of all 24 histograms. According to our experiments, extending the temporal scope to the symmetric five frames encoding does not improve performance either by its own or when concatenated with the suggested encoding.

Figure 5 describes the features extracted by the three MIP components of the temporal MIP on examples from CASIA-B and CASIA-C datasets, both on the downscaled frames and on the original size frames after details removal. The details removal variant is computed on the frames enlarged to their original size, thus produces significantly more features to describe the same action compared to standard MIP.

Fig. 4. Visualization of the Temporal MIP extension. Standard MIP encodes three successive frames, $t-1$, t and $t+1$ (solid arrows). Temporal MIP additionally encodes frames $t-1$, t and $t+2$ (dotted arrows), and frames $t-1$, t, and $t+2$ (dashed arrows). Frame t is emphasized in red.

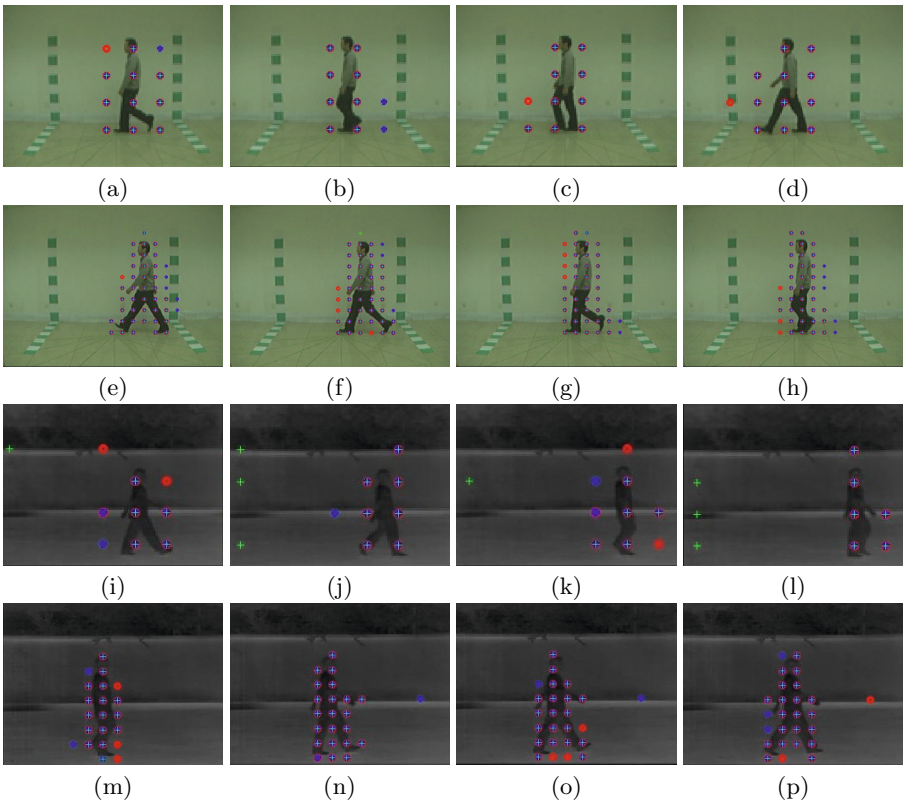


Fig. 5. Representation of the Temporal MIP local features on walking people from the CASIA datasets. Images (a)-(d) show temporal MIP features on a video taken from CASIA-B , (e)-(h) show the details removal variant on the same video. Images (i)-(l) show temporal MIP features on a video taken from CASIA-C, and (m)-(p) show the details removal variant. In the details removal variant, MIP is applied on the full sized frames and hence contains more features. Legend: green pluses - standard MIP features, blue stars - MIP features on frames $(t-2, t, t+1)$, red circles - MIP features on frames $(t-1, t, t+2)$.

6 Classification

Given a gallery set, each image is represented by a global descriptor. These descriptors are used to train a multiclass linear SVM classifier. For N different subjects (class labels), N binary classifiers are obtained in the One-vs-All scheme. Prediction of a new example is performed by extracting its global descriptor, applying all binary classifiers and choosing the subject whose matching classifier gains the highest confidence score.

7 Experiments

We demonstrate our method on the CASIA-B and CASIA-C datasets and on the recently published TUM-GAID dataset. These datasets are challenging, containing various walking styles such as walking in different paces, walking while wearing a coat and carrying a bag or wearing restrictive shoes. Variation in the time of recording given in the TUM-GAID dataset are not tested here.

We test the performance of our method for standard MIP and temporal MIP representations, both with and without confounding details removal, and compare to the results reported by other methods on these datasets.

Performance is evaluated by the classification accuracy – the rate of correct identification by the first match. Experimentally, in most cases our method is comparable or superior to the other approaches, and the temporal MIP and confounding details removal adjustments usually outperform the vanilla MIP classification.

7.1 CASIA-B

The CASIA-B dataset [39] is a large multi-view gait database, containing 124 subjects captured from 11 views. For each subject, three walking styles are recorded - six video clips of normal walk (NN), two of carrying a bag (BG), and two of wearing a coat (CL). CASIA-B was recorded in a controlled indoor environment, with no textured outfits. Therefore, the performance of the details removal MIP in this case is equivalent to a direct encoding of the frames in their original resolution with no filtering applied.

In this work, only recordings captured from a lateral viewpoint are considered. The protocols used for testing are described in Table 1. The first set of experiments follows the evaluation protocol suggested in [39]. It uses as gallery the first four normal walk (NN) sequences per subject and three probe sets, one

Table 1. The evaluation protocols for the CASIA-B dataset. Gallery and probe size represents the number of examples taken for each of the 124 subjects participating in the evaluation test. (a) first set of experiments, the protocol is defined in [39], (b) second set of experiments, the protocol is defined in [16]

Gallery	Probe
NN - first 4	NN - last 2
NN - first 4	BG - 2
NN - first 4	CL - 2

(a)

Gallery	Probe
NN - 5	NN -1
NN - 6	CL - 2
NN - 6	BG - 2
CL - 1	CL - 1
CL - 2	NN - 6
CL - 2	BG - 2
BG - 1	BG - 1
BG - 2	NN - 6
BG - 2	CL - 2

(b)

Table 2. Comparison on CASIA-B dataset from a lateral viewpoint. The model is trained on normal walking and tested separately on each of the walking styles. Left - comparison of the performance on the normal (NN) style probe, right - comparison of the performance on carrying a bag (BG) and wearing a coat (CL). (*) The *Robust* method [17] is trained on three examples per subject and tested on the remaining examples, differently from the protocol defined in Table 1

Method	**NN**
MIP	95.96
Temporal MIP	96.37
MIP + Detail removal	98.79
Temporal MIP + Detail removal	99.19
LBP-FLOW [16]	94
HWLD [32]	**100**
GEI+ nn [39]	97.6
GEI + LDA [11] (results from [4])	83.1
PSC [24]	97.7
FDEI - Wavelet [4]	90.3
FDEI - Frieze [4]	91.1
IDTW [38]	83.5

Method	**BG**	**CL**
MIP	87.9	55.64
Temporal MIP	88.3	57.66
MIP + Details removal	**98.38**	83.87
Temporal MIP + Details Removal	97.98	77.82
LBP-FLOW [16]	45.2	42.9
HWLD [32]	92.2	**96.5**
GEI+ nn [39]	32.7	52.0
GFI Fusion [3]	83.6	48.8
Cross-view [2]	78.3	44.0
Robust(*) [17]	91.9	78.0
PRWGEI [37]	93.1	44.4

Table 3. Comparison on CASIA-B dataset of all combinations of gallery and probe against LBP-FLOW, following the protocol specified in Table 1(b)

Gallery	**NN**			**BG**			**CL**		
Probe	**NN**	**BG**	**CL**	**NN**	**BG**	**CL**	**NN**	**BG**	**CL**
MIP	95.96	89.11	66.12	75	87.5	50.8	51.34	54.43	87.9
LBP-FLOW [16]	94	45.2	42.9	45.2	64.2	25	36.9	22.6	57.1

per each walking style. The second set of experiments follows the evaluation protocol in [16] and contains all gallery-probe combinations of walking styles.

Table 2 compares the performance on the first set of experiments. The results on the left refer to probe NN, and the results on the right refer to probes BG and CL. All compared methods except LBP-Flow [16] rely on silhouette extraction. Our method achieves good performance on the NN probe, and the details removal variants generalize well to the other walking styles, outperforming the other methods on the BG probe by $\sim 5\%$, and achieving the second best result on the CL probe.

Table 3 compares performance of standard MIP against LBP-Flow [16] for all combinations of walking styles per gallery and probe, following the evaluation protocol given in [16]. When the gallery and probe contain different walking styles, all existing sequences are used in both gallery and probe. When the gallery and probe share the same walking style, cross-validation is performed with one example per subject as the probe and the other examples in the gallery, and the average performance is reported. In all combinations, MIP outperform LBP-FLOW by a large gap.

Table 4. The evaluation protocol for CASIA-C dataset: (a) the gallery and probe are from the same walking style, (b) cross style experiments. The number of examples per subject taken as gallery and as probe is specified, for each of the 153 subjects participating. CV stands for cross-validation

Gallery	Probe	Remarks
fn - 3	fn - 1	4-fold CV
fs - 1	fs - 1	2-fold CV
fq - 1	fq - 1	2-fold CV
fb - 1	fb - 1	2-fold CV

(a)

Gallery		Probe		
fn - 4		fs - 2	fq - 2	fb - 2
fs - 2	fn - 4		fq - 2	fb - 2
fq - 2	fn - 4	fs - 2		fb - 2
fb - 2	fn - 4	fs - 2	fq - 2	

(b)

7.2 CASIA-C

The CASIA-C dataset [33] contains video of lateral view captured at night and recorded by a fixed low resolution infra-red camera. There are 153 subjects walking in four walking styles with 10 movies per subject: four movies for normal walking (fn), and two movies per each of the other walking styles – slow pace (fs), quick pace (fq) and carrying a bag (fb).

Table 4 summarizes the evaluation protocol used for CASIA-C dataset. In the experiments referring to gallery and probe that share the same walking style (within), the probe contains one example per subject and the other examples serve as the gallery. Each experiment is repeated with different probe examples for k times, where k is the number of examples per subject in the relevant walking style. We report the average accuracy on the k repetitions. In the experiments training on one walking style and evaluating on a different walking style (cross), all available sequences are used.

Table 5 shows the classification accuracy when training on normal walking and evaluating on all walking styles. The MIP variants outperform all compared methods, and the confounding details removal boosts performance on the bag carrying test set. Table 6 summaries the results when learning on the slow pace, quick pace and carrying a bag train sets, evaluated within the same walking style and on the other styles. MIP variants outperform the compared methods on most combinations.

7.3 TUM-GAID

The TUM-GAID [14] is a recently published dataset with 305 subjects, captured indoor from a lateral viewpoint. The movies were taken by a 3D-depth camera and provide matching audio. In this work we only use the 2D RGB images of the recorded subjects. For each subject, three walking styles are recorded - normal walking (N), carrying a backpack (B) and wearing coating shoes (S). A subset of 32 people is recorded again after a three months period in all walking styles (TN, TB, TS).

The evaluation protocol designed in [14] defines a test set containing 155 subjects. For recognition, the gallery consists of four normal walk recordings per each

Table 5. Results on CASIA-C dataset for a gallery containing normal walking style and evaluated on all probe sets. The first column refers to the normal walking probe. (*) The PSA results in [24] refer to a random subset of 50 subjects (out of 153 subjects)

Method	**Within**	**Cross**		
		fs	**fq**	**fb**
MIP	99.34	**95.09**	**98.69**	96.73
Temporal MIP	99.34	93.79	**98.69**	97.05
MIP+Details removal	99.34	92.15	98.36	99.02
Temporal MIP + Details Removal	99.34	92.16	**98.69**	**99.34**
WBP [23]	99.02	86.3	89.5	80.7
PSA(*) [24]	98	92	92	93
Gait curves [6]	91	65.4	69	25
Bag Of Gait [30]	**99.84**	91.23	95.78	89.82
Pseudo Shape [33]	98	82.4	91.8	24.4
GEI [39]	96	74	83	60
HTI [33]	94	85	88	51

Table 6. Results on the CASIA-C dataset. The top two rows refer to the gallery and probe walking styles respectively. (*) The PSA results in [24] refer to a random subset of 50 subjects (out of 153 subjects).

	Within			**Cross**								
Gallery	**fs**	**fq**	**fb**	**fs**			**fq**			**fb**		
Probe	**fs**	**fq**	**fb**	**fn**	**fq**	**fb**	**fn**	**fs**	**fb**	**fn**	**fs**	**fq**
MIP	99	**99.34**	99	**93.13**	89.54	**88.23**	95.75	**84.31**	92.48	92.97	83.98	90.52
Temporal MIP	**99.34**	**99.34**	**99.34**	91.17	87.25	85.94	96.95	83.98	**94.44**	93.95	**86.93**	91.5
MIP + Details Removal	99	**99.34**	**99.34**	87.41	66.33	80.07	97.05	62.41	88.88	96.73	84.31	91.83
Temporal MIP + Details Removal	**99.34**	**99.34**	**99.34**	85.78	66.33	78.43	**97.22**	62.41	91.83	**97.22**	85.29	**93.46**
WBP [23]	95	96	96	88	61	71	84	61	71	81	70	80
PSA(*) [24]	98	96	96		**93**							
Gait curves [6]	85	79.1	81									

Table 7. Evaluation protocol for the N, B and S probe sets from the TUMGAID dataset as defined in [14]. The number of examples per subject taken as gallery and as probe is specified for each of the 155 subjects

Gallery	**Probe**
N - first 4	N - last 2
N - first 4	B - 2
N - first 4	S - 2

of the 155 subjects and the probe is divided into six test sets, for each walking style and recording phase. The experiments conducted here use the N, B and S probe sets. Table 7 shows the evaluation protocol used for those probe sets.

Table 8. Results on the TUM-GAID dataset trained on normal walking and evaluated on three walking styles. N - normal walking, B - carrying a backpack and S - wearing coating shoes. All compared methods except for our method and GEI utilize depth information

Method	**N**	**B**	**S**
MIP	98.06	95.8	**97.42**
Temporal MIP	98.38	**97.42**	96.77
MIP + Details Removal	97.41	90.96	89.35
Temporal MIP + Details Removal	97.74	94.19	91.61
GEI (results from [13])	94.2	13.9	87.7
Depth-GHEI [13]	96.8	3.9	88.7
Depth-GEI [13]	99	40.3	96.1
GEV [13]	99.4	27.1	52.6
Unimodal RSM [10]	**100**	79	97
SVIM [35]	98.4	64.2	91.6

Table 8 compares our results to other methods. This comparison is challenging, as all methods apart from MIP and GEI [13] employ the depth information provided by the dataset.

MIP and MIP variants cope well with all walking styles. When normal walk is used for both training and testing, all presented methods show very good performance. The RSM method [10] achieves the best performance, utilizing the depth information to extract high quality silhouettes. When training on normal walk and testing on either (B) or the coating shoes probe (S), Mip and temporal MIP outperform all other methods. Temporal MIP gains the highest accuracy on the backpack carrying probe, while MIP wins temporal MIP by a small margin on the coating shoes probe (S).

Although the TUM-GAID dataset is captured indoor, it contains a challenging background of a brick wall nearby the subjects. Due to the lighting conditions, the subjects cast shadows on the wall, which follow them and vary in shape and direction.

When applying MIP, the shadow is encoded along with the movement, as shown in Figure 6(a) and Figure 3(e). Hence, the shaded area contributes motion patterns to the MIP encoding. Since the background contains repetitive strong edges and colored bricks, the filtering in the details removal pre-process does not eliminate these undesirable patterns that clearly reflects the brick edges, as shown in Figure 6(b).

Elimination of these edges is done by applying a Gaussian filter (3×3, $\sigma = 1$) on each frame after downscaling, and then upscaling the frame to the original size. Figure 6(c) demonstrates the new encoding, which focuses on the moving body while avoiding the misleading wall and shadow patterns.

The standard MIP encoding performs better on this dataset over the details removal MIP encoding. The reason might be the information found in the shadow,

(a) (b) (c)

Fig. 6. Detail removal preprocessing for TUMGAID dataset. (a) Low resolution MIP encoding shows the shaded area is encoded, creating motion patterns caused by the shade and the patterned wall. (b) After applying detail removal preprocessing (downscaling then upscaling again to the original frame size) misleading motion patterns that reflects the bricks pattern are still exists in the current shaded area. (c) the result of the new preprocessing flow using a Gaussian filtering to suppress the strong edges, now following mostly the moving body.

that is coded when no details removal is applied. Since all scenes in this dataset were recorded in the same location, in similar conditions and from the same viewpoint, the information encoded in the shaded area might contribute to identification.

8 Summary and Conclusions

Most methods applied to gait recognition involve a preprocessing step of silhouette extraction, making them sensitive to the silhouettes quality and unstable in unconstrained environments.

In this work, we examine the the Motion Interchange Patterns, designed to directly represent motion in unconstrained 2D videos, on gait recognition datasets. Following our observations, we suggest two adaptations of MIP to the task of gait recognition – a temporal extension of the encoded motion, and confounding details removal that enables the analysis of the frames in their original size without getting lost in confounding details.

Employing MIP is a step towards motion analysis that is perceptive enough to identify people from a distance, in real world sequences and under various appearances.

Acknowledgments. Portions of the research in this paper use the CASIA Gait Database collected by Institute of Automation, Chinese Academy of Sciences.

References

1. Ali, S., Shah, M.: Human action recognition in videos using kinematic features and multiple instance learning. TPAMI **32**(2), 288–303 (2010)
2. Bashir, K., Xiang, T., Gong, S.: Cross view gait recognition using correlation strength. In: BMVC, pp. 1–11 (2010)

3. Bashir, K., Xiang, T., Gong, S., Mary, Q.: Gait representation using flow fields. In: BMVC, pp. 1–11 (2009)
4. Chen, C., Liang, J., Zhao, H., Hu, H., Tian, J.: Frame difference energy image for gait recognition with incomplete silhouettes. Pattern Recognition Letters **30**(11), 977–984 (2009)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, CVPR. vol. 1, pp. 886–893. IEEE (2005)
6. DeCann, B., Ross, A.: Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. In: International Society for Optics and Photonics, SPIE Defense, Security, and Sensing, pp. 76670Q–76670Q (2010)
7. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proceedings of Ninth IEEE International Conference on Computer Vision, 2003, pp. 726–733 (2003)
8. Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR, pp. 1–8 (2008)
9. Gong, W., Sapienza, M., Cuzzolin, F.: Fisher tensor decomposition for unconstrained gait recognition. Training **2** 3 (2013)
10. Guan, Y., Wei, X., Li, C.T., Marcialis, G.L., Roli, F., Tistarelli, M.: Combining gait and face for tackling the elapsed time challenges. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2013)
11. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(2), 316–322 (2006)
12. Hassner, T., Itcher, Y., Kliper-Gross, O.: Violent flows: Real-time detection of violent crowd behavior. In: 2012 IEEE Computer Society Conference on CVPRW, pp. 1–6. IEEE (2012)
13. Hofmann, M., Bachmann, S., Rigoll, G.: 2.5 d gait biometrics using the depth gradient histogram energy image. In: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 399–403. IEEE (2012)
14. Hofmann, M., Geiger, J., Bachmann, S., Schuller, B., Rigoll, G.: The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. Journal of Visual Communication and Image Representation **25**(1), 195–206 (2014)
15. Hofmann, M., Rigoll, G.: Improved gait recognition using gradient histogram energy image. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 1389–1392. IEEE (2012)
16. Hu, M., Wang, Y., Zhang, Z., Zhang, D., Little, J.J.: Incremental learning for video-based gait recognition with lbp flow. IEEE Transactions on Cybernetics **43**(1), 77–89 (2013)
17. Iwashita, Y., Uchino, K., Kurazume, R.: Gait-based person identification robust to changes in appearance. Sensors **13**(6), 7884–7901 (2013)
18. Ke, Y., Sukthankar, R., Hebert, M.: Efficient visual event detection using volumetric features. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 1, pp. 166–173. IEEE (2005)
19. Kellokumpu, V., Zhao, G., Li, S.Z., Pietikäinen, M.: Dynamic texture based gait recognition. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1000–1009. Springer, Heidelberg (2009)

20. Kellokumpu, V., Zhao, G., Pietikäinen, M.: Human activity recognition using a dynamic texture based method. BMVC **1**, 2 (2008)
21. Kliper-Gross, O., Gurovich, Y., Hassner, T., Wolf, L.: Motion interchange patterns for action recognition in unconstrained videos. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 256–269. Springer, Heidelberg (2012)
22. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2046–2053 (2010)
23. Kusakunniran, W., Wu, Q., Li, H., Zhang, J.: Automatic gait recognition using weighted binary pattern on video. In: Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2009, pp. 49–54 (2009)
24. Kusakunniran, W., Wu, Q., Zhang, J., Li, H.: Pairwise shape configuration-based psa for gait recognition under small viewing angle change. In: 2011 8th IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 17–22. IEEE (2011)
25. Lam, T.H., Cheung, K.H., Liu, J.N.: Gait flow image: A silhouette-based gait representation for human identification. Pattern recognition **44**(4), 973–987 (2011)
26. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64**(2–3), 107–123 (2005)
27. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
28. Liu, J., Yang, Y., Saleemi, I., Shah, M.: Learning semantic features for action recognition via diffusion maps. Computer Vision and Image Understanding **116**(3), 361–377 (2012)
29. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (2002)
30. Qin, J., Luo, T., Shao, W., Chung, R., Chow, K.: A bag-of-gait model for gait recognition
31. Schindler, K., Van Gool, L.: Action snippets: How many frames does human action recognition require? In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
32. Sivapalan, S., Chen, D., Denman, S., Sridharan, S., Fookes, C.: Histogram of weighted local directions for gait recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 125–130. IEEE (2013)
33. Tan, D., Huang, K., Yu, S., Tan, T.: Efficient night gait recognition based on template matching. In: 18th International Conference on Pattern Recognition, ICPR 2006, vol. 3, pp. 1000–1003. IEEE (2006)
34. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2011)
35. Whytock, T., Belyaev, A., Robertson, N.M.: Dynamic distance-based shape features for gait recognition. Journal of Mathematical Imaging and Vision, pp. 1–13 (2014)
36. Yeffet, L., Wolf, L.: Local trinary patterns for human action recognition. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 492–497. IEEE (2009)

37. Yogarajah, P., Condell, J.V., Prasad, G.: P rw gei: Poisson random walk based gait recognition. In: 2011 7th International Symposium on Image and Signal Processing and Analysis (ISPA), pp. 662–667. IEEE (2011)
38. Yu, S., Tan, D., Huang, K., Tan, T.: Reducing the effect of noise on human contour in gait recognition. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 338–346. Springer, Heidelberg (2007)
39. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th International Conference on Pattern Recognition, ICPR 2006, vol. 4, pp. 441–444. IEEE (2006)

Micro-Facial Movements: An Investigation on Spatio-Temporal Descriptors

Adrian K. Davison[1(✉)], Moi Hoon Yap[1], Nicholas Costen[1], Kevin Tan[1], Cliff Lansley[2], and Daniel Leightley[1]

[1] Manchester Metropolitan University, Manchester M1 5GD, UK
{A.Davison,M.Yap,N.Costen,K.Tan,D.Leightley}@mmu.ac.uk
[2] The Emotional Intelligence Academy, Walkden M28 7BQ, UK
Cliff@eiacademy.co.uk

Abstract. This paper aims to investigate whether micro-facial movement sequences can be distinguished from neutral face sequences. As a micro-facial movement tends to be very quick and subtle, classifying when a movement occurs compared to the face without movement can be a challenging computer vision problem. Using local binary patterns on three orthogonal planes and Gaussian derivatives, local features, when interpreted by machine learning algorithms, can accurately describe when a movement and non-movement occurs. This method can then be applied to help aid humans in detecting when the small movements occur. This also differs from current literature as most only concentrate in emotional expression recognition. Using the CASME II dataset, the results from the investigation of different descriptors have shown a higher accuracy compared to state-of-the-art methods.

Keywords: Micro-movement detection · Facial analysis · Random forests · Support vector machines

1 Introduction

Detecting micro-facial movements (MFMs) is a new and challenging area of research in computer vision that has been inspired by work done by psychologists studying micro-facial expressions (MFEs) [7,12]. Facial expressions have strong scientific evidence suggesting they are universal rather than culturally defined [6]. When an emotional episode is triggered, there is an impulse that cannot be controlled which may induce one of the 7 universal facial expressions (happy, sad, anger, fear, surprise, disgust or contempt). When a person consciously realises that this facial expression is happening, the person may try to suppress the facial expression. Doing this can mask over the original facial expression and cause a transient facial change referred to as a MFE. The speed of these MFEs are high, typically less than 1/5th of a second. During experiments [6] where videos were recorded at 25 frames per second (fps), MFEs have been found to last 1/25th of a second.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 111–123, 2015.
DOI: 10.1007/978-3-319-16181-5_8

MFEs are not so straightforward that they can be interpreted as an emotion and require the context of when the movement occurred to understand whether the movement can be classed as an MFE or as an MFM. Both can be coded objectively using the Facial Action Coding System (FACS) [5], which defines muscle movements and intensity on the face with no emotional interpretation.

The process of detecting normal facial expressions in computer vision usually involves preprocessing, feature extraction and classification. Methods such as Support Vector Machines (SVM) or Random Forests (RF) [21,26] are used to classify and recognise an emotion. This process is similar for MFEs and MFMs, however the features used must be descriptive enough to detect a movement has occurred, because large movements of normal facial expressions usually have more descriptive features making them easier to detect.

Due to the problems described above, this paper extracts local features from image sequences using local binary patterns on three orthogonal planes (LBP-TOP) and Gaussian derivatives (GDs) to accurately determine that a micro-movement has occurred on a face within the dataset compared with an image sequence where no movement occurs (neutral expression). Using these features, two classifiers, SVM and RF, are investigated in how they classify the movements. From the results, a human interpreter would be able to see any movements they may miss, and it can help in interpreting what the movements may mean in the context of the situation.

The remainder of this paper is divided into the following sections; Section 2 discusses related work and approaches in current literature. Section 3 and 4 describe our investigation of detecting micro-facial movement against a neutral face and the results from experiments respectively. Finally, section 5 concludes this paper.

2 Related Work

Previous work in this field is limited, with current literature focusing on recognising what emotion has occurred, and not when a movement occurs.

Pfister et al. [15] use temporal interpolation with multiple kernel learning and RF classifiers on their own spontaneous micro-expression corpus (SMIC dataset) [11]. The authors classify a MFE into positive or negative categories depending on two annotators labelling based on subjects' self reported emotions. Polikovsky et al. [16] introduce another new dataset recorded at 200 frames per second (fps) and the face images are divided into regions created from manually selected points. Motion in each region is then calculated using a 3D-Gradient orientation histogram descriptor. Shreve et al. [19] propose an automatic method of detecting macro- and micro-expressions in long video sequences by utilising the strain on the facial skin as a facial expression occurs. The magnitude of the strain is calculated using the central difference method over the dense optical flow field observed in regions of the face. Wang et al. [23,24] use discriminant tensor subspace analysis and extreme learning machine as a novel way of recognising faces and MFEs. The authors take a grey-scaled facial image and treat it as a second

order tensor and adopt two-sided transformations to reduce dimensionality. Further, they use a tensor independent color space model to show performance of MFE recognition in a different colour space compared with RGB and grey-scale.

Local Binary Pattern (LBP) features [14] form labels for each pixel in an image by thresholding a 3x3 neighbourhood of each pixel with the centre value. The result is a binary number where if the outside pixels are equal to or greater than the centre pixel, it is assigned a 1, otherwise 0. The amount of labels will therefore be $2^8 = 256$ labels. This operator was extended to use neighbourhoods of different sizes [13]. Using a circular neighbourhood and bilinearly interpolating values at non-integer pixel coordinates allow any radius and number of pixels in the neighbourhood. The grey-scale variance of the local neighbourhood can be used as the complementary contrast measure.

As a further extension to local binary patterns (LBP) as a static texture descriptor, Zhao et al. [27] take the LBP in three orthogonal planes, these planes being the spatial and temporal planes (XY, XT ,YT). Originally for dynamic texture recognition, it was used alongside volume LBP to recognise facial expressions. However, unlike dynamic textures, the recognition of facial expressions was done by dividing the facial expression image sequence into blocks and computing the LBP for each block in each plane. These LBP features were then concatenated to form the final LBP-TOP feature histogram. The LBP-TOP histogram provides a robustness to pose and illumination changes, and as the images are split into blocks, the local features of the face better describe facial expressions than a global description of the whole face would.

The Gaussian function is a well-known algorithm and is usually referred to being a normal distribution. Ruiz-Hernandez et al. [18] use the second order derivative to extract blobs, bars and corners to eventually use the features to detect faces in a scene. GDs also provide a powerful feature set with scale and rotation invariant image description. However, when processing higher order derivatives, the feature selection becomes more sensitive to noise, and computationally expensive.

Classification in this area is well established. Random Forests [2] are an ensemble learning method used for classification. It uses many decision trees to find an average balance of votes to decide where a feature should be classified. As a supervised learning method, RF will require training from processed images from a dataset. Support Vector Machines [4] is another supervised learning algorithm which finds the optimal separating hyperplane to decide where to classify data. Both RF and SVM have been used in facial expression recognition [15,21,25,27] and also in other methods such as physical rehabilitation [10] and bioinformatics [20].

The investigation of this paper does not attempt to recognise MFEs as most others, and treats the problem as detecting whether a MFE has occurred compared with a sequence of images that does not contain any movement. These two classes can then be classified using RF and SVM. The potential application of this is to aid a person in detecting when the micro-movement has occurred, and then use this to interpret potential emotion.

3 Method

This section describes a method of differentiating between a MFM and a neutral expression. Normalisation is described by automatically using the centre point of the two eyes and affine transformation to rotate each face from CASME II [25], and then cropping each image to just the face itself. Finally, LBP-TOP and GD features are obtained and classified into either a MFM or neutral expression using RF and SVM.

3.1 Normalisation

Normalisation is applied to all sequences so that all the faces are in the same position based on a constant reference point, in this case, the midpoint between the eyes. Once the midpoint has been obtained, affine transformation is used to rotate the face so that all faces line up horizontally based on this point. The face of the sequences then needs to be cropped to remove the unnecessary background in each image.

To calculate the midpoint of the eyes, first the centre of both eyes are obtained automatically by using a Viola-Jones Haar cascade detector [22] to detect both the left and right eyes separately. Closed eye Haar detectors are available, however as the dataset does not include closed eyes, this has not been implemented. This creates a bounding box around both eyes which the centre point of an eye can then be extracted

$$(C_x, C_y) = \left(\frac{W}{2} + x, \frac{H}{2} + y\right) \tag{1}$$

where C is the centre of the eye, W is the width of the bounding box, and H is the height and x and y are the pixel locations of the top-left corner of the bounding box for the eye. Once the centre points are found for both the left and the right eye, this paper computes the midpoint of the eyes

$$(M_x, M_y) = \left(\frac{LC_x + RC_x}{2}, \frac{LC_y + RC_y}{2}\right) \tag{2}$$

where M is the midpoint between the eyes and LC and RC are the centres of the left and right eye respectively. Using the calculated points, it can be worked out how to apply affine transformation to all images. First the distance between the eyes in Eq. 3 is found and then the angle between the eyes is calculated in Eq. 4

$$(D_x, D_y) = (|RC_x - LC_x|, |RC_y - LC_y|) \tag{3}$$

$$\theta = \frac{\arctan(D_x, D_y)180}{\pi} \tag{4}$$

where D is the distance between the eyes and θ is the angle between the eyes. Using the extracted points, affine transform is used to align the eyes horizontally, ready to be processed.

3.2 Processing Images

Feature extraction begins by grey scaling each image sequence and dividing each image into 9x8 non-overlapping blocks, as proposed by Zhao et al. [27] as their best performing block size (see Fig. 1). This sequence then has a GD operator applied with σ (the standard deviation) being changed from 1-7 in each iteration once the whole database has been processed.

Fig. 1: Images are split into 9x8 blocks so each can be processed separately and obtain local features that are concatenated to form the overall global feature description

A Gaussian function is used as a blurring filter to smooth an image, lowering the high frequencies denoted as

$$G(x, y; \sigma) = e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{5}$$

To extract features such as blobs and corners from the face images, the first and second order derivatives [17] of the Gaussian function is calculated. The first order GD is defined as

$$G_x(x, y; \sigma) = \frac{\partial G(x, y; \sigma)}{\partial x} = -\frac{x}{\sigma^2} G(x, y; \sigma) \tag{6}$$

$$G_y(x, y; \sigma) = \frac{\partial G(x, y; \sigma)}{\partial y} = -\frac{y}{\sigma^2} G(x, y; \sigma) \tag{7}$$

where σ is the scaling element of the GD. The second order GD is defined as

$$G_{xx}(x, y; \sigma) = \left(\frac{x^2}{\sigma^4} - \frac{1}{\sigma^2} \right) G(x, y; \sigma) \tag{8}$$

$$G_{yy}(x, y; \sigma) = \left(\frac{y^2}{\sigma^4} - \frac{1}{\sigma^2} \right) G(x, y; \sigma) \tag{9}$$

$$G_{xy}(x, y; \sigma) = \frac{xy}{\sigma^4} G(x, y; \sigma) \tag{10}$$

the first and second order derivative features are then summed together to get the final GD feature and form a stronger feature representation of blobs, corners and other important features. LBP-TOP is then applied as follows: each block has the standard LBP operator applied [13] with α being the centre pixel and P being neighbouring pixels with a radius of R

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_\alpha) 2^p \tag{11}$$

where g_α is the grey value of the centre pixel and g_p is the grey value of the p-th neighbouring pixel around R. 2^p defines weights to neighbouring pixels and is used to convert the binary string pattern into a decimal. The sign function to determine the binary values assigned to the pattern is

$$s(x) = \{^{1, x \geq 0}_{0, x < 0} \tag{12}$$

where if x is greater than or equal to 0 then $s(x)$ is 1, otherwise 0. After the image has been assigned LBPs, the histogram can be calculated

$$H_i = \sum_{x,y} I\{f_l(x, y) = i\}, i = 0, ..., n - 1 \tag{13}$$

where $f_l(x, y)$ is the image labelled with LBPs. Completing this task on only the XY plane would be suitable for static images, however calculating the XY, XT and YT planes is required to gain a spatio-temporal view of the sequence of images, as expressions are much better described in the temporal domain than still frames [1]. Each plane has been divided into blocks and the LBP histograms extracted to be concatenated into the final feature histogram to be used in classification. For this method, the radius R was set to 3 and the neighbouring points P was set to 8. Fig. 2 shows a representation of creating the LBP-TOP features.

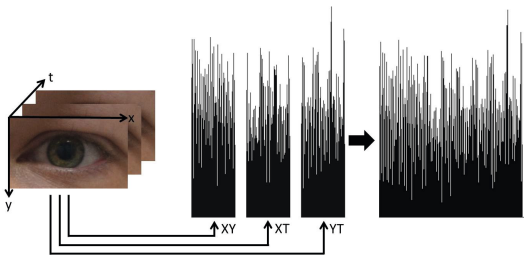


Fig. 2: LBP is calculated on every block in all three planes. Each plane is then concatenated to obtain the final LBP-TOP feature histogram

3.3 Classification

Two popular data classification methods, SVM and RF, will be used to classify between micro-movement and neutral faces within the whole of the CASME dataset. The results of the experiment will compare MFMs against neutral face sequences.

A RF model is constructed by using the bootstrap method to randomly generate a number of decision trees (n_{tree}), which are each provided with randomly selected samples of the training input and then all decision trees are combined into a decision forest. For each bootstrap, a random sample of the training data is used which determines the size of an un-pruned classification tree (*mtry*, default 3). Voting from all trees is used for classification, with the highest voted choice within the data to be selected.

The selected data is taken from the CASME II dataset and consists of micro-movement and neutral face sequences. The RF will determine the accuracy based on correctly classified labels against incorrectly classified labels. RF requires one parameter, n_{tree}, which sets the number of trees to grow. In this experiment the number of trees was set to 300. The software used to implement RF was randomForest Toolbox for Matlab [9].

SVM attempts to find a linear decision surface (hyperplane) that can separate the two classes and has the largest distance between support vectors (elements in data closest to each other across classes). If a linear surface does not exist, then the SVM maps the data into a higher dimensional space where a decision surface can be found. The kernel selected for SVM is the radial basis function (RBF) and will use the same movement and neutral data as RF to determine the accuracy based on the correctly classified labels against incorrectly classified labels.

There are two main parameters that will be selected: Parameter c is a user-defined parameter that controls the trade-off between model complexity and empirical error in SVM. In addition, the parameter γ determines the shape of the separating hyperplane in the RBF kernel. Selection of the optimised parameters was undertaken according to the method by Hsu et al. [8]. The classifier was trained on one subset (training data) and accuracy is tested with the introduction of the second subset (testing). The optimisation process was repeated for each of the possible parameter in exponential steps for both c and γ between 2^{-10} to 2^{10} and 2^{-3} to 2^{3} respectively. The software used to implement SVM is libSVM Toolbox for Matlab [3].

4 Experimental Results

To test this method's performance, combinations of image planes are used with temporal and spatial mixes. The testing data is set up to 50%, therefore if 30% is training the remaining 70% is used for testing. No data within the training set is used for testing to ensure all testing data is unseen. Each plane is tested using 100-fold cross-validation. Other literature [15,25,27] use leave-one-subject-out evaluation with data. This paper uses more or equal testing than training

Table 1: All results using the SVM classifier. Each plane used the the combination of LBP-TOP and GD features. The training percentage is displayed for each plane from 10% to 50%.

Plane	σ	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	1	51.20	47.10	43.30	39.90	36.20
XY	1	51.10	46.90	43.10	39.10	35.10
XTYT	1	51.90	47.90	44.20	40.00	36.80
YT	1	51.00	46.90	43.10	39.40	35.80
All Planes	1	52.50	48.70	44.80	40.90	37.50
XT	2	52.40	48.80	44.80	41.10	36.90
XY	2	52.10	48.10	44.30	40.20	36.40
XTYT	2	53.20	49.80	46.80	43.60	40.90
YT	2	52.50	48.80	45.10	41.30	37.70
All Planes	2	53.70	51.30	48.80	46.30	43.90
XT	3	52.30	48.50	44.60	41.00	37.20
XY	3	52.30	48.30	44.50	40.70	37.10
XTYT	3	53.20	50.20	47.50	44.70	41.70
YT	3	52.60	48.70	45.50	41.60	38.50
All Planes	3	54.20	52.00	50.10	48.30	46.20
XT	4	52.30	48.20	44.40	40.90	36.90
XY	4	52.40	48.30	44.60	40.70	37.10
XTYT	4	53.30	50.20	47.40	44.30	41.20
YT	4	52.40	48.70	45.30	41.90	38.50
All Planes	4	***54.30***	52.30	50.20	48.40	46.60
XT	5	49.82	46.33	42.47	39.56	36.21
XY	5	50.62	46.45	42.74	39.12	35.65
XTYT	5	51.51	47.36	44.03	40.18	37.00
YT	5	50.00	46.12	42.94	40.20	36.56
All Planes	5	52.28	48.26	44.32	40.62	36.40
XT	6	49.89	45.64	42.59	39.03	35.70
XY	6	50.47	46.04	42.47	38.66	35.02
XTYT	6	51.08	47.17	43.64	39.78	36.37
YT	6	49.84	45.86	42.38	38.95	36.11
All Planes	6	52.24	48.08	44.02	39.91	36.26
XT	7	49.62	45.40	41.87	38.27	34.93
XY	7	50.30	46.02	42.26	38.47	34.73
XTYT	7	50.81	46.79	43.36	39.43	36.13
YT	7	49.75	45.87	42.53	39.25	36.43
All Planes	7	52.14	48.01	43.91	39.83	36.09

to describe the robustness of this method compared to others in the literature. The dataset being used is the CASME II recorded at 200 fps with 35 Chinese participants with a mean age of 22.03 years.

For both RF and SVM the σ value for GDs goes from $1-7$. In RF the accuracy increases until the 5th value, where it peaks and begins to decrease, indicating that when $\sigma = 5$ the accuracy is at its highest. In SVM, the accuracy decreases as the σ value increases.

Table 1 shows the results from the SVM experiment and Table 2 shows results from the RF experiment. SVM and RF results vary considerably with the highest accuracy for SVM was 54.3% with training set to 10%. The accuracy

Table 2: All results using the RF classifier. Each plane used the the combination of LBP-TOP and GD features. The training percentage is displayed for each plane from 10% to 50%. The results for RF are significantly higher than SVM with results starting to plateau and decrease when $\sigma = 6$.

Plane	σ	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	1	59.00	63.00	65.60	67.00	70.60
XY	1	51.20	49.30	48.10	46.30	44.50
XTYT	1	57.60	61.20	63.80	66.00	68.00
YT	1	56.60	59.00	60.50	62.70	65.30
All Planes	1	55.80	57.60	58.60	60.00	60.70
XT	2	66.80	73.70	77.20	80.70	82.30
XY	2	51.80	49.80	48.20	45.60	43.90
XTYT	2	66.10	72.20	75.90	79.30	81.60
YT	2	64.90	70.70	75.00	77.70	80.30
All Planes	2	61.30	66.40	69.50	71.20	74.00
XT	3	74.30	82.80	85.90	88.50	90.10
XY	3	52.90	50.80	49.40	48.30	46.20
XTYT	3	73.90	81.80	85.40	87.90	89.20
YT	3	72.30	80.20	84.30	86.50	88.00
All Planes	3	68.10	74.70	78.60	81.30	83.80
XT	4	79.40	86.80	89.20	91.30	92.40
XY	4	53.10	51.70	50.10	48.40	46.60
XTYT	4	78.50	86.10	88.50	90.90	91.70
YT	4	77.80	84.60	87.40	89.00	90.80
All Planes	4	70.60	78.10	81.70	84.80	86.70
XT	5	78.80	86.50	89.50	91.20	92.50
XY	5	53.30	51.50	49.80	47.10	45.40
XTYT	5	79.30	86.70	89.20	91.40	***92.60***
YT	5	78.30	85.70	88.70	90.60	92.20
All Planes	5	71.70	79.00	82.50	84.80	87.30
XT	6	78.60	85.70	88.70	90.80	91.80
XY	6	52.80	50.60	48.30	46.90	44.50
XTYT	6	78.30	86.10	88.70	90.90	92.00
YT	6	78.40	84.90	87.60	90.00	91.40
All Planes	6	70.30	77.30	80.40	84.30	86.40
XT	7	75.40	83.00	85.90	88.40	89.80
XY	7	52.70	50.20	48.30	45.40	42.80
XTYT	7	77.40	83.90	87.10	89.10	90.60
YT	7	77.60	83.90	87.20	88.80	90.20
All Planes	7	69.20	75.60	79.00	81.00	84.00

gradually decreased as training increased. As the data is high-dimensional and values lie close together, SVM struggles to separate the data beyond chance. As RF uses a bootstrap method it is able to generate many classifiers (ensemble learning) and aggregate results to handle the data more appropriately, only ever choosing random samples and ignoring irrelevant descriptors. This gave the highest accuracy of **92.6%** in the XTYT plane with a standard deviation (STD) of 1.78.

By removing the spatial information and just using the temporal planes, classification results for RF are higher. In SVM the results did not vary considerably

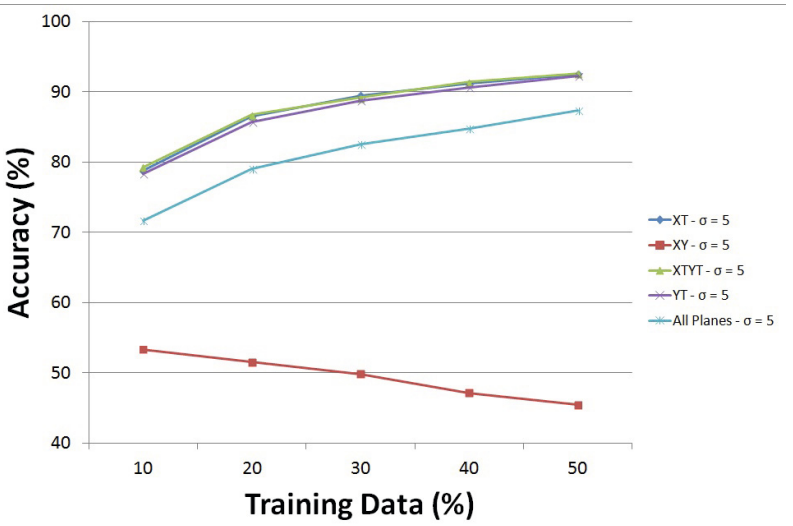


Fig. 3: Using RF, the accuracy of all planes where $\sigma = 5$. Notice XY decreases as training increases due to the lack of temporal information.

Table 3: All results using the SVM classifier when using only LBP-TOP features. The training percentage is displayed for each plane from 10% to 50%.

Plane	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	52.4	48.8	46.1	43	40.9
XY	51.9	48.1	43.9	40.1	36.7
XTYT	53.6	51.1	48.8	46.7	44.2
YT	53.1	50.6	47.9	45	43.1
All Planes	54.2	52.2	50.2	48.3	46.3

across planes, and the highest result was for all planes (54.3%, STD: 0.56) and the lowest being the XY plane alone (34.73%, STD: 2.6).

The highest results were found to be when the σ value was set to 5. Fig. 3 shows the gradual increase in accuracy as training is increased in all planes with a temporal element. A decrease was shown in just the XY plane, supporting that as more training is introduced, the XY plane acts as noise to any movement. This can also be seen when all planes are used and the accuracy is pushed lower than just the temporal planes.

SVM and RF were also used to classify the image sequences using only LBP-TOP features. The results in Table 3 show that all of the planes perform no much better than chance, if not lower, with accuracy decreasing as the amount of training data is introduced. SVM appears to perform similar to results with GD, and separating the features is difficult. Table 4 shows the results from RF using only LBP-TOP features. The accuracy for detecting movement increased significantly compared with SVM, however the highest result was lower than

Table 4: All results using the RF classifier when using only LBP-TOP features. The training percentage is displayed for each plane from 10% to 50%.

Plane	10% Train. Accuracy (%)	20% Train. Accuracy (%)	30% Train. Accuracy (%)	40% Train. Accuracy (%)	50% Train. Accuracy (%)
XT	60.4	64.3	66.6	69.4	71.4
XY	50.9	48.1	46	43.3	41.4
XTYT	58.8	61.7	63.3	65.5	67.8
YT	56.8	58.7	60.6	62.3	64
All Planes	56	57.8	58.5	59.6	59.9

when combined with GDs at **71.4%** when using 50% training and 50% testing data in the XT plane.

To the best of our knowledge, there has not been any results from purely detecting MFM when comparing with neutral faces and so a benchmark for comparing our results could not be found. Most previous work focuses on detecting the movements and classifying into distinct emotional categories and therefore include automatic interpretation based on the FACS equivalent muscle movements (i.e. happy would be movement in AU12).

5 Conclusion

This paper shows that the combination of LBP-TOP and GD features, classification with RF can perform significantly better than SVM when detection micro-movement against neutral, with a highest accuracy of 92.6%. The standard deviation of results is low indicating mean accuracies are consistent using cross-validation. This paper also shows that combining the higher order GD and LBP-TOP can represent the subtle temporal movement of the face well with RF. However, the features are unable to be split by the SVM hyperplane beyond chance.

When using spatial XY planes alone or combined with temporal planes, detection accuracy decreases, suggesting the XY plane is introducing noise to subtle movement. Our method specifically detects micro-movement against neutral faces, which has yet to become a well established method. Most current research detects the MFEs to then classify them into emotion categories.

Future work will look into how the data is represented for MFMs and MFEs, including exploring further methods of temporal feature selection and extraction for micro-movements and how best to discriminate clearly when a subtle movement occurs. Other work includes exploring unsupervised learning methods of classifying movement and non-movement instead of using supervised and computationally expensive methods that require training.

References

1. Bassili, J.N.: Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. Journal of Personality and Social Psychology **37**(11), 2049 (1979)

2. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
3. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**, 273–297 (1995)
5. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press (1978)
6. Ekman, P.: Emotions Revealed: Understanding Faces and Feelings. Phoenix (2004)
7. Ekman, P.: Lie catching and microexpressions. In: The Philosophy of Deception. Oxford University Press (2009)
8. Hsu, C.W., Chang, C.C., Lin, C.J., et al.: A practical guide to support vector classification (2003). https://www.cs.sfu.ca/people/Faculty/teaching/726/spring11/svmguide.pdf
9. Jaiantilal, A.: Random forest (regression, classification and clustering) implementation for matlab (2009). http://code.google.com/p/randomforest-matlab
10. Leightley, D., Darby, J., Li, B., McPhee, J., Yap, M.H.: Human activity recognition for physical rehabilitation. In: International Conference on Systems, Man, and Cybernetics (SMC), pp. 261–266 (2013)
11. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: Inducement, collection and baseline. In: FG (2013)
12. Matsumoto, D., Hwang, H.S.: Evidence for training the ability to read micro-expressions of emotion. Motivation and Emotion **35**, 181–191 (2011)
13. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24**, 971–987 (2002)
14. Ojala, T., Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996)
15. Pfister, T., Li, X., Zhao, G., Pietikainen, M.: Recognising spontaneous facial micro-expressions. In: ICCV (2011)
16. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In: ICDP (2009)
17. Romeny, B.M.H.: Gaussian derivatives. In: Front-End Vision and Multi-Scale Image Analysis. Springer, The Netherlands (2003)
18. Ruiz-Hernandez, J., Lux, A., Crowley, J.: Face detection by cascade of gaussian derivates classifiers calculated with a half-octave pyramid. In: International Conference on Automatic Face Gesture Recognition, pp. 1–6, September 2008
19. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro- and micro-expression spotting in long videos using spatio-temporal strain. In: FG (2011)
20. Statnikov, A., Wang, L., Aliferis, C.: A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics **9**(1), 319 (2008)
21. Tian, Y.L., Kanade, T., Cohn, J.F.: Facial expression analysis. In: Handbook of Face Recognition. Springer, New York (2005)
22. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
23. Wang, S.J., Chen, H.L., Yan, W.J., Chen, Y.H., Fu, X.: Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. Neural Processing Letters **39**, 25–43 (2014)

24. Wang, S.J., Yan, W.J., Li, X., Zhao, G., Fu, X.: Micro-expression recognition using dynamic textures on tensor independent color space. In: ICPR (2014)
25. Yan, W.J., Li, X., Wang, S.J., Zhao, G., Liu, Y.J., Chen, Y.H., Fu, X.: Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. PLoS ONE **9**, e86041 (2014)
26. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. PAMI **31**, 39–58 (2009)
27. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. PAMI **29**, 915–928 (2007)

Analysis of Sampling Techniques for Learning Binarized Statistical Image Features Using Fixations and Salience

Hamed Rezazadegan Tavakoli(✉), Esa Rahtu, and Janne Heikkilä

Center for Machine Vision Research, University of Oulu, Oulu, Finland
{hamed.rezazadegan,esa.rahtu,janne.heikkila}@ee.oulu.fi

Abstract. This paper studies the role of different sampling techniques in the process of learning *Binarized Statistical Image Features* (BSIF). It considers various sampling approaches including random sampling and selective sampling. The selective sampling utilizes either human eye tracking data or artificially generated fixations. To generate artificial fixations, this paper exploits salience models which apply to key point localization. Therefore, it proposes a framework grounded on the hypothesis that the most salient point conveys important information. Furthermore, it investigates possible performance gain by training BSIF filters on class specific data. To summarize, the contribution of this paper are as follows: 1) it studies different sampling strategies to learn BSIF filters, 2) it employs human fixations in the design of a binary operator, 3) it proposes an attention model to replicate human fixations, and 4) it studies the performance of learning application specific BSIF filters using attention modeling.

Keywords: Binary operators · Visual attention · Salience modeling

1 Introduction

The research on image descriptors is a well-studied area in computer vision. In general, image descriptors describe the visual characteristic (e.g., shape, color, texture, motion) of the image. They are the building blocks of many vision related tasks such as image retrieval, recognition tasks (e.g., texture, object, face), action recognition, facial expression analysis, and etc.

Today, the computer vision domain is replete with image descriptors. Some descriptors are more generic, e.g., SIFT [10], SURF [1], BRIEF [2], DAISY [18] and their variants, compared to other operators such as LBP [12], LPQ [13] which are mostly developed for class specific applications (e.g., texture classifications, and face recognition). Nonetheless, they are somehow linked by a common framework of Filtering, Labeling and Statistics (FLS) which provides a unique implementation for LBP and SIFT like features [4].

Adopting the concepts of [4], one can write the LBP operator as the thresholded-quantized-mapped response of a series of multi-directional filter banks.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 124–134, 2015.
DOI: 10.1007/978-3-319-16181-5_9

While traditionally the filters are hand tuned, intrigued to improve quality of filters, [9] proposed to learn the filters using image statistics in which the premise is that statistically learned filters convey image information better. Nonetheless, such an approach poses a new challenge by requiring effective training of the filters. Thus, this paper tries to seek a suitable answer by investigating the domain of salience modeling and visual attention. Initially, it exploits human fixations to train BSIF filters from natural image statistics in order to analyze possible relation between informative regions and training of filters.

Afterwards, motivated by the success of learning based methods, e.g. [16], in which a set of filters specific to a class category is learned, this paper explores learning the filters from application specific data sets and particular class categories. However, it faces a difficulty in using human fixations because there is no such a data set available. To compensate, it develops an attention model to replicate human fixations during the learning process.

Eventually, the performance of the sampling strategies is studied in several applications such as texture classification and face recognition. It will be demonstrated that learning of filters somehow benefits from selective sampling and the proposed framework for attention-based learning of filters improves the performance of face recognition.

1.1 Related Work

This paper targets domain of binary patterns such as LBP [12]. Such operators treat the relation of each pixel and its surrounding as a binary code string. Consequently, an image is represented by the probability distribution of binary code strings obtained in terms of histograms. Thus, the paper adopts the binarized statistical image features (BSIF) to investigate the role of underlying data set information in the process of learning statistical representations.

BSIF binarizes the response of a set of statistically learned filters with a threshold at zero, in which each filter response is in correspondence with a different filter. The filters are learned by maximizing the statistical independence of the filter responses using *Independent Component Analysis* (ICA) [6].

In a few words, given an image I and a filter w_i of size $l \times l$, the filter response is

$$s_i = w_i * I, \tag{1}$$

where s_i is the response of the i-th filter, and $*$ is the convolution operator. For a specific pixel $\mathbf{x}$, BSIF derives a binarized filter response such that $b_{i,\mathbf{x}} = 1$ if $s_i > 0$ at $\mathbf{x}$, otherwise $b_{i,\mathbf{x}} = 0$. Thus, in presence of n filters a binary string of length n describes each pixel.

BSIF learns the filters using independent component analysis. To this end, it forms a training set of image patches by taking random samples from natural images. Afterwards, it employs a canonical preprocessing step and performs *Principal Component Analysis* (PCA) to obtain dimension-reduced whitened data samples. Eventually, it utilizes a standard ICA algorithm [6] to obtain a set of linearly defined filters.

2 Fixations and BSIF

In order to learn the filters, BSIF requires several sampled image patches. It obtains them by randomly sampling image patches from natural images. Nonetheless, there are arguments and evidence that supports the fact that random sampling does not necessarily provide the best informative image patches. For instance [8] proposed taking image patch samples from the most salient regions to make descriptors in a recognition task and demonstrated the success of the attention based learning.

Intrigued to investigate application of informative regions in training of BSIF, the filters are learned using patches extracted around human fixation points on natural image statistics. The learning procedure is as follows: 1) The images are converted to grayscale, 2) The patches are selected around the fixation points of observers, 3) the DC-component (i.e., mean value) of each image patch is discarded, 4) The patches are dimension reduced and whitened, 5) the independent components are estimated. In mathematical terms, for an image patch, $\{x\}$, of size $l \times l$ centered at $\mathbf{x}$, one can apply ICA algorithm to estimate the independent components, i.e., the $n \times l^2$ filter matrix $\mathbf{W}$. The filter matrix includes n vectorized filters, w_i, of length l^2. Knowing that the all-in-one response of the filters on a patch can be formulated as $s = \mathbf{W}\{x\}$, one can write

$$s = \mathbf{U}\mathbf{z}, \tag{2}$$

where $z = \mathbf{V}\{x\}$, $\mathbf{U}$ is a $n \times n$ matrix which is estimated via ICA. The matrix $\mathbf{V}$ conveys the PCA whitening procedure which facilitates estimation of the orthogonal matrix $\mathbf{U}$ using the fact that $\mathbf{z} = \mathbf{U}^{-1}s$. Eventually, by estimating $\mathbf{V}$ and $\mathbf{U}$, it obtains $\mathbf{W} = \mathbf{U}\mathbf{V}$.

2.1 Fixations' Replicate

In order to boost the performance of the operator, one may suggest learning the filters tuned for a specific data set, e.g. learning the filters from face images for a face recognition task. In this context, the aforementioned methodology for learning filters has one disadvantage which is the requirement of human fixations. Access to reordered fixations on class specific data is not always possible due to expensive gathering procedures. To compensate, this section introduces an artificial mechanism of fixation selection. The mechanism relies on a salience map, which is obtained using natural image statistics, and application of inhibition of return (IOR) procedure in selection of most salient region.

To compute the salience map, the proposed framework utilizes the filters learned from the previous step and intensity of an image. For each filter, it employs the *Saliency Using Natural statistics* (SUN) [21] to derive a conspicuousness map. SUN defines bottom-up salience as $P(F)^{-1}$ in which F indicates w_i learned as described before. It approximates $P(F)$ as the generalized Gaussian distribution (GGD) estimate of unidimensional distributions such that $P(F = f) = \prod_i P(f_i)$, where f_i is the i-th element in f, and

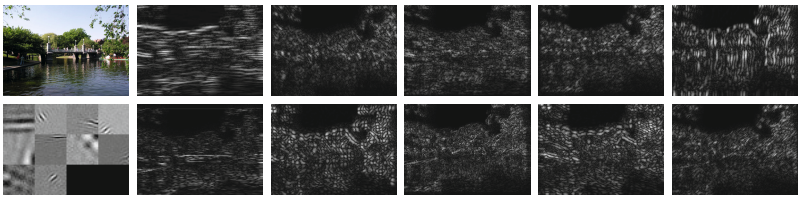


Fig. 1. ICA filter response, left column depicts an image and 10 ICA filters and on the right side the corresponding filter responses are visualized

$$P(f_i) = \frac{\theta_i}{2\sigma_i \Gamma(\theta_i^{-1})} \exp\left(-\left|\frac{f_i}{\sigma_i}\right|^{\theta_i}\right), \tag{3}$$

The discriminative power of ICA filters are even enhanced by nonlinear weighting of each dimension of f using GGD fit to their responses [15]. Fig. 1 depicts the conspicuousness maps obtained from 10 of the ICA filters. Traditionally these conspicuousness maps are combined with equal weights to derive a central saliency map (e.g. [8,21]). Contrarily, the proposed framework treats them as features and employs linear *Support Vector Machines* (SVM) to combine the conspicuousness maps and intensity features to produce a salience map. To this end, it learns a linear SVM on a groundtruth consisting of human fixation density maps in which top 10% salient regions form positive set and top 10% non-salient regions form negative set. Thus, given a training set of n points with the feature input $x_i \in \mathcal{R}_n$ and the corresponding target label $y_i \in \{-1, +1\}$, the SVM is defined as a linear scoring function with a prediction rule such that

$$\hat{y}(x_i) = \text{sign}(\omega^T x_i + \beta), \tag{4}$$

where β is the bias and ω is a weight vector. The weight vector ω is obtained via a minimization problem as follows

$$\begin{aligned} \min_{\omega}. \quad & \frac{1}{2}\omega^T\omega + \lambda \sum_{i=1}^{n} \zeta_i \\ \text{s.t.} \quad & \hat{y}(\omega^T x_i + b) = 1 - \zeta_i \\ & \zeta_i \geq 0 \quad i = 1 \ldots n \end{aligned} \tag{5}$$

where λ is a smoothing regularization parameter balancing the trade-off between error and margin. Consequently, the saliency map is defined as the score obtained by combining the features using ω. In other words, for a feature vector of f, the saliency Sal is defined as $Sal = \omega^T f$. Fig. 2 depicts saliency maps produced using the described technique.

To select fixations, the proposed method applies an inhibition of return (IOR) like mechanism. As depicted in Fig. 3, it implements an iterative scheme consisted of 1) it picks randomly among the salient locations, 2) it attenuates the salience map response at the selected fixation proportional to a Gaussian kernel. The procedure is repeated until enough number of fixations are obtained which

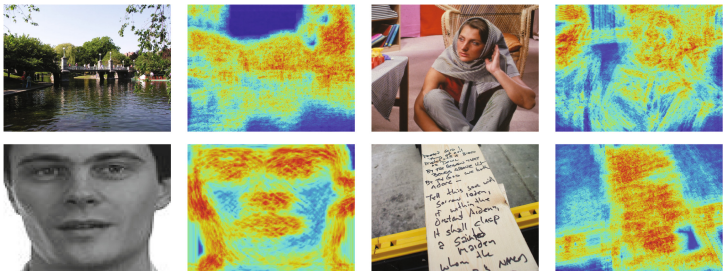


Fig. 2. Salience maps produced using the described technique. As depicted, more salient regions are somehow meaningful to human, e.g. eyes and mouth regions of the face.

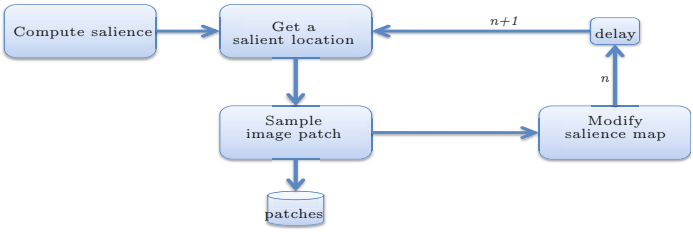


Fig. 3. Sampling image patches using an artificial fixation generation mechanism. For an image, a saliency map is generated and fixations are taken by considering the salient locations. Each time, a location among salient locations is selected randomly and its corresponding image patch is extracted. Afterwards, the salience map is modified and the current fixation location is attenuated to reflect its selection. The process continues over time until enough samples are taken.

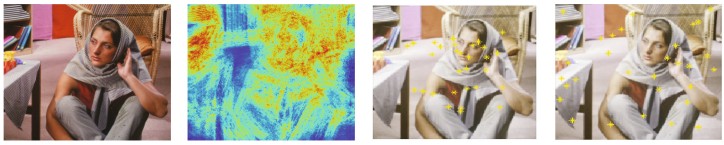


Fig. 4. Sampling using artificial fixations, from left to right, original image, saliency map, samples taken using artificial fixations, random sampling

replicate the human fixations. Fig. 4 visualizes samples taken by such a process, as depicted, samples taken using artificial fixations are concentrated on more meaningful parts of the image compared to random samples.

3 Experimental Analysis

This section assesses the aforementioned scenarios. The analysis covers experiments on texture and faces. Initially, it discusses the texture classification experiments. Afterwards, it continues with the experiments on face recognition which is followed by a discussion.

3.1 Texture Classification

The texture experiments assess two sampling strategies for learning the BISF filters. It compares filters learned from patches taken randomly with the filters learned from patches centered on human fixations. The filters are learned using natural images provided by MIT [7] database. It consists of 1,003 images along with the eye movement statistics, particularly fixations, of 15 viewers at a distance of 48cm. The image set includes natural indoor and outdoor images; each image is presented for 3 second. In order to learn the filters, the images are converted to gray-scale and 500,000 image patches are sampled either randomly or using human fixations. The image patches are of the sizes 3×3, 5×5, and 7×7, as bigger patch sizes are demonstrated not to perform well on the textures [9,20], which are learned at different number of bit levels (i.e. ICA filters) ranging from 5 to 11.

To perform texture analysis, this study utilizes CUReT [3], Outex [11] datasets. The Columbia-Utrecht (CUReT) dataset consists of 61 texture classes, each observed with almost 205 viewing and illumination combinations (more than 12,000 images in total). The categories include a variety of surfaces such as specular, diffuse, isotropic, and etc. The Outex database consists of several test suits. This study utilizes test suits TC_00002 and TC_00012. Each of them consists of 24 classes of texture, while TC_00002 has no rotation and contains only one illuminant, TC_00012 has three illuminants and considers 7 rotation orientations[1].

The classification procedure is chosen to be consistent with the protocols used in [9]. In other words, texture classification is carried out using nearest neighbor classifier in which the distance measure is $\mathcal{X}^2$ using l_1-normalized feature histograms. To classify the CUReT textures, the images are grouped into non-overlapping train and test sets and the procedure is repeated 100 times as described in [19]. The Outex experiments utilizes the provided partitions of [11].

Fig. 5 depicts the results of the two differently trained filters on the CUReT database. There seems to be a small difference between the two sampling approaches. Nonetheless, the filters learned using the fixation sampled images perform marginally better than randomly learned ones meanwhile achieving maximum accuracy of 96.6.

Fig. 6 visualizes the results of the Outex database. As depicted in 6(a), similarly the Outex TC_00002 results indicate slight improvement in training the filters using patches sampled at fixation points. On the other hand, the performance analysis of Outex TC_00012, showed in 6(b), reveals a 4% performance improvement using fixations to train the BSIF filters (5×5-7 bits performing 66.4% vs. 5×5-6 bits performing 64%).

Comparing the results on Outex TC_00012[2] with TC_00002 and CUReT conveys that *the selective training of filters boosts the performance of operator in*

[1] Please see: http://www.outex.oulu.fi/index.php?page=classification for detailed information on test suits.

[2] TC_00012 is difficult because it contains several rotations and illuminants.

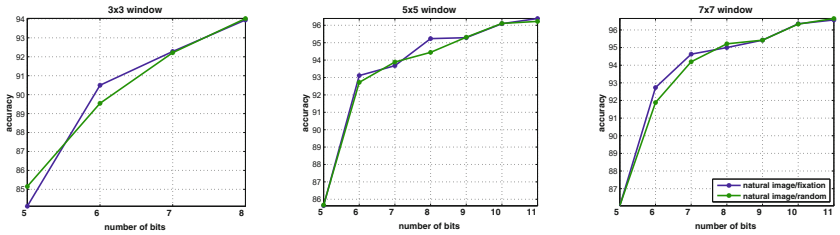
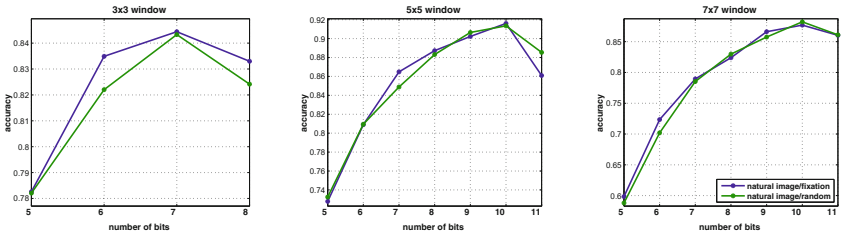
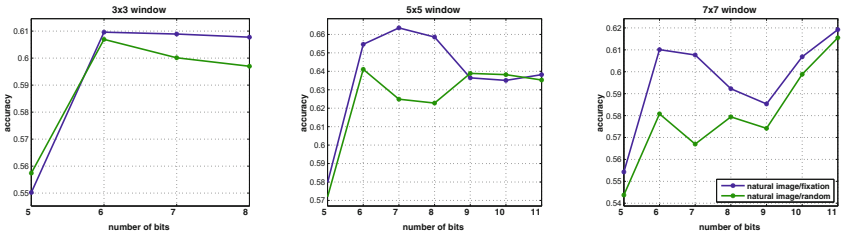


Fig. 5. CUReT, performance analysis of filters trained randomly compared to filters trained on human fixations



(a) TC_00002



(b) TC_00012

Fig. 6. Outex, performance analysis of filters trained randomly compared to filters trained on human fixations

handling data carrying more information. Thus, this study motivates the assessment of sampling strategies on more complicated scenarios and data. Intrigued to have a better understanding, this paper performs a series of analysis on face recognition task.

3.2 Face Recognition

This section considers face recognition task in order to study the role of sampling in training of BSIF in a more challenging task. It extends the sampling mechanism by incorporating faces in the learning process. To learn the filters from face images, it adopts a cropped version of the Labeled Faces in the Wild (LFW) [5], recognized as LFWcrop [17][3]. It consists of more than 13,000 images of faces

[3] Download link: http://conradsanderson.id.au/lfwcrop/

which are cropped to prevent the recognition by getting advantage of the background information. However, it does not have any eye tracking data available. Therefore, the artificial fixation generation scheme mentioned above is employed in order to select the location of each image patch selectively. Eventually, this section analyzes a set of 4 different filters: the filters learned on face data using random sampling and artificial fixation selection mechanism and the two filters applied in the texture analysis study. The same parameters and configurations applied in the learning of the filters on face data.

The experiments are carried out on the FERET database [14] using the frontal profile images. The images are partitioned into gallery (*fa*) and *fb* probe images. The gallery consists of 1196 images, and the probe consists of 1195 images with varying facial expressions. Fig. 7 depicts some of the face images. It is expected that the performance will be somehow related to the amount of information the filters would be able to encode and the data of the experiments.

Fig. 7. Sample images from FERET data base

The recognition procedure initially crops the images using the location of subjects' eyes to have the complete frontal face in the center of frame. Afterwards, the images are normalized to a canonical size of 128×128. It divides the face image into 8×8 non-overlapping rectangular regions and computes the BSIF descriptor independently for each segment. Concatenation of l_1 normalized descriptors makes an image descriptor. The classification uses nearest neighbor and $\mathcal{X}^2$ distance measure.

Fig. 8 depicts the results of the face recognition task. The 7×7 filters with 12 coding bits achieve the performance of 94.23%. The comparison of curves somehow expresses that *the number of coding bits (i.e. information) has a direct relationship with using selective sampling approach in the learning process of ICA filters*. It is worth-noting that while learning small filters does not benefit from training on class specific data, bigger windows and higher number of bits

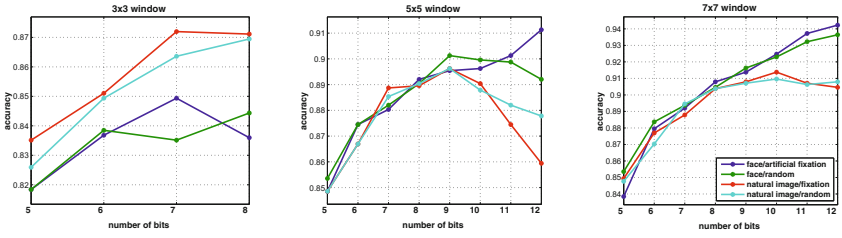


Fig. 8. Face recognition and sampling strategies using different window sizes

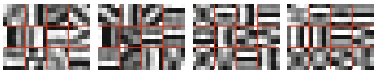
get advantage of such data. Nonetheless, the behavior of curves raises some questions which this study tries to address in the next section.

3.3 Discussion

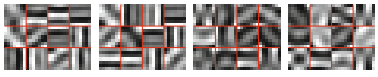
The variation in the curves depicting performance of texture and face recognition raises some questions. **Why there is a marginal contribution in adapting selective sampling for texture?** The texture often consists of simple repeating patterns which makes them difficult to discriminate. Nonetheless, the learning of such simple structures are somehow easily doable by having enough number of samples via ICA. As depicted in Fig. 4, the filters learned from natural image statistics consists of similar structures which are probably enough to represent the textures. Nonetheless, the significance of selective sampling becomes apparent in applying a rotation variant operator (i.e. BSIF) to rotated texture samples; referring to Fig. 6(b), one realizes that selectively trained filters perform 4% better than randomly trained filters in the task of the recognition of TC_00012 textures.

The sampling strategies and learning are not limited to selective sampling. Face recognition included filters learned on class specific trained filters, i.e. filters learned on faces. **Is there any benefit in training the filters on class specific data sets?** As depicted in Fig. 8, the maximum face recognition rate is achieved using the filters which are trained on class specific data and convey more information. In other words, learning ICA filters from class specific data becomes useful as the amount of information required to perform a task increases. To find grounds for such a behavior, Fig. 4 visualizes the ICA filters learned using various sampling techniques and data. As shown, in case of small filters of size 3×3, the filters learned on natural statistics present a structure similar to gradient filters, which effectively encodes almost any data in such a small neighborhood. Consequently, natural image statistics somehow perform better for that specific scale. Contrarily, as the filter size and number of bits increase, one can observe – e.g., from Fig. 9(c) – that the filters trained from class specific

(a) 3×3 Filters of 7 bits



(b) 5×5 Filters of 12 bits



(c) 7×7 Filters of 12 bits

Fig. 9. Visualization of the ICA filters learned using different sampling strategies, from left to right: randomly learned from natural images, learned from fixations using natural images, learned randomly from face images, and learned using artificial fixations from face images

data are absolutely different, probably reflecting the underlying data better. The performance even slightly improves using the artificial fixation based sampling meanwhile it seems that more complicated filters are learned using fixation.

What is the verdict? The conducted experiments reveals that there is a relation between the amount of information conveyed by the BSIF filters, sampling strategies and learning filters from class specific data. The conclusion is that *depending on the amount of information embedded in the data, reaching the optimum operation point may benefit from learning on class specific data and selective sampling.*

4 Conclusion

This study examined different sampling strategies for learning ICA filters used by BSIF operator. These strategies include random sampling and selective sampling. The study employed two techniques for taking the samples selectively, first it utilized fixation points on natural image statistics, second it developed an artificial fixation generation scheme to replicate human fixations in the process of learning the filters.

To generate artificial fixations, it proposed an attention model. The attention model derives a salience map using natural image statistics responses and linear support vector machine. Afterwards, it implements an inhibition of return mechanism to replicate the human fixations. Consequently, the proposed locations of image patches are more concentrated on meaningful areas of the image. The mechanism is particularly applied in the process of learning ICA filters for the task of face recognition. Eventually, the proposed mechanism is applied to replicate human fixations in the process of learning from face data.

The experiments suggest that using selective sampling and class specific data in learning the filters affects the performance of the BSIF operator. Nonetheless, the improvement is somehow dependent on the assigned task because it is affected by the the amount of information required to represent the image.

Acknowledgments. This work was supported by the Infotech Oulu doctoral program. The first author thanks Dr. Juho Kannala for sharing his code.

References

1. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding **110**(3), 346–359 (2008). Similarity Matching in Computer Vision and Multimedia
2. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: BRIEF: Computing a Local Binary Descriptor Very Fast. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(7), 1281–1298 (2012)
3. Dana, K., Ginneken, B., Nayar, S., Koenderink, J.: Reflectance and texture of real world surfaces. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 151–157 (1997)

4. He, C., Ahonen, T., Pietikainen, M.: A bayesian local binary pattern texture descriptor. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4, December 2008
5. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07–49. University of Massachusetts, Amherst, October 2007
6. Hyvärinen, A., Hurri, J., Hoyer, P.O.: Natural Image Statistics A probabilistic approach to early computational vision. Springer (2009)
7. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: IEEE International Conference on Computer Vision (ICCV) (2009)
8. Kanan, C., Cottrell, G.: Robust classification of objects, faces, and flowers using natural image statistics. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2472–2479, June 2010
9. Kannala, J., Rahtu, E.: Bsif: binarized statistical image features. In: ICPR (2012)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision **60**(2), 91–110 (2004)
11. Ojala, T., Mäenpää, T., Pietikäinen, M., Viertola, J., Kyllönen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: 16th International Conference on Pattern Recognition (2002)
12. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. **24**(7), 971–987 (2002)
13. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008. LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008)
14. Phillips, P., Wechsler, H., Huang, J., Rauss, P.J.: The feret database and evaluation procedure for face-recognition algorithms. Image and Vision Computing **16**(5), 295–306 (1998)
15. Shan, H., Cottrell, G.: Looking around the backyard helps to recognize faces and digits. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (2008)
16. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
17. Tistarelli, M., Nixon, M.S. (eds.): ICB 2009. LNCS, vol. 5558. Springer, Heidelberg (2009)
18. Tola, E., Lepetit, V., Fua, P.: DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(5), 815–830 (2010)
19. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. Int. J. Comput. Vision **62**(1–2), 61–81 (2005)
20. Ylioinas, J., Kannala, J., Hadid, A., Pietikäinen, M.: Learning local image descriptors using binary decision trees. In: WACV (2014)
21. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision **8**(7) (2008)

Facial Expression Analysis Based on High Dimensional Binary Features

Samira Ebrahimi Kahou(✉), Pierre Froumenty, and Christopher Pal

École Polytechique de Montréal, Université de Montréal, Montréal, Canada
{samira.ebrahimi-kahou,pierre.froumenty,christopher.pal}@polymtl.ca

Abstract. High dimensional engineered features have yielded high performance results on a variety of visual recognition tasks and attracted significant recent attention. Here, we examine the problem of expression recognition in static facial images. We first present a technique to build high dimensional, $\sim$ 60k features composed of dense Census transformed vectors based on locations defined by facial keypoint predictions. The approach yields state of the art performance at 96.8% accuracy for detecting facial expressions on the well known Cohn-Kanade plus (CK+) evaluation and 93.2% for smile detection on the GENKI dataset. We also find that the subsequent application of a linear discriminative dimensionality reduction technique can make the approach more robust when keypoint locations are less precise. We go on to explore the recognition of expressions captured under more challenging pose and illumination conditions. Specifically, we test this representation on the GENKI smile detection dataset. Our high dimensional feature technique yields state of the art performance on both of these well known evaluations.

Keywords: Facial expression recognition · Smile detection · High-dimensional feature · Census transformation · Deep learning · GENKI · CK+

1 Introduction

Local binary patterns (LBPs) [1] are well known texture descriptors that are widely used in a number of applications. LBP features have been found to be particularly effective for face related applications [2]. As an example, high dimensional features based on LBPs have yielded highly competitive results on the well known Labeled Faces in the Wild face verification evaluation [3,4].

We are interested here in recognizing facial expressions in static imagery. Facial expression analysis can be a particularly challenging problem, especially when using imagery taken under "in the wild" conditions as illustrated by the recent Emotion Recognition in the Wild Challenge [5]. Here we examine both controlled environment facial expression analysis and an "in the wild" problem through evaluations of our proposed method using the Extended Cohn-Kanade (CK+) database [6,7] and the GENKI-4K smile detection evaluation. The CK+ database is a widely used standard evaluation dataset containing acted

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 135–147, 2015.
DOI: 10.1007/978-3-319-16181-5_10

expressions. The expressions to be recognized are based on Ekman's six basic universal categories of: happiness, sadness, surprise, fear, anger, and disgust [8]. The GENKI-4K [9] dataset contains comparatively low resolution images harvested from the web.

We provide a number of technical contributions in this paper. First, we provide a formulation of high dimensional features that is different from other standard formulations. Our descriptor is a high dimensional feature vector in which each dimension consists of the bits derived from Census transformation. Features are obtained based on image patches centered on facial keypoints. We use a slight variant of LBPs known as the Census transform [10]. To the best of our knowledge this representation yields the highest known performance on CK+ using the same evaluation criteria as in [7].

We go on to adapt our technique to be more robust to inaccurately localized facial keypoints using a multi-resolution technique and local Fisher discriminant analysis (LFDA) [11] - a recently proposed extension to the widely used Fisher discriminant analysis technique. The issue of keypoint localization accuracy is particularly important when turning to the problem of recognition in the wild, but even in controlled environments there are well known degradations in performance when per subject keypoint training data is not used to fit a facial keypoint model. Turning to the problem of smile recognition using in the wild GENKI imagery, it is much harder to detect a large number of keypoints due to the quality and variability of the imagery. For the GENKI evaluation in particular we are however able to detect five keypoints reliably. Adapting our method to this setting, here again our proposed method yields the highest known performance of which we are aware on this well known evaluation.

The remainder of this manuscript is structured as follows: We provide a brief review of some other relevant work in section 2, but also discuss other relevant work throughout this document. In section 3 we present our novel feature extraction technique based on high dimensional binary features, multi-scale patches and discriminative dimensionality reduction. In section 4 we benchmark our high dimensional feature vector technique using CK+, examining experimentally the issue of facial landmark prediction quality, its impact on prediction performance and our motivations for extending our basic formulation to include multi-scale analysis and discriminative dimensionality reduction. We then provide our experiments on GENKI-4K, where we also compare directly with a state of the art convolutional neural network technique that does not rely on keypoints. We provide conclusions and additional discussion in section 5.

2 Other Relevant Work

A number of modern, state of the art approaches to expression detection are based on handcrafted features, such as: Local binary patterns or LBP features [1], Histograms of oriented gradients or HOG features [12], or Lowe's Scale-invariant feature transform (SIFT) descriptors [13]. For example, the influential work of Shan et al. [14] studied histograms of LBP features for facial expression

recognition. They introduced Boosted-LBP by using AdaBoost [15] for feature selection. Their experiments showed that LBP features are powerful for low resolution images. Dahmane et al. [16] built face representation based on histograms of HOG features from dense grids. Their representation followed by nonlinear SVM outperforms an approach based on *uniform* LBP. Other work has used SIFT features for facial expression analysis [17], yielding competitive results on CK+.

Techniques based on convolutional neural networks have also yielded state of the art performance for the task of emotion recognition, including top performing results on competitive challenges [18–20]. The CK+ data and classification tasks were introduced in Lucey et al. [7]. They provided both the additional facial examples that were used to extend the original Cohn-Kanade (CK) dataset of [6], yielding the combined dataset known as CK+ as well a number of experimental analyses. They provided a variety of baseline experiments and a state of the art result at the time in which they combine a landmark based representation (SPTS) and appearance features both before and after shape normalization using landmarks, which they refer to as CAPP features. They combine two different classifiers for landmarks and appearance using a logistic regression on the outputs of the classifiers. This procedure yields their best result with an average accuracy of 83.33%.

Jeni et al. [21] used shape only information for expression recognition experiments with CK+; however, they removed the sequences with noisy landmarks. The work of Sikka et al. [17] compares the performance for a variety of techniques on the CK+ expression recognition task, including the well known uniform LBP histogram technique in [14] which they state yields 82.38% ±2.34 average accuracy. They state that their own bag of words architecture yields 95.85% ±1.4 average per subject accuracy using a leave one subject out evaluation protocol. Other work has also explored the problem of smile detection using the GENKI-4K data. Jain et al. [22] report 92.97% accuracy using multi-scale gaussian derivatives combined with an SVM, but they removed ambiguous cases and images with serious illumination problems (423 removed faces). Shan et al. [23] report 89.70% ±0.45 using an Adaboost based technique; however, they manually labeled eye positions which is not practical for many applications. Liu et al. [24] report 92.26% ±0.81 accuracy and also provide the splits used for their evaluation. We therefore use their splits in our evaluation below to permit our technique to be directly compared to their results.

3 Our Models

In this section, we present our technique which we show later is capable of obtaining state of the art results on both the CK+ and GENKI evaluations. We also present a deep neural network approach for expression recognition that we shall use for additional comparisons on the GENKI evaluation.

3.1 High Dimensional Engineered Features

Our high dimensional feature approach is conceptually simple. We extract a form of local binary pattern known as the Census transform for each pixel found within small image patches, each centered on a facial keypoint. Unlike previous work which typically creates histograms of LBPs, here we create our feature vector by concatenating the bits for each pixel of the image patch into a binary vector. We also concatenate bits obtained from patches extracted at multiple scales centered on the keypoints. As far as we are aware this is different from previous uses of LBP techniques which have relied on histogramming operations. This high dimensional binary feature vector is then projected into a smaller dimensional space via principal component analysis (PCA), followed by a recently proposed variation of multiclass Fisher Discriminant Analysis (FDA) known as local FDA or LFDA [11]. The resulting vector is then used within a Support Vector Machine (SVM). There are a number of choices to be made throughout this processing and classification pipeline and we search over key subsets of these choices using cross validation techniques. We discuss the different steps of our procedure in more detail below.

The Census Transform. The Census transform [10] is computed as follows. If $\mathbf{p} = \{u, v\}$ is the index of a pixel and $I(\mathbf{p})$ is its intensity, define $\xi(\mathbf{p}, \mathbf{p}') = 1, \text{if } I(\mathbf{p}') < I(\mathbf{p})$; otherwise $\xi(\mathbf{p}, \mathbf{p}') = 0$. The Census transform simply concatenates the bits obtained from comparisons using a fixed ordering of pixels within spatial neighborhood around the pixel. The result is a bit string with ones representing the pixels that are less than the value of the central pixel. Using $\bigotimes$ to denote concatenation, the census transform for the pixel at location $\mathbf{p} = \{u, v\}$ is simply

$$I^c(\mathbf{p}) = \bigotimes_{j=-n}^{n} \bigotimes_{i=-m}^{m} \xi(I(u, v), I(u + i, v + j)), \tag{1}$$

typically computed using a window of size $(2m+1) \times (2n+1)$. In other words, for a given image patch the CT simply compares each pixel with the center pixel. If its value is greater than the center pixel's value it assigns 0 and 1 otherwise. Common window sizes are 3 and 5. In our experiment, we used 3 as the window size which allows the information to be stored in an 8-bit binary number if desired. The ability to store such descriptors using a binary encoding means that even if our descriptor is of extremely high dimension the information can be stored in a highly compact format. Various other operations using these types of binary descriptors can also be implemented very efficiently.

Keypoint Guided Feature Extraction. As outlined above, we construct our descriptors by cropping small patches out of the larger facial image, applying the Census transform to each pixel for each patch and concatenating the resulting bits into a high dimensional vector. In our experiment below, each scale yields 19,992 features for CK+ and 4,312 for GENKI, due to the different number of keypoints produced by different methods. Patches are extracted centered on

each landmark, excluding the face contour. The patches have two parameters that are optimized by cross validation: patch width, defined in proportion to face size and the patch size. The optimal values for our initial CK+ experiment for example were 2/5ths of the face size and 9 pixels in width respectively. Each cropped patch is also resized before computing the Census transform allowing us to control both the dimensionality and the size or spatial extent of the patch separately. We will also present experiments where we extend this approach by extracting patches at each keypoint at three different scales. Depending on the experiment this produces about 60k features.

To obtain keypoints there are a variety of automated placement techniques which can be applied depending on the circumstances. For example, the CK+ dataset comes with landmark positions that were estimated by fitting an Active Appearance Model (AAM) [25]. AAMs can yield state of the art performance when labeled keypoints have been provided to train models for each subject of interest. For our first set of experiments we use the landmarks provided with the CK+ data. However, AAMs yield poor performance when per subject training data is unavailable. In many real world situations it is impractical to label keypoints for each subject. For this reason there has been a great deal of recent activity focused towards improving alternative approaches that are not identity dependent. For our second CK+ experiments we use the structured max margin technique of [26]. For GENKI experiments we use the convolutional neural network cascade technique in [27].

Dimensionality Reduction. As we shall see in our experimental work, our high dimensional Census feature technique can yield encouraging results on the CK+ evaluation. However, Working with high dimensional vectors can be impractical for many applications. We therefore employ a two phase dimensionality reduction procedure based on an initial projection using PCA followed by LFDA [11]. LFDA obtains a discriminative linear projection matrix through minimizing an objective function of the same form as FDA. The underlying problem is therefore also equivalent to solving a generalized eigenvalue problem. More precisely, a projection matrix $\mathbf{M}$ is obtained from

$$\arg\max_{\mathbf{M}} \mathrm{Tr}\Big\{(\mathbf{M}^T\mathbf{S}_W\mathbf{M})^{-1}\mathbf{M}^T\mathbf{S}_B\mathbf{M}\Big\}, \tag{2}$$

where there are $i = 1, \ldots, n$ feature vectors $\mathbf{x}_i$ with class labels C_i, given by $c = 1, \ldots, n_c$ class indices, and

$$\mathbf{S}_W = \frac{1}{2}\sum_{i,j=1}^{n} \mathbf{W}_{i,j}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \tag{3}$$

which defines a *local* within-class scatter matrix using

$$\mathbf{W}_{i,j} = \begin{cases} \mathbf{A}_{i,j} & C_i = C_j = c \\ 0 & C_i \neq C_j, \end{cases} \tag{4}$$

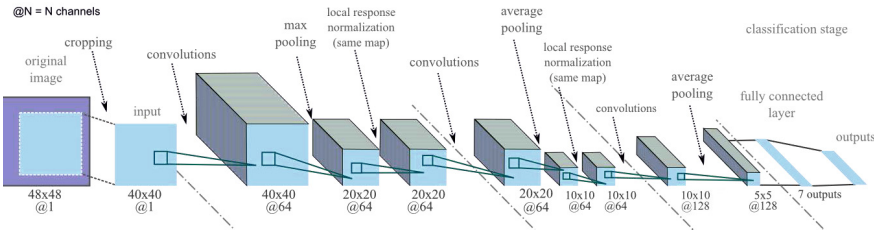


Fig. 1. The architecture of the convolutional neural network used in our experiments

and a *local* between-class scatter matrix defined by

$$\mathbf{S}_B = \frac{1}{2} \sum_{i,j=1} \mathbf{B}_{i,j} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T, \tag{5}$$

where

$$\mathbf{B}_{i,j} = \begin{cases} \mathbf{A}_{i,j} \left(\frac{1}{n} - \frac{1}{n_c}\right) & C_i = C_j = c \\ \frac{1}{n} & C_i \neq C_j, \end{cases} \tag{6}$$

and for both types of local scatter matrix one uses an affinity matrix $\mathbf{A}$ defined, for example by

$$\mathbf{A}_{i,j} = \exp(\|\mathbf{x}_i - \mathbf{x}_j\|^2). \tag{7}$$

3.2 A Deep Convolutional Neural Network Approach

We shall also compare with a deep convolutional neural network approach to expression recognition based on the framework presented in [28] which was used to win the recent ImageNet challenge. The particular architecture we used here for expression recognition is shown in Fig. 1. A similar deep neural network architecture and training approach for expression recognition in the wild was used in [18] to win the recent Emotion Recognition in the Wild Challenge [29] where the goal was to predict expressions in short clips from movies. In [18] the deep network was only trained on the Toronto Face Database TFD [30] - a large set of different standard expression datasets including Cohn-Kanade and a dataset mined from Google image search results [31] containing 35,887 images tagged with the corresponding emotion categories. In contrast for our GENKI experiments here we do not use additional training data.

Since this implementation and architectural variants of it have won a number of competitive challenges we believe the approach is representative of a state of the art deep neural network approach for expression recognition with wild imagery. We therefore use it here to provide a point of comparison for our GENKI experiments.

4 Experiments and Results

Here we provide two sets of experiments. First, we present experiments using the standard CK+ evaluation and our high dimensional feature technique. We examine in particular the sensitivity of our approach to keypoint localization quality, the results of which partly motivated the development of the multi-resolution extensions to our basic approach - making it more robust to inaccurate keypoints. We then present results for the smile detection problem using the GENKI-4K dataset, comparing with the deep convolutional neural network approach presented above.

For our last CK+ experiment with noisy keypoints and for our GENKI experiment we apply our full approach in which multi-scale patches are extracted and feature descriptors are reduced in dimensionality using LFDA. The dimensionality reduction is applied on a per patch basis. For PCA we search in the region of dimension reductions that capture 95% of the variance. For LFDA we search in the region of reductions that reduce the final output to 5-20% of the original dimensionality. It is interesting to note that the multi-scale descriptor has about 60k dimensions for our CK+ experiment and is reduced to about 6k dimensions.

4.1 Experiments on CK+

The CK+ database [6,7] is a widely used benchmark for evaluating emotion recognition techniques. It is perhaps more precise to characterize the emotion recognition task using CK+ as facial expression recognition since the majority of sequences were acted. The evaluation includes image sequences with 6 basic expressions. Each sequence starts with a neutral face and ends with an image showing the most exaggerated variation of a given expression. CK+ has large variation in gender, ages and ethnicity. The database consists of 593 image sequences of 123 different subjects and covers both spontaneous and acted expressions. Only one expression "Happy" is spontaneous and it's because some actors smiled during video recordings. CK+ dataset includes labels for expressions, landmarks and labels for the Facial Action Coding System (FACS). We focus here on the expression recognition task.

We use the CK+ data in our work to benchmark and evaluate our approach on a standard dataset before tackling data that is of principal interest to our work in which expressions are exhibited by subjects in natural and spontaneous situations. We begin by placing our high dimensional feature technique in context with the state of the art by showing the complete result of Lucey et al.'s top performing SPTS+CAPP technique discussed in more detail in our literature review [7]. To evaluate our technique performance when high precision keypoints are not available we then show the impact of using realistic keypoint predictions from the keypoint predictor in [26].

High Dimensional Binary Feature Vectors. For our first experiment here we created a high dimensional binary vector from densely sampled keypoint locations as discussed in section 3. We give the resulting vector to a linear support vector machine using the implementation in [32]. We perform leave one subject

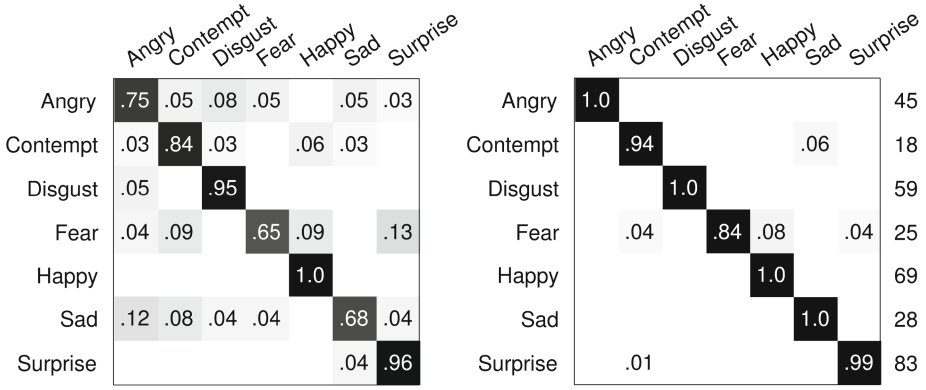


Fig. 2. (**left**) A confusion matrix for expression detection from the SPTS + CAPP result of Lucey et al. [7]. The average per class recognition rate was 83.3%. The matrix is row normalized as in [7]. (**right**) The confusion matrix for expression detection on CK+ using our high dimensional binary features. The average per class accuracy is 96.8%. The overall average accuracy is 98.2%. We give the number of examples per class in the column on the right.

out experiments and optimize hyperparameters using an inner cross validation procedure within the training set. Results are shown in Fig. 2 (right). We are aware of no other published result with higher performance. The best result of which we are aware on CK+ also gives an accuracy of 96% [21]; however, they exclude five subjects from their evaluation. Table 1 provides comparison of our results to other methods.

The Impact of Noisy Keypoints. As we have discussed, in many practical situations it is not possible to obtain highly accurate keypoints such as those possible when using an AAM trained on labeled examples of each subject. For this reason we perform the same experiment above but using the keypoint detector of [26]. As seen in Fig. 3 (left), there is a drop in performance (i.e. 90.0% vs 96.8%), but it is not as dramatic as one might expect due in part to the improved quality for subject independent keypoint predictions afforded by [26].

The Impact of Multiscale Patches. We then evaluated the hypothesis that the use of multiscale patches centered on each keypoint could make the approach more robust to keypoint localization errors. The result of this experiment is shown in Fig. 3 (right). While we cannot recover the original performance, we do see a slight boost in performance over the original single resolution technique.

4.2 Smile Detection Experiments

The GENKI-4K dataset [9,33] consists of 4,000 facial images labelled with pose and smile content. The images are relatively low resolution and in jpeg for-

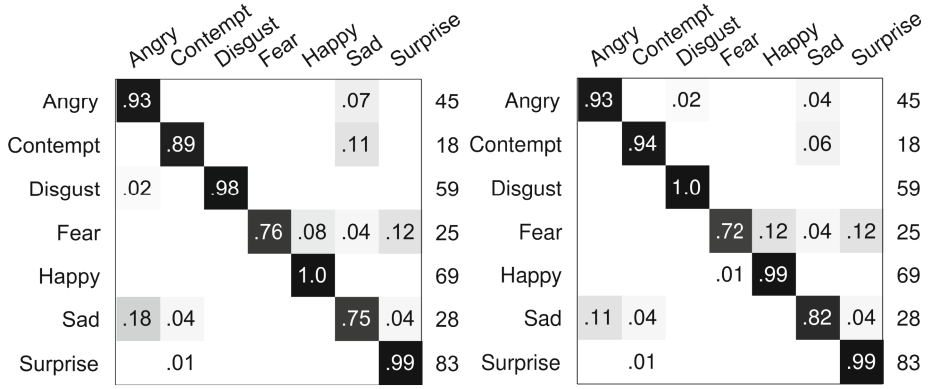


Fig. 3. (**left**) Confusion matrix for expression detection on CK+ using our high dimensional binary features, but based on less accurate keypoints. The average per class accuracy is 90.0%. The overall average accuracy is 93.4%. (**right**) The average per class accuracy when using our multi-scale strategy increases to 91.3% as does the average accuracy, which increases to 94.5%.

Table 1. CK+ Experiments: Comparison and summary

Method	%
Lucy et al. (2010) [average accuracy] using a landmark based representation and appearance features [7]	83.33
Sikka et al. (2012) [average accuracy] LBP histogram architecture [14,17]	82.38
Sikka et al. (2012) [average per subject accuracy] bag of words [17]	95.85
Our technique [average accuracy], accurate keypoints	96.8
Our technique [average class accuracy], accurate keypoints	98.2
Our technique [average accuracy], noisy keypoints	94.5

mat. This dataset has large variations in pose, illumination and ethnicity. We extracted faces from the original images using a combination of the opencv's Haar cascade face detection [34] and the convolutional neural network cascade of [27]. Where these detectors failed to detect any face, we just kept the original.

The resolution of imagery in this dataset was such that we were only able to detect a set of 5 keypoints reliably for our high dimensional feature technique. In order to cover the whole face we computed 6 more points located between eyes, mouth corners and the nose. We provide a comparison with the convolutional neural network (Convnet) architecture discussed in section 3.2, which does not rely on keypoints. For both our high dimensional feature technique and our ConvNet experiments we split the dataset into 4 equal folds using the precise splits defined in [24].

For each experiment with the convolutional neural network, we used random cropping with a 4-pixel border for 48×48 images. Also images were flipped

horizontally with a probability of 0.5 at each epoch. The model with no pre-processing yielded 91.5% 1-fold accuracy. We explored preprocessing with isotropic smoothing [35,36], yielding 91.5%, and histogram equalization on the grayscale imagery, which yielded 91.7%. From these experiments we found that these preprocessing options did not alter performance in a substantial way. We therefore ran a full four fold experiment using grayscale faces with no pre-processing at 96×96 pixel resolution, which yielded 92.97% ±0.71 accuracy.

Using our complete high dimensional feature technique consisting of both the initial feature construction and including the use of multi-resolution patches and the local fisher discriminant analysis step, followed by the application of an SVM with radial basis function kernel for the final classification, we were able to achieve 93.2% ±0.92 average accuracy. We place our results here in context with prior work in Table 2.

Table 2. GENKI-4K Experiments (Accuracies)

Method	%
Shan et al. (2012), using an Adaboost based technique; however, they manually labeled eye positions [23]	89.70
Jain et al. (2013), using multi-scale Gaussian derivatives combined with an SVM; however, they removed ambiguous cases & images with serious illumination problems (423 faces removed) [22]	92.97
Liu et al. (2013), using HOG features and SSL [24]	92.29
Liu et al. (2013), with only labeled data	91.85
Our ConvNet at 48 × 48 pixel resolution (no preprocessing)	91.5
Our ConvNet at 96 × 96 pixel resolution (±0.71)	93.0
Our high dimensional LBP technique (±0.92)	93.2

5 Final Conclusions and Discussion

It is important to emphasize that traditionally LBP based techniques have used histogramming operations to create underlying feature representations. In contrast, in our work we do not compute histograms and use bits directly. For example previous work [17] has given an accuracy of 82.38% on CK+ for a traditional LBP approach using histograms computed on grid locations defined by a face bounding box using a boosted SVM classification approach. Since we use LFDA to learn a discriminative reduced dimensionality space, our work thus also blurs the lines between traditional notions of engineered feature representations and learned representations. Since we use LBP-like descriptors defined by keypoint locations, in a sense we also blur the lines between keypoint vs. non-keypoint based representations. We hope that our results here will help motivate further work exploring other alternative approaches using LBP descriptors as underlying input representations.

Acknowledgments. We thank Yoshua Bengio, Pascal Vincent, Ian Goodfellow, David Warde-Farley, Mehdi Mirza and David Krueger for helpful discussions. We also thank NSERC and Ubisoft for their support.

References

1. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
3. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 3025–3032. IEEE Computer Society, Washington, DC (2013)
4. Lu, C., Tang, X.: Surpassing human-level face verification performance on LFW with gaussianface. In: Technical report arXiv:1404.3840 (2014)
5. Sikka, K., Dykstra, K., Sathyanarayana, S., Littlewort, G., Bartlett, M.: Multiple kernel learning for emotion recognition in the wild. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI 2013, pp. 517–524. ACM, New York (2013)
6. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)
7. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101 (2010)
8. Eisert, P., Girod, B.: Analyzing facial expressions for virtual conferencing. IEEE Computer Graphics and Applications, 70–78 (1998)
9. GENKI-4K: The MPLab GENKI Database, GENKI-4K Subset. http://mplab.ucsd.edu
10. Zabih, R., Woodfill, J.: Non-parametric local transforms for computing visual correspondence. In: Eklundh, J.-O. (ed.) ECCV 1994. LNCS, vol. 801, pp. 151–158. Springer, Heidelberg (1994)
11. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. The Journal of Machine Learning Research **8**, 1027–1061 (2007)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893 (June 2005)
13. Lowe, D.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157 (1999)
14. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image Vision Comput. **27**(6), 803–816 (2009)

15. Freund, Y., Schapire, R.E.: A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, Paul M.B. (ed.) EuroCOLT 1995. LNCS, vol. 904, pp. 23–37. Springer, Heidelberg (1995)
16. Dahmane, M., Meunier, J.: Emotion recognition using dynamic grid-based HoG features. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pp. 884–888 (March 2011)
17. Sikka, Karan, Wu, Tingfan, Susskind, Josh, Bartlett, Marian: Exploring bag of words architectures in the facial expression domain. In: Fusiello, Andrea, Murino, Vittorio, Cucchiara, Rita (eds.) ECCV 2012 Ws/Demos, Part II. LNCS, vol. 7584, pp. 250–259. Springer, Heidelberg (2012)
18. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, c., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., Mirza, M., Jean, S., Carrier, P.L., Dauphin, Y., Boulanger-Lewandowski, N., Aggarwal, A., Zumer, J., Lamblin, P., Raymond, J.P., Desjardins, G., Pascanu, R., Warde-Farley, D., Torabi, A., Sharma, A., Bengio, E., Côté, M., Konda, K.R., Wu, Z.: Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI 2013, pp. 543–550. ACM, New York (2013)
19. Tang, Y.: Deep learning using support vector machines. CoRR abs/1306.0239 (2013)
20. Rifai, S., Bengio, Y., Courville, A., Vincent, P., Mirza, M.: Disentangling factors of variation for facial expression recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 808–822. Springer, Heidelberg (2012)
21. Jeni, L., Takacs, D., Lorincz, A.: High quality facial expression recognition in video streams using shape related information only. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2168–2174 (November 2011)
22. Jain, V., Crowley, J.: Smile detection using multi-scale gaussian derivatives. In: 12th WSEAS International Conference on Signal Processing, Robotics and Automation, Cambridge, United Kingdom (February 2013)
23. Shan, C.: Smile detection by boosting pixel differences. Trans. Img. Proc. **21**(1), 431–436 (2012)
24. Liu, M., Li, S., Shan, S., Chen, X.: Enhancing expression recognition in the wild with unlabeled reference data. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 577–588. Springer, Heidelberg (2013)
25. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision **60**(2), 135–164 (2004)
26. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886 (June 2012)
27. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 3476–3483. IEEE Computer Society, Washington, DC (2013)
28. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS, pp. 1106–1114 (2012)
29. Dhall, A., Goecke, R., Joshi, J., Wagner, M., Gedeon, T.: Emotion recognition in the wild challenge 2013. In: ACM ICMI (2013)

30. Susskind, J., Anderson, A., Hinton, G.: The toronto face database. Technical report, UTML TR 2010–001, University of Toronto (2010)
31. Carrier, P.L., Courville, A., Goodfellow, I.J., Mirza, M., Bengio, Y.: FER-2013 Face Database. Technical report, 1365, Université de Montréal (2013)
32. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011). http://www.csie.ntu.edu.tw/cjlin/libsvm
33. Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M., Movellan, J.: Toward practical smile detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(11), 2106–2111 (2009)
34. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, vol. 1, pp. I-511–I-518 (2001)
35. Štruc, V., Pavešić, N.: Gabor-based kernel partial-least-squares discrimination features for face recognition. Informatica **20**(1), 115–138 (2009)
36. Štruc, V., Pavešić, N.: Photometric normalization techniques for illumination invariance, pp. 279–300. IGI-Global (2011)
37. Dollár, P.: Piotr's Image and Video Matlab Toolbox (PMT). http://vision.ucsd.edu/pdollar/toolbox/doc/index.html

Weight-Optimal Local Binary Patterns

Felix Juefei-Xu(✉) and Marios Savvides

Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
Felixu@cmu.edu, msavvide@ri.cmu.edu

Abstract. In this work, we have proposed a learning paradigm for obtaining weight-optimal local binary patterns (WoLBP). We first reformulate the LBP problem into matrix multiplication with all the bitmaps flattened and then resort to the Fisher ratio criterion for obtaining the optimal weight matrix for LBP encoding. The solution is closed form and can be easily solved using one eigen-decomposition. The experimental results on the FRGC ver2.0 database have shown that the WoLBP gains significant performance improvement over traditional LBP, and such WoLBP learning procedure can be directly ported to many other LBP variants to further improve their performances.

Keywords: Local binary patterns (LBP) · Weight-optimal local binary patterns (WoLBP)

1 Introduction

In the field of computer vision and pattern recognition, local binary patterns (LBP) and its variants have been widely used throughout many applications. The LBP has gained its prominence due to its discriminative power and computational simplicity. The simple yet very efficient operator labels the pixels of an image (patch) by thresholding the neighborhood of each pixel and converts the result as a binary number.

The LBP was invented in 1992, with the idea that two-dimensional textures can be described by two complementary local measures: pattern and contrast [21]. By separating pattern information from contrast, invariance to monotonic gray scale changes can be obtained. The first published work using LBP for face recognition is done by Ahonen et al. in 2004 [1], where they divided the face image into several regions from which the LBP features are extracted and concatenated into an enhanced feature vector to be used as a face descriptor. Following this, LBP and its variants have been widely used in the field of biometrics for face recognition [27], face detection [3], facial expression recognition [24], gender classification [25], and iris recognition [26]. Recently, Pietikäinen et al. [21] have summarized the state-of-the-art LBP in the field of computer vision and pattern recognition. More face recognition and analysis related work using LBP variants can be found in [4–13].

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 148–159, 2015.
DOI: 10.1007/978-3-319-16181-5_11

Recently, many efforts are devoted to learning optimal local binary patterns for various applications. For example, Lei et al. [14] (an extended work of [15]) have learned discriminant face descriptor by first learning the discriminant image filters; second, soft determining the optimal neighborhood sampling strategy; and third, statistically constructing the dominant patterns. Their method is iterative and relies on 2D-LDA type of formulation, which is quite computationally expensive. Shan [23] uses AdaBoost to select discriminative LBP features for gender classification. Liao et al. [17] again applies AdaBoost to select the most effective uniform multi-scale block LBP for enhanced face recognition. Only this time, the computation is done based on average values of block subregions, instead of individual pixels. Maturana et al. [18] consider the following method. Within any square neighborhood given by r, there are $(2r+1)^2-1$ possible pixel comparisons. They wish to select a subset $\mathbf{n}$ of those comparisons of size S that maximizes the discriminability of the output histograms. To achieve this, an iterative heuristic approach called stochastic hill climbing is adopted for obtaining an approximate solution, since the exact solution is intractable due to the combinatorial nature of the problem.

The related work is either relying on boosting algorithm for the selection of the optimal LBP features or iterative method for solving optimization with heavy computational cost. In this work, however, we propose an inexpensive, closed-form solution for learning weight-optimal local binary patterns (WoLBP), which can be easily extended to many LBP variants and should lead to performance boost. For this very reason, we only benchmark our proposed WoLBP against traditional LBP implementation.

2 Weight-Optimal Local Binary Patterns

In this section, we will first review the formulation of traditional local binary patterns and then detail the formulation of the proposed weight-optimal local binary patterns.

2.1 Traditional Local Binary Patterns

We start by formulating the traditional LBP operator first introduced by Ojala et al. [19]. The basic idea of this approach is demonstrated in Figure 1. Here we have shown both the 3×3 patch and 5×5 patch. All neighbors that have values higher than the value of the center pixel are given value 1 and 0 otherwise. The binary numbers associate with the neighbors are then read sequentially to form an binary bit string. The equivalent of this binary number (usually converted to decimal) may be assigned to the center pixel to characterize the local texture.

The LBP texture for center point (x_c, y_c) can be represented as:

$$LBP(x_c, y_c) = \sum_{n=0}^{L-1} s(i_n - i_c)2^n \tag{1}$$

where i_n denotes the intensity of the n^{th} surrounding pixel, i_c denotes the intensity of the center pixel, L is the length of the sequence, and $s = 1$ if $i_n \geq i_c$, otherwise, $s = 0$. In the case of a $N \times N$ neighborhood, there are $N^2 - 1$ surrounding pixels, so the bit string is of length $N^2 - 1$.

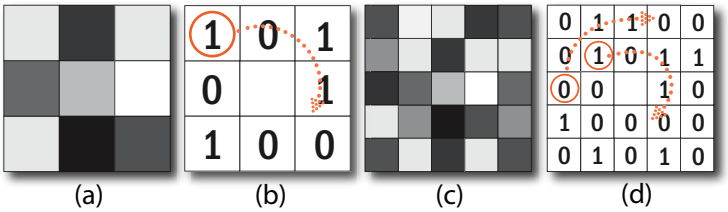


Fig. 1. (a) 3×3 neighborhood, (b) LBP encoding, the 8-bit representation for the center pixel is 10110010 and 178 in decimal, (c) 5×5 neighborhood, and (d) LBP encoding using both radii around the center pixel

During the formulation of the LBP feature, there are many knobs one can play with and result in totally different LBPs. For example, the ordering of the bit string matters if it is converted to a decimal number, the choice of the pivot point (center point), and the choice of bases. More discussions can be found in [13,22].

Varying Base. One can vary the base used for forming the decimal number from the bit string. Instead of using base 2 for conversion as is universally adopted [21], fractional bases (e.g., 1.6, 0.76) or other integer bases (e.g., 3, 4) can also be used. Unleashing the restriction of using only base 2 for decimal conversion, much more diversity can be achieved when encoding LBPs.

Varying Pivot/Center. In the case of 3×3 neighborhood, the center pixel for thresholding neighbors is usually the physical center of the neighborhood. However, one can vary the center in a larger neighborhood as shown in Figure 2. Each pivot (thresholding center) gives different bit string, so varying the center will also provide much more diversity.

Varying Ordering. If the neighborhood size and the thresholding center are both fixed, different ordering of the neighbors (or the weighting of each bit) gives different decimal outputs. One can easily vary the ordering of the neighbors as shown in Figure 2, and thus lead to different formulation of the LBPs.

2.2 Weight-Optimal Local Binary Patterns

All the possible variations mentioned above can be determined empirically, for instance, the choice of center point, the base, and the ordering of the neighbors.

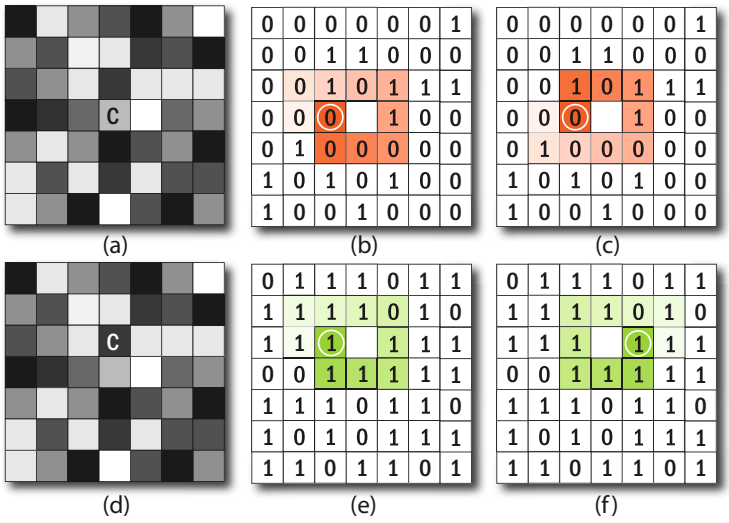


Fig. 2. Example of the LBP encoding scheme [13]: (a) 7×7 neighborhood with the thresholding center denoted as C , (b,c) 2 ordering schemes of (a), (d) 7×7 neighborhood with different thresholding center, and (e,f) 2 possible ordering schemes of (d). Different choices of the thresholding center and ordering schemes result in different LBP code.

In this work, we propose to reformulate the problem of LBP encoding by using a learning framework for obtaining the optimal weights.

First, we need to reformulate the LBP encoding problem into matrix multiplication. The traditional way of encoding LBP feature is to use a 3×3 window to scan through the entire image. At each 3×3 patch, perform the encoding using Equation 1. However, such formulation is neither efficient, nor provides insight towards an optimal weight learning scheme.

Instead of scanning through the entire image using small window and compare the neighborhood values to its center point, a simple convolution of the image with 8 difference masks, followed by simple binarization can achieve the same goal. As shown in Figure 3, we can use 8 difference masks of size 3×3 to convolve with the face image. The 8 resulting bitmaps are shown around the original face image. The traditional LBP is simply a weighted sum of all the bitmaps using the weight vector $\mathbf{w} = [2^7, 2^6, 2^5, 2^4, 2^3, 2^2, 2^1, 2^0]$. Therefore, the reformulation of the LBP can be shown as:

$$\mathbf{y} = \sum_{i=1}^{8} \sigma(\mathbf{h}_i * \mathbf{f}) \cdot \mathbf{w}_i \tag{2}$$

where $\mathbf{f} \in \quad ^d$ is the original image, $\mathbf{h}_i$'s are the difference masks, σ is the binarization operator, and $\mathbf{y} \in \quad ^d$ is the resulting LBP image. Note that only the binarization is non-linear operation.

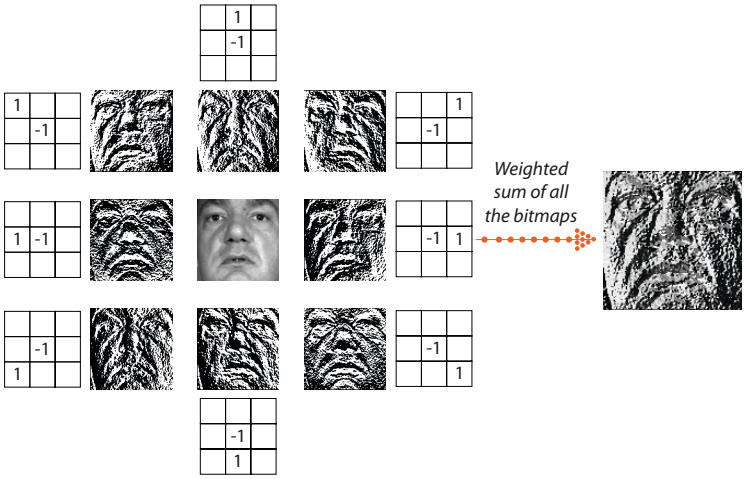


Fig. 3. Reformulation of the traditional LBP encoding using convolution

Now, we are one step closer towards the formulation of WoLBP. In the context of N training images from K classes, we re-arrange them in the following way: the N training images are vectorized and becomes one column in the data matrix $\mathbf{F} \in {}^{d\times N}$, and for each image in $\mathbf{F}$ we apply convolutional mask $\mathbf{h}_1$ to obtain the first bitmap $\mathbf{X}_1 \in {}^{d\times N}$. Then we repeat for $\mathbf{h}_2$ to $\mathbf{h}_8$ to obtain $\mathbf{X}_2$ to $\mathbf{X}_8$. Stacking all $\mathbf{X}_i$'s would give us the new bitmap matrix $\mathbf{X} \in {}^{8d\times N}$. The weight vector $\mathbf{w}$ is now re-written as a weight matrix $\mathbf{\Omega} \in {}^{d\times 8d}$, where $\mathbf{\Omega} = [\mathbf{\Omega}_1, \mathbf{\Omega}_2, \mathbf{\Omega}_3, \mathbf{\Omega}_4, \mathbf{\Omega}_5, \mathbf{\Omega}_6, \mathbf{\Omega}_7, \mathbf{\Omega}_8]$, and $\mathbf{\Omega}_i = \mathbf{w}_i \cdot \mathbf{I}$. In this way, the LBP image for all the N training images can be found in $\mathbf{Y} \in {}^{d\times N}$ using:

$$\mathbf{Y} = \mathbf{\Omega}\mathbf{X} \tag{3}$$

As shown in Figure 4, the weight matrix of the traditional LBP is a horizontal stacking of 8 diagonal matrices, each is a multiple of the identity matrix, and the multiple is defined by the weight vector $\mathbf{w}$. One generalization is as follows. For each of the 8 diagonal weight matrices $\mathbf{\Omega}_i$, we allow the diagonal to take d different values corresponding to the d dimensions of each bitmap. An even further generalization is allowing $\mathbf{\Omega}$ to be a full matrix, as shown in Figure 5, and when multiplied with the bitmap matrix $\mathbf{X}$, the generated LBP image can be somewhat optimal.

Here, the objective of the optimization is to make the LBP images $\mathbf{Y}$ have the best class separation, and thus lead to better classification performance. Fisher ratio is one way to characterize the class separability by simultaneously maximizing the between-class scatter and minimizing the within-class scatter. Note that the only non-linear part within the LBP formulation, binarization, has been taken care of by stacking all the bitmaps in the matrix $\mathbf{X}$, and a linear method is sufficient to learn an optimal weight matrix $\mathbf{\Omega}$.

So we are trying to solve for the following optimization:

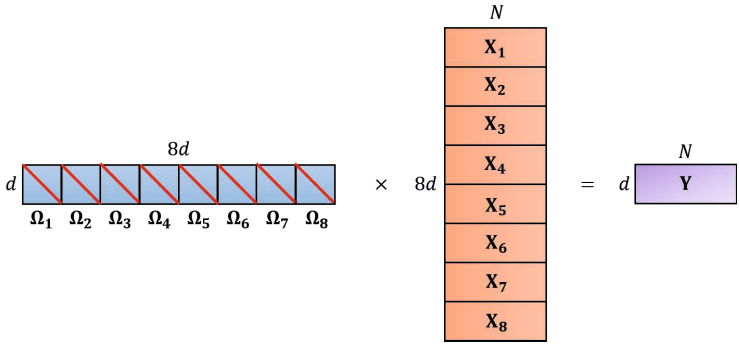


Fig. 4. Traditional LBP in matrix multiplication form

$$\text{maximize} \frac{|\mathbf{S}_b^{\mathbf{Y}}|}{|\mathbf{S}_w^{\mathbf{Y}}|} = \underset{\mathbf{\Omega}}{\text{maximize}} \frac{|\mathbf{\Omega}\mathbf{S}_b^{\mathbf{X}}\mathbf{\Omega}^\top|}{|\mathbf{\Omega}\mathbf{S}_w^{\mathbf{X}}\mathbf{\Omega}^\top|} \tag{4}$$

whose optimality can be found by solving the eigenvalue problem:

$$\mathbf{\Omega}\mathbf{\Lambda} = \left((\mathbf{S}_w^{\mathbf{X}})^{-1}\mathbf{S}_b^{\mathbf{X}}\right)\mathbf{\Omega} \tag{5}$$

where

$$\mathbf{S}_w^{\mathbf{X}} = \sum_{i=1}^{K} \sum_{j \in C_i} (\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^\top \tag{6}$$

$$\mathbf{S}_b^{\mathbf{X}} = \sum_{i=1}^{K} (\mu_i - \mu)(\mu_i - \mu)^\top \tag{7}$$

where μ_i are the mean vectors of all the $\mathbf{x}_i$'s belonging to class i (denoted as C_i), and μ is the global mean vector. $\mathbf{x}_1 \ldots \mathbf{x}_N$ are the columns of bitmap matrix $\mathbf{X}$.

Solving Equation 5 would give the optimal weight matrix $\mathbf{\Omega}$ which leads to the highest Fisher ratio for the LBP image matrix $\mathbf{Y}$. The optimal weight matrix can be seen as a linear transformation matrix that reduces the dimensionality from $8d$ to d. Please note that this WoLBP learning procedure is different from regular Linear Discriminant Analysis (LDA) because in LDA, a transformation matrix $\mathbf{W}$ is learned to reduce the dimensionality of $\mathbf{Y}$ from d to d' where $d' < d$. Whereas in WoLBP procedure, the learning is restricted to feature encoding which maps the dimension from $8d$ to d on the bitmap matrix $\mathbf{X}$. In short, we have carried out feature encoding learning in WoLBP, not subspace learning for images.

3 Experiments

In this section, the effectiveness of the proposed WoLBP is validated in the context of face recognition. We detail the database used in the experiments first, and then the experimental setup and results.

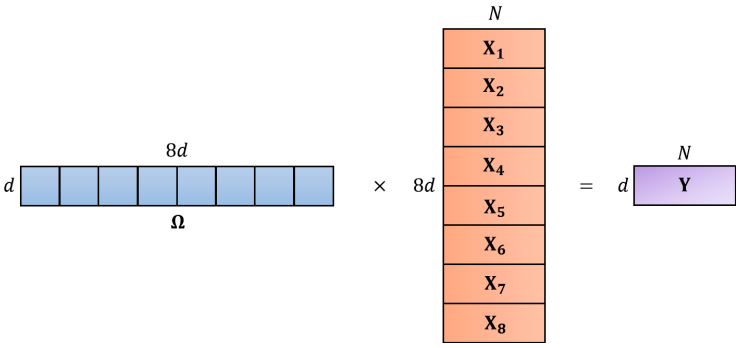


Fig. 5. Allowing $\mathbf{\Omega}$ to be a full matrix and form the WoLBP

3.1 Database

In this work, we utilize the largest frontal face database that is publicly available: NIST's Face Recognition Grand Challenge (FRGC) ver2.0 database [20] to validate the effectiveness of our proposed method, which is also adopted in [17].

The FRGC database is collected at the University of Notre Dame. Each subject session consists of controlled and uncontrolled still images. The controlled full frontal facial images were taken under two lighting conditions under studio setting with two facial expressions. The uncontrolled images were taken in various locations such as hallways, atria, or outdoors under varying illumination settings also with two expressions, smiling and neutral, as shown in Figure 6.

Fig. 6. Examples from the FRGC ver2.0 database: (a1,a2) controlled and uncontrolled still, (b1,b2) cropped full face

The FRGC ver2.0 database has three components: First, the generic **training** set is typically used in the training phase to extract features. It contains both controlled and uncontrolled images of 222 subjects, and a total of 12,776 images. Second, the **target** set represents the people that we want to find. It has 466 different subjects, and a total of 16,028 images. Last, the **probe** set represents the unknown images that we need to match against the **target** set. It contains the same 466 subjects as in target set, with half as many images for each person as in the target set, bringing the total number of probe images to 8,014. All

the **probe** subjects that we are trying to identify are present in the **target** set. **FRGC Experiment 1** is the largest experiment in the FRGC protocol which involves over 256 million face matching comparisons.

One of the latest trends in face recognition community seems to be working on unconstrained dataset such as the LFW [2], with pose variations, occlusions, expression variations, and illumination variations. Though many algorithms have been proposed that can perform fairly well on such datasets, given the complexity of many of these algorithms, it remains unclear as to what underlying objective each of them aim to achieve in the context of unconstrained face matching. Although success on the LFW framework has been very encouraging, there has been a paradigm shift towards the role of such large unconstrained datasets. It has been suggested that the unconstrained face recognition problems can be decoupled into subtasks where one such factor is tackled at a time [16]. Therefore in this work, we focus on a more constrained face recognition paradigm where many such unconstrained factors have been marginalized out already. The findings of this paper can be easily ported towards unconstrained cases where the proposed feature descriptors can further improve the performance of unconstrained face recognition.

3.2 Experimental Setup

In our experiments, we follow the NIST's **FRGC Experiment 1** protocol which involves 1-to-1 matching of 16,028 target images to themselves (∼256 million pair-wise face match comparisons). The WoLBP training is carried out on the generic training set. After obtaining the optimal weight matrix $\mathbf{\Omega}$, it is applied on all the images in the target set. In this experiment, we do not resort to any subspace learning algorithms. The normalized cosine distance (NCD) measurement is adopted to compute similarity matrix between target set images:

$$d(\mathbf{x}, \mathbf{y}) = \frac{-\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}. \tag{8}$$

Compared to other commonly used distance measurement such as ℓ_1-norm, ℓ_2-norm, and the Mahalanobis distance, NCD exhibits the best result.

The result of each algorithm is a similarity matrix with the size of $16,028 \times 16,028$ whose entry SimM_{ij} is the NCD between the feature vector of target image i and target image j. The performance of WoLBP and traditional LBP is analyzed using verification rate (VR) at 0.1% (0.001) false accept rate (FAR), equal error rate (EER), and receiver operating characteristic (ROC) curves.

3.3 Experimental Results

The VR at 0.1% FAR and EER are shown in Table 1. ROC curves are shown in Figure 7 for 32×32 and 64×64 image size respectively. As can be seen, the

WoLBP performs significantly higher than the traditional LBP which has hard-coded encoding scheme. We have also found the same trend on other frontal face databases such as CMU Multi-PIE and YaleB+ database. However, the scale of these databases are no comparison with the FRGC ver2.0 database which involves more than 256 million face matches. Therefore, we do not report the results and ROC curves for those databases for the sake of brevity.

Table 1. VR at 0.1% FAR and EER for the FRGC evaluation

	32×32		64×64	
	VR at 0.1% FAR	***EER***	***VR at 0.1% FAR***	***EER***
WoLBP	0.807	0.040	0.801	0.042
LBP	0.516	0.131	0.496	0.137
Pixel	0.349	0.170	0.350	0.167

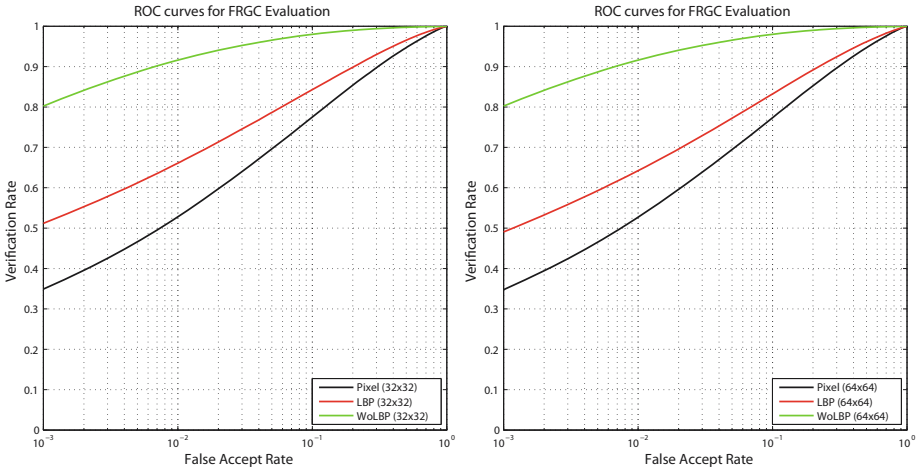


Fig. 7. ROC curves on FRGC face images of size 32×32 (left) and 64×64 (right)

3.4 Discussions

The optimality discussed in this work is solely determined by Fisher ratio. Of course, there can be other optimally obtained local binary patterns via other criteria. It is also worth noted that in order to make Fisher ratio work properly, the homoscedasticity property has to hold, meaning the training data from different classes should all be uni-modal Gaussian distributed with equal covariance. For natural images, this is most likely true. However, readers are encouraged to check

the homoscedasticity property when applying the WoLBP technique discussed in this work.

4 Conclusions

In this work, we have proposed a learning paradigm for obtaining weight-optimal local binary patterns (WoLBP). We first re-formulate the LBP problem into matrix multiplication with all the bitmaps flattened and then resort to the Fisher ratio criterion for obtaining the optimal weight matrix for LBP encoding. The solution is closed form and can be easily solved using one eigen-decomposition. The experimental results have shown that the WoLBP gains significant performance improvement over traditional LBP, and such WoLBP learning procedure can be directly ported to many other LBP variants to further improve their performances.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
2. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07–49, Univ. of Massachusetts, Amherst, October 2007
3. Jin, H., Liu, Q., Lu, H., Tong, X.: Face detection using improved lbp under bayesian framework. In: Proc. 3rd Int'l Conf. on Image and Graphics, pp. 306–309, December 2004
4. Juefei-Xu, F., Cha, M., Heyman, J.L., Venugopalan, S., Abiantun, R., Savvides, M.: Robust local binary pattern feature sets for periocular biometric identification. In: 2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), pp. 1–8. IEEE (2010)
5. Juefei-Xu, F., Cha, M., Savvides, M., Bedros, S., Trojanova, J.: Robust periocular biometric recognition using multi-level fusion of various local feature extraction techniques. In: IEEE 17th International Conference on Digital Signal Processing (DSP) (2011)
6. Juefei-Xu, F., Luu, K., Savvides, M., Bui, T.D., Suen, C.Y.: Investigating age invariant face recognition based on periocular biometrics. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7. IEEE (2011)
7. Juefei-Xu, F., Pal, D.K., Savvides, M.: Hallucinating the full face from the periocular region via dimensionally weighted k-svd. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW). IEEE (2014)
8. Juefei-Xu, F., Savvides, M.: Can your eyebrows tell me who you are?. In: 2011 5th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1–8. IEEE (2011)

9. Juefei-Xu, F., Savvides, M.: Unconstrained periocular biometric acquisition and recognition using cots ptz camera for uncooperative and non-cooperative subjects. In: 2012 IEEE Workshop on Applications of Computer Vision (WACV), pp. 201–208. IEEE (2012)
10. Juefei-Xu, F., Savvides, M.: An augmented linear discriminant analysis approach for identifying identical twins with the aid of facial asymmetry features. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2013)
11. Juefei-Xu, F., Savvides, M.: An image statistics approach towards efficient and robust refinement for landmarks on facial boundary. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2013)
12. Juefei-Xu, F., Savvides, M.: Facial ethnic appearance synthesis. In: European Conference on Computer Vision (ECCV) Workshops. Springer (2014)
13. Juefei-Xu, F., Savvides, M.: Subspace based discrete transform encoded local binary patterns representations for robust periocular matching on nists face recognition grand challenge. IEEE Transactions on Image Processing (2014)
14. Lei, Z., Pietikainen, M., Li, S.: Learning discriminant face descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(2), 289–302 (2014)
15. Lei, Z., Yi, D., Li, S.: Discriminant image filter learning for face recognition with local binary pattern like representation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2512–2517, June 2012
16. Leibo, J.Z., Liao, Q., Poggio, T.: Subtasks of unconstrained face recognition. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP) (2014)
17. Liao, S.C., Zhu, X.X., Lei, Z., Zhang, L., Li, S.Z.: Learning multi-scale block local binary patterns for face recognition. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 828–837. Springer, Heidelberg (2007)
18. Maturana, D., Mery, D., Soto, A.: Learning discriminative local binary patterns for face recognition. In: 2011 IEEE International Conference on Automatic Face Gesture Recognition and Workshops (FG 2011), pp. 470–475, March 2011
19. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51–59 (1996)
20. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Computer Society Conf. on Computer Vision and Pattern Recognition CVPR. vol. 1, pp. 947–954, June 2005
21. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer Vision Using Local Binary Patterns. Springer (2011)
22. Savvides, M., Juefei-Xu, F.: Image matching using subspace-based discrete transform encoded local binary patterns (09 2013). http://www.google.com/patents/US20140212044
23. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recognition Letters **33**, 431–437 (2012)
24. Shan, C., Gong, S., McOwan, P.W.: Robust facial expression recognition using local binary patterns. In: IEEE Int'l Conf. on Image Processing ICIP. vol. 2, pp. 370–376, September 2005

25. Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L.: Gender classification based on boosting local binary pattern. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 194–201. Springer, Heidelberg (2006)
26. Sun, Z., Tan, T., Qiu, X.: Graph matching iris image blocks with local binary pattern. In: Zhang, D., Jain, A.K. (eds.) ICB 2005. LNCS, vol. 3832, pp. 366–372. Springer, Heidelberg (2005)
27. Zhang, H., Zhao, D.: Spatial histogram features for face detection in color images. In: Aizawa, K., Nakamura, Y., Satoh, S. (eds.) PCM 2004. LNCS, vol. 3331, pp. 377–384. Springer, Heidelberg (2004)

Some Faces are More Equal than Others: Hierarchical Organization for Accurate and Efficient Large-Scale Identity-Based Face Retrieval

Binod Bhattarai[1(✉)], Gaurav Sharma[2], Frédéric Jurie[1], and Patrick Pérez[2]

[1] GREYC, CNRS UMR 6072, Université de Caen Basse-Normandie, Caen, France
binod.bhattarai@unicaen.fr
[2] Technicolor, Rennes, France

Abstract. This paper presents a novel method for hierarchically organizing large face databases, with application to efficient identity-based face retrieval. The method relies on metric learning with local binary pattern (LBP) features. On one hand, LBP features have proved to be highly resilient to various appearance changes due to illumination and contrast variations while being extremely efficient to calculate. On the other hand, metric learning (ML) approaches have been proved very successful for face verification 'in the wild', *i.e.* in uncontrolled face images with large amounts of variations in pose, expression, appearances, lighting, *etc.* While such ML based approaches compress high dimensional features into low dimensional spaces using discriminatively learned projections, the complexity of retrieval is still significant for large scale databases (with millions of faces). The present paper shows that learning such discriminative projections locally while organizing the database hierarchically leads to a more accurate and efficient system. The proposed method is validated on the standard Labeled Faces in the Wild (LFW) benchmark dataset with millions of additional distracting face images collected from photos on the internet.

1 Introduction

In the present paper, we address the task of identity-based face retrieval: given a query face image, retrieve the face(s) of the same person from a large database of known faces with large changes in face appearances due to pose, expression, illumination, etc. This task finds numerous applications, particularly in indexing and searching large video archives and surveillance videos and in controlling access to resources.

Many appearance features, based on highly localized pixel neighborhoods, have been proposed in the recent literature [1–4]. All of them attempt to capture the statistics of local pixel neighborhoods using either histograms [1,2,4] or with higher order statistics [3]. While the more expressive features add some extra performance, Local Binary Patterns (LBP) are attractive because of their

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 160–172, 2015.
DOI: 10.1007/978-3-319-16181-5_12

extreme computational efficiency. Such efficiency is especially desirable in the case of limited computational capability *e.g.* embedded systems (see comparisons for LBP computation times on different architectures [5]), or in that of very large datasets *e.g.* millions of faces. In the present paper, we propose to use LBP features as our feature descriptor for the task of large scale identity based face retrieval.

Metric learning based approaches [6–8] have shown that learned low dimensional discriminative projections can be applied for the task of comparing faces with good performances. Such metric learning can be seen as a *global* approach where a single linear projection is learned to discriminate all types of faces. Recently, the SVM-KNN method of Zhang *et al.* [9] has demonstrated (for visual classification task) that learning a collection of *local* (linear) discriminative models leads to better performance. Also, recent Kumar *et al.*'s attribute-based works on facial analysis [10,11] hint towards the presence of local modes in the (attribute transformed) space of faces. In the same way, Verma *et al.* [12] proposed a novel framework to learn similarity metrics using class taxonomies, showing that nearest neighbor classifiers using the learned metrics get improved performance over the best discriminative methods. Inspired by these previous works, we propose to organize large face databases hierarchically using locally and discriminatively learned projections. More concretely, we propose a semi-supervised hierarchical clustering algorithm, alternating between the two steps of (i) learning local projections and (ii) clustering for splitting the faces into sets of more localized regions in face space. Intuitively, we expect such a hierarchical setup to capture coarse differences, *e.g.* gender, at the top levels and then specialize the different projections at the bottom levels to finer differences between the faces. Fig. 1 gives an overview of our approach in contrast to traditional metric learning. One big difference with [10,11] or [12] is that our approach does not need any face taxonomy nor predefined set of attributes. Both are automatically discovered.

In the following, we set the context for our work in §2 and then describe our approach in detail in §3. We discuss our approach in relation to the most closely related works in §3.1. We then give qualitative and quantitative experimental results validating our approach in §4 and conclude the paper in §5.

2 Context and Related Works

Comparing face images of different persons with large variations in appearance, pose, illumination, *etc.*, is a challenging problem. Locally computed features like Local Binary Patterns (LBP), Local Ternary Patterns (LTP) and Local quantized patterns (LQP) have been quite successful to address these kinds of problems [2,13,14]. One of the recent state-of-art methods [15] on Labeled Faces in the Wild (LFW) [16], the most challenging face verfication dataset, computes very high dimensional LBP (of dimension as high as 100k). In the recent years, several other variants of LBP have been introduced for different computer vision tasks (*e.g.* [17–20]). In this paper, we use the standard LBP descriptor for a good efficiency and performance trade-off.

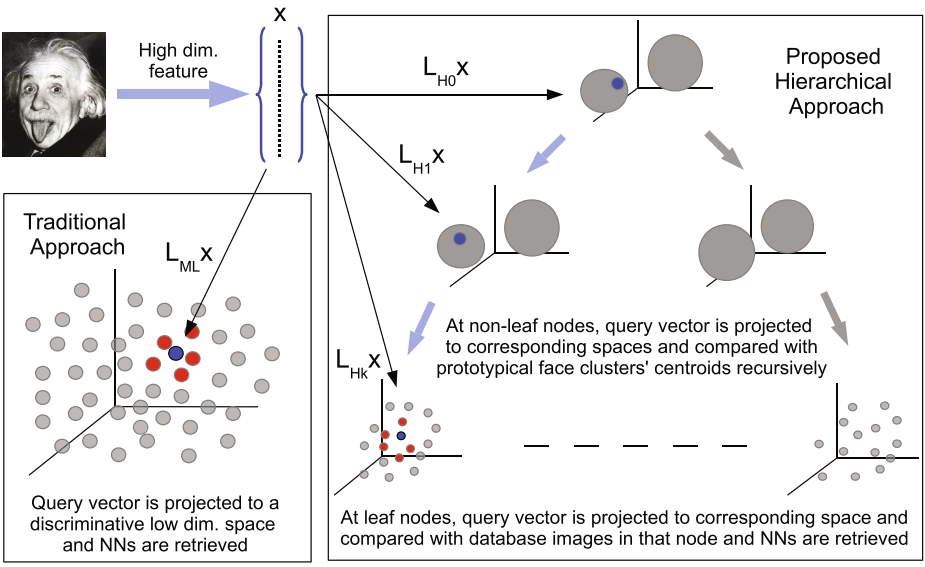


Fig. 1. Principle of the proposed method, in contrast with the traditional metric learning based approach. While the traditional approach learns a single projection (L_{ML}) the proposed approach works hierarchically and learns different projection matrices (L_{H_n}) for different nodes. See §3 for details.

Many other recent papers address the problem with novel approaches, *e.g.* discriminative part-based approach by Berg and Belhumeur [21], probabilistic elastic model by Li *et al.* [22], Fisher vectors with metric learning by Simonyan *et al.* [7], novel regularization for similarity metric learning by Cao *et al.* [23], fusion of many descriptors using multiple metric learning by Cui *et al.* [24], deep learning by Sun *et al.* [25], method using fast high dimensional vector multiplication by Barkan *et al.* [26] or robust feature set matching for partial face recognition by Weng *et al.* [27]. Many of the most competitive approaches on LFW combine different features, *e.g.* [6,28,29] and/or use external data, *e.g.* [10,30].

Metric learning has been recently shown to give good results on very diverse computer vision tasks [31–35]. We refer the reader to Bellet *et al.* [36] for an excellent survey on Metric Learning. More specifically, methods based on metric learning have been reported to improve accuracy for face verification, either on static images [6–8,23] or on videos [37]. The key idea is to learn a Mahalanobis like metric of the form $D^2_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top M(\mathbf{x}_i - \mathbf{x}_j)$, parametrized by the symmetric positive semi-definite (PSD) matrix M, to compare any two faces (described with some features) $\mathbf{x}_i$ and $\mathbf{x}_j$. The learning is based on optimizing some loss function which penalizes high distance between positives and small distance between negative pairs (see [36] for a survey of different metric learning methods/objectives). Since maintaining M as PSD is usually computationally expensive, M is often factorized as $M = L^\top L$. Then the problem can be seen as a linear embedding problem where the features are embedded in the row space

of L and compared with the Euclidean distance there:

$$D_L^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top L^\top L(\mathbf{x}_i - \mathbf{x}_j) = \|L\mathbf{x}_i - L\mathbf{x}_j\|_2^2. \quad (1)$$

Local metric learning, *e.g.* learning a metric as a function of input vector, has also been studied [38]. However, this is expensive, specially in large scale as comparison with every instance will require projecting the query with a different matrix *vs.* only one projection in the case of a global metric.

Closely related to our work, hierarchically organized (metric) learning systems have also been explored in the past, *e.g.* the works by Hwang *et al.* [39], Deng *et al.* [40], Zheng *et al.* [41], Verma *et al.* [12]. However, they assume the presence of a taxonomy (most often a natural semantic taxonomy), while here we do not assume any such information. Our method is also related to clustering in general and with side information in particular [42–45], the side information here being in the form of (sparse) pairwise *must-link* and *must-not-link* constraints. The goal of many of these works is to learn a metric to improve the performance of clustering with an implicit assumption that the constraints relate directly to the clusters. While in the current work, the metric learning with constraints relates to a first level of embedding which can be thought of a person identity space and then the clustering is done in such identity space. So, unlike previous works, it will be normal in our approach that two must-not-link vectors (faces of different persons) get assigned to same cluster as long as these different people share similar facial traits.

We are interested in the problem of comparing faces using learned metrics. In particular, we are interested in identity-based face retrieval with a focus on accuracy and efficiency of the setup for large-scale scenarios, *i.e.* with hundreds of thousands of distractors. As such, in addition to the above mentioned works on facial analysis, our method is also related to the SVM-KNN method of Zhang *et al.* [9] and to works on large scale image retrieval using product quantization of Jégou *et al.* [46]. We postpone discussing our method in the context of these methods to §3.1, after describing our method in the next section.

3 Approach

We work in the semi-supervised scenario where we have some annotated training pairs $\mathcal{A} = \{(\mathbf{x}_i, \mathbf{x}_j), y_{ij}\}$ with $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^D$ being features for face examples i, j resp. (*e.g.*, Local Binary Patterns [1]) and $y_{ij} = 1$ if the image pairs are of the same person and $y_{ij} = -1$ otherwise. We propose to learn a hierarchical organization of the faces for efficient face retrieval. Note that we assume the annotations are sparse, in the sense that only a very small fraction of pairs in the database is annotated.

We aim at exploiting the similarities between faces of different persons. In our hierarchical layout, we would like to first split the faces into groups based on coarse appearance similarities, *e.g.* gender, and then, at finer level, we would like to learn to discriminate between finer details in coarsely similar faces. We

now discuss the case of a binary tree but the method could be applied for arbitrary k-ary trees. We start by taking all the faces into one node and learn a discriminative subspace using margin maximizing metric learning: we minimize a logistic loss function using the recently proposed Pairwise Constrained Principal Components (PCCA) [8] approach. In particular, we solve the following optimization,

$$\min_{L} E(L) = \sum_{\{(i,j)\}} \ell_\beta \left(y_{i,j} (\mathcal{D}_L^2(\mathbf{x}_i, \mathbf{x}_j) - 1) \right), \tag{2}$$

where $\ell_\beta(x) = \frac{1}{\beta} \log(1 + e^{\beta x})$ is the generalized logistic loss,

$$\mathcal{D}_L^2(\mathbf{x}_i, \mathbf{x}_j) = \|L(\mathbf{x}_i - \mathbf{x}_j)\|_2^2 \tag{3}$$

is the distance in the row space of the projection matrix L and sum is taken over all labeled face pairs. The intuition of such metric learning formulation is that we would like to find a subspace (parametrized by the projection matrix L) where the distance between the positive pairs is small and that between the negative pairs is large.

We then obtain the projected features $X_p = LX$, where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ is the matrix of all face features in the database, and use k-means to cluster X_p into two clusters in the projected space. By doing this we hope to cluster the faces based on relatively coarse similarities. Once we have the clustering, we create two child nodes of the root containing only the faces from the two clusters respectively. We then repeat the process at each of the child nodes, working with faces in the current node only. At each node we save the indices of the faces which belong to the node along with the current projection matrix and cluster centroids (for the non-leaf nodes). We continue the process until a certain depth, which is a free parameter, is achieved. Algorithm 1 gives the pseudocode for the learning algorithm.

Once the hierarchical structure is built, the retrieval for a new query face is done by traversing the tree with the following decision rule at each node: if it is a non-leaf node, project the face into its subspace and compare with the centroids and move to the closest child node (recall there is a child node for every cluster). If it is a leaf node, then project the face to its subspace and compare with all the faces in that node (projected onto the same subspace) and return the list of the nearest neighbors. Fig. 1 gives an illustration of the retrieval process.

3.1 Relation with Closely Related Works

Recently, Zhang *et al.* [9] proposed the SVM-KNN method, which for a test example creates on-the-fly a local discriminative support vector machine (SVM) classifier, based on its nearest neighbors. The motivation is that a complex non-linear decision boundary could be approximated with a piece-wise linear decision boundary. Also recently, many works based on 'local' comparisons, *e.g.* attribute based works of Kumar *et al.* [10,11] where the faces are represented as vectors of confidences for the presence of attribute like long hairs, open mouth, *etc.*,

Algorithm 1. Learning local metrics and organizing face database hierarchically.

Input: (i) Set of face features $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$, (ii) Sparse pairwise annotation $\mathcal{A}$, (iii) Height of the tree h, (iv) Dimensions of local projection subspaces at different depths/levels $\{D_0, \ldots, D_h\}$
Initialize: $n \leftarrow 0$, $\texttt{idxs} \leftarrow (1, \ldots, N)$, $\texttt{tree} \leftarrow \emptyset$
$\texttt{queue}.\text{add}(n, \texttt{idxs})$ {Tree construction in a breadth-first manner}
while $n < 2^h - 1$ **do**
  $n, \texttt{idxs} \leftarrow \texttt{queue}.\text{pop}()$
  $\ell \leftarrow \lceil \log_2 n \rceil$ {Current level/depth}
  $L_n \leftarrow \texttt{learn\_metric}(X[:, \texttt{idxs}], \mathcal{A}[\texttt{idxs}], D_\ell)$
  if $\ell < h$ **then**
    $C_1, C_2 \leftarrow \texttt{cluster}(LX[:, \texttt{idxs}], 2)$
    $\texttt{idxs}_1, \texttt{idxs}_2 \leftarrow \texttt{cluster\_assign}(X[:, \texttt{idxs}], C_1, C_2)$
    $\texttt{queue}.\text{add}(n + 1, \texttt{idxs}_1)$
    $\texttt{queue}.\text{add}(n + 2, \texttt{idxs}_2)$
  else
    $C_1, C_2 \leftarrow \emptyset$
  end if
  $\texttt{tree}.\text{add\_node}(\{n, L_n, \texttt{idxs}, C_1, C_2\})$
end while

have been shown to be important. We could imagine that the faces with such attributes would occupy a local region (or perhaps manifold) in the full face space and, thus, the success of such facial analysis system motivates us to work locally in the face space. Also, the success of SVM-KNN reassures us of the merit of a local strategy. In our case, such locality is automatically discovered in a data driven way. In the upper levels of the tree, the Vonoroi cells, corresponding to the clustering in the respective discriminative spaces of the nodes, can be interpreted as such local regions where the faces are similar in a coarse way, *e.g.* one node could be of female faces *vs.* another of that of males. While as we go down the levels we expect such differences to become more and more subtle. We show later that qualitative results support our intuition. Hence, we could hope that concentrating on a local region (towards the bottom of the tree) where faces differ very slightly could help us discriminate better, perhaps even at a cheaper cost.

Another closely related but complementary stream of work is that of product quantization by Jégou *et al.* [46]. They propose to learn, in an unsupervised fashion, very compact binary codes to represent images and do very fast nearest neighbor retrieval at large scale. The key point is that they assume/expect the feature space to be Euclidean. However, face retrieval by directly comparing the image representations with Euclidean distance is not optimal and learning a Mahalanobis metric or equivalently a projection is required. Upon projecting the faces to such a space, Euclidean distance can be used and hence product quantization can be applied. As we have already discussed before, the proposed method can be seen as learning different projections for different local regions, we could use different product quantizations in corresponding different local

regions found by the proposed method. Hence, the proposed method and product quantization are complementary to each other.

Finally, it worth comparing our approach to the recent work of Verma *et al.* [12], who proposed a framework for learning hierarchical similarity metrics using class taxonomies. Interestingly, they show that nearest neighbor classifiers using the learned metrics get improved performance over Euclidean distance-based k-NN and over discriminative methods. Our approach bears similarity with [12] as we also learn a hierarchy of similarity metrics. However, a notable difference is that our approach does not require any taxonomy. This is a big advantage as defining a taxonomy of faces would be more than challenging. Providing sufficient training annotations (*i.e.* sufficient number of faces for each level of the hierarchy) would be another complication.

4 Experimental Results

Metric Used. We are interested in the task of identity based face retrieval, *i.e.* given a query face images, retrieving face(s) of the same person from a large database of known face images. Our objective is to find the same person and so, for us, it suffices if at least one of the retrieved faces is of the same person. In the ideal case, the top ranked retrieved face would be of the same person, but it would make a practical system if the correct face is ranked in the top n images, for a small value of n, as they can be manually verified by an operator. Hence, we propose to evaluate the method for k-call@n [47] (with $k = 1$): the metric is 1 if at least k of the top n retrieved results are relevant. We average this over our queries and report the mean 1-call@n.

Database and Query Set. We use the aligned version [28] of the Labeled Faces in the Wild (LFW) database by Huang *et al.* [16]. The dataset has more than 13000 images of over 4000 persons. In addition to LFW, for large-scale experiments, we add up to one million distractor faces that were obtained by crawling Flickr.com and retaining face detection with high confidences. We select the persons/identities in LFW which have at least five example images and randomly sample one image each from them to use as our query set. We use all the LFW images except the query set to learn our system. The results are reported as the mean performance (1-call@n) over all the queries. All the evaluation is done with LFW annotations and, as the distractor images are from personal image collections from the internet while LFW images are that of well-known/celibrities, it is assumed that the distractors do not have the same identities as the query images.

Image Description. To describe the images we use the Local Binary Pattern (LBP) descriptors of Ojala *et al.* [1]. We use grayscale images and centre crop them to size 170×100 pixels and do not do any other preprocessing. We use the publicly available `vlfeat` [48] library for LBP, with cell size parameter set to 10, of dimension 9860 for a face image.

Baseline Parameter. To set the dimension of the baseline projection matrix we did preliminary experiments, with the standard protocol of LFW dataset, with values in $\{16, 32, 64, 128\}$ and found the performance (verification on LFW test set) saturated for d greater than 32. Hence we fixed the projection dimension to 32.

Tree Parameters. We fixed the learned tree to be a binary tree and also fixed the dimension of the projection at successive levels to differ by a multiplicative factor of 2. Thus, the two parameters for the proposed hierarchical organization are the tree depth and the starting projection dimension. We report experiments with depths of 3 and 4, and with starting projection dimension of 128 and 256, leading to leaf nodes with dimensions 32 (same as baseline) in two cases and 16 (half of baseline) in one case. We discuss further in the §4.2.

4.1 Qualitative Results

Fig. 2 shows some example images from the 16 nodes obtained with a tree of depth 4. The clusters shown correspond to the ordering of the leaf nodes at the bottom, *i.e.* every odd cluster and its next neighbor were grouped together in the previous level in the tree and so on. We can note how similar faces are grouped together successively in the different levels of the tree. Cluster 1–12 are predominantly male faces, cluster 13–16 are females. Cluster 15 seems to specialize to females with bangs (hair over the forehead) and 14 on short hair and smiling females. Cluster 2 seems to have bald (or with very little hair) males who wear glasses while cluster 11 has males with smiling faces. With such semantically interpretable visual qualitative results, we conclude that the method seems to perform an attribute-based clustering.

4.2 Quantitative Results

Fig. 3 shows the performances of the baseline *vs*. the proposed method for three different configurations of (i) starting projection dimension 128 with tree depth 3, denoted '128-d3', (ii) starting projection dimension 128 with tree depth 4, denoted '128-d4', and (iii) starting projection dimension 256 with tree depth 4, denoted '256-d4'.

We note that the different configurations of the proposed method give different time complexities. The 128-d3 and 256-d4 trees have leaf node projection dimensions of 32 (same as baseline) with 4 and 8 leaf nodes respectively while the 128-d4 tree has a projection dimension of 16 with 8 nodes. The time complexity for the proposed method depends on (i) projection and Euclidean distance computation with two centroids at non-leaf nodes (repeated $(h-1)$ times, where h is the height of the tree) and (ii) projection and Euclidean distance computation with all the database vectors in leaf nodes. The leaf nodes have about the same number of database vectors and hence a tree with same leaf node projection dimension (of 32) as baseline but with 4 (8) nodes is expected to be $4\times$ ($8\times$) faster than baseline as the bottleneck in large-scale scenario is the computation of Euclidean distances with a large number of (compressed) database vectors.

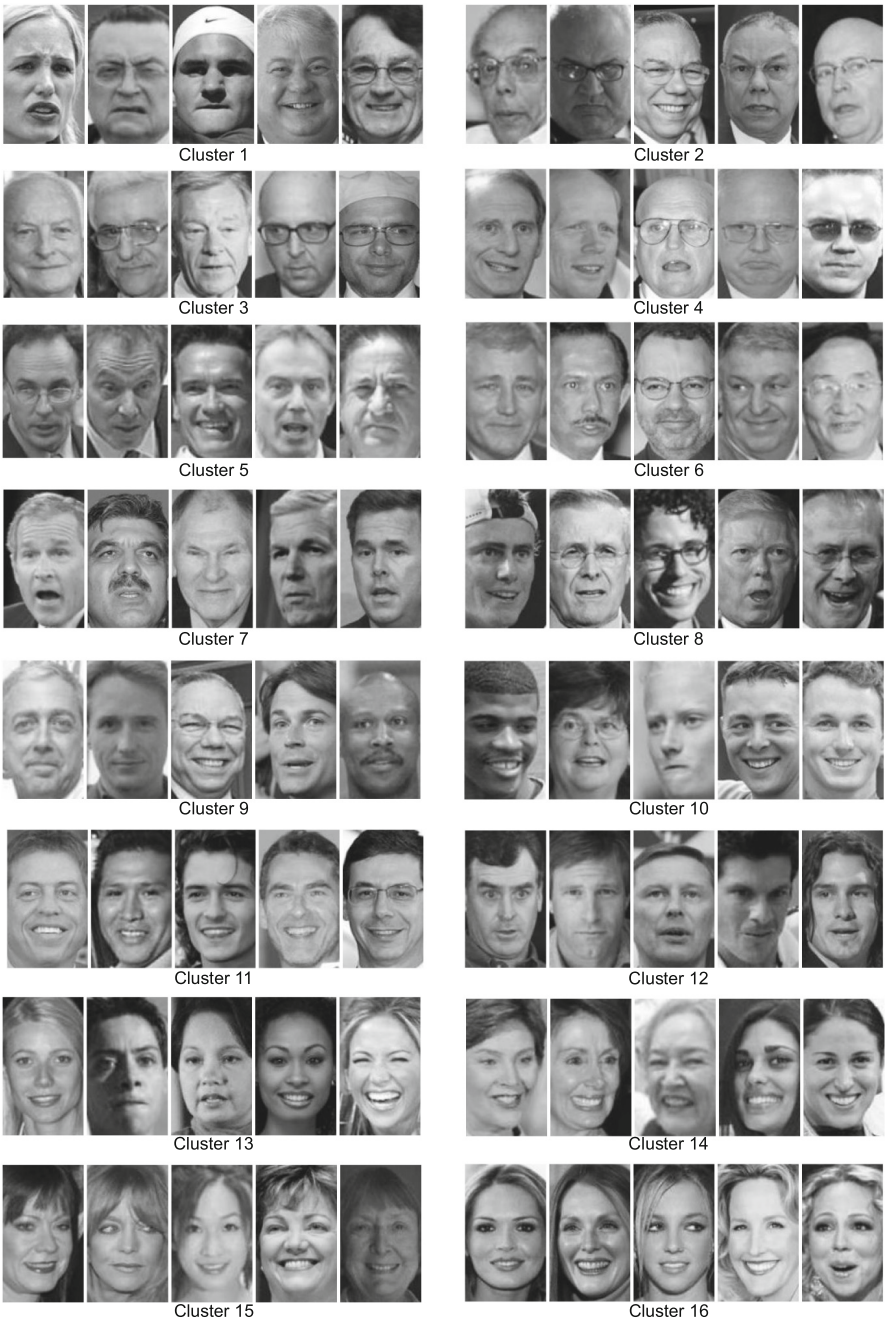


Fig. 2. Visualization the clustering obtained at leaf nodes for a tree of depth 4. The clusters are ordered from left to right and top to bottom, *i.e.* top eight (bottom eight) clusters together form the left (right) node at the first split. Images are randomly selected.

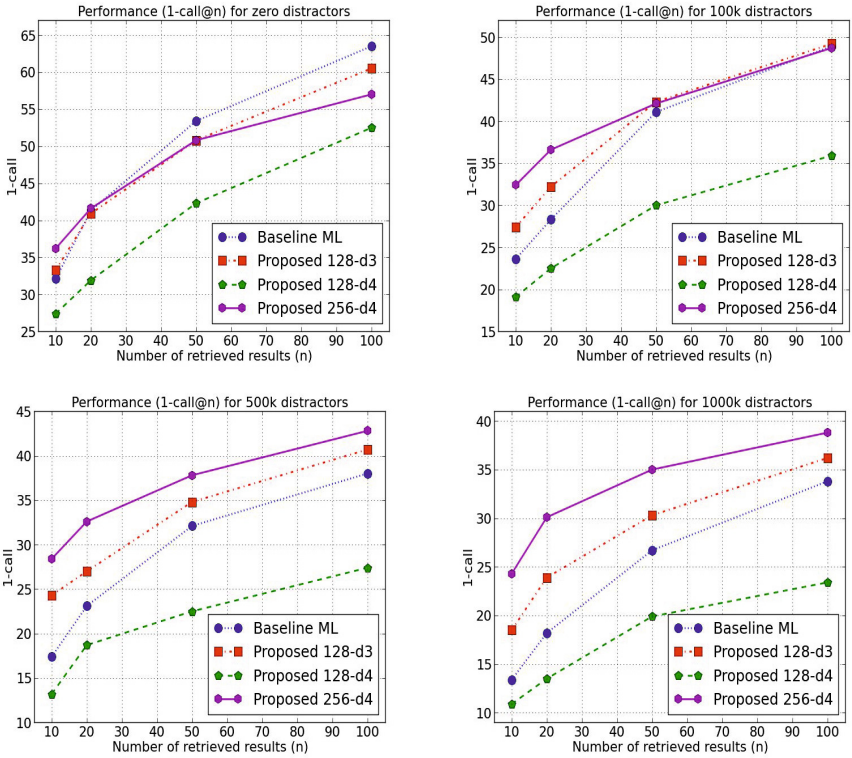


Fig. 3. The performance of the baseline method and that of the proposed method for three different combinations of parameters (starting projection dimension and tree depth) for different numbers of distractors (0, 100k, 500k and 1m) at different operating points

We observe that as more and more distractors are added the proposed method performs better. In the presence of large number of distractors, 100 nearest neighbor are expected to lie in a smaller region around the query points and hence an explanation for the better performance of the method could be that it is better adapted to local neighborhood. In the zero distractor case, we observe that the proposed method is better in the case of small n, *i.e.* it is able to do relatively better retrieval when smaller neighborhoods are considered, while the baseline performs better when n is large and hence larger neighborhoods are considered. The success of the method in the presence of a large number of disctractors underlines the need for locally adapted metrics for identity based face retrieval, especially in a large scale scenario.

Time complexity. The proposed method is expected to be faster in the large scale setting where the number of vectors in the database is greater than the feature dimension. In that case the cost of projecting the query becomes negligible compared to the cost of computing the nearest neighbors in the projected

space. Assuming the database vectors uniformly occupy the leaf nodes, a tree with N leaves is then expected to give an N fold speed-up. We carried out all our experiments on a computer with Intel Xeon 2.8 GHz CPU running linux. Empirically we obtain speedups of about 2.8×, 5.9× and 10.2× for trees with $4, 8$ and 16 nodes respectively, with our unoptimized Python implementation for the experiments with one million distractors, with all computations being timed with data in RAM.

5 Conclusions

We presented a method for accurate and efficient identity based face retrieval, which relies on a hierarchical organization of the face database. The method is motivated by the recent works on local learning of discriminative decision boundaries and of metrics, and works based on attributes. We showed quantitatively that organizing faces hierarchically, with automatically learned hierarchy, leads to an attribute based clustering of faces. Further, we showed quantitatively that the method is capable of better retrieval at a better time complexity compared to the baseline method in large-scale setting.

References

1. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24**(7), 971–987 (2002)
2. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing **19**(6), 1635–1650 (2010)
3. Sharma, G., ul Hussain, S., Jurie, F.: Local higher-order statistics (LHS) for texture categorization and facial analysis. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 1–12. Springer, Heidelberg (2012)
4. Hussain, S.U., Triggs, B.: Feature sets and dimensionality reduction for visual object detection. In: BMVC (2010)
5. López, M.B., Nieto, A., Boutellier, J., Hannuksela, J., Silvén, O.: Evaluation of real-time LBP computing in multiple architectures. Journal of Real-Time Image Processing 1–22 (2014)
6. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: ICCV (2009)
7. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: BMVC (2013)
8. Mignon, A., Jurie, F.: PCCA: A new approach for distance learning from sparse pairwise constraints. In: CVPR (2012)
9. Zhang, H., Berg, A.C., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: CVPR (2006)
10. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV (2009)

11. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Describable visual attributes for face verification and image search. PAMI **33**(10), 1962–1977 (2011)
12. Verma, N., Mahajan, D., Sellamanickam, S., Nair, V.: Learning hierarchical similarity metrics. In: CVPR (2012)
13. Ahonen, T., Hadid, A., Pietikäinen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
14. Hussain, S.U., Napoléon, T., Jurie, F., et al.: Face recognition using local quantized patterns. In: BMVC (2012)
15. Chen, D., Cao, X., Wen, F., Sun, J.: Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In: CVPR (2013)
16. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07–49, University of Massachusetts, Amherst (October 2007)
17. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer vision using local binary patterns, vol. 40. Springer (2011)
18. Heikkilä, M., Pietikäinen, M., Schmid, C.: Description of interest regions with center-symmetric local binary patterns. In: Kalra, P.K., Peleg, S. (eds.) ICVGIP 2006. LNCS, vol. 4338, pp. 58–69. Springer, Heidelberg (2006)
19. Xie, X.: A review of recent advances in surface defect detection using texture analysis techniques. Electronic Letters on Computer Vision and Image Analysis **7**(3), 1–22 (2008)
20. Ojansivu, V.: Blur invariant pattern recognition and registration in the Fourier domain. PhD thesis (2009)
21. Berg, T., Belhumeur, P.N.: POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: CVPR (2013)
22. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic matching for pose variant face verification. In: CVPR (2013)
23. Cao, Q., Ying, Y., Li, P.: Similarity metric learning for face recognition. In: ICCV (2013)
24. Cui, Z., Li, W., Xu, D., Shan, S., Chen, X.: Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In: CVPR (2013)
25. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. In: ICCV (2013)
26. Barkan, O., Weill, J., Wolf, L., Aronowitz, H.: Fast high dimensional vector multiplication face recognition. In: ICCV (2013)
27. Weng, R., Lu, J., Hu, J., Yang, G., Tan, Y.P.: Robust feature set matching for partial face recognition. In: ICCV (December 2013)
28. Wolf, L., Hassner, T., Taigman, Y.: Similarity scores based on background samples. In: Zha, H., Taniguchi, R., Maybank, S. (eds.) ACCV 2009, Part II. LNCS, vol. 5995, pp. 88–97. Springer, Heidelberg (2010)
29. Nguyen, H.V., Bai, L.: Cosine similarity metric learning for face verification. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 709–720. Springer, Heidelberg (2011)
30. Berg, T., Belhumeur, P.N.: Tom-vs-pete classifiers and identity-preserving alignment for face verification. In: BMVC (2012)
31. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D., Ridgeway, G.: Learning a Mahalanobis metric from equivalence constraints. JMLR **6**(6) (2005)
32. Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.: Information-theoretic metric learning. In: ICML (2007)

33. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: CVPR (2007)
34. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: NIPS (2006)
35. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning with application to clustering with side-information. In: NIPS (2003)
36. Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data. arXiv.org (2013)
37. Cinbis, R.G., Verbeek, J., Schmid, C.: Unsupervised metric learning for face identification in TV video. In: ICCV (2011)
38. Wang, J., Kalousis, A., Woznica, A.: Parametric local metric learning for nearest neighbor classification. In: NIPS (2012)
39. Hwang, S.J., Grauman, K., Sha, F.: Semantic kernel forests from multiple taxonomies. In: NIPS (2012)
40. Deng, J., Berg, A.C., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: CVPR (2011)
41. Zheng, L., Li, T.: Semi-supervised hierarchical clustering. In: ICDM (2011)
42. Zeng, H., Song, A., Cheung, Y.M.: Improving clustering with pairwise constraints: a discriminative approach. Knowledge and Information Systems **36**(2), 489–515 (2013)
43. Sublemontier, J., Martin, L., Cleuziou, G., Exbrayat, M.: Integrating pairwise constraints into clustering algorithms: optimization-based approaches. In: ICDMW (2011)
44. Wang, X., Davidson, I.: Flexible constrained spectral clustering. In: SIGKDD (2010)
45. Jain, A.K.: Data clustering: 50 years beyond k-means. PRL **31**(8), 651–666 (2010)
46. Jegou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. PAMI **33**(1), 117–128 (2011)
47. Chen, H., Karger, D.R.: Less is more: probabilistic models for retrieving fewer relevant documents. In: Special Interest Group in Information Retrieval (2006)
48. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008). http://www.vlfeat.org/

On the Effects of Illumination Normalization with LBP-Based Watchlist Screening

Ibtihel Amara[1], Eric Granger[1(✉)], and Abdenour Hadid[2]

[1] École de Technologie Supérieure, Université du Québec, Montreal, Canada
[2] Center for Machine Vision Research, University of Oulu, Oulu, Finland
eric.granger@etsmtl.ca

Abstract. Still-to-video face recognition (FR) is an important function in several video surveillance applications like watchlist screening, where faces captured over a network of video cameras are matched against reference stills belonging to target individuals. Screening of faces against a watchlist is a challenging problem due to variations in capturing conditions (e.g., pose and illumination), to camera inter-operability, and to the limited number of reference stills. In holistic approaches to FR, Local Binary Pattern (LBP) descriptors are often considered to represent facial captures and reference stills. Despite their efficiency, LBP descriptors are known as being sensitive to illumination changes. In this paper, the performance of still-to-video FR is compared when different passive illumination normalization techniques are applied prior to LBP feature extraction. This study focuses on representative retinex, self-quotient, diffusion, filtering, means de-noising, retina, wavelet and frequency-based techniques that are suitable for fast and accurate face screening. Experimental results obtained with videos from the Chokepoint dataset indicate that, although Multi-Scale Weberfaces and Tan and Triggs techniques tend to outperform others, the benefits of these techniques varies considerably according to the individual and illumination conditions. Results suggest that a combination of these techniques should be selected dynamically based on changing capture conditions.

Keywords: Illumination normalization · Local binary patterns · Face screening · Still-to-video face recognition · Video surveillance

1 Introduction

In watchlist screening applications, systems for still-to-video FR are increasingly employed to automatically detect the presence of target individuals of interest for enhanced public security. Accurate and timely responses are required to recognize faces captured under semi-controlled or uncontrolled conditions, as found at various security checkpoint entries, inspection lanes, portals, etc. Under these conditions, face captures incorporate variations due to ambient illumination,

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 173–188, 2015.
DOI: 10.1007/978-3-319-16181-5_13

pose, expressions, occlusion, scale, resolution and blur [2,21], and the performance of FR systems tend to deteriorate. Despite these challenges, it is generally possible to exploit spatiotemporal information extracted from video streams to improve system robustness and accuracy [4,11].

Recent developments in image analysis and recognition have shown that the Local Binary Patterns (LBP) [14] provide a simple yet powerful approach to represent faces for human computer interaction, biometric recognition, surveillance and security, etc. [1,16]. LBP is a gray-scale invariant texture operator which labels each pixel of an image by thresholding its neighborhood pixels with the intensity value of the center pixel. The resulting LBP labels can be regarded as local primitives such as curved edges, spots, flat areas, etc. The histogram of these labels over facial image can be then used as a face descriptor. Given its discriminative power, tolerance to monotonic grey-scale changes, and computational efficiently, LBP has become a well-established technique in FR[1], and has inspired many recent extensions and new research on related methods.

However, it is well known that LBP and other variants are sensitive to severe illumination changes. Variations in facial appearance caused by changes in ambient illumination conditions play an important role in the performance of any FR system applied to video surveillance. It has been shown that face images of different individuals appear more similar than images of the same individual under severe illumination variations [18].

Several techniques have been proposed in the literature for illumination invariant FR [17]. Zou et al. [25] presented a survey of techniques to manage variations in face appearance due to illumination changes using passive and active approaches. Passive approaches focus on the visible spectrum images, where face appearance has been altered by illumination variations, while active ones employ active imaging techniques to capture face images under consistent illumination conditions, or images of illumination invariant modalities.

Among passive techniques, some are specialized at either the pre-processing, the feature extraction, or the classification level [18]. At the pre-processing level, normalization techniques seek to transform facial images such that facial variations induced by illumination are removed. These approaches can be adapted for use with any FR algorithm. Techniques at the feature extraction level seek to achieve illumination invariance by using features or representations that are stable under different illumination conditions. However, some empirical studies have shown that no descriptor can ensure illumination invariant FR in the presence of severe illumination changes. Finally, classification level techniques compensate for the illumination based on the type of face model or classifier employed for FR. Assumptions regarding the effects of illumination on the face model or classifier are employed in counter measures to obtain illumination invariance.

In this study, the performance of several illumination normalization techniques is compared for representation of face captures in still-to-video FR systems using LBP descriptors, as seen in many watchlist screening applications. This empirical study focuses on passive techniques applied at the pre-processing

[1] See LBP bibliography at http://www.cse.oulu.fi/MVG/LBP_Bibliography

level, and compares the performance of a basic FR system that uses representative retinex, self-quotient, diffusion, filtering, means de-noising, retina, wavelet and frequency-based techniques in term of ROC and Precision Recall performance. The benefits of these approaches are assessed using faces captured in the Chokepoint video data set, with individuals walking through an array of cameras located above different portals.

The rest of this paper is organized as follows. Section 2 describes the application focus of this paper which is face screening in video surveillance. Then, Section 3 gives an introduction to the popular LBP approach to face recognition. Section 4 discusses different methods for illumination normalization. The experimental results are presented in Section 5 while a conclusion is given in Section 6.

2 Face Screening in Video Surveillance

Watchlist screening is an important application for decision support in video surveillance systems. It involves still-to-video FR according to the following steps [3]. During enrollment to a watchlist, the segmentation process isolates the regions of interest (ROIs) from reference still images (mugshots) that were previously captured under controlled conditions. Features are extracted and assembled into a discriminant and compact ROI patterns to design facial models[2]. These features are often image-based (e.g., LBP descriptors) or pattern recognition-based (e.g., PCA projections).

During operations, a video stream is captured using some video surveillance camera, and segmentation isolates the ROIs corresponding to faces captured in successive frames. A tracker is often initialized when an emergent ROI is detected far from other faces, and a track is defined to follow the movement or expression of distinct faces across consecutive frames using appearance, position and motion information. Features are extracted into ROI pattern for matching against the facial models of individuals enrolled to the watchlist. A positive prediction is produced if a matching score surpasses an individual-specific threshold. Finally, the decision function combines the tracks and classification predictions in order to recognize the most likely individuals in the scene.

Systems for still-to-video FR are typically modeled in terms of independent detection problems, each one implemented using a template matcher or classifier. These individual-specific detectors are designed with reference face samples from target and non-target individuals (from a cohort or the background model). The advantages of modular architectures with individual-specific detectors include the ease with which face models may be added, updated and removed from the systems, and the possibility of specializing pre-processing, feature extraction, matching and decision thresholds to each specific individual [5,15].

[2] A *facial model* of an individual is defined as a set of one or more reference ROI patterns (used for a template matching system), or parameters estimated from reference ROI patterns (for a classification system).

The performance of state-of-the-art FR systems applied to video surveillance is limited by the difficulty in recognizing facial regions from video streams under semi-controlled and uncontrolled capture conditions (e.g., at inspection lanes, portals and checkpoint entries, in cluttered free-flow scenes at airports or casinos). In particular, performance is severely affected by the variations in ambient illumination, pose, expression, occlusion, scale, resolution, blur and ageing. Still-to-video FR is particularly challenging because very few reference samples are typically available for enrollmentof a person to the system, and because of camera inter-operability – ROIs captured with still cameras (during enrollment) have different properties than those captured with video cameras (during operations). In pattern recognition literature, the situation where only one reference sample is available for system design are often referred to as a "single sample per person" (SSPP) or "one sample training" problem. Techniques specialized for SSPP in FR include multiple face representations, synthetic face generation, and enlarging the training set using an auxiliary set [7]. Note that the still-to-video FR systems from the literature assume that the single face reference is consistent and representative of individuals captures in operational conditions.

3 LBP-Based Face Recognition

The LBP texture analysis operator, introduced by Ojala et al. [14], is defined as a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood. It is a powerful means of texture description and among its properties in real-world applications are its discriminative power, computational simplicity and tolerance against monotonic gray-scale changes.

The original LBP operator forms labels for the image pixels by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. Fig. 1 shows an example of an LBP calculation. The histogram of these $2^8 = 256$ different labels can then be used as a texture descriptor.

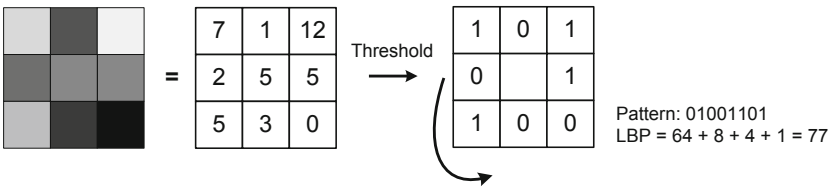


Fig. 1. The basic LBP operator

The operator has been extended to use neighborhoods of different sizes. Using a circular neighborhood and bilinearly interpolating values at non-integer pixel coordinates allow any radius and number of pixels in the neighborhood. The notation (P, R) is generally used for pixel neighborhoods to refer to P sampling

points on a circle of radius R. The calculation of the LBP codes can be easily done in a single scan through the image. The value of the LBP code of a pixel (x_c, y_c) is given by:

$$\mathrm{LBP}_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \tag{1}$$

where g_c corresponds to the gray value of the center pixel (x_c, y_c), g_p refers to gray values of P equally spaced pixels on a circle of radius R, and s defines a thresholding function as follows:

$$s(x) = \begin{cases} 1, \text{ if } x \geq 0; \\ 0, \text{ otherwise.} \end{cases} \tag{2}$$

Another extension to the original operator is the definition of so called *uniform patterns*. This extension was inspired by the fact that some binary patterns occur more commonly in texture images than others. A local binary pattern is called uniform if the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa when the bit pattern is traversed circularly. In the computation of the LBP labels, uniform patterns are used so that there is a separate label for each uniform pattern and all the non-uniform patterns are labeled with a single label. This yields to the following notation for the LBP operator: $\mathrm{LBP}^{u2}_{P,R}$. The subscript represents using the operator in a (P, R) neighborhood. Superscript $u2$ stands for using only uniform patterns and labeling all remaining patterns with a single label.

Each LBP label (or code) can be regarded as a micro-texton. Local primitives which are codified by these labels include different types of curved edges, spots, flat areas etc. The occurrences of the LBP codes in the image are collected into a histogram. The classification is then performed by computing histogram similarities. For an efficient representation, facial images are first divided into several local regions from which LBP histograms are extracted and concatenated into an enhanced feature histogram.

It is known that LBP is sensitive to severe illumination changes. As a consequence, several attempts have been made to overcome this sensitivity. For instance, Tan and Triggs [19] developed a very effective preprocessing chain for compensating illumination variations in face images. It is composed of gamma correction, difference of Gaussian (DoG) filtering, masking (optional) and equalization of variation. This approach has been very successful in LBP-based face recognition under varying illumination conditions. When using it for the original LBP, the last step (i.e. equalization of variations) can be omitted due to LBPs invariance to monotonic gray scale changes.

Aiming at reducing the sensitivity of the image descriptor to illumination changes, a Bayesian LBP (BLBP) was developed by He et al.[6]. This operator is formulated in a Filtering, Labeling and Statistic framework for texture descriptors. In the framework, the local labeling procedure, which is a part of many popular descriptors such as LBP and SIFT, can be modeled as a probability and optimization process. This enables the use of more reliable prior and

Table 1. Illumination normalization techniques studied in this paper

Family	Specific Technique
Retinex	Adaptive Single-Scale Retinex (ASSR), Large and Small-Scale Features (LSSF)
Self Quotient	Multi-Scale Self Quotient (MSSQ)
Diffusion	Isotropic Diffusion (ID), Modified Anisotropic Diffusion (MAD)
Filter	Tan and Triggs (TT)
Gradient	Multi-Scale Weberfaces (MSW)
Mean Denoising	Adaptive Non Local Means (ANLM)
Retina	Retina Modeling (RM)
Wavelet	Wavelet Denoising (WD)
Frequency	Homomorphic

likelihood information, and reduces the sensitivity to noise. The BLBP operator pursues a label image, when given the filtered vector image, by maximizing the joint probability of two images.

Liao et al.[9] noticed that adding a small offset value for comparison in LBP-like methods is not invariant under scaling of intensity values. The intensity scale invariant property of a local comparison operator is very important for example in background modeling, because illumination variations, either global or local, often cause sudden changes of gray scale intensities of neighboring pixels simultaneously, which would approximately be a scale transform with a constant factor. Therefore, a Scale Invariant Local Ternary Pattern (SILTP) operator was developed for dealing with the gray scale intensity changes in complex background. Assuming linear camera response, The SILTP feature is invariant if the illumination is suddenly changed from darker to brighter or vice versa. Besides, SILTP is robust when a soft shadow covers a background region, because the soft cast shadow reserves the background texture information but tends to be darker than the local background region with a scale factor. A downside of the methods mentioned above using one or two thresholds is that the methods are not strictly invariant to local monotonic gray level changes as the original LBP. The feature vector lengths of these operators are also longer.

In order to deal with strong illumination variations, Li et al. developed an active approach combining near-infrared (NIR) imaging with local binary pattern features and AdaBoost learning [8]. The invariance of LBP with respect to monotonic gray level changes makes the NIR images illumination invariant. For instance, the method achieved a verification rate of a FAR=1% on their NIR database with 870 subjects.

4 Illumination Normalization

Changes in ambient illumination, and the resulting variations to facial appearance, are known to significantly deteriorate the performance of FR systems. Accordingly, several techniques have been proposed for illumination invariant FR [17]. Zou et al. [25] presented a survey of techniques according to passive and active approaches. *Passive approaches* focus on the visible spectrum images where face appearance has been altered by illumination variations. They

include illumination variation modelling, illumination invariant features, photometric normalisation, and 3D morphable model techniques. In contrast, *active approaches* employ active imaging techniques to obtain face images captured under consistent illumination condition, or images of illumination invariant modalities. Additional devices (optical filters, active illumination sources or specific sensors) are usually involved to actively obtain different modalities of face images that are insensitive to or independent of illumination change. Those modalities include 3D face information and face images in those spectra other than visible spectra, such as thermal infrared image and near-infrared hyperspatial image.

Passive approaches fall under three main types of techniques to produce illumination invariant facial images – those applied at the pre-processing, feature extraction and classification levels [18]. Pre-processing techniques seek to produce (prior to feature extraction) facial images without facial variations caused by illumination. They compensate for the illumination within any FR system, since no prior assumptions influence feature extraction or classification procedures. They may also be computationally simple, and effective at achieving illumination invariant FR. Feature extraction techniques seek to compensate for appearance variations in facial images using descriptors or representations that are stable under different illumination conditions. However, different empirical studies with LBP, Gabor wavelet-based features, and other descriptors have shown that none of these can ensure illumination invariant FR given severe illumination changes [10]. Classification-level techniques compensate for illumination changes according to the type of face model or classifier employed in the FR system. First, some assumptions regarding the effects of illumination on face models or classification procedure are made, and then based on these assumptions, counter measures are undertaken to obtain illumination invariant face models or illumination insensitive classification procedures. Managing the effects of illumination at the feature extraction level is debatable, while classification level techniques may impose difficult requirements on design data. Although they may provide the more efficient approach to illumination invariant FR, large training set must usually be acquired under a number of lighting conditions and are, furthermore, also computationally expensive.

In this paper, we focus our empirical study on passive techniques for illumination normalization at the pre-processing level. Table 1 presents the specific techniques from the literature that are considered in our study. A more detailed description of these techniques may be found in [18]. They are selected because they are the newer and more representative techniques from different families, e.g., retinex, diffusion, wavelet, frequency-based techniques.

5 Experimental Analysis

5.1 Dataset and Experimental Protocol:

To compare the performance achieved by a still-to-video FR system using different illumination normalization techniques prior to LBP, Chokepoint video

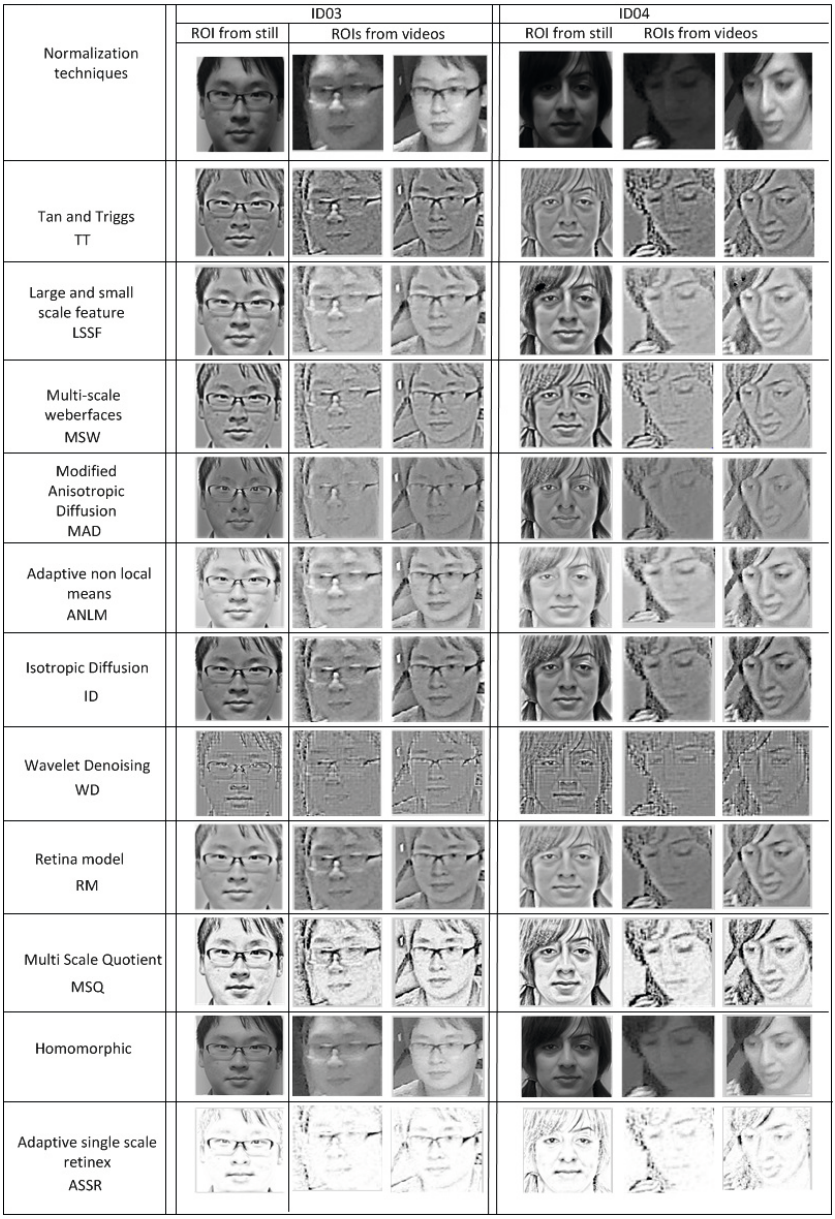


Fig. 2. Examples of face images obtained after illumination normalization is applied to ROIs in stills and videos from individuals ID03 and ID04

dataset [24] has been employed. An array of three cameras is installed above several portals (natural choke points for pedestrian traffic) to capture 25 individuals walking through in a natural way. Videos are challenging for still-to video

FR since faces are captured under semi-controlled conditions, with changes in illumination, pose, scale, blur and occlusion. All 48 video sequences from the center camera, in both entering and leaving cases of Chokepoint have been considered. Cameras have a frame rate of 30 fps and the image resolution is 800 x 600 pixels.

Prior to each replication, 5 persons are randomly selected as target watchlist individuals, where just one reference still image (high-quality neutral mug-shot) is available to design each face model. These reference stills are used a priori to design templates for this FR system. The remaining individuals are used in the testing phase as non-target subjects. The enrollment of each target individual involves isolating a ROI from the reference still image using the Viola-Jones face detection algorithm, and converting the ROI into grey scale, and then cropping it to a common size of 48x48 pixels to limit processing time. For each watchlist individual, the 11 illumination normalization techniques selected for this study (see Table 1) are used to represent the reference ROI using the INface Toolbox[3] [22,23]. At this level, 12 representations of one ROI are created in which 11 represent the normalized ROI in terms of illumination and 1 represents the original ROI (without application of illumination normalization techniques). These representations are shown for individual ID03 and ID04 of Chokepoint in Figure 2.

A division into 3x3 = 9 uniform non-overlapping patches of 16x16 pixels is performed on each ROI representations after illumination normalization. With patch-based methods, facial ROIs are divided into several overlapping or non-overlapping regions called patches, and then features are extracted locally from each patch for recognition purposes. Some specialized decision fusion techniques have been introduced in [13,20] for patch-based FR. In this paper, a uniform pattern of 59 LBP features is extracted from each patch, normalized to range between 0 and 1, and assembled into a ROI pattern of 531 features for matching. The latter are then stored as a template into a gallery. The enrollment phase produces a template gallery with 12 different templates per watchlist person (the original image plus 11 normalized images).

During the testing or operational phase, frames undergo the same processing steps as for enrollment. For each normalization technique, an ROI pattern extracted from a video frame is compared with the corresponding template of the 5 watchlist individuals. Template matching is performed with the Euclidian distance, and produces matching scores.

To assess the transaction-level performance, *receiver operating characteristic (ROC)* space is considered. A ROC curve displays the proportion of target ROIs that are correctly detected as individual of interest over the total number of target ROIs in the sequence, the true positive rate (tpr), as a function of the proportion of non-target (imposter) ROI detected as individual of interest over the total number of non-target ROIs, the false positive rate (fpr). The area under ROC curve (AUC) provides a global scalar measure that can be interpreted as the probability of classification over the range of tpr and fpr. Due to imbalance

[3] http://luks.fe.uni-lj.si/sl/osebje/vitomir/face_tools/INFace/

Table 2. Average pAUC(5%) performance (with standard deviation) for each watchlist individual with illumination normalization techniques

Illumination Normalisation	ID # of Watchlist Individuals					
	ID03	ID04	ID07	ID09	ID12	Average
Entering Videos						
No Normalization	0.66±0.04	0.96±0.01	0.72±0.02	0.84±0.03	0.91±0.02	0.82±0.02
Adaptive Single Scale Retinex	0.65±0.04	0.90±0.01	0.54±0.02	0.76±0.05	0.90±0.01	0.75±0.02
Large and Small Scale Features	0.72±0.06	0.89±0.03	0.69±0.02	0.89±0.03	0.92±0.03	0.82±0.03
Multi Scale Self-Quotient	0.69±0.04	0.88±0.03	0.67±0.05	0.87±0.02	0.93±0.02	0.81±0.03
Isotropic Diffusion	0.69±0.06	0.86±0.03	0.70±0.01	0.90±0.01	0.97±0.01	0.82±0.02
Modified Anisotropic Diffusion	0.74±0.05	0.85±0.03	0.74±0.02	0.80±0.03	0.94±0.02	0.81±0.03
Tan & Triggs	0.74±0.03	0.86±0.03	0.71±0.02	0.88±0.04	0.92±0.03	0.82±0.03
Multi Scale Weberfaces	0.82±0.02	0.83±0.03	0.73±0.03	0.88±0.05	0.91±0.03	0.83±0.03
Adaptive Non-Local Means	0.71±0.02	0.89±0.02	0.66±0.03	0.69±0.04	0.84±0.03	0.76±0.02
Retina Modeling	0.73±0.05	0.85±0.03	0.69±0.02	0.90±0.03	0.91±0.05	0.82±0.03
Wavelet Denoising	0.66±0.03	0.89±0.02	0.54±0.03	0.83±0.02	0.87±0.01	0.76±0.02
Homomorphic	0.62±0.04	0.94±0.01	0.73±0.02	0.81±0.05	0.91±0.01	0.80±0.02
Leaving Videos						
No Normalization	0.67±0.08	0.91±0.03	0.79±0.02	0.91±0.02	0.94±0.02	0.84±0.03
Adaptive Single Scale Retinex	0.73±0.03	0.89±0.02	0.66±0.01	0.89±0.02	0.92±0.01	0.82±0.01
Large and Small Scale Features	0.78±0.03	0.94±0.02	0.54±0.02	0.94±0.02	0.96±0.01	0.83±0.02
Multi Scale Self - Quotient	0.74±0.03	0.82±0.07	0.74±0.02	0.92±0.01	0.93±0.02	0.83±0.03
Isotropic Diffusion	0.82±0.03	0.83±0.05	0.75±0.02	0.91±0.02	0.95±0.01	0.85±0.02
Modified Anisotropic Diffusion	0.78±0.02	0.89±0.02	0.64±0.02	0.93±0.01	0.94±0.01	0.83±0.01
Tan & Triggs	0.80±0.03	0.93±0.01	0.61±0.03	0.97±0.01	0.96±0.01	0.85±0.01
Multi-Scale Weberfaces	0.85±0.03	0.92±0.01	0.73±0.02	0.95±0.01	0.95±0.01	0.88±0.01
Adaptive Non-Local Means	0.74±0.04	0.94±0.02	0.71±0.02	0.86±0.01	0.95±0.01	0.84±0.02
Retina Modeling	0.77±0.03	0.93±0.01	0.55±0.03	0.96±0.01	0.96±0.01	0.83±0.01
Wavelet Denoising	0.71±0.02	0.91±0.02	0.66±0.02	0.87±0.01	0.92±0.01	0.81±0.01
Homomorphic	0.65±0.03	0.90±0.02	0.78±0.01	0.87±0.02	0.91±0.01	0.82±0.01

between target and non-target ROI captures, the precision-recall (PROC) space is also considered to measure the performance. Recall is the *tpr* and the precision is the ratio of correctly detected target ROIs to all target ROIs. The AUPR measures system performance based on targets ROI patterns given an imbalance between target (minority) and non-targets (majority) proportions. In trajectory-level analysis, a tracking module is employed to regroup ROIs captured for a same person over successive frames and to accumulate positive decisions for each person over time. Accumulated predictions are then compared to a detection threshold for a final recognition score. In this paper, we show the matching scores linked to ROI patterns of each person appearing in the scene w.r.t each face model.

5.2 Results and Discussion

Results in Tables 2 and 3 present the average transaction-level performance (pAUC(5%) and AUPR) for each watchlist individual obtained by applying the 11 illumination normalization techniques over all entering and leaving videos of Chokepoint. Based on overall results, MSW and TT techniques tend to outperform the others with both entering and leaving videos.

It can however be observed that the results vary significantly according to the watchlist individual and to capturing conditions (sequence and portals). For

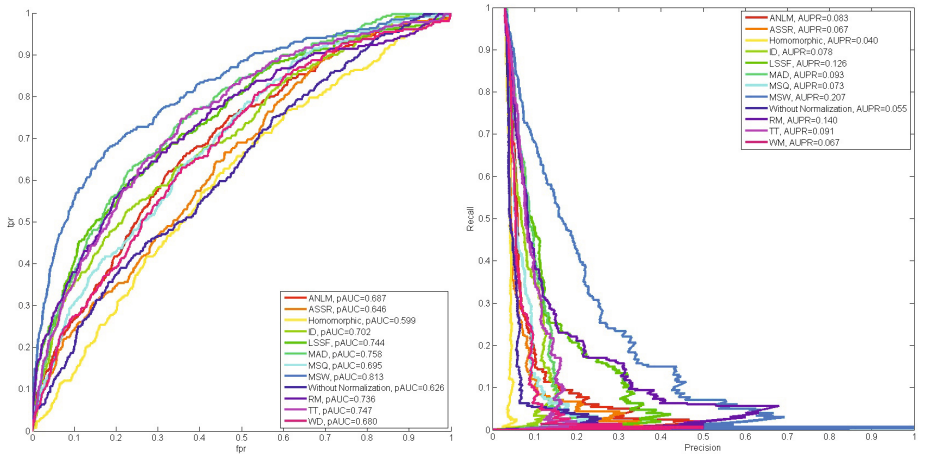
Table 3. Average AUPR performance (with standard deviation) for each watchlist individual with illumination normalization techniques

Illumination Normalisation	ID # of Watchlist Individuals					
	ID03	ID04	ID07	ID09	ID12	Average
Entering Videos						
Without Normalization	0.06±0.01	0.64±0.07	0.16±0.03	0.30±0.08	0.60±0.08	0.35±0.05
Adaptive Single Scale Retinex	0.09±0.03	0.28±0.03	0.06±0.08	0.17±0.04	0.46±0.07	0.21±0.05
Large and Small Scale features	0.18±0.06	0.40±0.04	0.14±0.03	0.51±0.08	0.63±0.01	0.37±0.04
Multi Scale Self-Quotient	0.12±0.05	0.45±0.07	0.15±0.02	0.31±0.04	0.57±0.09	0.32±0.05
Isotropic Diffusion	0.13±0.04	0.31±0.05	0.11±0.01	0.35±0.05	0.74±0.05	0.33±0.04
Modified Anisotropic Diffusion	0.13±0.04	0.34±0.07	0.17±0.02	0.28±0.06	0.68±0.09	0.32±0.05
Tan & Triggs	0.16±0.06	0.37±0.05	0.16±0.02	0.59±0.10	0.64±0.10	0.38±0.06
Multi-Scale Weberfaces	0.25±0.07	0.37±0.05	0.19±0.03	0.58±0.10	0.57±0.11	0.39±0.07
Adaptive Non-Local Means	0.11±0.04	0.51±0.06	0.14±0.02	0.07±0.01	0.53±0.07	0.27±0.04
Retina Modeling	0.22±0.07	0.32±0.05	0.16±0.02	0.63±0.09	0.66±0.10	0.40±0.06
Wavelet Denoising	0.08±0.02	0.37±0.05	0.06±0.01	0.14±0.02	0.32±0.06	0.19±0.03
Homomorphic	0.05±0.01	0.65±0.06	0.18±0.04	0.18±0.05	0.47±0.08	0.30±0.04
Leaving Videos						
Without Normalization	0.19±0.06	0.43±0.07	0.23±0.03	0.57±0.08	0.66±0.07	0.42±0.06
Adaptive Single Scale Retinex	0.14±0.02	0.26±0.06	0.11±0.01	0.41±0.04	0.58±0.03	0.30±0.03
Large and Small Scale features	0.22±0.03	0.49±0.07	0.07±0.01	0.67±0.06	0.71±0.06	0.43±0.04
Multi Scale Self-Quotient	0.11±0.01	0.31±0.09	0.19±0.04	0.59±0.05	0.66±0.04	0.40±0.04
Isotropic Diffusion	0.26±0.05	0.27±0.07	0.21±0.03	0.58±0.06	0.65±0.07	0.40±0.05
Modified Anisotropic Diffusion	0.16±0.03	0.29±0.04	0.08±0.01	0.60±0.07	0.61±0.08	0.35±0.04
Tan & Triggs	0.29±0.05	0.35±0.05	0.10±0.01	0.81±0.04	0.78±0.05	0.47±0.04
Multi-Scale Weberfaces	0.39±0.05	0.34±0.05	0.18±0.03	0.79±0.04	0.78±0.05	0.50±0.04
Adaptive Non-Local Means	0.22±0.05	0.58±0.09	0.17±0.04	0.37±0.04	0.69±0.03	0.41±0.05
Retina Modeling	0.23±0.03	0.38±0.05	0.09±0.02	0.75±0.06	0.68±0.06	0.43±0.04
Wavelet Denoising	0.09±0.01	0.37±0.08	0.09±0.01	0.30±0.03	0.46±0.06	0.26±0.03
Homomorphic	0.10±0.02	0.41±0.08	0.18±0.02	0.44±0.07	0.54±0.08	0.33±0.05

instance, with individual ID04, applying illumination normalization decreases system performance compared to the results without any normalization (see Figure 3(c) and (d)). In contrast, with individual ID03, the pAUC(5%) and AUPR are significantly higher when normalization are applied, specially with the MSW technique (see Figure 3(a) and (b)). Figure 4 displays face representations of individuals ID03 and ID04.

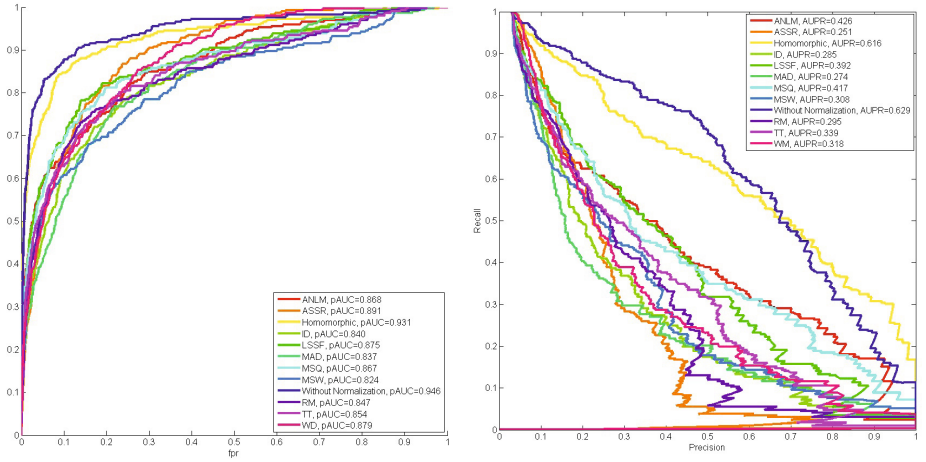
Figures 5 and 6 present an example of trajectory-level analysis with accumulated scores from each target and non-target subject ROIs over time when compared to the template for ID03 and ID04 individuals, respectively. They show matching scores along with measures of brightness and sharpness [12] with MSW and TT normalization associated with each ROI captures in Chokepoint video P1E-S1-C2. In Figure 5, the performance of the FR system that uses MSW normalization yields the best target vs non-target discrimination, although this tends to vary along with brightness and sharpness measures. In Figure 6, the scores are already very high for individual ID04, and normalization only improves non-target scores. This reduced the target vs non-target discrimination, and the overall ROC and Precision-Recall space performance. In this last case, there is no benefit to applying a normalization technique.

In most cases, there is at least one normalization technique that provides an improvement over the case without normalization. Given the diversity of



(a) ROC curves, ID03

(b) Precision-Recall curves, ID03



(c) ROC curves, ID04

(d) Precision-Recall curves, ID04

Fig. 3. Transaction-level performance obtained with individuals ID03 and ID4 after using different illumination techniques

approaches, results suggest that the scores obtained from a set of normalization techniques could be combined through fusion to achieve a higher level of accuracy and robustness. Since there is a correlation between brightness and scores achieved through normalization, a combination of these techniques should be selected dynamically based on changing capture conditions.

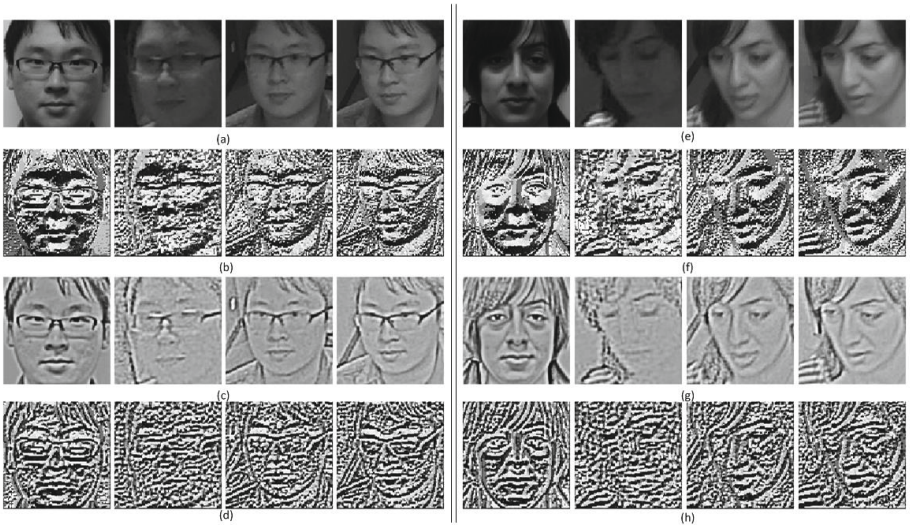


Fig. 4. Face representations of individuals ID03 and ID04. (a) The original ROIs of ID03 (mug-shot and 3 from video captures). (b) LBP projection of the original ROI of ID03. (c) Normalization of ROIs for ID03 using MSW. (d) LBP projection of the MSW normalized images of ID03. (e) The original ROIs of ID03 (mug-shot and 3 from video captures). (f) LBP projection of the original ROI of ID04. (g) Normalization of ROIs for ID04 using MSW. (h) LBP projection of the MSW normalized images of ID04.

6 Conclusion

The popular LBP-based approach to face analysis is known to be sensitive to severe illumination changes. Based on this observation, our study investigated the effect on performance of representative passive illumination normalization techniques for representation of face captures in watchlist screening with LBP. Watch-list screening is an important application for decision support in video surveillance systems.

Extensive experimental analysis on videos from the benchmark Chokepoint dataset indicated that the benefit of different techniques varies considerably according to the individual and illumination conditions. This suggests that a combination of these techniques should be selected dynamically based on changing capture conditions. Overall, the Multi-Scale Weberfaces and Tan and Triggs techniques tend to provide the most interesting results compared to other techniques.

Techniques in this study compensate for illumination changes at the preprocessing level, and may be computationally simple and effective at achieving illumination invariant FR. However, a common challenges among all theses techniques is that performance depends heavily on their implementation, and on the suitable selection of their parameters that must be set empirically. In this study,

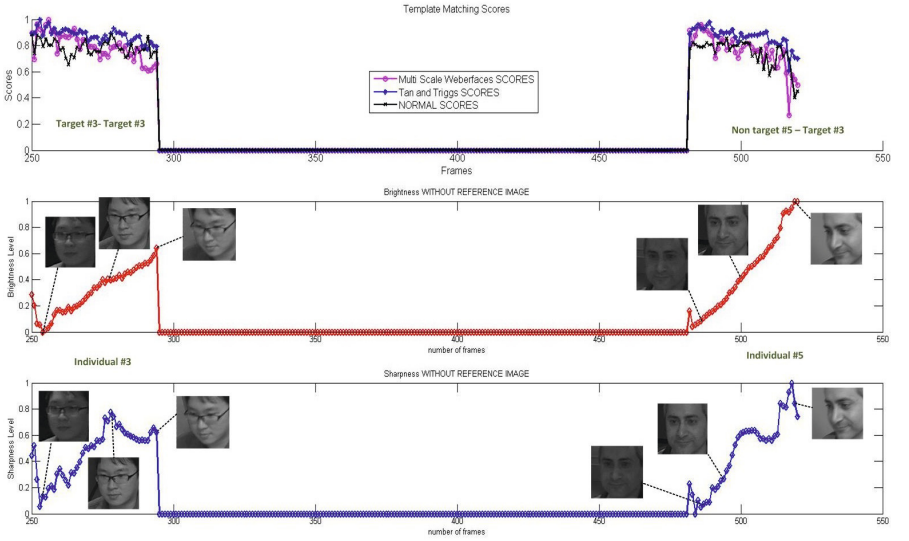


Fig. 5. Trajectory-level analysis for individual ID03 – matching scores and brightness and sharpness levels over time

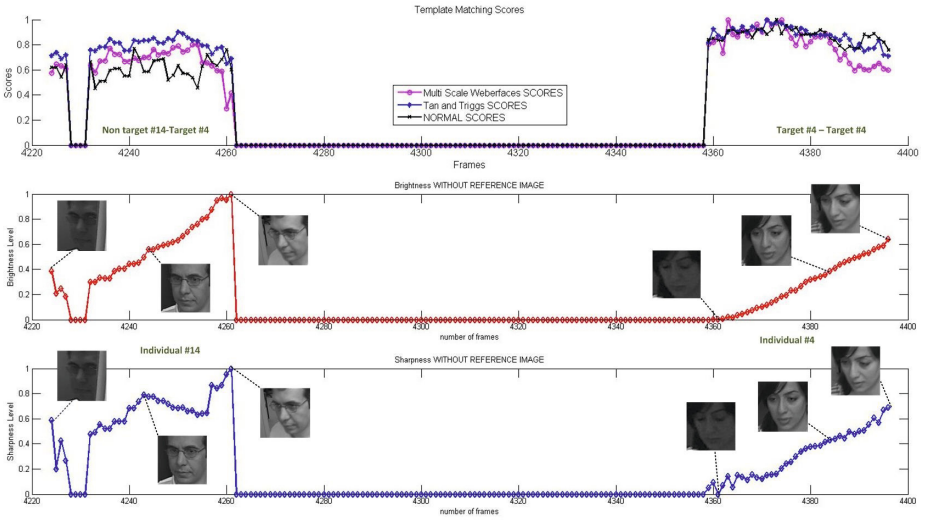


Fig. 6. Trajectory-level analysis for individual ID04 – matching scores and brightness and sharpness levels over time

results were produced using default setting from the authors of respective techniques.

References

1. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. TPAMI **28**(12), 2037–2041 (2006)
2. Barr, J.R., Bowyer, K.W., Flynn, P.J., Biswas, S.: Face recognition from video: A review. International Journal of Pattern Recognition and Artificial Intelligence **26**(05) (2012)
3. Chellappa, R., Sinha, P., Phillips, P.J.: Face recognition by computers and humans. Computer **43**(2), 46–55 (2010)
4. Dewan, M., Granger, E., Roli, F., Sabourin, R., Marcialis, G.L.: A comparison of adaptive appearance methods for tracking faces in video surveillance. In: The 5th International Conference on Imaging for Crime Detection and Prevention, December 16–17 (2013)
5. Ekenel, H.K., Stallkamp, J., Stiefelhagen, R.: A video-based door monitoring system using local appearance-based face models. CVIU **114**(5), 596–608 (2010)
6. He, C., Ahonen, T., Pietikainen, M.: A bayesian local binary pattern texture descriptor. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4 (December 2008)
7. Kan, M., Shan, S., Su, Y., Xu, D., Chen, X.: Adaptive discriminant learning for face recognition. Pattern Recognition **46**(9), 2497–2509 (2013)
8. Li, S.Z., Chu, R., Liao, S., Zhang, L.: Illumination Invariant Face Recognition Using Near-Infrared Images. IEEE T. PAMI **29**(4), 627–639 (2007)
9. Liao, S., Zhao, G., Kellokumpu, V., Pietikainen, M., Li, S.: Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In: 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1301–1306 (June 2010)
10. Marcel, S., Rodriguez, Y., Heusch, G.: On the recent use of local binary patterns for face authentication. International Journal of Image and Video Processing, Special Issue on Facial Image Processing, 469–481 (2007)
11. Matta, F., Dugelay, J.L.: Person recognition using facial video information: a state of the art. Journal of Visual Languages and Computing **20**(3), 180–187 (2009)
12. Nasrollahi, K., Moeslund, T.B.: Face quality assessment system in video sequences. In: Schouten, B., Juul, N.C., Drygajlo, A., Tistarelli, M. (eds.) BIOID 2008. LNCS, vol. 5372, pp. 10–18. Springer, Heidelberg (2008)
13. Nikan, S., Ahmadi, M.: Human face recognition under occlusion using lbp and entropy weighted voting. In: ICPR, pp. 1699–1702. IEEE (2012)
14. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. TPAMI **24**(7), 971–987 (2002)
15. Pagano, C., Granger, E., Sabourin, R., Gorodnichy, D.O.: Detector ensembles for face recognition in video surveillance. In: IJCNN, pp. 1–8. IEEE (2012)
16. Pietikäinen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer Vision Using Local Binary Patterns. Springer (2011)
17. Sharma, A., Kaushik, V.D., Gupta, P.: Illumination invariant face recognition. In: Huang, D.-S., Bevilacqua, V., Premaratne, P. (eds.) ICIC 2014. LNCS, vol. 8588, pp. 308–319. Springer, Heidelberg (2014)

18. Struc, V., Pavesic, N.: Performance evaluation of photometric normalization techniques for illumination invariant face recognition. In: Zhang, Y. (ed.) Advances in Face Image Analysis: Techniques and Technologies. IGI Global, Hershey (2011)
19. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. In: Zhou, S.K., Zhao, W., Tang, X., Gong, S. (eds.) AMFG 2007. LNCS, vol. 4778, pp. 168–182. Springer, Heidelberg (2007)
20. Topcu, B., Erdogan, H.: Decision fusion for patch-based face recognition. In: ICPR, pp. 1348–1351. IEEE (2010)
21. De-la Torre, M., Granger, E., Sabourin, R., Gorodnichy, D.O.: Partially-supervised learning from facial trajectories for face recognition in video surveillance. Information Fusion. Doi:10.1016/j.inffus.2014.05.006 (2014, in Press)
22. Štruc, V., Pavešić, N.: Gabor-based kernel partial-least-squares discrimination features for face recognition. Informatica (Vilnius) **20**(1), 115–138 (2009)
23. Štruc, V., Pavešić, N.: IGI Global (2011)
24. Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C.: Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In: IEEE Biometrics Workshop, Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 81–88. IEEE (June 2011)
25. Zou, X., Kittler, J., Messer, K.: Illumination invariant face recognition: A survey. In: First IEEE International Conference on Biometrics: Theory, Applications, and Systems, 2007, pp. 1–8 (2007)

W09 - Visual Object Tracking Challenge

The Visual Object Tracking VOT2014 Challenge Results

Matej Kristan[1(✉)], Roman Pflugfelder[2], Aleš Leonardis[3], Jiri Matas[4], Luka Čehovin[1], Georg Nebehay[2], Tomáš Vojíř[4], Gustavo Fernández[2], Alan Lukežič[1], Aleksandar Dimitriev[1], Alfredo Petrosino[5], Amir Saffari[6], Bo Li[7], Bohyung Han[8], CherKeng Heng[7], Christophe Garcia[9], Dominik Pangeršič[1], Gustav Häger[10], Fahad Shahbaz Khan[10], Franci Oven[1], Horst Possegger[11], Horst Bischof[11], Hyeonseob Nam[8], Jianke Zhu[12], JiJia Li[13], Jin Young Choi[14], Jin-Woo Choi[15], João F. Henriques[16], Joost van de Weijer[17], Jorge Batista[16], Karel Lebeda[18], Kristoffer Öfjäll[10], Kwang Moo Yi[19], Lei Qin[20], Longyin Wen[21], Mario Edoardo Maresca[5], Martin Danelljan[10], Michael Felsberg[10], Ming-Ming Cheng[22], Philip Torr[22], Qingming Huang[23], Richard Bowden[18], Sam Hare[24], Samantha YueYing Lim[7], Seunghoon Hong[8], Shengcai Liao[21], Simon Hadfield[18], Stan Z. Li[21], Stefan Duffner[9], Stuart Golodetz[22], Thomas Mauthner[11], Vibhav Vineet[22], Weiyao Lin[13], Yang Li[12], Yuankai Qi[23], Zhen Lei[21], and ZhiHeng Niu[7]

[1] University of Ljubljana, Ljubljana, Slovenia
{matej.kristan,luka.cehovin}@fri.uni-lj.si,
{alan.lukezic,frenk.oven}@gmail.com, {ad7414,dp3698}@student.uni-lj.si
[2] Austrian Institute of Technology, Vienna, Austria
{Roman.Pflugfelder,Georg.Nebehay.fl,Gustavo.Fernandez}@ait.ac.at
[3] University of Birmingham, Birmingham, UK
ales.leonardis@fri.uni-lj.si
[4] Czech Technical University, Prague, Czech Republic
Jiri.matas@cmp.felk.cvut.cz
[5] Parthenope University of Naples, Naples, Italy
petrosino@uniparthenope.it, mariomaresca@hotmail.it
[6] Affectv Limited, London, UK
amir@ymer.org
[7] Panasonic R&D Center, Singapore, Singapore
{libohit,hengcherkeng235,yueying53,niuzhiheng}@gmail.com
[8] POSTECH, Pohang, Korea
{bhhan,maga33}@postech.ac.kr
[9] LIRIS, Lyon, France
{christophe.garcia,stefan.duffner}@liris.cnrs.fr
[10] Linköping University, Linköping, Sweden
hager.gustav@gmail.com,
{fahad.khan,kristoffer.ofjall,martin.danelljan,michael.felsberg}@liu.se
[11] Graz University of Technology, Graz, Austria
{possegger,bischof,mauthner}@icg.tugraz.at
[12] Zhejiang University, Hangzhou, China
{jkzhu,liyang89}@zju.edu.cn
[13] Shanghai Jiao Tong University, Shanghai, China
{lijijia,wylin}@sjtu.edu.cn

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 191–217, 2015.
DOI: 10.1007/978-3-319-16181-5_14

[14] ASRI Seoul National University, Gwanak, Korea
jychoi@snu.ac.kr
[15] Electronics and Telecommunications Research Institute, Daejeon, Korea
jwc@etri.re.kr
[16] University of Coimbra, Coimbra, Portugal
{henriques,batista}@isr.uc.pt
[17] Universitat Autonoma de Barcelona, Barcelona, Spain
joost@cvc.uab.es
[18] University of Surrey, Surrey, UK
{k.lebeda,r.bowden}@surrey.ac.uk
[19] EPFL CVLab, Lausanne, Switzerland
kwang.yi@epfl.ch
[20] ICT CAS, Beijing, China
qinlei@ict.ac.cn
[21] Chinese Academy of Sciences, Beijing, China
{lywen,scliao,szli,zlei}@nlpr.ia.ac.cn
[22] University of Oxford, Oxford, UK
cmm.thu@qq.com, philip.torr@eng.ox.ac.uk, stuart.golodetz@ndcn.ox.ac.uk
[23] Harbin Institute of Technology, Harbin, China
qingming.huang@vipl.ict.ac.cn
[24] Obvious Engineering Limited, London, UK
sam@samhare.net

Abstract. The Visual Object Tracking challenge 2014, VOT2014, aims at comparing short-term single-object visual trackers that do not apply pre-learned models of object appearance. Results of 38 trackers are presented. The number of tested trackers makes VOT 2014 the largest benchmark on short-term tracking to date. For each participating tracker, a short description is provided in the appendix. Features of the VOT2014 challenge that go beyond its VOT2013 predecessor are introduced: (i) a new VOT2014 dataset with full annotation of targets by rotated bounding boxes and per-frame attribute, (ii) extensions of the VOT2013 evaluation methodology, (iii) a new unit for tracking speed assessment less dependent on the hardware and (iv) the VOT2014 evaluation toolkit that significantly speeds up execution of experiments. The dataset, the evaluation kit as well as the results are publicly available at the challenge website (http://votchallenge.net).

Keywords: Performance evaluation · Short-term single-object trackers · VOT

1 Introduction

Visual tracking has received a significant attention over the last decade largely due to the diversity of potential applications which makes it a highly attractive research problem. The number of accepted motion and tracking papers in high profile conferences, like ICCV, ECCV and CVPR, has been consistently high

in recent years (∼40 papers annually). For example, the primary subject area of twelve percent of papers accepted to ECCV2014 was motion and tracking. The significant activity in the field is also reflected in the abundance of review papers [22,23,29,40,43,44,65] summarizing the advances published in conferences and journals over the last fifteen years.

The use of different datasets and inconsistent performance measures across different papers, combined with the high annual publication rate, makes it difficult to follow the advances made in the field. Indeed, in computer vision fields like segmentation [18,19], optical-flow computation [3], change detection [24], the ubiquitous access to standard datasets and evaluation protocols has substantially contributed to cross-paper comparison [56]. Despite the efforts invested in proposing new trackers, the field suffers from a lack of established evaluation methodology.

Several initiatives have been put forward in an attempt to establish a common ground in tracking performance evaluation. Starting with PETS [66] as one of most influential performance analysis efforts, frameworks have been presented since with focus on surveillance systems and event detection, e.g., CAVIAR[1], i-LIDS [2], ETISEO[3], change detection [24], sports analytics (e.g., CVBASE[4]), faces, e.g. FERET [50] and [31], and the recent long-term tracking and detection of general targets[5] to list but a few.

This paper discusses the VOT2014 challenge organized in conjunction with the ECCV2014 Visual object tracking workshop and the results obtained. The challenge considers single-camera, single-target, model-free, causal trackers, applied to short-term tracking. The *model-free* property means that the only supervised training example is provided by the bounding box in the first frame. The *short-term* tracking means that the tracker does not perform re-detection after the target is lost. Drifting off the target is considered a failure. The *causality* means that the tracker does not use any future frames, or frames prior to re-initialization, to infer the object position in the current frame. In the following we overview the most closely related work and then point out the contributions of VOT2014.

1.1 Related Work

Recently, several attempts have been made towards benchmarking the class of trackers considered in this paper. Most notable are the online tracking benchmark (OTB) by Wu et al. [62] and the experimental survey based on Amsterdam Library of Ordinary Videos (ALOV) by Smeulders et al. [53]. Both benchmarks compare a number of recent trackers using the source code obtained from the original authors. All trackers were integrated into their experimental environment by

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[2] http://www.homeoffice.gov.uk/science-research/hosdb/i-lids
[3] http://www-sop.inria.fr/orion/ETISEO
[4] http://vision.fe.uni-lj.si/cvbase06/
[5] http://www.micc.unifi.it/LTDT2014/

the benchmark authors themselves and both report carefully setting the parameters. Nevertheless, it is difficult to guarantee equal quality of the parameter setting since, for some trackers, the operation requires thorough understanding.

The OTB [62] contains a dataset containing 50 sequences and annotates each sequence globally with eleven visual attributes. Sequences are not per-frame annotated. For example, a sequence has the "occlusion" attribute if the target is occluded anywhere in the sequence. The evaluation kit with pre-integrated trackers is publicly available. However, in our experience, the integration of third-party trackers into this kit is not straightforward due to a lack of standardization of the input/output communication between the tracker and the evaluation kit.

The ALOV [53] benchmark provides an impressive dataset with 315 sequences annotated with thirteen visual attributes. A drawback of this dataset is that some sequences contain cuts and ambiguously defined targets such as fireworks.

OTB [62] evaluates trackers using two measures: *precision* score and *success* score. Precision score represents the percentage of frames for which the center-distance error (e.g., [33,51]) is below 20 pixels. However, this threshold is strongly affected by the object size, which makes this particular measure quite brittle. A normalized center error measured during successful tracks may be used to alleviate the object size problem, however, the results in [53] show that the trackers do not differ significantly under this measure which makes it less appropriate for tracker comparison. The success plot represents the percentage of frames for which the overlap measure (e.g., [39,58]) exceeds a threshold, with respect to different thresholds. The area under the success plot is taken as an overall success measure. Čehovin et al. [58] have recently shown that this is simply an average overlap computed over the sequence. Alternatively, F-score based on Pascal overlap (threshold 0.5) is proposed in ALOV [53]. Note that the F-score based measure was originally designed for object detection. The threshold 0.5 is also rather high and there is no clear justification of why exactly this threshold should be used to compare trackers [62]. The ALOV [53] proposes an original approach to visualize tracking success. For each tracker, a performance measure is calculated per-sequence. These values are ordered from highest to lowest, thus obtaining a so-called survival curve and a test of statistical significance of differences is introduced to compare these curves across trackers. Special care has to be taken in interpreting the differences between these curves, as the orderings differ between trackers.

Both, the OTB and ALOV initialize the trackers at the beginning of the sequence and let them run until the end. While such a setup significantly simplifies the evaluation kit, it is not necessarily appropriate for short-term tracker evaluation, since short-term trackers are not required to perform re-detection. Therefore, the values of performance measures become irrelevant after the point of tracking failure, which significantly distorts the value of globally computed performance measure. The results are reported with respect to visual attributes in OTB and ALOV for in-depth analysis. However, most visual phenomena do not usually last throughout the entire sequence. For example, consider a tracker that performs poorly on a sequence with attribute occlusion according to a

globally calculated performance measure. This might be interpreted as poor performance under occlusion, but actual occlusion might occur at the end of the sequence, while the poor performance is in fact due to some other effects occurring at the beginning of the sequence.

Collecting the results from the existing publications is an alternative for benchmarking trackers. Pang et al. [48] have proposed a page-rank-like approach to data-mine the published results and compile unbiased ranked performance lists. However, as the authors state in their paper, the proposed protocol is not appropriate for creating ranks of the recently published trackers due to the lack of sufficiently many publications that would compare these trackers.

The most closely related work is the recent visual object tracking challenge, VOT2013 [36]. The authors of that challenge provide the evaluation kit, a fully annotated dataset and an advanced performance evaluation methodology. In contrast to related benchmarks, the goal of VOT2013 was to have as many experiments as possible performed by the original authors of trackers while the results were analyzed by the VOT2013 committee. VOT2013 introduced several novelties in benchmarking short-term trackers: The evaluation kit is cross-platform, allowing easy integration with third-party trackers, the dataset is per-frame annotated with visual attributes and a state-of-the-art performance evaluation methodology was presented that accounts for statistical significance of the results on all measures. The results were published in a joint paper with over 50 co-authors [36], while the evaluation kit, the dataset, the tracking outputs and the code to reproduce all the results are made freely-available from the VOT2013 homepage[6].

1.2 The VOT2014 Challenge

The VOT2014 follows the VOT2013 challenge and considers the same class of trackers. The organisers of VOT2014 provided an evaluation kit and a dataset for automatic evaluation of the trackers. The evaluation kit records the output bounding boxes from the tracker, and if it detects tracking failure, re-initializes the tracker. The authors attending the challenge were required to integrate their tracker into the VOT2014 evaluation kit, which automatically performed a standardized experiment. The results were analyzed by the VOT2014 evaluation methodology.

Participants were expected to submit a single set of results per tracker. Participants who have investigated several trackers submitted a single result per tracker. Changes in the parameters did not constitute a different tracker. The tracker was required to run with fixed parameters on all experiments. The tracking method itself was allowed to internally change specific parameters, but these had to be set automatically by the tracker, e.g., from the image size and the initial size of the bounding box, and were not to be set by detecting a specific test sequence and then selecting the parameters that were hand-tuned to this sequence. Further details are available from the challenge sequence[7].

[6] http://www.votchallenge.net/vot2013/
[7] http://www.votchallenge.net/vot2014/participation.html

The VOT2014 Improves on VOT2013 in Several Aspects:

- A new fully-annotated dataset is introduced. The dataset is per-frame annotated with visual properties, while the objects are annotated with rotated bounding boxes to more faithfully denote the target position.
- Unlike in VOT2013, trackers can predict the target position as a rotated bounding box as well.
- A new evaluation system is introduced that incorporates direct communication with the tracker [59] and offers faster execution of experiments and is backward compatible with VOT2013.
- The evaluation methodology from VOT2013 is extended to take into account that while the difference in accuracy of pair of trackers may be statistically significant, but negligibly small from perspective of ground truth ambiguity.
- A new unit for tracking speed is introduced that is less dependant on the hardware used to perform experiments.
- All accepted trackers are required to outperform the reference NCC tracker provided by the VOT2014 evaluation kit.
- A new web-based system for interactive exploration of the competition results has been implemented.

The remainder of this paper is structured as follows. In Section 2, the new dataset is introduced. The methodology is presented in Section 3, the main results are discussed in Section 4 and conclusions are drawn in Section 5.

2 The VOT2014 Dataset

VOT2013 noted that a big dataset does not necessarily mean richness in visual properties and introduced a dataset selection methodology to compile a dataset that includes various real-life visual phenomena, while containing a small number of sequences to keep the time for performing the experiments reasonably low. We have followed the same methodology in compiling the VOT2014 dataset. Since the evaluation kit for VOT2014 is significantly more advanced than that of VOT2013, we were able to increase the number of sequences compared to VOT2013, while still keeping the time for experiments reasonably low.

The dataset was prepared as follows. The initial pool included 394 sequences, including sequences used by various authors in the tracking community, the VOT2013 benchmark [36], the recently published ALOV dataset [53], the Online Object Tracking Benchmark [62] and additional, so far unpublished, sequences. The set was manually filtered by removing sequences shorter than 200 frames, grayscale sequences, sequences containing poorly defined targets (e.g., fireworks) and sequences containing cuts. Ten global attributes were automatically computed for each of the 193 remaining sequences. In this way each sequence was represented as a 10-dimensional feature vector. Sequences were clustered in an unsupervised way using affinity propagation [21] into 12 clusters. From these, 25 sequences were manually selected such that the various visual phenomena like, occlusion, were still represented well within the selection.

The relevant objects in each sequence are manually annotated by bounding boxes. Most sequences came with axis-aligned bounding boxes placed over the target. For most frames, the axis-aligned bounding boxes approximated the target well with large percentage of pixels within the bounding box (at least $> 60\%$) belonging to the target. Some sequences contained elongated, rotating or deforming targets and these were re-annotated by rotated bounding boxes.

As in the VOT2013, we have manually or semi-manually labeled each frame in each selected sequence with five visual attributes that reflect a particular challenge in appearance attribute: (i) occlusion, (ii) illumination change, (iii) motion change, (iv) size change, (v) camera motion. In case a particular frame did not correspond to any of the five degradations, we denoted it as (vi) neutral. In the following we will use the term *attribute sequence* to refer to a set of frames with the same attribute pooled together from all sequences in the dataset.

3 Performance Measures and Evaluation Methodology

As in VOT2013, the following two weakly correlated performance measures are used due to their high level of interpretability [58]: (i) accuracy and (ii) robustness. The accuracy measures how well the bounding box predicted by the tracker overlaps with the ground truth bounding box. On the other hand, the robustness measures how many times the tracker loses the target (fails) during tracking. A sfailure is indicated when the overlap measure becomes zero. To reduce the bias in robustness measure, the tracker is re-initialized five frames after the failure and ten frames after re-initialization are ignored in computation to further reduce the bias in accuracy measure [34]. Trackers are run 15 times on each sequence to obtain a better statistics on performance measures. The per-frame accuracy is obtained as an average over these runs. Averaging per-frame accuracies gives per-sequence accuracy, while per-sequence robustness is computed by averaging failure rates over different runs.

Apart from accuracy and robustness, the tracking speed is also an important property that indicates practical usefulness of trackers in particular applications. While accuracy and robustness results can be made comparable across different trackers by using the same experiments and dataset, the speed measurement depends on the programming language, implementation skills and most importantly, the hardware used to perform the experiments. To reduce the influence of hardware, the VOT2014 introduces a new unit for reporting the tracking speed. When an experiment is conducted with the VOT2014 evaluation kit, the kit benchmarks the machine by measuring the time required to perform a maximum pixel value filter on a grayscale image of size 600×600 with a 30×30 pixel window. The benchmark filter operation was coded in C by the VOT2014 committee. The VOT tracking speed is then reported by dividing the measured tracking time with the time required for the filtering operation. Thus the speed is reported in equivalent filter operations (EFO) which are defined by the VOT2014 evaluation kit.

3.1 Evaluation Methodology

To address the unequal representation of the attributes in the sequences, the two measures are calculated only on the subset of frames in the dataset that contain that attribute (attribute subset). The trackers are ranked with respect to each measure separately on each attribute. The VOT2013 recognized that subsets of trackers might be performing equally well and this should be reflected in the ranks. Therefore, for each i-th tracker a set of equivalent trackers is determined. The corrected rank of the i-th tracker is obtained by averaging the ranks of these trackers including the considered tracker. The final ranking is obtained by averaging the ranks.

The equivalency of trackers is determined in VOT2013 by testing for the statistical significance of difference in performance of pairs of trackers. Separate statistical tests are applied for accuracy and robustness. The VOT2013 acknowledged that statistical significance of performance differences does not directly imply a practical difference [16], but did not address that. The practical difference is a level of difference that is considered negligibly small. This level can come from the noise in annotation, the fact that multiple ground truth annotations might be equally valid, or simply from the fact that very small differences in trackers are negligible from a practical point of view.

The VOT2014 extends the methodology by introducing tests of practical difference on tracking accuracy. In VOT2014, a pair of trackers is considered to perform equally well in accuracy if their difference in performance is not statistically significant or if it fails the practical difference test.

Testing for Practical Difference: Let $\phi_t(i)$ and $\phi_t(j)$ be the accuracies of the i-th and the j-th tracker at the t-th frame and let $\mu(i) = \frac{1}{T}\sum_{t=1}^{T}\phi_t(i)$ and $\mu(j) = \frac{1}{T}\sum_{t=1}^{T}\phi_t(j)$ be the average accuracies calculated over a sequence of T frames. The trackers are said to perform differently if the difference of their averages is greater than a predefined threshold γ, i.e., $|\mu(i) - \mu(j)| > \gamma$, or, by defining $d_t(i,j) = \phi_t(i) - \phi_t(j)$, expanding the sums and pulling the threshold into the summation, $\frac{1}{T}|\sum_{t=1}^{T} d_t(i,j)/\gamma| > 1$. In VOT2014, the frames $t = 1:T$ actually come from multiple sequences, and γ values may vary over frames. Therefore, in VOT2014, a pair of trackers passes the test for practical difference if the following relation holds

$$\frac{1}{T}|\sum_{t=1}^{T} d_t(i,j)/\gamma_t| > 1, \tag{1}$$

where γ_t is the practical difference threshold corresponding to t-th frame.

Estimation of Practical Difference Threshold: The practical difference strongly depends on the target as well as the number of free parameters in the annotation model (i.e., in our case a rotated bounding box). Ideally a per-frame estimate of γ would be required for each sequence, but that would present a significant undertaking. On the other hand, using a single threshold for entire

Fig. 1. Examples of diversity of bounding box annotations for different images

dataset is too restrictive as the properties of targets vary across the sequences. A compromise can be taken in this case by computing one threshold per sequence. We propose selecting M frames per sequence and have J expert annotators place the bounding boxes carefully K times on each frame. In this way $N = K \times J$ bounding boxes are obtained per frame. One of the bounding boxes can be taken as a possible ground truth and $N - 1$ overlaps can be computed with the remaining ones. Since all annotations are considered "correct", any two overlaps should be considered equivalent, therefore the difference between these two overlaps is an example of negligibly small difference. By choosing each of the bounding boxes as ground truth, $M(N((N-1)^2 - N + 1))/2$ samples of differences are obtained per sequence. The practical difference threshold per sequence is estimated as the average of these values.

4 Analysis and Results

4.1 Estimation of Practical Difference Thresholds

The per sequence practical difference thresholds were estimated by the following experiment. For each sequence of the dataset, we identified four frames with axis-aligned ground-truth bounding boxes. The annotators were presented with two images side by side. The first image showed the first frame with overlaid ground-truth bounding box. This image served as a guidance on which part of the object should be annotated and was kept visible throughout the annotation of the four frames from the same sequence. These frames were displayed in the second image and the annotator was asked to place an axis-aligned bounding box on the target in each one. The process of annotation was repeated by each annotator three times. See Figure 1 In this setup a set of 15960 samples of differences was obtained per sequence and used to compute the practical difference threshold as discusses in Section 3.1.

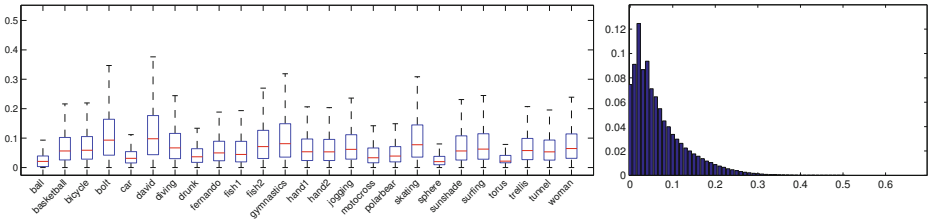


Fig. 2. Box plots of differences per sequence (left) and distribution of differences over entire dataset (right)

Figure 2 shows boxplots of difference distributions w.r.t. sequences and a distribution over entire dataset. It is clear that the threshold on practical difference varies over the sequences. For the sequences containing rigid objects, the practical difference threshold is small (e.g., ball) and becomes large for sequences with deformable/articulated objects (e.g., bolt).

4.2 The VOT2014 Experiments

The VOT2014 challenge includes the following two experiments:

- Experiment 1: This experiment runs a tracker on all sequences in the VOT2014 dataset by initializing it on the ground truth bounding boxes.
- Experiment 2: This experiment performs Experiment 1, but initializes with a noisy bounding box. By a noisy bounding box, we mean a randomly perturbed bounding box, where the perturbation is in the order of ten percent of the ground truth bounding box size.

In Experiment 2 there was a randomness in the initialization of the trackers. The bounding boxes were randomly perturbed in position and size by drawing perturbations uniformly from $\pm 10\%$ interval of the ground truth bounding box size, while the rotation was perturbed by drawing uniformly from ± 0.1 radians. All the experiments were automatically performed by the evaluation kit[8]. A tracker was run on each sequence 15 times to obtain a better statistic on its performance. Note that it does not make sense to perform Experiment 1 multiple times for the deterministic trackers. In this case, the evaluation kit automatically detects whether the tracker is deterministic and reduces the number of repetitions accordingly.

4.3 Trackers Submitted

Together 33 entries have been submitted to the VOT2014 challenge. Each submission included the binaries/source code that was used by the VOT2014 committee for results verification. The VOT2014 committee additionally contributed

[8] https://github.com/vicoslab/vot-toolkit

5 baseline trackers. For these, the default parameters were selected, or, when not available, were set to reasonable values. Thus in total 38 trackers were included in the VOT2014 challenge. In the following we briefly overview the entries and provide the references to original papers. For the methods that are not officially published, we refer to the Appendix A instead.

Several tracker explicitly decomposed target into parts. These ranged from key-point-based trackers CMT (A.32), IIVTv2 (A.6), Matrioska (A.11) and its derivative MatFlow (A.13) to general part-based trackers LT-FLO (A.10), PT+ (A.27), LGT (A.33), OGT (A.30), DGT (A.31), ABS (A.2), while three trackers applied flock-of-trackers approaches FoT (A.22), BDF (A.12) and FRT (A.34). Several approaches were applying global generative visual models for target localization: a channel blurring approach EDFT (A.4) and its derivative qwsEDFT (A.3), GMM-based VTDMG (A.7), scale-adaptive mean shift eASMS (A.21), color and texture-based ACAT (A.20), HOG correlation-based SAMF (A.9), NCC based tracker with motion model IMP-NCC (A.15), two color-based particle filters SIR-PF (A.1) and IPRT (A.18), a compressive tracker CT (A.35) and intensitiy-template-based pca tracker IVT (A.36). Two trackers applied fusion of flock-of-trackers and mean shift, HMM-TxD (A.23) and DynMS (A.26). Many trackers were based on discriminative models, i.e., boosting-based particle filter MCT (A.8), multiple-instance-learning-based tracker MIL (A.37), detection-based FSDT (A.29) while several applied regression-based techniques, i.e., variations of online structured SVM, Struck (A.16), aStruck (A.5), ThunderStruck (A.17), PLT_13 (A.14) and PLT_14 (A.19), kernelized-correlation-filter-based KCF (A.28), kernelized-least-squares-based ACT (A.24) and discriminative correlation-based DSST (A.25).

4.4 Results

The results are summarized in Table 1 and visualized by the AR rank plots [36,58], which show each tracker as a point in the joint accuracy-robustness rank space (Figure 3 and Figure 4). For more detailed rankings and plots please see the VOT2014 results homepage. At the time of writing this paper, the VOT committee was able to verify some of the submitted results by re-running parts of the experiments using the binaries of the submitted trackers. The verified trackers are denoted by * in Table 1. The AR rank plots for baseline experiment (Experiment 1) and noise experiment (Experiment 2) are shown in Figure 3, while per-visual-attribute ranking plots for the baseline experiment are shown in Figure 4.

In terms of accuracy, the top performing trackers on both experiments, starting with best performing, are DSST, SAMF and KCF (Figure 3). Averaging together the accuracy and robustness, the improvement of DSST over the other two is most apparent at *size change* and *occlusion* attributes (Figure 4). For the noise experiment, these trackers remain the top performing, but the difference in accuracy is very small. In terms of robustness, the top performing trackers on the baseline experiment are PLT_13, PLT_14, MatFlow and DGT. These trackers come from two classes of trackers. The first two, PLT_13 and

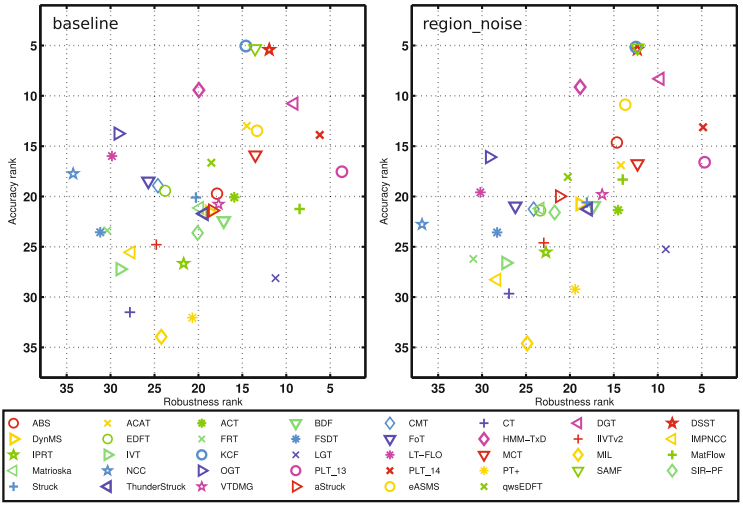


Fig. 3. The accuracy-robustness ranking plots with respect to the two experiments. Tracker is better if it resides closer to the top-right corner of the plot.

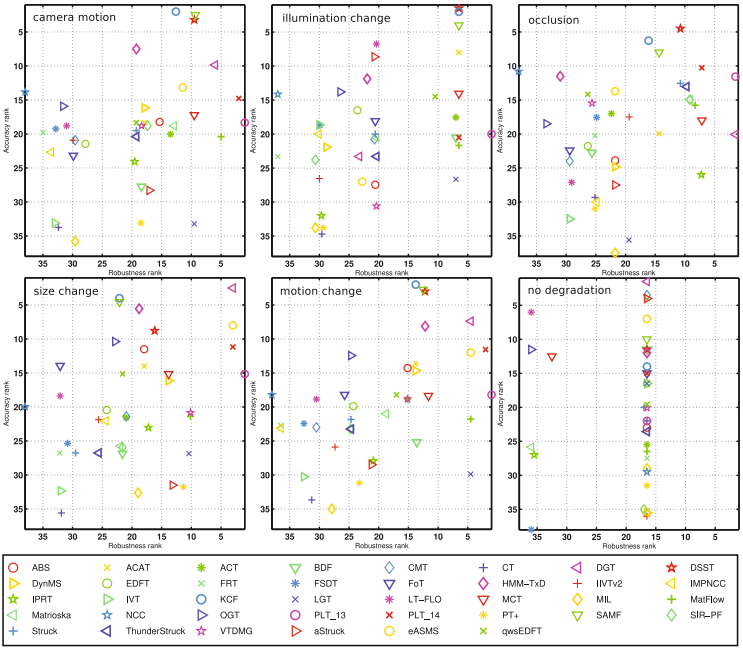


Fig. 4. The accuracy-robustness ranking plots of Experiment 1 with respect to the six sequence attributes. The tracker is better if it resides closer to the top-right corner of the plot.

PLT_14 are extensions of the Struck [25] tracker that apply histogram backprojection as feature selection strategy in SVM training. The second two trackers are part-based trackers that apply different types of parts. MatFlow is extension of Matrioska [42] which applies a ORB/SURF keypoints and robust voting and matching techniques. On, the other hand, DGT decomposes target into parts by superpixels and applies graph-matching techniques to perform association of parts across the frames. The DGT is generally well ranked with respect to different visual properties, however, it significantly drops in performance during illumination changes (Figure 3). In the second experiment with initialization noise, MatFlow drops in ranks and the fourth-top tracker becomes the MCT which applies a holistic discriminative model and a motion model with particle filter. From Figure 4, we can see that a large majority of trackers, including NCC, performed equally well on frames denoted as *neutral* in terms of robustness, but differed quite significantly in accuracy.

The entries included several trackers from the same class. The top-performing trackers in accuracy, DSST, SAMF and KCF, formulate tracking as a ridge regression problem for correlation filter learning and apply HOG [13] in their visual model. The DSST is an extension of the MOSSE [5] that uses grayscale in addition to HOG, while SAMF and KCF seem to be extensions of [27] that address scale change. The similarity in design is reflected in the AR-rank plots as they form tight clusters in baseline as well as noise experiment. The PLT_13 and PLT_14 are also from the same class of trackers. The PLT_13 is the winner of the VOT2013 challenge [36] which does not adapt the target size, while the PLT_14 is an extension of PLT_13 that adapts the size as well. Interestingly, the PLT_14 does improve in accuracy compared to PLT_13, but sacrifices the robustness. In the noise experiment the PLT_14 is still outperforms the PLT_13 in accuracy, but the difference in robustness is reduced. MatFlow is an extension of Matrioska that applies a flock-of-trackers variant BDF. At a comparable accuracy ranks, the MatFlow by far outperforms the original Matrioska in robustness. The boost in robustness ranks might be attributed to addition of BDF, which is supported by the fact that BDF alone outperforms in robustness the FoT and trackers based on variations of FoT, i.e., aStruck, HMMTxD and dynMS. This speaks of resiliency to outliers in flock selection in BDF. Two trackers combine color-based mean shift with flow, i.e., dynMS and HMMTxD and obtain comparable ranks in robustness, however, the HMMTxD achieves a significantly higher accuracy rank, which might be due to considerably more sophisticated tracker merging scheme in HMMTxD. Both methods are outperformed in robustness by the scale-adaptive mean shift eASMS that applies motion prediction and colour space selection. While this version of mean shift performs quite well over a range of visual attributes, the performance drops in ranks drastically for *occlusion* and *illumination change*. The entries contained the original Struck and two variations, ThunderStruck and aStruck. ThunderStruck is a CUDA-speeded-up Struck and performs quite similarly to the original Struck in baseline and noise experiment. The aStruck applies the flock-of-trackers for scale adaptation in Struck and improves in robustness on the baseline experiment, but is ranked lower in the noise experiment.

Table 1. Ranking results. The top, second and third lowest average ranks are shown in red, blue and green respectively. The R_Σ column displays a joined ranking for both experiments, which were also used to order the trackers. The trackers that have been verified by the VOT committee are denoted by the asterisk *****.

	baseline			region_noise					
	R_A	R_R	R	R_A	R_R	R	R_Σ	**Speed**	**Impl.**
DSST*	5.41	11.93	**8.67**	5.40	12.33	8.86	**8.77**	7.66	Matlab & Mex
SAMF*	*5.30*	13.55	*9.43*	*5.24*	12.30	**8.77**	*9.10*	1.69	Matlab & Mex
KCF*	**5.05**	14.60	9.82	**5.17**	12.49	*8.83*	9.33	24.23	Matlab & Mex
DGT	10.76	9.13	9.95	8.31	9.73	9.02	9.48	0.23	C++
PLT_14*	13.88	*6.19*	10.03	13.12	*4.85*	8.99	9.51	62.68	C++
PLT_13	17.54	**3.67**	10.60	16.60	**4.67**	10.63	10.62	*75.92*	C++
eASMS*	13.48	13.33	13.40	10.88	13.70	12.29	12.85	13.08	C++
HMM-TxD*	9.43	19.94	14.69	9.12	18.83	13.98	14.33	2.08	C++
MCT	15.88	13.52	14.70	16.75	12.30	14.52	14.61	1.45	C, C++
ACAT	12.99	14.49	13.74	16.90	14.20	15.55	14.65	3.24	unknown
MatFlow	21.25	8.49	14.87	18.33	13.99	16.16	15.51	19.08	C++
ABS	19.72	17.88	18.80	14.63	14.65	14.64	16.72	0.62	Matlab & Mex
ACT	20.08	15.91	18.00	21.36	14.53	17.94	17.97	18.26	Matlab
qwsEDFT	16.65	18.53	17.59	18.07	20.24	19.15	18.37	3.88	Matlab
LGT*	28.12	11.22	19.67	25.25	9.08	17.17	18.42	1.23	Matlab & Mex
VTDMG	20.77	17.70	19.24	19.81	16.33	18.07	18.65	1.83	C++
BDF	22.42	17.12	19.77	20.91	17.29	19.10	19.44	46.82	C++
Struck	20.11	20.29	20.20	20.60	18.08	19.34	19.77	5.95	C++
DynMS*	21.54	18.75	20.14	20.76	18.84	19.80	19.97	3.21	Matlab & Mex
ThunderStruck	21.71	19.35	20.53	21.26	17.92	19.59	20.06	19.05	C++
aStruck*	21.41	18.40	19.90	19.98	21.19	20.59	20.24	3.58	C++
Matrioska	21.15	19.86	20.50	21.19	23.39	22.29	21.40	10.20	unknown
SIR-PF	23.62	20.09	21.86	21.58	21.74	21.66	21.76	2.55	Matlab & Mex
EDFT	19.43	23.80	21.61	21.39	23.37	22.38	22.00	4.18	Matlab
OGT	13.76	29.15	21.45	16.09	29.16	22.63	22.04	0.39	unknown
CMT*	18.93	24.61	21.77	21.26	24.13	22.69	22.23	2.51	Python, C++
FoT*	18.48	25.70	22.09	20.96	26.21	23.58	22.84	**114.64**	C++
LT-FLO	15.98	29.84	22.91	19.59	30.20	24.90	23.90	1.10	Matlab
IPRT	26.68	21.68	24.18	25.54	22.73	24.14	24.16	14.69	C, C++
IIVTv2	24.79	24.79	24.79	24.61	22.97	23.79	24.29	3.67	C++
PT+	32.05	20.68	26.37	29.23	19.41	24.32	25.34	49.89	C++
FSDT	23.55	31.17	27.36	23.58	28.29	25.93	26.65	1.47	C++
IMPNCC	25.56	27.66	26.61	28.28	28.32	28.30	27.45	8.37	Matlab
IVT*	27.23	28.92	28.07	26.60	27.29	26.95	27.51	2.35	Matlab & Mex
FRT*	23.38	30.38	26.88	26.21	30.99	28.60	27.74	3.09	C++
NCC*	17.74	34.25	26.00	22.78	36.83	29.80	27.90	6.88	Matlab
CT*	31.51	27.79	29.65	29.66	26.94	28.30	28.98	6.29	C++
MIL*	33.95	24.22	29.09	34.61	24.87	29.74	29.41	1.94	C++

Note that majority of the trackers submitted to VOT2014 are fairly competitive trackers. This is supported by the fact that the trackers, that are often used as baseline trackers, NCC, MIL, CT, FRT and IVT, occupy the bottom-left part of the AR rank plots. Obviously these approaches vary in accuracy and robustness and are thus spread perpendicularly to the bottom-left-to-upper-right diagonal of AR-rank plots. In both experiments, the NCC is the least robust tracker. In summary, as in VOT2013 [36], the most robust tracker over individual visual properties remains the PLT_13 (A.14). This tracker is surpassed by far in combined accuracy-robustness rank by the trackers DSST (A.25), SAMF (A.9) and KCF (A.28), of which the DSST (A.25) outperforms the other two in robustness. According to the average ranks, the DSST (A.25) is thus the winner of VOT2014.

The VOT2014 evaluation kit also measured the times required to perform a repetition of each tracking run. For each tracker, the average tracking speed was estimated from these measurements. Table 1 shows the tracking speed per frame in the EFO units, introduced in Section 3. Note that the times for the Matlab trackers included an overhead required to load the Matlab environment, which depends mostly depends on hard drive reading speed which was measured during the evaluation. Table 1 shows adjusted times that accounted for this overhead. While one has to be careful with speed interpretation, we believe that these measurements still give a good comparative estimate of the trackers practical complexity. The trackers that stand out are the FoT and PLT_13, achieving speeds in range of around 100 EFO units (C++ implementations). To put this into perspective, a C++ implementation of a NCC tracker provided in the toolkit processes the VOT2014 dataset with an average of 220 frames per second on a laptop with an Intel Core i5 processor, which equals to approximately 80 EFO units.

5 Conclusions

This paper reviewed the VOT2014 challenge and its results. The challenge contains a annotated dataset of sequences in which targets are denoted by rotated bounding boxes to aid a precise analysis of the tracking results. All the sequences are labelled per-frame with attributes denoting various visual phenomena. The challenge also introduces a new *Matlab/Octave* evaluation kit for fast execution of experiments, proposes a new unit for measuring tracker speed, and extends the VOT2013 performance evaluation methodology to account for practical equivalence of tracker accuracy. The dataset, evaluation kit and VOT2014 results are publicly available from the challenge webpage.

The results of VOT2014 indicate that a winner of the challenge according to the average results is the DSST (A.25) tracker. The results also show that trackers tend to specialize either for robustness or accuracy. None of the trackers consistently outperformed the others by all measures at all sequence attributes. One class of trackers that consistently appears at the top of ranks are large

margin regression-based trackers which apply global visual models[9], while the other class of trackers is the part-based trackers in which the target is considered as a set of parts or keypoints.

The main goal of VOT is establishing a community-based common platform for discussion of tracking performance evaluation and contributing to the tracking community with verified annotated datasets, performance measures and evaluation toolkits. Following the very successful VOT2013, VOT2014 was the second attempt towards this. Our future work will be focused on revising the evaluation kit, dataset, performance measures, and possibly launching challenges focused to narrow application domains, depending on the feedbacks and interest expressed from the community.

Acknowledgments. This work was supported in part by the following research programs and projects: Slovenian research agency projects J24284, J23607 and J2-2221 and European Union seventh framework programme under grant agreement no 257906. Jiri Matas and Tomas Vojir were supported by CTU Project SGS13/142/OHK3/2T/13 and by the Technology Agency of the Czech Republic project TE01020415 (V3C – Visual Computing Competence Center).

A Submitted Trackers

In this appendix we provide a short summary of all trackers that were considered in the VOT2014 competition.

A.1 Sequential Importance Re-Sampling Particle Filter (SIR-PF)

D. Pangeršič (*dp3698@student.uni-lj.si*)

SIR-PF tracker makes Particle Filter approach more robust on sequences with fast motion and illumination changes. To do that, the tracker changes RGB data into YCbCr data and it generates a background model used by Comaniciu et al. [11]. The tracking task is done by using a window adaptation approach and a reference histogram adaptation to perform the matching between candidate objects.

A.2 Appearance-Based Shape-Filter (ABS)

H. Possegger, T. Mauthner, H. Bischof
({*possegger, mauthner, bischof*}*@icg.tugraz.at*)

ABS tracker relies on appearance and shape cues for tracking. In particular, a histogram-based pixel-wise foreground is modelled to create a filter capturing discriminative object areas. This model combined with colour gradient templates to capture the object shape, allows to efficiently localize the object using mean shift tracking. ABS employs graph cut segmentation based on the pixel-wise foreground probabilities to adapt changes of object scales.

[9] We consider the Structured SVM as regression from image intensities to image displacements.

A.3 Power Updated Weighted Comparison Enhanced Distribution Field Tracker (qwsEDFT)

K. Öfjäll, M. Felsberg ({*kristoffer.ofjall, michael.felsberg*}*@liu.se*)

A model matching approach where the tracked model is represented by a channel distribution field. Previous approaches such as DFT [52] and EDFT [20] do not exploit the possibilities of the model representation. The qwsEDFT tracker features a power update scheme and a standard deviation weighted comparison.

A.4 Enhanced Distribution Fields for Tracking (EDFT)

M. Felsberg (*michael.felsberg@liu.se*)

The EDFT is a novel variant of the DFT tracker as proposed in [52]. EDFT derives an enhanced computational scheme by employing the theoretic connection between averaged histograms and channel representations. For further details, the interested reader is referred to [20].

A.5 Scale Adaptative Struck Tracker (aStruck)

A. Lukežič, L. Čehovin (*alan.lukezic@gmail.com, luka.cehovin@fri.uni-lj.si*)

aStruck is a combination of optical-flow-based tracker and the discriminative tracker Struck [25]. aStruck uses low-level cues such as optical flow to handle significant scale changes. Besides, a framework akin to the FoT [60] tracker is utilized to robustly estimate the scale changes using the sparse Lucas-Kanade [41] pyramidal optical flow at points placed at a regular grid.

A.6 Initialization Insensitive Visual Tracker Version 2 (IIVTv2)

K. Moo Yi, J. Y. Choi (*kwang.yi@epfl.ch, jychoi@snu.ac.kr*)

IIVTv2 is an implementation of the extended version of the initialization insensitive tracker [63]. The change from the original version include motion prior calculated from optical flow [54], normalization of the two proposed saliency weights in [63], inclusion of recent features in the feature database, and location based initialization of SURF [4] feature points.

A.7 Visual Tracking with Dual Modeling through Gaussian Mixture Modeling (VTDMG)

K. M. Yi, J. Y. Choi (*kwang.yi@epfl.ch, jychoi@snu.ac.kr*)

VTDMG is an extended implementation of the method presented in [64]. Instead of using simple Gaussian modelling, VTDMG uses mixture of Gaussians. Besides, VTDMG models the target object and the background simultaneously and finds the target object through maximizing the likelihood defined using both models.

A.8 Motion Context Tracker (MCT)

S. Duffner, C. Garcia ({*stefan.duffner, christophe garcia*}*@liris.cnrs.fr*)

The Motion Context Tracker (MCT) is a discriminative on-line learning classifier based on Online Adaboost (OAB) which is integrated into the model collecting negative training examples for updating the classifier at each video frame. Instead of taking negative examples only from the surroundings of the object region or from specific distracting objects, MCT samples the negatives from a contextual motion density function in a stochastic manner.

A.9 A Kernel Correlation Filter Tracker with Scale Adaptive and Feature Integration (SAMF)

Y. Li, J. Zhu ({*liyang89, jkzhu*}*@zju.edu.cn*)

SAMF tracker is based on the idea of correlation filter-based trackers [5,15,26,27] with aim to improve the overall tracking capability. To tackle the problem of the fixed template size in kernel correlation filter tracker, an effective scale adaptive scheme is proposed. Moreover, features like HoG and colour naming are integrated together to further boost the overall tracking performance.

A.10 Long Term Featureless Object Tracker (LT-FLO)

K. Lebeda, S. Hadfield, J. Matas, R. Bowden
({*k.lebeda, s.hadfield*}*@surrey.ac.uk, matas@cmp.felk.cvut.cz, r.bowden@surrey.ac.uk*)

LT-FLO is designed to track texture-less objects. It significantly decreases reliance on texture by using edge-points instead of point features. The tracker also has a mechanism to detect disappearance of the object, based on the stability of the gradient in the area of projected edge-points. The reader is referred to [37] for details.

A.11 Matrioska

M. E. Maresca, A. Petrosino ({*mariomaresca, petrosino*}*@uniparthenope.it*)

Matrioska [42] decomposes tracking into two separate modules: detection and learning. The detection module can use multiple key point-based methods (ORB, FREAK, BRISK, SURF, etc.) inside a fallback model, to correctly localize the object frame by frame exploiting the strengths of each method. The learning module updates the object model, with a growing and pruning approach, to account for changes in its appearance and extracts negative samples to further improve the detector performance.

A.12 Best Displacement Flow (BDF)

M. E. Maresca, A. Petrosino ({*mariomaresca, petrosino*}*@uniparthenope.it*)

Best Displacement Flow is a new short-term tracking algorithm based on the same idea of Flock of Trackers [60] in which a set of local tracker responses are robustly combined to track the object. BDF presents two main contributions: (i) BDF performs a clustering to identify the Best Displacement vector which is used to update the object's bounding box, and (ii) BDF performs a procedure named Consensus-Based Reinitialization used to reinitialize candidates which were previously classified as outliers.

A.13 Matrioska Best Displacement Flow (MatFlow)

M. E. Maresca, A. Petrosino ({*mariomaresca, petrosino*}*@uniparthenope.it*)

MatFlow enhances the performance of the first version of Matrioska [42] with response given by aforementioned new short-term tracker BDF (see A.12). By default, MatFlow uses the trajectory given by Matrioska. In the case of a low confidence score estimated by Matrioska, MatFlow corrects the trajectory with the response given by BDF. Matrioska's confidence score is based on the number of key points found inside the object in the initialization.

A.14 Single Scale Pixel Based LUT Tracker (2013) (PLT_13)

C. Heng, S. YueYing Lim, Z. Niu, B. Li
({*hengcherkeng235, yueying53, niuzhiheng, libohit*}*@gmail.com*)

PLT runs a classifier at a fixed single scale for each test image, to determine the top scoring bounding box which is then the result of object detection. The classifier uses a binary feature vector constructed from colour, greyscale and gradient information. To select a small set of discriminative features, an online sparse structural SVM [25] is used. For more details, the interested reader is referred to [36].

A.15 Improved Normalized Cross-Correlation Tracker (IMPNCC)

A. Dimitriev (*ad7414@student.uni-lj.si*)

This tracker improves the NCC tracker [7] in three ways: (i) by using a non-constant adaptation, the template is updated with new information; (ii) scale changes are handled by running an sliding window for the original image and two resized ones choosing the maxima of them; (iii) a Kalman Filter [30] is also used to smooth the trajectory and reduce drift. This improved tracker was based on the code of the original NCC tracker supplied with the VOT 2013 toolkit [35].

A.16 Struck

S. Hare, A. Saffari, P. H. S. Torr
(*sam@samhare.net, amir@ymer.org, philip.torr@eng.ox.ac.uk*)

Struck [25] presents a framework for adaptive visual object tracking based on structured output prediction. By explicitly allowing the output space to express the needs of the tracker, need for an intermediate classification step is avoided. The method uses a kernelized structured output support vector machine (SVM), which is learned online to provide adaptive tracking.

A.17 ThunderStruck

S. Hare, A. Saffari, S. Golodetz, V. Vineet, M. Cheng, P. H. S. Torr
(*sam@samhare.net, amir@ymer.org, sgolodetz@gxstudios.net, vibhav.vineet@gmail.com, cmm.thu@qq.com, philip.torr@eng.ox.ac.uk*)

ThunderStruck is a CUDA-based implementation of the Struck tracker presented by Hare et al. [25]. As with the original Struck, tracking is performed using a structured output SVM. On receiving a new frame, the tracker predicts a bounding box for the object in the new frame by sampling around the old object position and picking the location that maximises the response of the current SVM. The SVM is then updated using LaRank [6]. A support vector budget is used to prevent the unbounded growth in the number of support vectors that would otherwise occur during tracking.

A.18 Iterative Particle Repropagation Tracker (IPRT)

J.-W. Choi (*jwc@etri.re.kr*)

IPRT is a particle filter based tracking method inspired by colour-based particle filter [47,49] with the proposed iterative particle re-propagation. Multiple HSV colour histograms with $6 \times 6 \times 6$ bins are used as an observation model. In order to reduce the chance of tracker drift, the states of particles are saved before propagation. If tracker drift is detected, particles are restored and re-propagated. The tracker drift is detected by a colour histogram similarity measure derived from the Bhattacharyya coefficient.

A.19 Size-Adaptive Pixel Based LUT Tracker (2014) (PLT_14)

C. Heng, S. YueYing Lim, Z. Niu, B. Li
(*{hengcherkeng235, yueying53, niuzhiheng, libohit}@gmail.com*)

PLT_14 tracker is an improved version of PLT tracker used in VOT 2013 [36], with size adaptation for the tracked object. PLT_14 uses discriminative pixel features to compute the scanning window score in a tracking-by-detection framework. The window score is 'back projected' to its contributing pixels. For each pixel, the pixel score is computed by summing the back projected scores of the windows that use this pixel. This score contributes to estimate which pixel belongs to the object during tracking and determine a best bounding box.

A.20 Augment Color Attributes Tracker (ACAT)

L. Qin, Y. Qi, Q.g Huang
(*qinlei@ict.ac.cn, {yuankai.qi, qingming.huang}@vipl.ict.ac.cn*)

Augment Color Attributes Tracker is based on the method of Colour Attributes Tracker (CAT) [15]. Colour features used in CAT is just colour. CAT extends CSK tracker [26] to multi-channel colour features and it also augments CAT by including texture features and shape features.

A.21 Enhanced Scale Adaptive MeanShift (eASMS)

T. Vojíř, J. Matas (*{vojirtom, matas}@cmp.felk.cvut.cz*)

eASMS tracker is a variation of the scale adaptive mean-shift [10–12]. It enhances its performance by utilizing background subtraction and motion prediction to allow the mean-shift procedure to converge in presence of high background clutter. The eASMS tracker also incorporates automatic per-frame selection of colour space (from pool of the available ones, e.g. HSV, Luv, RGB).

A.22 Flock of Trackers (FoT)

T. Vojíř, J. Matas (*{vojirtom, matas}@cmp.felk.cvut.cz*)

The Flock of Trackers (FoT) [60] is a tracking framework where the object motion is estimated from the displacements or using a number of local trackers covering the object. Each local tracker is attached to a certain area specified in the object coordinate frame. The FoT object motion estimate is robust due to the combination of local tracker motions.

A.23 Hidden Markov Model Fusion of Tracking and Detection (HMM-TxD)

T. Vojíř, J. Matas (*{vojirtom, matas}@cmp.felk.cvut.cz*)

The HMM-TxD tracker is a novel method for fusing diverse trackers by utilizing a hidden Markov model (HMM). The HMM estimates the changes in individual tracker performance, its state corresponds to a binary vector predicting failure of individual trackers. The proposed approach relies on a high-precision low-recall detector that provides a source of independent information for a modified Baum-Welch algorithm that updates the Markov model. Two trackers were used in the HMM-TxD: Flock of Trackers [60] estimating similarity and scale adaptive mean-shift tracker [10–12].

A.24 Adaptive Color Tracker (ACT)

M. Danelljan, F. S. Khan, M. Felsberg, J. van de Weijer
(*{fmartin.danelljan, fahad.khan, michael.felsberg}@liu.se, joost@cvc.uab.es*)

The Adaptive Color Tracker (ACT) [15] extends the CSK tracker [26] with colour information. ACT tracker contains three improvements to CSK tracker: (i) A temporally consistent scheme for updating the tracking model is applied instead of training the classifier separately on single samples, (ii) colour attributes [61] are applied for image representation, and (iii) ACT employs a dynamically adaptive scheme for selecting the most important combinations of colours for tracking.

A.25 Discriminative Scale Space Tracker (DSST)

M. Danelljan, G. Häger, F. S. Khan, M. Felsberg
(*fmartin.danelljan@liu.se, hager.gustav@gmail.com,*
{fahad.khan, michael.felsberg}@liu.se)

The Discriminative Scale Space Tracker (DSST) [14] extends the Minimum Output Sum of Squared Errors (MOSSE) tracker [5] with robust scale estimation. The MOSSE tracker works by training a discriminative correlation filter on a set of observed sample grey scale patches. This correlation filter is then applied to estimate the target translation in the next frame. The DSST additionally learns a one-dimensional discriminative scale filter, that is used to estimate the target size. For the translation filter, the intensity features employed in the MOSSE tracker is combined with a pixel-dense representation of HOG-features.

A.26 Dynamic Mean Shift (DynMS)

Franci Oven, Matej Kristan (*frenk.oven@gmail.com, matej.kristan@fri.uni-lj.si*)

DynMS is a Mean Shift tracker [9] with an isotropic kernel bootstrapped by a flock-of-features (FoF) tracker. The FoF tracker computes a sparse Lucas Kanade flow [41] and uses MLESAC [55] with similarity transform to predict the target position. The estimated states of the target are merged by first moving to estimated location of FoF and then using Mean Shift to find the object.

A.27 Pixeltrack+ (PT+)

S. Duffner, C. Garcia (*{stefan.duffner, christophe garcia}@liris.cnrs.fr*)

Pixeltrack+ is based on the Pixeltrack tracking algorithm [17]. The algorithm uses two components: a detector that makes use of the generalised Hough transform with pixel-based descriptors, and a probabilistic segmentation method based on global models for foreground and background. The original Pixeltrack method [17] has been improved to cope with varying scale by estimating the objects size based on the current segmentation.

A.28 Kernelized Correlation Filter (KCF) Tracker (KCF)

J. F. Henriques, J. Batista (*{henriques, batista}@isr.uc.pt*)

This tracker is basically a Kernelized Correlation Filter [27] operating on simple HOG features. The KCF is equivalent to a Kernel Ridge Regression trained with

thousands of sample patches around the object at different translations. The improvements over the previous version [27] are multi-scale support, sub-cell peak estimation and replacing the model update by linear interpolation with a more robust update scheme [15].

A.29 Adaptive Feature Selection and Detection Based Tracker (FSDT)

J. Li, W. Lin ({*lijijia, wylin*}*@sjtu.edu.cn*)

FSDT is a tracking-by detection method that exploits the detection results to modify the tracker in the process of tracking. The detection part maintains a variable features pool where features are added or deleted as frames are processed. The tracking part implements a rough estimation of object tracked mainly by the velocity of objects. Afterwards, detection results are used to modify the rough tracked object position and to generate the final tracking result.

A.30 Online Graph-Based Tracking (OGT)

H. Nam, S. Hong, B. Han ({*namhs09, maga33, bhhan*}*@postech.ac.kr*)

OGT [45] is an online Orderless Model-Averaged tracking (OMA) [28]. OGT uses an unconventional graphical model beyond chain models, where each node has a single outgoing edge but may have multiple incoming edges. In this framework, the posterior is estimated by propagating multiple previous posteriors to the current frame along the identified graphical model, where the propagation is performed by a patch matching technique [32] as in [28]. The propagated densities are aggregated by weighted Bayesian model averaging, where the weights are determined by the tracking plausibility.

A.31 Dynamic Graph Based Tracker (DGT)

L. Wen, Z. Lei, S. Liao, S. Z. Li (*lywen, zlei, scliao, szli*}*@nlpr.ia.ac.cn*)

DGT is an improvement of the method proposed in [8]. The tracking problem is formulated as a matching problem between the target graph G(V;E) and the candidate graph G0(V0;E0). SLIC algorithm is used to oversegment the searching area into multiple parts (superpixels), and exploit the Graph Cut approach to separate the foreground superpixels from background superpixels. An affinity matrix based on motion, appearance and geometric constraints is built to describe the reliability of the matchings. The optimal matching from candidate superpixels is found from the affinity matrix applying the spectral technique [38]. The location of the target is voted by a series of the successfully matched parts according to their matching reliability.

A.32 Consensus-Based Matching and Tracking (CMT)

G. Nebehay, R. Pflugfelder ({*Georg.Nebehay.fl, Roman.Pflugfelder*}*@ait.ac.at*)

The CMT tracker [46] is a key point-based method in a combined matching-and-tracking framework. To localise the object in every frame, each key point casts votes for the object center. A consensus-based scheme is applied for outlier detection in the voting behaviour. By transforming votes based on the current key point constellation, changes of the object in scale and rotation are considered. The use of fast key point detectors and binary descriptors allows the current implementation to run in real-time.

A.33 Local-Global Tracking (LGT)

L. Čehovin, M. Kristan, A. Leonardis
({*luka.cehovin, matej.kristan, ales.leonardis*}*@fri.uni-lj.si*)

The core element of LGT is a coupled-layer visual model that combines the target global and local appearance by interlacing two layers. By this coupled constraint paradigm between the adaptation of the global and the local layer, a more robust tracking through significant appearance changes is achieved. The reader is referred to [57] for details.

A.34 Fragment Tracking (FRT)

VOT 2014 Technical Committee

The FRT tracker [1] represents the model of the object by multiple image fragments or patches. The patches are arbitrary and are not based on an object model. Every patch votes on the possible positions and scales of the object in the current frame, by comparing its histogram with the corresponding image patch histogram. We then minimize a robust statistic in order to combine the vote maps of the multiple patches. The algorithm overcomes several difficulties which cannot be handled by traditional histogram-based algorithms like partial occlusions or pose change.

A.35 Compressive Tracking (CT)

VOT 2014 Technical Committee

The CT tracker [67] uses an appearance model based on features extracted from the multi-scale image feature space with data-independent basis. It employs non-adaptive random projections that preserve the structure of the image feature space of objects. A very sparse measurement matrix is adopted to efficiently extract the features for the appearance model. Samples of foreground and background are compressed using the same sparse measurement matrix. The tracking task is formulated as a binary classification via a naive Bayes classifier with online update in the compressed domain.

A.36 Incremental Learning for Robust Visual Tracking (IVT)

VOT 2014 Technical Committee

The idea of the IVT tracker [51] is to incrementally learn a low-dimensional subspace representation, adapting online to changes in the appearance of the target. The model update, based on incremental algorithms for principal component analysis, includes two features: a method for correctly updating the sample mean, and a forgetting factor to ensure less modelling power is expended fitting older observations.

A.37 Multiple Instance Learning Tracking (MIL)

VOT 2014 Technical Committee

MIL [2] is a tracking-by-detection approach. MIL uses Multiple Instance Learning instead of traditional supervised learning methods and shows improved robustness to inaccuracies of the tracker and to incorrectly labeled training samples.

References

1. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: CVPR, vol. 1, pp. 798–805. IEEE Computer Society (June 2006)
2. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. **33**(8), 1619–1632 (2011)
3. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vision **92**(1), 1–31 (2011)
4. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
5. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: Comp. Vis. Patt. Recognition (2010)
6. Bordes, A., Bottou, L., Gallinari, P., Weston, J.: Solving multiclass support vector machines with larank. In: Proceedings of the 24th International Conference on Machine Learning (ICML) (2007)
7. Briechle, K., Hanebeck, U.D.: Template matching using fast normalized cross correlation. In: Aerospace/Defense Sensing, Simulation, and Controls, International Society for Optics and Photonics, pp. 95–102 (2001)
8. Cai, Z., Wen, L., Yang, J., Lei, Z., Li, S.Z.: Structured visual tracking with dynamic graph. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part III. LNCS, vol. 7726, pp. 86–97. Springer, Heidelberg (2013)
9. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(5), 603–619 (2002)
10. Comaniciu, D., Ramesh, V., Meer, P.: The variable bandwidth mean shift and data-driven scale selection. In: Int. Conf. Computer Vision, vol. 1, pp. 438–445 (2001)
11. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(5), 564–577 (2003)
12. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Comp. Vis. Patt. Recognition, vol. 2, pp. 142–149 (2000)
13. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Comp. Vis. Patt. Recognition, vol. 1, pp. 886–893 (June 2005)
14. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference BMVC (2014)
15. Danelljan, M., Khan, F.S., Felsberg, M., Van de Weijer, J.: Adaptive color attributes for real-time visual tracking. In: 2014 Conference on Computer Vision and Pattern Recognition CVPR (2014)
16. Demšar, J.: On the appropriateness of statistical tests in machine learning. In: Workshop on Evaluation Methods for Machine Learning ICML (2008)
17. Duffner, S., Garcia, C.: Pixeltrack: a fast adaptive algorithm for tracking non-rigid objects. In: Proceedings of the International Conference on Computer Vision (ICCV), pp. 2480–2487 (2013)
18. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge - a retrospective. Int. J. Comput. Vision (2014)

19. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
20. Felsberg, M.: Enhanced distribution field tracking using channel representations. In: Vis. Obj. Track. Challenge VOT 2013, In conjunction with ICCV 2013 (2013)
21. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
22. Gabriel, P., Verly, J., Piater, J., Genon, A.: The state of the art in multiple object tracking under occlusion in video sequences. In: Proc. Advanced Concepts for Intelligent Vision Systems, pp. 166–173 (2003)
23. Gavrila, D.M.: The visual analysis of human movement: A survey. Comp. Vis. Image Understanding **73**(1), 82–98 (1999)
24. Goyette, N., Jodoin, P.M., Porikli, F., Konrad, J., Ishwar, P.: Changedetection.net: a new change detection benchmark dataset. In: CVPR Workshops, pp. 1–8. IEEE (2012)
25. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.J.V. (eds.) Int. Conf. Computer Vision, pp. 263–270. IEEE (2011)
26. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the circulant structure of tracking-by-detection with kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 702–715. Springer, Heidelberg (2012)
27. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Trans. Pattern Anal. Mach. Intell. **1**(3), 125–141 (2014)
28. Hong, S., Kwak, S., Han, B.: Orderless tracking through model-averaged posterior estimation. In: Proceedings of the International Conference on Computer Vision (ICCV) (2013)
29. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. IEEE Trans. Systems, Man and Cybernetics, C **34**(30), 334–352 (2004)
30. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Engineering **82**, 34–45 (1960)
31. Kasturi, R., Goldgof, D.B., Soundararajan, P., Manohar, V., Garofolo, J.S., Bowers, R., Boonstra, M., Korzhova, V.N., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 319–336 (2009)
32. Korman, S., Avidan, S.: Coherency sensitive hashing. In: Proceedings of the International Conference on Computer Vision (ICCV) (2011)
33. Kristan, M., Perš, J., Perše, M., Kovačič, S.: Closed-world tracking of multiple interacting targets for indoor-sports applications. Comput. Vision Image Understanding **113**(5), 598–611 (2009)
34. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Cehovin, L., Nebehay, G., Fernandez, G., Vojir, T.: The vot2013 challenge: overview and additional results. In: Computer Vision Winter Workshop (2014)
35. Kristan, M., Čehovin, L.: Visual Object Tracking Challenge (VOT2013) Evaluation Kit. Visual Object Tracking Challenge (2013)
36. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Čehovin, L., Nebehay, G., Fernandez, G., Vojir, T., Gatt, A., Khajenezhad, A., Salahledin, A., Soltani-Farani, A., Zarezade, A., Petrosino, A., Milton, A., Bozorgtabar, B., Li,

B., Chan, C.S., Heng, C., Ward, D., Kearney, D., Monekosso, D., Karaimer, H.C., Rabiee, H.R., Zhu, J., Gao, J., Xiao, J., Zhang, J., Xing, J., Huang, K., Lebeda, K., Cao, L., Maresca, M.E., Lim, M.K., Helw, M.E., Felsberg, M., Remagnino, P., Bowden, R., Goecke, R., Stolkin, R., Lim, S.Y., Maher, S., Poullot, S., Wong, S., Satoh, S., Chen, W., Hu, W., Zhang, X., Li, Y., Niu, Z.: The visual object tracking VOT 2013 challenge results. In: ICCV Workshops, pp. 98–111 (2013)
37. Lebeda, K., Bowden, R., Matas, J.: Long-term tracking through failure cases. In: Vis. Obj. Track. Challenge VOT 2013, In conjunction with ICCV 2013 (2013)
38. Leordeanu, M., Hebert, M.: A spectral technique for correspondence problems using pairwise constraints. In: Proceedings of the International Conference on Computer Vision (ICCV), vol. 2, pp. 1482–1489 (2005)
39. Li, H., Shen, C., Shi, Q.: Real-time visual tracking using compressive sensing. In: Comp. Vis. Patt. Recognition, pp. 1305–1312. IEEE (2011)
40. Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A.R., Van den Hengel, A.: A survey of appearance models in visual object tracking. arXiv:1303.4803 [cs.CV] (2013)
41. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Imaging Understanding Workshop, pp. 121–130 (1981)
42. Maresca, M.E., Petrosino, A.: Matrioska: A multi-level approach to fast tracking by learning. In: Proc. Int. Conf. Image Analysis and Processing, pp. 419–428 (2013)
43. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Comp. Vis. Image Understanding **81**(3), 231–268 (2001)
44. Moeslund, T.B., Hilton, A., Kruger, V.: A survey of advances in vision-based human motion capture and analysis. Comp. Vis. Image Understanding **103**(2–3), 90–126 (2006)
45. Nam, H., Hong, S., Han, B.: Online graph-based tracking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 112–126. Springer, Heidelberg (2014)
46. Nebehay, G., Pflugfelder, R.: Consensus-based matching and tracking of keypoints for object tracking. In: IEEE Winter Conference on Applications of Computer Vision (March 2014)
47. Nummiaro, K., Koller-Meier, E., Van Gool, L.: Color features for tracking non-rigid objects. Chinese J. Automation **29**(3), 345–355 (2003)
48. Pang, Y., Ling, H.: Finding the best from the second bests - inhibiting subjective bias in evaluation of visual tracking algorithms. In: Int. Conf. Computer Vision (2013)
49. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: Proc. European Conf. Computer Vision, vol. 1, pp. 661–675 (2002)
50. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**(10), 1090–1104 (2000)
51. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Comput. Vision **77**(1–3), 125–141 (2008)
52. Sevilla-Lara, L., Learned-Miller, E.G.: Distribution fields for tracking. In: Comp. Vis. Patt. Recognition, pp. 1910–1917. IEEE (2012)
53. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual Tracking: an Experimental Survey. TPAMI (2013)
54. Tomasi, C., Kanade, L.: Detection and tracking of point features. Carnegie Mellon University, Tech. rep. (1991)
55. Torr, P.H., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding **78**(1), 138–156 (2000)

56. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: Comp. Vis. Patt. Recognition, pp. 1521–1528. IEEE (2011)
57. Čehovin, L., Kristan, M., Leonardis, A.: Robust Visual Tracking using an Adaptive Coupled-layer Visual Model. TPAMI **35**(4), 941–953 (2013)
58. Čehovin, L., Kristan, M., Leonardis, A.: Is my new tracker really better than yours?. In: IEEE WACV 2014 (2014)
59. Čehovin, L.: Trax: Visual tracking exchange protocol (April 2014)
60. Vojir, T., Matas, J.: Robustifying the flock of trackers. In: Comp. Vis. Winter Workshop, pp. 91–97. IEEE (2011)
61. Van de Weijer, J., Schmid, C., Verbeek, J.J., Larlus, D.: Learning color names for real-world applications. IEEE Transaction in Image Processing **18**(7), 1512–1524 (2009)
62. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: a benchmark. In: Comp. Vis. Patt. Recognition (2013)
63. Yi, K.M., Jeong, H., Heo, B., Chang, H.J., Choi, J.Y.: Initialization-insensitive visual tracking through voting with salient local features. In: 2013 IEEE International Conference on Computer Vision ICCV, pp. 2912–2919 (2013)
64. Yi, K.M., Jeong, H., Kim, S.W., Choi, J.Y.: Visual tracking with dual modeling. In: Proceedings of the 27th Conference on Image and Vision Computing New Zealand, IVCNZ 2012, pp. 25–30 (2012)
65. Yilmaz, A., Shah, M.: Object tracking: A survey. Journal ACM Computing Surveys **38**(4) (2006)
66. Young, D.P., Ferryman, J.M.: PETS Metrics: On-line performance evaluation service. In: ICCCN 2005 Proceedings of the 14th International Conference on Computer Communications and Networks, pp. 317–324 (2005)
67. Zhang, K., Zhang, L., Yang, M.-H.: Real-Time compressive tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)

Weighted Update and Comparison for Channel-Based Distribution Field Tracking

Kristoffer Öfjäll(✉) and Michael Felsberg

Computer Vision Laboratory Department of Electrical Engineering,
Linköping University, Linköping, Sweden
kristoffer.Ofjall@liu.se

Abstract. There are three major issues for visual object trackers: model representation, search and model update. In this paper we address the last two issues for a specific model representation, grid based distribution models by means of channel-based distribution fields. Particularly we address the comparison part of searching. Previous work in the area has used standard methods for comparison and update, not exploiting all the possibilities of the representation. In this work we propose two comparison schemes and one update scheme adapted to the distribution model. The proposed schemes significantly improve the accuracy and robustness on the Visual Object Tracking (VOT) 2014 Challenge dataset.

1 Introduction and Related Work

For online appearance-based object tracking, there are three primary concerns: how to represent the object to be tracked (model), how to find the object in a new frame (search/comparison) and finally how to update the model given the information obtained from the new frame (update). These are not independent, choosing one component influences the choice of the other two. There are other approaches to tracking, such as using a classifier for discriminating the target object from the background, however, only template-based methods will be considered here.

Several different categories of target models for representing the tracked object have been proposed in literature. One obvious appearance-based representation of the object is by means of an image patch cut out from the first frame according to the bounding box defining the object to be tracked. The locations of the object in the following frames are estimated by finding patches best corresponding to this target patch, employing some suitable distance function. Letting this simple model be linearly updated after every frame leads to a first order (weighted mean) model. A natural extension is a second order (Gaussian) approximation, where also the variance of each pixel is estimated.

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-16181-5_15) contains supplementary material, which is available to authorized users. Videos can also be accessed at http://www.springerimages.com/videos/978-3-319-16180-8.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 218–231, 2015.
DOI: 10.1007/978-3-319-16181-5_15

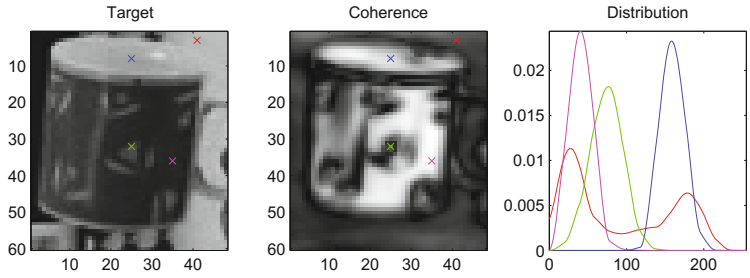


Fig. 1. Target and target model representation at the end of the VOT2013 cup-sequence. Left: found target patch. Middle: coherence of the target model (black: low, white: high), see Sect. 4.2. Right: represented pixel value distributions for a selection of points marked in left and middle images. Large coherence correspond to static pixel values on the tracked object and narrow distributions (blue, magenta). Low coherence correspond to background pixels (red, multimodal distribution) and varying pixels on the target (green, single wide mode).

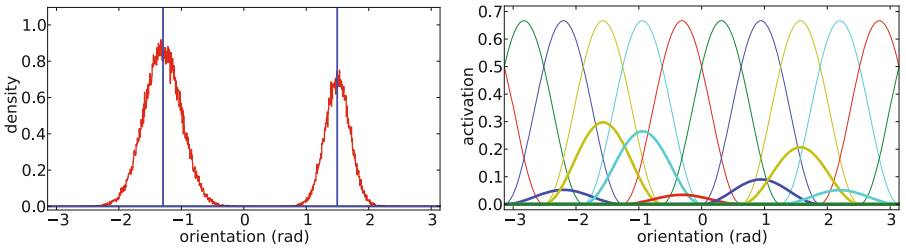
Another approach is to represent the full distribution of values within the target patch, illustrated in Fig. 1. Such a tracker, Distribution Field Tracking, DFT, was proposed by Sevilla et al. [13] where histograms are used for representing distributions. However, as was shown by Felsberg [4], replacing the histograms with channel representations [6] increases tracker performance, resulting in the Enhanced Distribution Field Tracker, EDFT.

In both cases, the model update is performed by a linear convex combination and the comparison uses an L_1 norm. However, the distribution view of the channel representation allows for other types of comparisons and update schemes compared to the direct pixel value representation. These possibilities were not used in previously proposed trackers.

In this work we evaluate a novel update scheme and novel comparison methods, exploiting the potential of the channel representation. We restrict ourselves to online methods implying: *i*) the tracking system should be causal, frames are made available to the tracker sequentially one by one and tracking results should be provided for one frame before the next frame is presented, and *ii*) the computational demands of the tracker, per frame, should not increase with sequence length. Further, the proposed trackers will be evaluated and compared to the baseline tracker from which they originate. Thorough comparisons to other state of the art trackers are available through the VOT 2014 Challenge[1].

As the ideas of channel representations may not be generally known, a brief introduction is presented in Sect. 2. The general tracker framework and target model representation is presented in Sect. 3. These sections also serve the purpose of introducing the notation used. The main contributions of the paper are presented in Sections 4 and 5. In Sect. 6, the effect of using the proposed meth-

[1] *http://votchallenge.net/vot2014/.*



(a) (b)

Fig. 2. Illustration of a channel representation for orientation data. (a) the orientation data (density in red) stems from two modes with different variance and probability mass. The blue lines indicate the mode estimates as extracted from the channel representation. (b) the channel representation of the distribution of orientation data. The thin plots indicate the kernel functions (the channels) and the bold lines illustrate the corresponding channel coefficients as weighted kernel functions.

ods in the tracker is evaluated. Sect. 7 concludes the paper. A video illustrating the approach is available as supplementary material[2].

2 Channel Representations

This section provides a brief introduction to channel representations at an intuitive level, since these methods will be required for our proposed contributions in Sections 4 and 5. Readers unfamiliar with these methods are referred to more comprehensive descriptions in literature [2,3,6] for details.

2.1 Channel Encoding

Channel representations have been proposed in 2000 [6]. The idea is to encode image features (e.g. intensity, orientation) in a vector of soft quantization levels, the channels. An example is given in Fig. 2, where orientation values are encoded.

Readers familiar with population codes [10,14], soft/averaged histograms [12], or Parzen estimators will find similarities. The major difference is that channel representations are very efficient to encode (because of the regular spacing of the channels) and decode (by applying frame theory [5]).

This computational efficiency allows for computing channel representations at each image pixel or for small image neighborhoods, as used in channel smoothing [2] as a variant of bilateral filtering [8], and tracking using distribution fields [4].

The kernel function, $b(\cdot)$, is defined to be non-negative, smooth and has compact support. In this paper, $\cos^2$ kernels with bandwidth parameter h are used:

$$b(\xi) = \frac{2}{3}\cos^2(\pi\xi/h) \qquad \text{for } |\xi| < h/2 \quad \text{and 0 otherwise.} \tag{1}$$

[2] Also available at *http://users.isy.liu.se/cvl/ofjall/vot2014.mp4*.

The components of the *channel vector* $\mathbf{x} = (x_1, x_2 \ldots, x_K)^T$ are obtained by shifting the kernel function K times with increments $h/3$. The range of the variable to be binned, ξ, together with the spacing of bins, v, determine the number of required kernel functions $K = (\max(\xi) - \min(\xi))/v + 2$. In most cases $v \gg 1$ such that K is of moderate size. The smooth kernel of the channel representation reduces the quantization effect compared to histograms by a factor of up to 20 in practice [2]. This allows reduction of the computational load by using fewer bins or to increase the accuracy for the same number of bins.

2.2 Robust Decoding

Using channel decoding [5], the modes of a channel representation can be obtained. Decoding is not required for the operation of the tracker, however concepts from the decoding are required for presenting the proposed coherence measure. Decoding is used for visualization of the target model in the supplementary video. Since $\cos^2$-channels establish a tight frame, the local maximum is obtained using three orthogonal vectors [5] $\mathbf{w}_1 \propto (\ldots, 0, 2, -1, -1, 0, \ldots)^T, \mathbf{w}_2 \propto (\ldots, 0, 0, 1, -1, 0, \ldots)^T, \mathbf{w}_3 \propto (\ldots, 0, 1, 1, 1, 0, \ldots)^T$ and

$$r_1 \exp(i2\pi\hat{\xi}/h) = (\mathbf{w}_1 + i\mathbf{w}_2)^T\mathbf{x} \qquad r_2 = \mathbf{w}_3^T\mathbf{x} \tag{2}$$

where i denotes the imaginary unit, $\hat{\xi}$ is the estimate (modulo an integer shift determined by the position of the three non-zero elements in $\mathbf{w}_k$, the *decoding window*), and r_1, r_2 are two confidence measures. The decoding window is chosen to maximize r_2 when only one mode is decoded. In particular, when decoding a channel representation with only one encoded value ξ, it can be shown that $\hat{\xi} = \xi$ if ξ is within the representable range of the channel representation [5]. For a sequence of single encoded values, the channel vector traces out a third of a circle with radius r_1 within each decoding window, however, a comprehensive description of this geometrical interpretation is out of scope.

3 General Tracking Framework and Representation

The general tracker framework is not different from DFT [13] and EDFT [4] and is briefly presented here, further details are available in [4,13]. In the first frame, the given bounding box of the object to be tracked is cut from the image. The intensity image patch of the target is channel encoded pixel wise using $K = 15$ channels, generating an I by J by K array denoted $\mathbf{C}$, where I and J are the height and width of the supplied bounding box. The 3D arrays generated from two channel encoded images are illustrated in Fig. 3.

In the next frame, the target representation $\mathbf{C}$ is compared to channel encoded patches (denoted $\mathbf{D}_{mn}$) from the new frame, where m and n represent a shift of the selected patch over the image. Gradient descent is used to find a minimum of a given comparison function, $d(\mathbf{C}, \mathbf{D}_{mn})$, with respect to the shift (m, n). Finally, the target representation is updated, $\mathbf{C} \leftarrow \mathbf{g}(\mathbf{C}, \mathbf{D}_{mn})$ and tracking continues in the next frame.

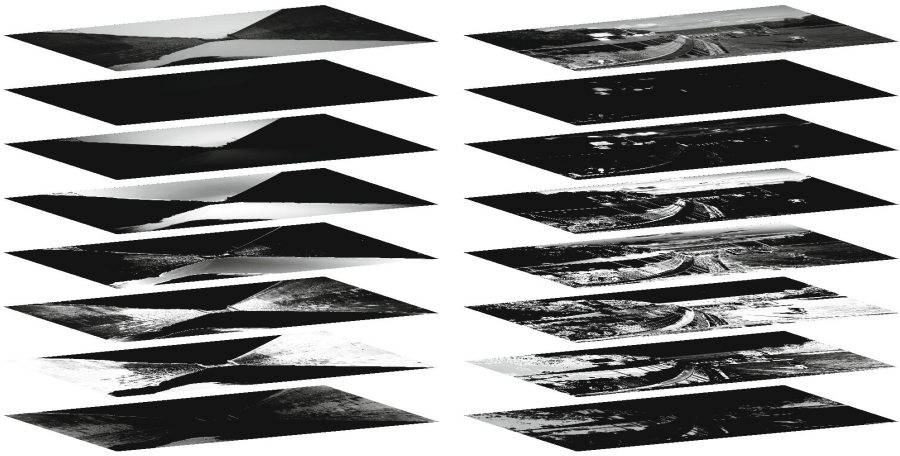


Fig. 3. Illustration of a pixel wise channel representation (with $K = 7$) of two images of canals. The top planes show the grayscale images while the lower seven planes indicate the activation of each channel (black: no activation, white: full activation). The lowest plane represents low image values (dark) while the seventh plane represents high image values (light).

Prior to comparison, i.e. calculation of $d(\mathbf{C}, \mathbf{D}_{mn})$, the channel planes of $\mathbf{C}$ and $\mathbf{D}_{mn}$ are smoothed. This was shown to increase the size of the basin of attraction for the correct solution [13]. Also, as in DFT and EDFT, a simple motion model (constant velocity model in the image plane) is used for initializing the gradient descent.

The main contribution of this work is a generalized model update function, $\mathbf{g}(\mathbf{C}, \mathbf{D}_{mn})$ and two proposals for the comparison function, $d(\mathbf{C}, \mathbf{D}_{mn})$. Earlier work has used a linearly weighted update, $\mathbf{g}(\mathbf{C}, \mathbf{D}_{mn}) = (1 - \gamma)\mathbf{C} + \gamma\mathbf{D}_{mn}$, and the IJK dimensional L_1 norm for comparison. The function $\mathbf{g}$ is a 3D array valued function of two 3D arrays. In this work, multiplication of a 3D array with a scalar is taken to be multiplication of each element in the array with the scalar, similar to regular matrix-scalar multiplication. Further, $[\cdot]_{ijk}$ denotes the element at index i, j, k and $[\cdot]_{ij}$ denotes the channel vector (with K coefficients) corresponding to pixel i, j in the bounding box.

4 Target Model Comparison

As previously mentioned, previous work has used the L_1 norm extended to 3D arrays for comparison. However, as is visualized in Fig. 1, the target model representation contains (after a few frames) a representation of the distribution

of values of each pixel within the bounding box. This should be exploited in the comparison function.

Since objects to be tracked are rarely rectangular, background pixels will be present in the bounding box. These pixels will generally vary more than the pixels on the object and such background pixels may disturb the tracker. This leads to a hypothesis that a weighted norm where the influence of inconsistent pixels is reduced, will improve the tracking results. Further, there may be areas of the tracked object which frequently change appearance, a weighted norm should also put more emphasis on parts of the tracked object showing more static appearance.

Two approaches will be presented. The first approach uses the reciprocals of the standard deviations of the represented distributions, which can be obtained directly from the channel coefficients. The second approach uses the *coherence*, which will be defined later.

4.1 Moments of Channel Representations

It can be shown that the average of several channel vectors, encoding values drawn from a specific distribution, tend to (up to scale) the probability density function convolved with the basis function, evaluated at channel centers[3](a sampled kernel density estimate) [5]. In this section, results based on a slightly different view of the distribution representation are presented. Here, the channel vector is assumed to represent a distribution, however, it is not necessarily the distribution from which a set of encoded values are drawn.

Let $b_k(\xi) \geq 0 \ \forall \ \xi$ be a set of regularly spaced channel basis functions normalized such that $\int_{-\infty}^{\infty} b_k(\xi)\,\mathrm{d}\xi = 1$, without loss of generality[4], and let $a_k \geq 0$ be the channel coefficients representing the distribution, $p(\xi)$, of a pixel,

$$p(\xi) = \sum_{k=1}^{K} a_k b_k(\xi). \tag{3}$$

Let the coefficients be normalized such that $\sum_{k=1}^{K} a_k = 1$, from which $p(\xi) \geq 0 \ \forall \ \xi$ and $\int_{-\infty}^{\infty} p(\xi)\,\mathrm{d}\xi = 1$ follow. Let the random variable $X : P(X < z) = \int_{-\infty}^{z} p(\xi)\,\mathrm{d}\xi$, then expectations of functions $g(X)$ become scalar products with the channel coefficient vector since

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(\xi)p(\xi)\,\mathrm{d}\xi = \sum_{k=1}^{K} a_k \int_{-\infty}^{\infty} g(\xi)b_k(\xi)\,\mathrm{d}\xi = \sum_{k=1}^{K} a_k g_{b_k} \tag{4}$$

with $g_{b_k} = \int_{-\infty}^{\infty} g(\xi)b_k(\xi)\,\mathrm{d}\xi$ (note: independent of the channel coefficients a_k).

[3] Assuming symmetric channels and that the support of the density function is within the representable range of the channel representation.

[4] Conventionally, the basis functions and channel vectors are normalized differently, however, rescaling of the basis functions is compensated by a scaling factor and the channel vectors can be normalized beforehand.

Let μ and σ^2 denote the mean and variance of the represented distribution. For the mean, $g(X) = X$ and $\mu = E[X] = \sum_{k=1}^{K} a_k \mu_{b_k}$ with basis function means $\mu_{b_k} = \int_{-\infty}^{\infty} \xi b_k(\xi)\,\mathrm{d}\xi$, which for symmetric kernels coincide with channel centers. For the variance, $g(X) = (X-\mu)^2$ and

$$\begin{aligned}
\sigma^2 = E\left[(X-\mu)^2\right] &= \sum_{k=1}^{K} a_k \int_{-\infty}^{\infty} \left(\xi^2 - 2\mu\xi + \mu^2\right) b_k(\xi)\,\mathrm{d}\xi = \\
&= \sum_{k=1}^{K} a_k \left(\underbrace{\int_{-\infty}^{\infty} \xi^2 b_k(\xi)\,\mathrm{d}\xi}_{=\sigma_b^2 + \mu_{b_k}^2} - 2\mu \underbrace{\int_{-\infty}^{\infty} \xi b_k(\xi)\,\mathrm{d}\xi}_{=\mu_{b_k}} + \mu^2 \underbrace{\int_{-\infty}^{\infty} b_k(\xi)\,\mathrm{d}\xi}_{=1} \right) = \\
&= \sigma_b^2 \underbrace{\sum_{k=1}^{K} a_k}_{=1} + \sum_{k=1}^{K} a_k \mu_{b_k}^2 - 2\mu \underbrace{\sum_{k=1}^{K} a_k \mu_{b_k}}_{=\mu} + \mu^2 \underbrace{\sum_{k=1}^{K} a_k}_{=1} = \\
&= \sigma_b^2 - \mu^2 + \sum_{k=1}^{K} a_k \mu_{b_k}^2
\end{aligned} \tag{5}$$

where $\sigma_b^2 = \int_{-\infty}^{\infty} (\xi - \mu_{b_k})^2 b_k(\xi)\,\mathrm{d}\xi \;\forall k$. Hence the mean and variance (and thus the standard deviation) of a channel represented distribution can be obtained through scalar products of channel coefficients and weight vectors. Further, these weight vectors only depend on the chosen channel basis functions and can be calculated in advance. The weighted comparison function thus is

$$d(\mathbf{C},\ \mathbf{D}) = \sum_{i,j,k} \frac{1}{\sigma_{ij}} \left| [\mathbf{C}]_{ijk} - [\mathbf{D}]_{ijk} \right| \tag{6}$$

where each σ_{ij} is the estimated standard deviation of each channel vector $[\mathbf{C}]_{ij}$ in the target model. The sum is over all pixels in the bounding box and all channel coefficients.

4.2 Coherence

For combinations of multiple channel encoded measurements of an entity, two properties characterizing the combined channel vector are of interest. Here we refer to them as *evidence* and *coherence*.

Evidence is what is referred to as r_2 in Sect. 2.2, the L_1 norm of the decoding window. When combining channel vectors by addition, r_2 is proportional to the number of samples accumulated within the current decoding window.

Coherence, which we define as r_1^2/r_2^2, is a measure of the consistency of samples resulting in a mode, see Fig. 2, where the right mode has higher coherence than the left mode. Coherence as just defined is a property related to a specific decoding window, and we define the coherence of a full channel vector as the

coherence of the strongest mode. The strongest mode is defined as the decoding window with largest evidence [5].

Motivation of the Definition of Coherence. Several norms and properties of channel encoded entities have been proposed and evaluated in the literature [5,7], however, coherence has not previously been suggested, although it has been suggested for the structure tensor [1]. For notational and conceptual clarity and without loss of generality, basis functions are assumed to be centered at integer positions in this section ($h = 3$).

As shown in [5] and indicated in Sect. 2.2, decoding of $\cos^2$ channel vectors (determining estimates of ξ) are carried out according to

$$\begin{pmatrix} r_1 \cos(\frac{2\pi}{3}(\xi - l)) \\ r_1 \sin(\frac{2\pi}{3}(\xi - l)) \\ r_2 \end{pmatrix} = \begin{pmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \\ 1 & 1 & 1 \end{pmatrix} \mathbf{x}_l \tag{7}$$

where l selects the decoding window and $\mathbf{x}_l$ is the corresponding three elements from the channel vector $\mathbf{x}$ to be decoded. It follows that when all elements in $\mathbf{x}_l$ are equal, $r_1 = 0$ and decoding is ambiguous. When the values within the decoding window are such that r_1 is large, the estimate of ξ is less dependent on small perturbations of the channel coefficients, however, the absolute value of r_1 varies with the scaling of the channel coefficients.

The proposed coherence measure, $\mathrm{coh}(\cdot)$, can be expressed as

$$\mathrm{coh}(\mathbf{x}_l) = \frac{r_1^2}{r_2^2} = \frac{1}{\mathbf{1}^{\mathrm{T}} \mathbf{x}_l \mathbf{x}_l^{\mathrm{T}} \mathbf{1}} \mathbf{x}_l^{\mathrm{T}} \begin{pmatrix} 4 & -2 & -2 \\ -2 & 4 & -2 \\ -2 & -2 & 4 \end{pmatrix} \mathbf{x}_l \tag{8}$$

with $\mathbf{1} = (1\ 1\ 1)^{\mathrm{T}}$ and where the last equality follows from (7). It can easily be verified that $\mathrm{coh}(\mathbf{x}_l) = 0$ when decoding is ambiguous and $\mathrm{coh}(\mathbf{x}_l) = 1$ for a single encoded value or for a combination of encodings of the same value. Further, $\mathrm{coh}(\alpha \mathbf{x}_l)$ is independent of scale ($\alpha > 0$) and, $\mathrm{coh}(\mathbf{x}_l)$ decreases monotonically with a wider distribution of the encoded values within the decoding window. These results build upon properties of the $\cos^2$ kernel, namely that for any value ξ within the representable range of a channel representation, the L_1 and L_2 norms of the corresponding channel vector are constant [5] (and specifically independent of the position of ξ with respect to channel centers). These properties do not hold for Gaussian or B-spline kernels.

Using coherence weighting gives a proposed comparison function

$$d(\mathbf{C},\ \mathbf{D}) = \sum_{i,j,k} |[\mathbf{C}]_{ijk} - [\mathbf{D}]_{ijk}|\ \ (\mathrm{coh}([\mathbf{C}]_{ij}) + \kappa) \tag{9}$$

where $[\cdot]_{ijk}$ denotes the element at index i, j, k and $[\cdot]_{ij}$ denotes the channel vector (with K coefficients) corresponding to pixel i, j in the bounding box. As defined earlier, the coherence of a full channel vector is the coherence of the decoding window corresponding to the strongest mode. $\kappa \geq 0$ is a parameter

representing the trust level of the coherence estimate. The sum is over all pixels in the bounding box and all channel coefficients.

For single-mode distributions, the coherence is inversely related to the variance of the distribution where wide distributions generate low coherence and vice versa. However, for multi-modal distributions, variance is generally large as it is a global property of the distribution. On the contrary, coherence may still be large (corresponding to low variance of the strongest mode) as it is a local property of the individual mode.

5 Target Model q-Update

In this section, the proposed target model update is presented. Here, a discrete time index t is used such that $\mathbf{C}_t$ is the target model obtained after applying all updates up to and including frame t. Similarly, $\mathbf{D}_t$ is the encoded bounding box found in frame t (the best match found by the tracking framework, hence removing the translation subscripts m, n from Sect. 3).

Previous approaches to channel (or distribution field) tracking have used a linear convex combination update

$$\mathbf{C}_t = (1-\gamma)\mathbf{C}_{t-1} + \gamma\mathbf{D}_t \tag{10}$$

with a learning rate parameter $0 < \gamma < 1$. This parameter also determines the forgetting factor $(1-\gamma)$. This update rule is also applicable to image based target representations, where $\mathbf{C}_t$ becomes a weighted mean of the target found in the last frames. However, for the channel based target representation, non-linear update rules are allowed as the update operates on channel coefficients and not directly on intensity values. We propose a power update rule

$$\mathbf{C}_t = \left((1-\gamma)\mathbf{C}_{t-1}^q + \gamma\mathbf{D}_t^q\right)^{\frac{1}{q}} \tag{11}$$

where array exponentiation is to be taken element-wise. The power function is strictly monotonic for positive arguments and thus the order of the channel coefficients is not affected. This bears some resemblance to α-divergences of distributions [11], however, the use is different.

All coefficients in $\mathbf{D}_t$ are non-negative and bounded by the maximum channel activation, $\max_\xi b(\xi)$. Also, from (11) follows that $\mathbf{C}_t \leq \max(\mathbf{C}_{t-1}, \mathbf{D}_t)$ (element wise), ensuring that all elements will remain bounded.

Increasing q shifts the weight towards the larger of each two corresponding elements in $\mathbf{C}_{t-1}$ and $\mathbf{D}_t$. If $[\mathbf{D}_t]_{ijk} > [\mathbf{C}_{t-1}]_{ijk}$, i.e. the current training sample is dominating, increased q leads to faster adaptation to new information. On the other hand, if $[\mathbf{D}_t]_{ijk} < [\mathbf{C}_{t-1}]_{ijk}$, increased q leads to slower forgetting. Increasing γ on the other hand, leads to faster learning and faster forgetting. Using both q and γ, learning rate and forgetting rate can be set independently. Letting $q \to \infty$, (11) becomes $\mathbf{C}_t = \max(\mathbf{C}_{t-1}, \mathbf{D}_t)$, i.e. learning is immediate and the model never forgets. The linear update is a special case ($q = 1$).

Note that for $\mathbf{C}_t$ to become true sampled kernel density estimates of the pixel values, a few more conditions have to be fulfilled in addition to the normalization requirements previously mentioned. In particular, a time dependent learning rate $\gamma = 1/t$ (ensuring equally weighted samples) and $q = 1$ is required. Using a fixed learning rate, more emphasis is given to more recent samples, which usually is beneficial in practice however, with processes not approximately stationary over longer time periods.

6 Experiments

Trackers enhanced with the proposed update scheme and the weighted comparison functions are implemented (in MATLAB) and compared to previous trackers on the VOT2014 challenge benchmark, according to the rules of, and using the evaluation framework provided by the challenge [9].

In the following, DFT and EDFT refer to the previously published trackers by Sevilla et al. [13], and Felsberg [4], respectively. NCC is an example normalized cross correlation tracker distributed with the evaluation framework. Trackers using the proposed q-update scheme are prefixed with a q, and followed by an indication of the value of q. Infinite q is denoted by *max*, a special case as the q-update tend to the max-operation for increasing q. Trackers using the proposed coherence weighted comparison are prefixed with a w and trackers using the proposed standard deviation weighted comparison are prefixed with $w\sigma$. Unmarked trackers use the L_1 norm comparison. For coherence weighting, the parameter κ was set to 2. For all trackers, learning rate γ is set to 0.05 and 15 channels are used.

Three performance measures are available, these are briefly presented here. For the comprehensive version, we refer to [9]. *Accuracy* is the ratio of the joint area of tracker output and ground truth and the union of the two, averaged over each sequence (larger is better). *Robustness* is the reset count, the evaluated tracker is reset as soon as there is no overlap between tracker output and ground truth (smaller is better). *Speed* is the average framerate of the tracker (larger is better).

Two experiments are performed. In the first, denoted *baseline*, each tracker is initialized using the ground truth bounding box of the first frame. In the second experiment, *region noise*, each tracker is initialized with the ground truth bounding box with a random offset. In the second experiment each sequence is evaluated 15 times with different offsets and the mean is reported by the VOT evaluation framework. The results for the baseline experiment are presented in table 1, and the results for the region noise experiment are presented in table 2. For each tracker, both the average and median score over all sequences are presented.

For the baseline experiment (table 1), all channel-based trackers outperform the tracker based on normalized cross correlation (NCC) in accuracy and robustness. For evaluation of the proposed extensions, the trackers using these will be compared to the baseline channel-based tracker (EDFT). Introducing the

Table 1. Summarized results for the baseline experiment, comparison to competing methods (best scores in boldface)

	Mean			Median		
Method	Accuracy	Robustness	Speed	Accuracy	Robustness	Speed
NCC	0.467	2.960	**14.8**	0.423	2.0	**11.5**
DFT	0.531	2.200	6.3	0.534	2.0	6.9
EDFT	0.521	1.840	10.0	0.528	2.0	10.8
qEDFT (q=2)	0.525	2.000	10.6	0.534	2.0	10.8
qEDFT (q=3)	0.536	1.720	7.0	0.541	**1.0**	6.8
qEDFT (q=4)	0.547	1.720	7.1	0.553	**1.0**	7.0
qEDFT (q=5)	0.552	1.720	7.2	0.560	**1.0**	7.1
qEDFT (q=6)	0.540	1.920	6.6	0.553	**1.0**	6.4
wEDFT	0.523	2.040	6.8	0.536	2.0	6.5
qwEDFT (q=2)	0.544	1.560	7.1	0.535	**1.0**	7.1
qwEDFT (q=3)	0.547	1.600	5.5	0.539	**1.0**	5.4
qwEDFT (q=4)	0.550	1.560	5.6	**0.565**	**1.0**	5.5
qwEDFT (q=5)	0.554	1.640	5.1	0.561	**1.0**	5.2
qwEDFT (q=6)	**0.558**	1.920	5.4	0.561	**1.0**	5.4
qwEDFT (q=7)	**0.558**	1.640	5.4	0.560	**1.0**	5.5
maxwEDFT	0.545	1.960	6.4	0.538	2.0	6.0
qwσEDFT (q=2)	0.522	1.400	8.8	0.534	**1.0**	9.1
qwσEDFT (q=3)	0.522	1.440	6.7	0.533	**1.0**	7.0
qwσEDFT (q=4)	0.540	**1.360**	6.2	0.560	**1.0**	6.2
qwσEDFT (q=5)	0.545	1.520	6.5	0.549	**1.0**	6.6
qwσEDFT (q=6)	0.541	1.480	6.8	0.558	**1.0**	7.1
qwσEDFT (q=7)	0.545	1.600	6.9	0.555	**1.0**	7.2
maxwσEDFT	0.547	1.960	7.3	0.549	2.0	7.0

non-linear update (qEDFT) increases accuracy and slightly increases robustness (decreasing failure rate) for increasing q up to $q = 5$. For $q = 6$, performance decreases slightly. Only using the proposed weighted comparison (wEDFT), robustness decreases slightly while accuracy stays similar to EDFT.

The best performance is achieved by combining the non-linear update with the weighted comparison. Using non-linear update and coherence weighted comparison (qwEDFT with $q = 4$), mean accuracy increases more than 5% and mean robustness is 15% better than EDFT. For larger q, accuracy increases further while the robustness degrades. The corresponding standard deviation weighted methods perform slightly inferior to the best methods (the coherence weighted) in terms of mean accuracy. However, the best robustness is achieved by a standard deviation weighted method (qwσEDFT with $q = 4$). In general, accuracy seem to improve with larger q while the best robustness is achieved for q close to 4. For median accuracy, $q = 4$ gives the best performance for both coherence weighted trackers and standard deviation weighted trackers, with better results for coherence weighting.

For the region noise experiments (table 2), accuracy generally increases with increasing q while the best robustness is achieved for $q = 4$ for the standard devi-

Table 2. Summarized results for the region noise experiment, comparison to competing methods (best scores in boldface)

	Mean			Median		
Method	Accuracy	Robustness	Speed	Accuracy	Robustness	Speed
NCC	0.456	2.973	**14.0**	0.414	1.8	**12.0**
DFT	0.493	2.389	6.0	0.512	2.4	5.9
EDFT	0.486	1.973	10.1	0.486	1.9	10.5
qEDFT (q=2)	0.492	2.059	10.3	0.488	1.9	10.8
qEDFT (q=3)	0.497	2.000	6.9	0.518	1.8	6.8
qEDFT (q=4)	0.498	2.032	6.7	0.492	1.7	6.6
qEDFT (q=5)	0.502	2.008	6.7	0.512	**1.3**	6.7
qEDFT (q=6)	0.499	2.093	6.4	0.521	1.6	6.5
wEDFT	0.489	2.088	6.2	0.492	1.9	6.3
qwEDFT (q=2)	0.501	1.835	6.7	0.500	1.9	6.8
qwEDFT (q=3)	0.508	1.819	5.4	0.520	1.6	5.4
qwEDFT (q=4)	0.509	**1.747**	5.1	0.502	1.5	4.9
qwEDFT (q=5)	0.515	1.819	5.2	0.499	1.5	5.1
qwEDFT (q=6)	0.516	1.837	5.2	0.530	1.5	5.2
qwEDFT (q=7)	0.514	1.923	5.1	0.515	1.5	5.1
maxwEDFT	0.514	2.163	6.3	0.500	2.0	6.2
qwσEDFT (q=2)	0.500	2.029	8.6	0.520	1.5	8.8
qwσEDFT (q=3)	0.502	1.832	6.6	0.510	1.5	6.6
qwσEDFT (q=4)	**0.521**	1.893	7.1	**0.534**	1.6	6.7
qwσEDFT (q=5)	0.506	1.803	6.5	0.515	1.7	6.4
qwσEDFT (q=6)	0.510	1.787	6.8	0.517	1.7	6.7
qwσEDFT (q=7)	0.511	1.795	6.6	0.517	1.8	6.4
maxwσEDFT	0.516	2.109	7.9	0.529	2.0	7.8

ation weighted tracker and for $q = 6$ for the coherence weighted tracker. Contrary to the baseline experiments, in the region noise experiments the coherence weighted methods perform best with respect to robustness while the standard deviation weighted methods perform best with respect to accuracy.

In table 3, the results for each sequence for three trackers are presented. A comprehensive description of the sequences themselves is available at the VOT challenge site[5]. Both proposed trackers outperform the EDFT tracker with respect to accuracy on 15 out of 25 sequences. On four sequences the EDFT tracker outperforms the proposed trackers and in three cases, performance is equal among the three trackers. With respect to robustness, all three trackers perform equal on 18 out of 25 sequences. The improvement compared to EDFT with respect to robustness is largest on the sequences where EDFT performs worst. On the *hand2* sequence, EDFT looses track of the object seven times while the proposed qwσEDFT tracker looses track of the object three times.

No parameters have been changed from those used in the baseline trackers, with the exception of the newly introduced parameter q. Since q and the learning

[5] *http://votchallenge.net/vot2014/dataset.html*

Table 3. Detailed *baseline* experiment results for three trackers. Best scores in bold.

	EDFT		qwσEDFT (q=4)		qwEDFT (q=4)	
	accuracy	**robustness**	**accuracy**	**robustness**	**accuracy**	**robustness**
ball	0.51	0	0.57	0	**0.59**	0
basketball	0.56	3	0.57	**1**	**0.59**	3
bicycle	**0.44**	0	0.43	0	0.43	0
bolt	0.51	3	0.56	3	0.56	3
car	0.53	1	0.53	1	0.53	1
david	0.68	0	0.71	0	**0.72**	0
diving	0.16	3	0.16	3	0.16	3
drunk	**0.51**	0	0.49	0	0.50	0
fernando	0.40	2	0.40	2	**0.43**	2
fish1	0.38	4	0.40	4	**0.42**	4
fish2	0.28	6	0.30	**5**	**0.32**	6
gymnastics	**0.55**	2	0.54	2	0.53	2
hand1	0.59	2	0.56	0	**0.60**	0
hand2	0.42	7	0.44	**3**	0.44	6
jogging	0.79	2	0.80	2	0.80	2
motocross	0.18	3	**0.23**	4	0.20	3
polarbear	0.53	0	0.58	0	**0.59**	0
skating	0.61	1	0.61	1	**0.62**	1
sphere	0.62	1	0.69	0	**0.71**	0
sunshade	0.65	3	0.70	1	**0.71**	1
surfing	0.85	0	0.89	0	**0.90**	0
torus	**0.82**	0	0.77	0	0.80	0
trellis	0.51	2	0.52	1	**0.56**	1
tunnel	0.31	0	0.31	0	0.31	0
woman	0.61	1	**0.72**	1	0.69	1

rate γ together determine the effective learning and forgetting rates of the final tracker, a further increase in performance should be possible by jointly optimizing these parameters. Also, by avoiding recomputation, primarily of weights in the search phase, an increase in framerate should be possible. Currently the proposed extensions slows down the tracker to the level of the DFT tracker. Implementing the trackers in C++ should allow video rates on the sequences.

As a final remark, a selection of different comparison functions were evaluated such as L_2, variance weighed L_2 and Hellinger distance. However, these performed inferior to the weighted L_1 norms.

7 Conclusion

In the present work, we have addressed two significant parts of a tracking system, comparison and model update. We have proposed a generalized update rule (q-update) and two weighted comparison functions (coherence weighted and reciprocal standard deviation weighted). The proposals aim to exploit the distribution representation of the target model. On the VOT challenge benchmark,

trackers extended with these proposals showed significant increase in tracking performance. We thus conclude that the proposed methods better utilize the possibilities of the model representation since these proposed methods rely on properties of the channel representation that do not hold for image representations or mean/variance (Gaussian approximation) representations.

Acknowledgements.. This work has been supported by SSF through a grant for the project CUAS, by VR through a grant for the project ETT, through the Strategic Area for ICT research CADICS, and ELLIIT.

References

1. Bigün, J., Granlund, G.H.: Optimal orientation detection of linear symmetry. In: Proceedings of the IEEE First International Conference on Computer Vision, London, Great Britain, pp. 433–438, June 1987
2. Felsberg, M., Forssén, P.E., Scharr, H.: Channel smoothing: Efficient robust smoothing of low-level signal features. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(2), 209–222 (2006)
3. Felsberg, M., Larsson, F., Wiklund, J., Wadströmer, N., Ahlberg, J.: Online learning of correspondences between images. IEEE Transactions on Pattern Analysis and Machine Intelligence (2013)
4. Felsberg, M.: Enhanced distribution field tracking using channel representations. In: IEEE ICCV Workshop on Visual Object Tracking Challenge (2013)
5. Forssén, P.E.: Low and Medium Level Vision using Channel Representations. Ph.D. thesis, Linköping University, Sweden (2004)
6. Granlund, G.H.: An Associative Perception-Action Structure Using a Localized Space Variant Information Representation. In: Proceedings of Algebraic Frames for the Perception-Action Cycle (AFPAC), Germany, September 2000
7. Johansson, B., Elfving, T., Kozlov, V., Censor, Y., Forssén, P.E., Granlund, G.: The application of an oblique-projected landweber method to a model of supervised learning. Mathematical and Computer Modelling **43**, 892–909 (2006)
8. Kass, M., Solomon, J.: Smoothed local histogram filters. In: ACM SIGGRAPH 2010 papers, SIGGRAPH 2010, pp. 100:1–100:10. ACM, New York (2010). http://doi.acm.org/10.1145/1833349.1778837
9. Kristan, M., Čehovin, L., Vojir, T., Nebehay, G.: Visual object tracking challenge 2014 evaluation kit. http://votchallenge.net/vot2014/download/vot2014-guidelines.pdf
10. Pouget, A., Dayan, P., Zemel, R.S.: Inference and computation with population codes. Annu. Rev. Neurosci. **26**, 381–410 (2003)
11. Rényi, A.: On measures of entropy and information. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, pp. 547–561. University of California Press, Berkeley (1961)
12. Scott, D.W.: Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. Annals of Statistics **13**(3), 1024–1040 (1985)
13. Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: IEEE Computer Vision and Pattern Recognition (2012)
14. Zemel, R.S., Dayan, P., Pouget, A.: Probabilistic interpretation of population codes. Neural Computation **10**(2), 403–430 (1998)

Exploiting Contextual Motion Cues for Visual Object Tracking

Stefan Duffner(✉) and Christophe Garcia

Université de Lyon, CNRS INSA-Lyon, LIRIS, UMR5205, F-69621 Lyon, France
stefan.duffner@liris.cnrs.fr

Abstract. In this paper, we propose an algorithm for on-line, real-time tracking of arbitrary objects in videos from unconstrained environments. The method is based on a particle filter framework using different visual features and motion prediction models. We effectively integrate a discriminative on-line learning classifier into the model and propose a new method to collect negative training examples for updating the classifier at each video frame. Instead of taking negative examples only from the surroundings of the object region, or from specific distracting objects, our algorithm samples the negatives from a contextual motion density function. We experimentally show that this type of learning improves the overall performance of the tracking algorithm. Finally, we present quantitative and qualitative results on four challenging public datasets that show the robustness of the tracking algorithm with respect to appearance and view changes, lighting variations, partial occlusions as well as object deformations.

Keywords: Object tracking · Adaptive particle filter · Motion cues

1 Introduction

We consider the problem of automatically tracking a single arbitrary object in a video, where the algorithm is initialised in the first frame from a bounding box around the object to track. No prior knowledge about appearance, shape, or motion of the objects or the environment is used. Also, we focus here on *on-line* tracking, *i.e.*at each time step, only past and present but no future information is used. Applications for on-line visual object tracking are numerous, including, for example, video indexation, Human-Computer or Human-Robot Interaction, video-surveillance, traffic monitoring, or autonomous driving.

In real-world scenarios, this problem is challenging as the object to track may change considerably its appearance, shape, size, and pose in the image (like the articulated human body for example). Furthermore, the object can be partially occluded by itself, other objects, or the environment. The object may also move abruptly or in unpredictable ways. Finally, the environment, *i.e.*the image background, may change considerably and rapidly in videos from moving cameras and be affected by varying illumination.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 232–243, 2015.
DOI: 10.1007/978-3-319-16181-5_16

Recent works [1,2,7,9,13] propose to tackle this problem by a tracking-by-detection framework, where a discriminative detector is trained with object and background samples. At each frame of the video, this detector is applied in a search window to estimate the current position of the object, and the model is updated using this estimate. The advantage of this approach is that no specific motion model needs to be designed and parameterised, and the output is deterministic. Classical tracking algorithms are based on recursive Bayesian filters like Kalman filters or particle filters [12,17,19]. These methods are able to estimate the posterior state distribution of the tracked object and allow for maintaining several state hypotheses. Usually, they explicitly integrate motion models used to predict the next object state by defining a probabilistic transition function independent from the image observations. Some particle filter techniques use some more advanced motion models, like [15], *i.e.*an optical flow-like dense parametric motion estimator with an affine model to propose new state values, as we propose in this paper. Also similar to this paper, parametric motion models have been used to estimate background (*i.e.*camera) motion [6] and segment the object region from the background, *e.g.*[24].

Other recently proposed approaches have also included this type of contextual motion information. For example, Yang *et al.*[23] introduced a method that, throughout a video, continuously discovers objects that move in the same direction as the tracked object by performing a motion correlation analysis. These auxiliary objects help to support and improve tracking by performing inference in a star-structured graphical model that includes their state. Spatial context has also been exploited by using supporters, *i.e.*other objects or feature points around the target in the image. Grabner *et al.*[8], for example, extended the well-known Implicit Shape Model by detecting feature points in the image that have a correlated motion with the target. These supporters are matched from frame to frame and their relative displacement vectors are updated on-line. Also, Wen *et al.*[21] proposed a method that detects supporters (here called contributors), *i.e.*interest points within a local neighbourhood around the target, in order to improve the tracking performance. Similarly, the approach proposed by Sun *et al.*[18] tracks "helper" objects using an on-line Adaboost detector, initialised manually at the first frame. Their relative position is learnt on-line and used to predict the target object's position. Finally, Dinh *et al.*[3] proposed a method using supporters as well as distractors, which are objects with similar appearance to the target. The distractors help to avoid confusion of the tracker with other similar objects in the scene, and they can possibly be used to reason about the objects' mutual occlusion. Supporters are not used directly for the target's state estimation but only to disambiguate between the target and its distractors. Hong *et al.*[10] recently proposed an approach based on the L1 tracker [13] that deals with distractors by automatically learning a metric not only between positive and negative examples but also within the collected negative examples, effectively replacing the originally proposed Euclidean distance.

The disadvantage with using supporting and distracting objects is that several objects need to be detected and tracked, which can be computationally expensive especially with a larger number of objects. Moreover, the success or

failure of data association or, in some methods, matching local features points in successive video frames, heavily depends on the type of object to track and the surrounding background. This process can be error-prone and, in some situations, may rather harm the overall tracking performance. Finally, modelling the spatial, temporal, or appearance-based pairwise relationships between objects and/or interest points can lead to a combinatorial explosion and make the inference on the state space difficult.

To alleviate this problem, in this work, we propose a probabilistic method that dynamically updates the foreground and background model depending on distracting objects or image regions in the scene background. This contextual appearance information is extracted from moving image regions and used to train on-line a discriminative binary classifier that, in each video frame, detects the image region corresponding to the object to track.

Traditionally, these discriminative on-line classifiers used in tracking-by-detection approaches learn negative examples extracted from the image region surrounding the current target object region. This choice is motivated by the fact that the object will move only slightly from one frame to the other w.r.t. the background or other objects, and by computational speed. In contrast, our method uses a stochastic sampling process to extract negative examples from image regions that move. We call these: *contextual motion cues* (see Fig. 1). In that way, regions that correspond to possibly distracting objects are detected efficiently and early, *i.e.*without them having to be inside a search window and without scanning the whole image at each point in time. The contributions of this paper are the following:

- a method for on-line learning of a discriminative classifier using stochastic sampling of negative examples from contextual motion cues in videos,
- the integration of this incremental discriminative model in an efficient adaptive particle filter framework combining effectively several visual cues,
- a thorough evaluation on difficult public benchmarks experimentally showing the performance increase from this type of online learning as well as an improvement over state-of-the-art tracking methods.

2 Tracking Algorithm

Our tracking algorithm is based on a recursive Bayesian framework implemented with a particle filter:

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) = \frac{1}{C}\, p(\mathbf{Y}_t|\mathbf{X}_t) \times \int_{\mathbf{X}_{t-1}} p(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1})\, \mathrm{d}\mathbf{X}_{t-1}\ , \qquad (1)$$

where C is a normalisation constant, $\mathbf{Y}_{1:t}$ are observations from time 1 to t, and $\mathbf{X}_t$ denotes the state at time t. Before describing the main contribution of the paper in section 3, for the sake of completeness, we will first describe the main elements of this model.

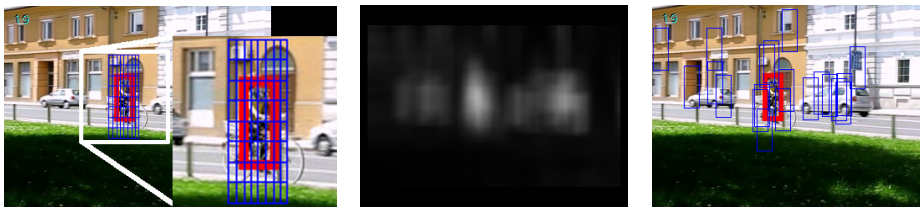


Fig. 1. Illustration of different sampling strategies of negative examples (blue). *Left:* traditional sampling at fixed positions within a search window around the object (red). *Middle:* the motion probability density function m (Eq. 11). *Right:* the proposed negative sampling from m.

2.1 Object State Representation and Inference

The state $\mathbf{X} = (x, y, v_x, v_y, s, e) \in \mathbb{R}^6$ of the object to track is described by an upright bounding box defined by the object's centre (x, y) in the image, its speed (v_x, v_y), scale (s), and eccentricity (e), *i.e.*the ratio of height and width. The state $\mathbf{X}_0$ is initialised manually (for each particle) by a bounding box around the object in the first frame. Then, for each video frame, the particle filter performs its classical steps of *predicting* particles $\mathbf{X}^{(i)}$ sampled from the proposal distribution $q(\mathbf{X}_t|\mathbf{X}_{t-1})$ and *updating* their weights according to the observation likelihood, state dynamics and proposal (see Section 2.2): $w_i = p(\mathbf{Y}_t|\mathbf{X}_t)\frac{p(\mathbf{X}_t|\mathbf{X}_{t-1})}{q(\mathbf{X}_t|\mathbf{X}_{t-1})}$, for each particle $i \in 1..N$. At the end of each iteration, the observation likelihood model parameters are updated using the mean particle of the posterior distribution $p(\mathbf{X}_x|\mathbf{Y}_{1:t})$, and systematic resampling is performed.

2.2 State Dynamics and Proposal Function

The state dynamic model $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is defined for each individual component of $\mathbf{X}$. The position and speed components of the object are described by a mixture of a first-order auto-regressive model with additive Gaussian noise and a uniform distribution allowing for small "jumps" coming from the proposal function (Eq. 2). A simple first order model is used for the scale and eccentricity parameters, s and e.

In order to cope with fairly complex motion of arbitrary objects in videos from a possibly moving camera, we use a proposal function composed of a mixture of three distributions:

$$q(\mathbf{X}_t|\mathbf{X}_{t-1}) = \beta_m p(\mathbf{X}_t|\mathbf{X}_{t-1}) + \beta_f p_f(\mathbf{X}_t|\mathbf{X}_{t-1}) + \beta_d p_d(\mathbf{X}_t|\mathbf{X}_{t-1}) \;, \qquad (2)$$

where β_m, β_f and β_d define the mixture weights, and $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is the state dynamics model. The function

$$p_f(\mathbf{X}_t|\mathbf{X}_{t-1}) = \mathcal{N}(\mathbf{X}_{t-1} + \mathbf{d}; 0, \mathbf{\Sigma}^f) \qquad (3)$$

predicts the new state by performing a parametric robust motion estimation of the image region defined by $\mathbf{X}_t$ like in [14]. The output of this multi-level

estimation is the differential vector $\mathbf{d}$ which updates position and scale. The last term:

$$p_d(\mathbf{X}_t|\mathbf{X}_{t-1}) = \mathcal{N}(\mathbf{X}^d; 0, \mathbf{\Sigma}^d) \tag{4}$$

uses the output $\mathbf{X}^d$ of a detector (see Section 3) that has been trained on-line and that is applied in the neighbourhood around $\mathbf{X}_t$ to predict the new object position and scale (as in [16]). See Section 4 for a summary of parameter values.

2.3 Observation Likelihood

The observation likelihood function $p(\mathbf{Y}|\mathbf{X})$ is a geometric mean of three distributions corresponding to different visual cues described in the following:

$$p(\mathbf{Y}_t|\mathbf{X}_t) = \left(p_H(\mathbf{Y}_t|\mathbf{X}_t)\, p_S(\mathbf{Y}_t|\mathbf{X}_t)\, p_T(\mathbf{Y}_t|\mathbf{X}_t)\right)^{1/3} . \tag{5}$$

Histogram Likelihood Ratio. The histogram likelihood function is defined as a ratio of foreground and background likelihoods:

$$p_H(\mathbf{Y}_t|\mathbf{X}_t) = \frac{p_{FG}(\mathbf{Y}_t|\mathbf{X}_t)}{p_{BG}(\mathbf{Y}_t|\mathbf{X}_t)} , \tag{6}$$

where

$$p_{FG}(\mathbf{Y}_t|\mathbf{X}_t) = \exp\left(-\lambda_{FG} \sum_{r=1}^{9} (D^2[h_t^*(r), h(r, \mathbf{X}_t)])\right) , \tag{7}$$

is the foreground likelihood defined over a grid of 3×3 regions r. D computes the Bhattacharyya distance between the HSV histograms h_t extracted from state $\mathbf{X}_t$ and the respective reference histograms h_t^* initialised from the first frame, and λ_{FG} is a constant. Similarly, the background likelihood:

$$p_{BG}(\mathbf{Y}_t|\mathbf{X}_t) = \exp\left(-\lambda_{BG}(D^2[\hat{h}_t^*, \hat{h}(\mathbf{X}_t)])\right) , \tag{8}$$

is computed over the image region surrounding the object's bounding box.

Global Colour Segmentation Likelihood. In addition to the more local colour models with one histogram per object part, we also use a global colour histogram model based on a pixel-wise colour segmentation of foreground and background. To this end, as above, HSV colour histograms with separate colour and greyscale bins are extracted, one inside the current bounding box of the object, and one around it. Then a probabilistic soft-segmentation is performed computing the probability $p(c_i|z_i)$ of each pixel i inside a search window belonging to the foreground $c = 1$ or background $c = 0$ given its colour z_i.

Then, the likelihood function is defined as:

$$p_S(\mathbf{Y}_t|\mathbf{X}_t) = \frac{\exp(-\lambda_S S_{FG}(\mathbf{X}_t)^2)}{\exp(-\lambda_S S_{BG}(\mathbf{X}_t)^2)} , \tag{9}$$

where λ_S is a constant, S_{FG} is the proportion of foreground pixels, *i.e.*for which $p(c = 1|z) > 0.5$, *inside* the object's bounding box and S_{BG} is the proportion of foreground pixels *outside* the bounding box.

Texture Likelihood. The likelihood $p_T(\mathbf{Y}|\mathbf{X})$ is based on the (greyscale) texture of the object to track. A discriminative classifier is trained at the first frame using the object region as positive and the background regions as negative examples. Then, the classifier is updated at each iteration collecting positive and negative examples from the foreground and background respectively (see Section 3). We use the On-line Adaboost classifier presented by Grabner *et al.*[7] that uses Haar-like features, but any other on-line classifier could be used as well.

The likelihood is based on the detector's confidence $c_D \in [0, 1]$ for the image patch defined by $\mathbf{X}_t$:

$$p_D(\mathbf{Y}_t|\mathbf{X}_t) = \exp(-\lambda_D(1 - c_D)^2) \; . \tag{10}$$

3 Model Adaptation with Contextual Cues

As mentioned earlier, in the particle filter, we use a binary discriminative classifier based on the On-line Adaboost (OAB) algorithm [7] for proposing new particles (Eq. 4) as well as for evaluating the observation likelihood (Eq. 10). The classifier is trained with the first video frame using the image patch inside the object's bounding box as a positive example and surrounding patches within a search window as negative examples. Then, the authors propose to update the classifier at each tracking iteration using the same strategy for extracting positive and negative examples. We refer to [7] for details on the model.

3.1 Background Sampling

We propose to sample negative examples from image regions that contain motion and thus likely correspond to moving objects (see Fig. 1). The idea is that these regions may distract the tracker at some point in time. Therefore it is preferable to learn these negative examples as early as possible, *i.e.*as soon as they appear in the scene. To this end, we first compensate for camera motion between two consecutive frames using a classical parametric motion estimation approach [14]. We apply a three-parameter model to estimate the translation and scale of the scene, and then compute the intensity differences for each pixel with its corresponding pixel in the previous frame. This gives an image $M(x, y)$ approximating the amount of motion present at each position (x, y) of the current frame of the video. We then transform this image into a probability density function (PDF) $m(x, y)$ over the 2-dimensional image space:

$$m(x, y) = Z^{-1} \sum_{(u,v)\in\Omega(x,y)} M(u, v) \; , \tag{11}$$

where $\Omega(x, y)$ defines an image region of the size of the bounding box of the object being tracked, centred at (x, y), and Z is a constant normalising the density function to sum up to 1. Thus, $m(x, y)$ represents the relative amount of motion inside the region centred at (x, y). Finally, N^- image positions (x, y) are sampled from this PDF corresponding to rectangles centred at (x, y), where, statistically, regions with high amount of motion are sampled more often than static image regions. This process is illustrated in Fig. 1.

3.2 Classifier Update

The N^- image patches corresponding to the sampled regions as well as the positive example coming from the mean particle of the tracker are then used to update the classifier. In this case, the OAB method needs a balanced number of positives and negatives, thus the positive example is used N^- times, alternating positive and negative updates.

The advantage of sampling positions from these motion cues is that we don't need to care about explicitly detecting, initialising, tracking, and eventually removing a certain number of distracting objects at each point in time. Note that, we could also sample regions of different scales but as scale does not change rapidly in most videos the benefit of this would be relatively small. Note also that the PDF could as well include appearance similarity with the tracked target. However, this would considerably increase the computational complexity.

4 Experiments

4.1 Parameters

The following tracking parameters that have been used for all the experiments:

$\hat{\Sigma}$	$\bar{\Sigma}$	$\Sigma^{f/p}$	β_m	β_f	β_d	λ_{FG}	λ_{BG}	λ_S	λ_D
$(7,7)$	$(0.001, 0.001)$	$(1, 1, 10^{-4}, 10^{-4})$	0.7	0.2	0.1	120	36	0.1	10

The variances for x and y values are scaled by $\frac{w}{200}$, w being the current width of the bounding box. We should highlight that only 100 particles have been used throughout all experiments. This turns out to be sufficient due to our effective proposal and discriminative likelihood functions.

4.2 Datasets

We performed a quantitative evaluation on 4 challenging public tracking datasets:

Babenko[1] [2] contains 8 videos of objects that undergo mostly rigid deformations and some rather large lighting variations and partial occlusions. Most of the videos are in grey-scale format (except "David", "Girl", and "Face Occl. 1").

Non-rigid objects[2] is a more challenging dataset composed of 11 videos showing moving objects that undergo considerable rigid and non-rigid deformations.

VOT2013[3] is the Visual Object Tracking (VOT) Benchmark 2013 [11] containing 16 videos that show a large variability in terms of camera motion, illumination change, occlusion, object size, and motion. Four of these sequences (David, diving, face, jump) are also part of the first or second dataset.

VOT2014[3] contains 25 challenging videos including eight from VOT2013.

[1] http://vision.ucsd.edu/~bbabenko/project_miltrack.shtml
[2] http://lrs.icg.tugraz.at/research/houghtrack/
[3] http://votchallenge.net/

Table 1. Babenko sequences: percentage of correctly tracked frames with fixed negative sampling, sampling from motion, combined fixed+random, and fixed+motion sampling

	fixed	fixed+rand.	motion	fixed+mot.
David	**62.6**	61.9	60.6	61.5
Sylvester	49.8	59.6	**81.6**	78.7
Girl	67.9	44.4	73.0	**74.2**
Face Occlusions 1	98.4	**100.0**	**100.0**	**100.0**
Face Occlusions 2	97.7	95.8	95.0	**98.4**
Coke	88.8	92.5	92.9	**93.2**
Tiger 1	**60.3**	58.9	59.7	59.7
Tiger 2	90.4	93.2	**97.3**	**97.3**
average	76.99	75.80	82.51	**82.88**

Table 2. Non-rigid object sequences: percentage of correctly tracked frames with fixed negative sampling, sampling from motion, combined fixed+random, and fixed+motion sampling

	fixed	fixed+rand.	motion	fixed+mot.
Cliff-dive 1	**100.0**	**100.0**	**100.0**	**100.0**
Motocross 1	75.9	84.7	94.1	**99.1**
Skiing	98.1	89.6	96.4	**99.2**
Mountain-bike	**100.0**	**100.0**	**100.0**	**100.0**
Cliff-dive 2	51.8	70.2	63.3	**73.8**
Volleyball	99.9	88.5	**100.0**	99.9
Motocross 2	**100.0**	**100.0**	**100.0**	**100.0**
Transformer	91.1	92.9	**94.4**	91.5
Diving	75.0	76.0	70.5	**77.4**
High Jump	52.5	59.8	**69.7**	66.6
Gymnastics	88.9	**99.1**	**99.1**	**99.1**
average	84.83	87.35	89.76	**91.5**

Note that long-term tracking datasets like LTDT2014 are not suitable for evaluating our approach as these videos contain longer periods of full occlusion which requires the algorithm to be able to re-detect the tracked object after occlusion.

4.3 Evaluation

We performed several experiments with different evaluation protocols. For the first two datasets we evaluated the robustness of the proposed algorithm by measuring the proportion of correctly tracked frames. A frame is counted as correct, if the tracking accuracy $A = \frac{R_T \cap R_{GT}}{R_T \cup R_{GT}}$ is greater than a threshold, where R_T is the rectangle corresponding to the mean particle from the tracking algorithm, and R_{GT} is the ground truth rectangle surrounding the object. We set the threshold to 0.1 in order not to penalise fixed-size, fixed-ratio trackers in our comparison. For every experiment and video sequence, the proposed algorithm has been run 5 times and the average result is reported.

For the VOT datasets, we used the evaluation protocol of the VOT2013 benchmark, which measures accuracy and robustness. For evaluating the accuracy, the measure A, defined above, is used. The robustness is measured in terms of number of tracking failures, where trackers are re-initialised after failures.

Table 3. Results of the proposed algorithm on the VOT2013 dataset

	accuracy			**robustness**		
	baseline	region-noise	greyscale	baseline	region-noise	greyscale
average	0.597	0.579	0.590	0.458	0.417	0.867

Table 4. Overall ranking result with the VOT2013 dataset. Only the first 6 out of 28 ranks are shown. The numbers represent the actual average ranking.

baseline		region-noise		greyscale	
PLT	4.96	PLT	3.58	PLT	3.96
MCT	**6.62**	**MCT**	**5.08**	FoT [20]	4.75
FoT [20]	8.25	CCMS	8.33	**MCT**	**6.25**
EDFT [4]	9.5	FoT [20]	9.04	EDFT [4]	7.5
CCMS	9.54	LGT++ [22]	9.04	GSDT [5]	9.5
LGT++ [22]	10.2	EDFT [4]	9.08	LGT++ [22]	9.58

Table 5. Results of the proposed algorithm on the VOT2014 dataset

	accuracy		**robustness**	
	baseline	region-noise	baseline	region-noise
ball	0.58	0.42	0.40	0.20
basketball	0.54	0.52	1.53	1.67
bicycle	0.54	0.52	0.00	0.00
bolt	0.58	0.62	0.60	1.13
car	0.66	0.62	0.00	0.00
david	0.70	0.65	0.00	0.00
diving	0.38	0.36	1.27	1.20
drunk	0.55	0.53	0.00	0.07
fernando	0.32	0.33	2.07	3.00
fish1	0.33	0.31	1.67	1.73
fish2	0.31	0.28	3.40	4.13
gymnastics	0.53	0.51	1.07	1.20
hand1	0.46	0.40	1.53	1.93
hand2	0.35	0.34	6.20	6.40
jogging	0.75	0.69	0.93	1.00
motocross	0.45	0.47	1.27	2.00
polarbear	0.70	0.64	0.00	0.00
skating	0.55	0.47	0.33	0.80
sphere	0.75	0.82	0.00	0.00
sunshade	0.63	0.58	0.00	0.00
surfing	0.71	0.71	0.00	0.00
torus	0.53	0.51	1.13	1.47
trellis	0.59	0.53	1.00	1.07
tunnel	0.35	0.41	0.33	0.53
woman	0.58	0.61	0.00	0.13
average	**0.54**	**0.51**	**0.99**	**1.19**

Every video sequence is evaluated 15 times and the average results are reported. In addition to this "baseline" experiment, there are two other experiments using the same data. In the "region-noise" experiment the initial bounding box is randomly, slightly shifted for each run, and in the "greyscale" experiment, each video is transformed into greyscale format. See [11] for more details.

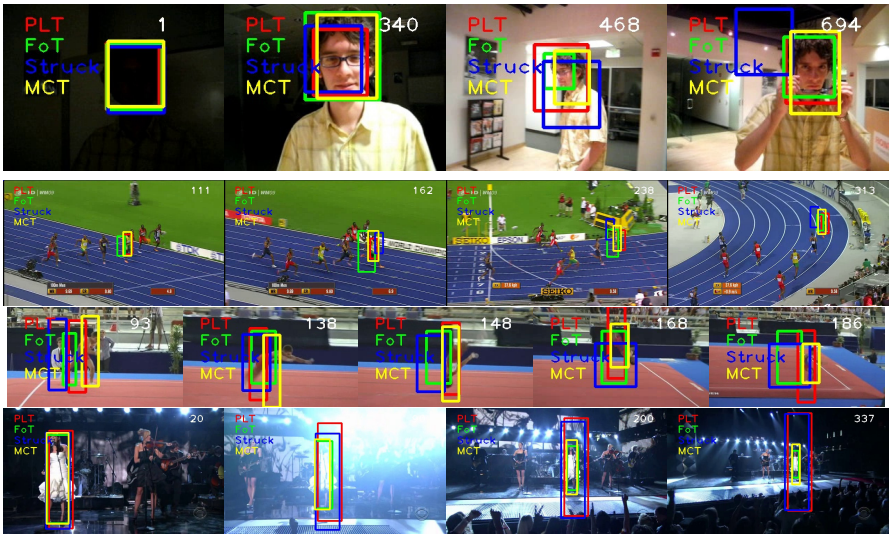


Fig. 2. Tracking results for PLT, FoT, Struck, and MCT on the sequences "David", "Bolt", "Gymnastics", and "Singer" (VOT2013/2014). Tracking is very robust to partial occlusions, illumination changes, deformations, pose or other appearance changes. The second last example shows some difficulties of MCT to adapt to different aspect ratios. And the last example illustrates the problem of drastic size change for single-scale trackers like Struck and PLT.

4.4 Results

In the first experiments, we evaluated four different strategies for the collection of negative examples of the discriminative OAB classifier (*c.f.* section 3):

fixed: N^- negatives are taken from fixed positions around the positive example inside the search window, which is twice the size of the object's bounding box.
fixed+random: $N^-/2$ examples are taken from fixed position (as for "fixed"), and $N^-/2$ examples are sampled from random image positions.
motion: N^- negative examples are sampled from the contextual motion distribution m (Eq. 11).
fixed+motion: $N^-/2$ examples are taken from fixed positions, and $N^-/2$ examples are sampled from the contextual motion distribution.

In any case, the negative examples do not overlap more than 70% with the positive ones in the image.

Table 1 and 2 show the results for the first two datasets in terms of the percentage of correctly tracked frames.

In most cases, the sampling of negative examples from the contextual motion PDF, *i.e.*"motion" and "fixed+motion", improves the tracking performance. For the Babenko sequences, the improvement is smaller because there are not many other moving objects that can distract the tracker. On average, the best strategy

is "fixed+motion", with a relative improvement of around 7.5%. We use this strategy for the following experiments and call the overall tracking algorithm "Motion Context Tracker" (MCT).

We further evaluated MCT with the VOT2013 dataset using the protocol of the VOT challenge and comparing it with 27 other state-of-the-art tracking methods. Table 3 shows the average accuracy and robustness with the three different experiments explained above: baseline, region-noise, and greyscale.

Table 4 lists the top 6 ranks for each experiment, combining accuracy and robustness. The results of MCT are very competitive, being the second-best method for baseline and region-noise and third-best for greyscale. Only, one method, the Pixel-based LUT Tracker (PLT), is consistently outperforming MCT on this dataset. It is an optimisation of the tracker called "Struck" [9], currently unpublished but some explanation can be found in [11]. Note that, PLT is a single-scale tracker and it uses different feature sets for greyscale and colour video sequences.

Table 5 shows the accuracy and robustness results for the VOT2014 dataset.

Finally, Fig. 2 shows some qualitative tracking results on some of the video sequences. One can see that the algorithm is very robust to changes in object appearance, illumination, pose as well as complex motion, and partial occlusions. The algorithm runs at around 4fps (or with a single-scale OAB detector: at 20fps) for a frame size of 320×240 on an Intel Xeon 3.4GHz not counting the initialisation phase and screen display.

5 Conclusions

We presented a new efficient particle filter-based approach for tracking arbitrary objects in videos. The method combines generative and discriminative models, by effectively integrating an online learning classifier. We propose a new method to train this classifier that samples the position of negative examples from contextual motion cues instead of a fixed region around the tracked object. Our extensive experimental results show that this procedure improves the overall tracking performance. Further, the proposed tracking algorithm gives state-of-the-art results on four different challenging tracking datasets, effectively dealing with large object shape and appearance changes, as well as complex motion, varying illumination conditions and partial occlusions.

References

1. Avidan, S.: Ensemble tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(2), 261–271 (2007)
2. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: Proc. of the International Conference on Computer Vision and Pattern Recognition, December 2009
3. Dinh, T., Vo, N., Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: Proc. of the Computer Vision and Pattern Recognition (2011)
4. Felsberg, M.: Enhanced distribution field tracking using channel representations. In: Visual Object Tracking Challenge (VOT 2013), ICCV (2013)

5. Gao, J., Xing, J., Hu, W., X., Z.: Graph embedding based semi-supervised discriminative tracker. In: Visual Object Tracking Challenge (VOT 2013), ICCV (2013)
6. Gengembre, N., Pérez, P.: Probabilistic color-based multi-object tracking with application to team sports. Tech. Rep. 6555, INRIA (2008)
7. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: Proc. of the British Machine Vision Conference (2006)
8. Grabner, H., Matas, J., Van Gool, L., Cattin, P.: Tracking the invisible: Learning where the object might be. In: Proc. of the Computer Vision and Pattern Recognition, vol. 3, pp. 1285–1292 (2010)
9. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: Proc. of the International Conference on Computer Vision (2011)
10. Hong, Z., Mei, X., Tao, D.: Dual-Force Metric Learning for Robust Distracter-Resistant Tracker. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 513–527. Springer, Heidelberg (2012)
11. Kristan, M., Cehovin, L., Pflugfelder, R., Nebehay, G., Fernandez, G., Matas, J., et al.: The Visual Object Tracking VOT 2013 challenge results. In: Proc. of the International Conference on Computer Vision (Workshops) (2013)
12. Maggio, E.: Adaptive multifeature tracking in a particle filtering framework. IEEE Trans. on Circuits and Systems for Video Technology 17(10), 1348–1359 (2007)
13. Mei, X., Ling, H.: Robust visual tracking and vehicle classification via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence 33(11), 2259–72 (2011)
14. Odobez, J.M., Bouthemy, P.: Robust multiresolution estimation of parametric motion models. Journal of Visual Communication and Image Representation 6(4), 348–365 (1995)
15. Odobez, J.M., Gatica-Perez, D., Ba, S.O.: Embedding motion in model-based stochastic tracking. IEEE Trans. on Image Processing **15**(11), 3514–3530 (2006)
16. Okuma, K., Taleghani, A., de Freitas, N., Little, J.J., Lowe, D.G.: A Boosted Particle Filter: Multitarget Detection and Tracking. In: Pajdla, T., Matas, J.G. (eds.) ECCV 2004. LNCS, vol. 3021, pp. 28–39. Springer, Heidelberg (2004)
17. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-Based Probabilistic Tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
18. Sun, Z., Yao, H., Zhang, S., Sun, X.: Robust visual tracking via context objects computing. In: Proc. of the International Conference Image Processing, pp. 509–512, September 2011
19. Čehovin, L., Kristan, M., Leonardis, A.: Robust visual tracking using an adaptive coupled-layer visual model. IEEE Trans. on Pattern Analysis and Machine Intelligence 35(4), 941–953 (2013)
20. Vojíř, T., Matas, J.: Robustifying the flock of trackers. In: Computer Vision Winter Workshop, pp. 91–97 (2011)
21. Wen, L., Cai, Z., Lei, Z., Yi, D., Li, S.: Robust online learned spatio-temporal context model for visual tracking. IEEE Trans. on Image Processing 23(2) (2013)
22. Xiao, J., Stolkin, R., Leonardis, A.: An enhanced adaptive coupled-layer lgtracker++. In: Visual Object Tracking Challenge (VOT 2013), ICCV (2013)
23. Yang, M., Wu, Y., Hua, G.: Context-aware visual tracking. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(7), 1195–1209 (2009)
24. Zhang, G., Jia, J., Xiong, W., Wong, T.T., Heng, P.A., Bao, H.: Moving object extraction with a hand-held camera. In: Proc. of the International Conference on Computer Vision, pp. 1–8 (2007)

Clustering Local Motion Estimates for Robust and Efficient Object Tracking

Mario Edoardo Maresca and Alfredo Petrosino(✉)

Department of Science and Technology, University of Naples Parthenope, Naples, Italy
mariomaresca@hotmail.it, alfredo.petrosino@uniparthenope.it

Abstract. We present a new short-term tracking algorithm called Best Displacement Flow (BDF). This approach is based on the idea of 'Flock of Trackers' with two main contributions. The first contribution is the adoption of an efficient clustering approach to identify what we term the 'Best Displacement' vector, used to update the object's bounding box. This clustering procedure is more robust than the median filter to high percentage of outliers. The second contribution is a procedure that we term 'Consensus-Based Reinitialization' used to reinitialize trackers that have previously been classified as outliers. For this reason we define a new tracker state called 'transition' used to sample new trackers in according to the current inlier trackers.

Keywords: Visual object tracking · Optical flow · Motion-based · Texture-less tracking

1 Introduction

The main challenge of an object tracking system is the difficulty to handle the appearance changes of the target object. The appearance changes can be caused by intrinsic changes such as pose, scale and shape variation and by extrinsic changes such as illumination, camera motion, camera viewpoint, and occlusions.

For instance, our approach Matrioska [13], while ranking closely to one of the best performing tracker EDFT [4] (see the Accuracy-Robustness plot shown in Figure 1 for the trackers that joined the VOT2013 challenge [10]), was not able to rank better due to failures on some sequences. Indeeed, as Figure 2 shows, Matrioska fails on sequences such as *hand* and *torus* mainly due to two factors: (i) texture-less objects and (ii) non-rigid transformations, resulting in low values for the Accuracy and Robustness, as reported in Table 1.

To model such variability, various approaches have been proposed, such as: updating a low dimensional subspace representation [15], MIL based [1] and template or patch based. Other approaches are reported in recent surveys ([10], [23] and [18]), and specifically [2–7,11,12,14,16,17,19,20,22,24–26].

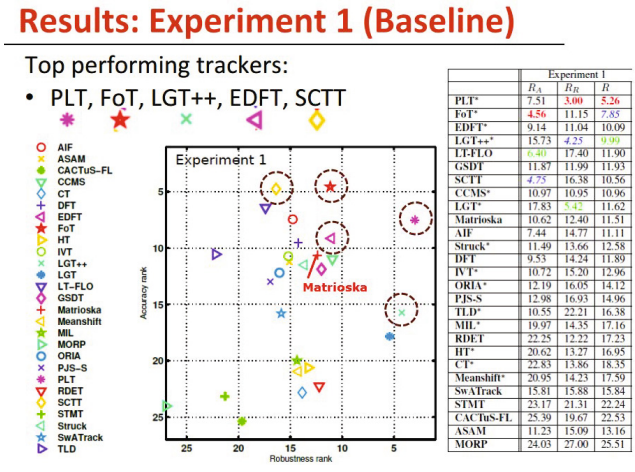
In this paper we introduce a new short-term tracking algorithm named *Best Displacement Flow* (BDF), that is aimed to avoid the Matrioska's failure cases.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 244–253, 2015.
DOI: 10.1007/978-3-319-16181-5_17

To achieve a better robustness over texture-less objects we adopt a different visual representation: Matrioska is based on a sparse representation with the use of point features, whereas BDF adopts a dense approach represented by local trackers that cover the entire object.

BDF is inspired by the Flock of Features ([9], [20]) where a set of displacements, estimated by local trackers, are robustly combined to localize the target object. We propose different contributions and we show how this approach reaches state-of-the-art performance for sequences in which a re-detector module is not required.

The main contributions, i.e. the clustering procedure and the consensus-based reinitialization, are discussed in sections 2.2 and 2.3, respectively.



	Experiment 1		
	R_A	R_R	R
PLT*	7.51	3.00	5.26
FoT*	4.56	11.15	7.85
EDFT*	9.14	11.04	10.09
LGT++*	15.73	4.25	9.99
LT-FLO	6.40	17.40	11.90
GSDT	11.87	11.99	11.93
SCTT	4.75	16.38	10.56
CCMS*	10.97	10.95	10.96
LGT*	17.83	5.42	11.62
Matrioska	10.62	12.40	11.51
AIF	7.44	14.77	11.11
Struck*	11.49	13.66	12.58
DFT	9.53	14.24	11.89
IVT*	10.72	15.20	12.96
ORIA*	12.19	16.05	14.12
PJS-S	12.98	16.93	14.96
TLD*	10.55	22.21	16.38
MIL*	19.97	14.35	17.16
RDET	22.25	12.22	17.23
HT*	20.62	13.27	16.95
CT*	22.83	13.86	18.35
Meanshift*	20.95	14.23	17.59
SwATrack	15.81	15.88	15.84
STMT	23.17	21.31	22.24
CACTuS-FL	25.39	19.67	22.53
ASAM	11.23	15.09	13.16
MORP	24.03	27.00	25.51

Fig. 1. The Accuracy-Robustness plot of VOT2013 challenge

Fig. 2. Snapshots of the *hand* and *torus* sequences showing typical Matrioska failure cases: texture-less objects and non-rigid transformations

Table 1. Matrioska's results on hand, torus and diving sequences of the VOT2013 dataset

	accuracy	**robustness**	**speed (fps)**
hand	0.37	7.00	24.81
torus	0.26	8.00	16.25
diving	0.32	4.00	14.00
iceskater	0.48	4.00	11.49

2 Best Displacement Flow (BDF)

In the following sections we will describe our tracking approach for short-term sequences. A short-term tracker is an algorithm able to track an unknown object for short sequences in which the target object is visible through the entire sequence, and it usually does not have a re-detector module (if the object goes out of the scene the tracker will drift).

Our approach called Best Displacement Flow is inspired by (in order of publication) Flock of Features [9], Median Flow [8] and Flock of Trackers ([20], [21]) where a set of displacements, estimated by local trackers, are robustly combined to localize the target object. The name of our approach, BDF, remarks the most important difference between our tracker and the other approaches: we apply a clustering procedure over all local trackers estimates to filter outliers instead of using the median filter. The biggest cluster identifies what we term the *best displacement* vector used to update the position of the target bounding box.

The following sections describe in detail the main components of our system: the multi-size initialization (section 2.1), the clustering procedure (section 2.2) and the consensus-based reinitialization (section 2.3).

2.1 Multi-size Initialization

The initialization is the first step of our approach. Unlike other approaches, which use the same patch size for each tracker (both MedianFlow [8] and Flock of Trackers [20] use a single grid with a fixed cell size), we allow the initialization of local trackers with different patch sizes, as Figure 3 shows. To estimate the optical flow we use the Block Matching algorithm, i.e. each patch is used as a template to find the displacement that optimizes a cost function in the following frame.

For this reason the patch size becomes an important factor, hence the use of patches with different sizes ensures a greater robustness. Note that we do not constraint the trackers position inside a cell (i.e. the local trackers can freely move inside the object bounding box).

2.2 Displacement Clustering

Each local tracker, after the initialization in the first frame, estimates the displacement that optimizes a cost function (usually the SSD or the NCC) using the block matching algorithm for the optical flow estimation.

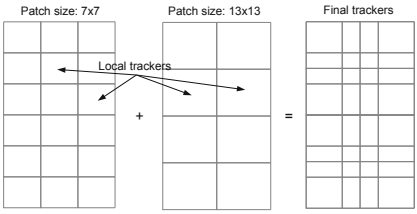


Fig. 3. The final trackers are obtained by superimposing grids with different patch sizes. In this case two grids with a size of 7x7 and 13x13 pixel. Using patches with different sizes ensure a greater robustness over different appearances of objects.

Once every tracker estimated its displacement vector (i.e. the optical flow) we need to filter each possible outlier. The median filter is robust up to 50% of outliers and this can represent a limitation in many challenging sequences. For this reason we employ a clustering procedure that produces good results even in presence of a greater percentage of outliers. The only exception to this rule is represented by rotational motion of the object, only in this case the median filter is better suited for inlier/outlier filtering.

Figure 4 shows this process: to efficiently cluster all displacements each tracker votes its displacement in the accumulator space. After all votes have been casted, the bucket with most votes identifies what we call the *best displacement* vector $\boldsymbol{\beta}$. Note that this process is equivalent to the hierarchical clustering using the infinity norm $\|\mathbf{d}\|_\infty = \max\{|d_1|, \ldots, |d_n|\}$ and a cut-off threshold of 1 but it is much more efficient. In this illustrative scenario the median filter would not produce a good results due to a high percentage of outliers (8 trackers out of 10 are outliers). Note that we use the infinity norm and not the Euclidean norm because: (i) it is more efficient and (ii) the accumulator space is partitioned into squares.

Furthermore, to improve the clustering process we assign a weight for every tracker based on its past performances (i.e. the weight is increased each time the tracker response agrees with the best displacement vector) that is used to cast a weighted vote in the accumulator space. The best displacement $\boldsymbol{\beta}$ is used to shift the center of the bounding box as follows: $\mathbf{O^{bb}_{t+1}} = \mathbf{O^{bb}_{t}} + \boldsymbol{\beta}_{t+1}$ where $\mathbf{O^{bb}}$ represents the center of the bounding box.

2.3 Consensus-Based Reinitialization

After the clustering procedure, each tracker response $\boldsymbol{\Delta_i}$ is compared to the best displacement vector $\boldsymbol{\beta}$. If their distance is greater than a threshold δ_s we set the tracker state $State(t_i)$ to outlier as follows:

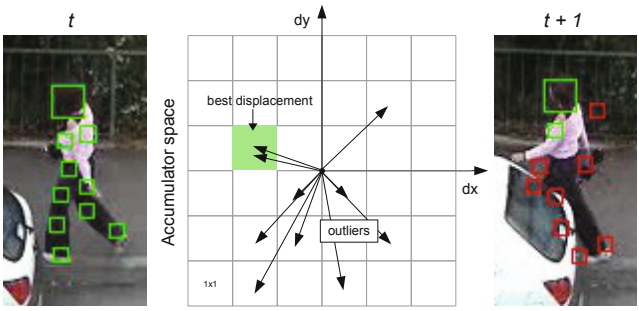


Fig. 4. Each tracker votes its displacement in the accumulator space. The most voted bucket identifies the *best displacement* vector used to update the bounding box.

$$State(t_i) = \begin{cases} inlier & \text{if } \|\boldsymbol{\Delta_i} - \boldsymbol{\beta}\|_\infty < \delta_s \\ outlier & \text{otherwise} \end{cases} \tag{1}$$

where δ_s is equal to 7. Once a tracker state is outlier it will not be used in the following frames to cast new displacement votes. For this reason we need a procedure to reinitialize the trackers when the number of inliers falls under a certain threshold δ_n (we set δ_n to 25% of the total number of trackers).

The *consensus-based reinitialization*, for every outlier tracker, performs two steps: (i) reinitializes the default position of the tracker inside the current bounding box and (ii) sets the state of the tracker to *transition*.

When a tracker state is equal to transition it will not contribute to the clustering procedure. The transition state indicates that the tracker has been reinitialized and it needs to be validated.

This validation is based on the consensus with the current inlier trackers, i.e. a tracker whose state is transition can be promoted to inlier if its response agrees (see equation 1) with the best displacement vector for at least δ_t frames following its reinitialization (we set δ_t to 3 frames) otherwise it is classified again as outlier.

Figure 5 shows the state diagram of this process, note that a tracker state, at any given time, can be either inlier or outlier or transition.

In the first frame all trackers are initialized as inliers. When the distance between a tracker displacement and the best displacement $\boldsymbol{\beta}$ is greater than a threshold δ_s, the tracker state is set to outlier and it will not be used again until the reinitialization. When the number of inliers falls under a threshold δ_n, the consensus-based reinitialization sets the state to transition to every outlier tracker. Only the transition trackers that agree for at least δ_t frames with the current inliers are promoted to the inlier state.

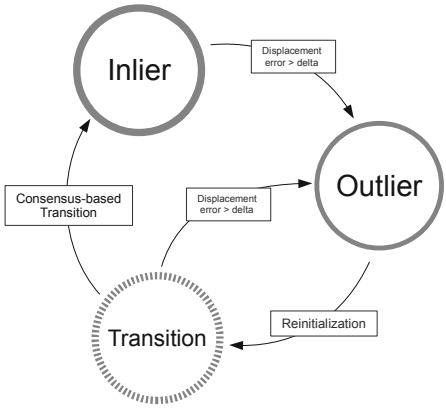


Fig. 5. The three tracker states, displayed in a state diagram

3 Quantitative Evaluation

In this section we evaluate our approach with benchmark sequences that are commonly used in the literature with the VOT2014 evaluation kit. The kit performs

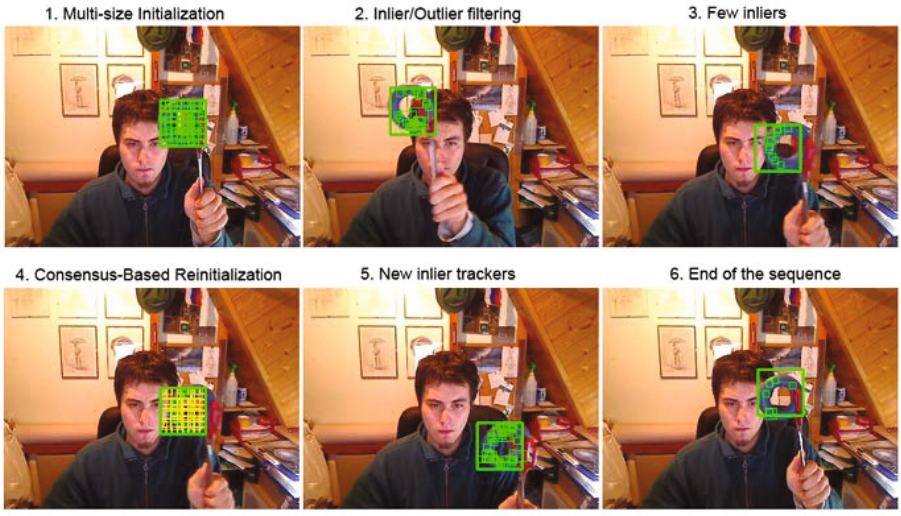


Fig. 6. BDF on *torus* sequence

Table 2. Results for tracker *BDF*

	Baseline			Region Noise		
	accuracy	**robustness**	**speed (FPS)**	**accuracy**	**robustness**	**FPS**
ball	0.52	2.00	177.21	0.52	2.80	182.02
basketball	0.56	2.00	99.82	0.47	2.33	100.47
bicycle	0.46	1.00	157.77	0.48	1.00	171.23
bolt	0.47	5.00	86.87	0.40	5.20	88.79
car	0.41	1.00	106.23	0.42	1.00	132.56
david	0.70	0.00	94.95	0.66	0.00	105.12
diving	0.29	2.00	91.68	0.30	2.13	99.40
drunk	0.53	1.00	76.59	0.49	0.80	83.82
fernando	0.42	1.00	53.48	0.40	1.47	53.78
fish1	0.29	2.00	112.43	0.28	2.67	123.92
fish2	0.23	5.00	81.10	0.17	5.53	89.39
gymnastics	0.57	1.00	62.77	0.50	1.53	67.57
hand1	0.55	1.00	89.23	0.55	1.07	92.99
hand2	0.48	1.00	87.22	0.46	1.00	85.14
jogging	0.75	2.00	117.94	0.62	1.13	113.35
motocross	0.41	0.00	64.53	0.39	1.13	91.79
polarbear	0.53	0.00	62.58	0.52	0.00	67.95
skating	0.57	2.00	69.53	0.49	1.20	75.06
sphere	0.36	0.00	109.03	0.62	0.20	110.91
sunshade	0.75	0.00	108.44	0.69	0.00	110.73
surfing	0.49	0.00	181.57	0.43	0.13	185.70
torus	0.61	0.00	66.08	0.63	0.27	78.06
trellis	0.48	0.00	124.56	0.45	0.20	163.76
tunnel	0.29	0.00	56.89	0.28	0.33	68.15
woman	0.61	1.00	68.84	0.61	1.07	73.63
Average	0.49	1.20	96.29	0.47	1.37	104.6

two experiments: Experiment "Baseline" and Experiment "Region Noise". Both the experiments are evaluated with two metrics: (i) accuracy and (ii) failures.

Accuracy is the mean overlap computed only over the valid frames on multiple trials. Failures indicate the number of times the algorithm drifted (i.e. the overlap between the tracker bounding box and the ground truth bounding box is equal to zero).

The overlap ϕ_i, given the ith frame, is defined as:

$$\phi_i = \frac{A^T \cap A^{GT}}{A^T \cup A^{GT}}$$

where A^T and A^{GT} represent the tracker bounding box and the ground truth bounding box.

As show in Table 2 BDF is able to get accuracy values of 0.49 for Baseline and 0.47 for Region Noise, while robustness values in average of 1.20 for Baseline

and 1.37 for Region Noise. We tested our C++ implementation on an Intel i7-920 processor, getting FPS of 96.29 for Baseline and 104.6 for Region Noise.

As an example, Figure 6 illustrates the Best Displacement Flow tracking the object in the *torus* sequence. In the first frame all trackers are initialized with three different patch sizes. The clustering procedure, in the following frames, identifies the *best displacement* vector that is used to: (i) update the bounding box and (ii) filter inlier/outlier trackers. When the number of inlier trackers (represented with red squares) falls under a threshold δ_n, the outlier trackers are reinitialized in their default position with *transition* state (represented with yellow squares). Only the trackers that agree with the current inliers, for at least δ_t frames, are promoted to the inlier state.

Best Displacement Flow is an optical-flow based tracker, hence it fails when the optical-flow estimation doesn't return a good result. The failure cases include: total occlusions and very large displacements between consecutive frames. The failures of *bicycle*, *basketball*, *car*, *fernando*, *fish2*, *jogging* and *woman* are due to total occlusions, whereas the failures of *bolt*, *fish1*, *fish2*, *gymnastics* and *skating* are due to abrupt appearance changes between consecutive frames.

4 Conclusions

In this paper we introduced a new short-term tracking algorithm called Best Displacement Flow (BDF) that tracks an object by robustly combining a set of local tracker estimates. We introduced two main contributions: (i) a clustering procedure to identify the *best displacement* vector and (ii) a *consensus-based reinitialization* to sample new trackers in according to the current inliers using a third state called *transition*.

Our approach reaches state-of-the-art performance and it is more robust than the median filter-based approaches in challenging sequences. Regarding future developments, it would be interesting to extend our approach by adding a re-detector module for handling situations such as: total occlusion and object out of the camera view.

References

1. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning (2011)
2. Chen, W., Cao, L., Zhang, J., Huang, K.: An adaptive combination of multiple features for robust tracking in real scene. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 129–136, December 2013
3. Duffner, S., Garcia, C.: PixelTrack: a fast adaptive algorithm for tracking non-rigid objects. In: International Conference on Computer Vision (ICCV 2013), Proceedings of the International Conference on Computer Vision, pp. 2480–2487, December 2013. http://liris.cnrs.fr/publis/?id=6293
4. Felsberg, M.: Enhanced distribution field tracking using channel representations. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 121–128, December 2013

5. Gao, J., Xing, J., Hu, W., Zhang, X.: Graph embedding based semi-supervised discriminative tracker. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 145–152, December 2013
6. Godec, M., Roth, P.M., Bischof, H.: Hough-based tracking of non-rigid objects. In: Proc. Int. Conf. on Computer Vision (2011)
7. Heng, C.K., Yokomitsu, S., Matsumoto, Y., Tamura, H.: Shrink boost for selecting multi-lbp histogram features in object detection. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3250–3257, June 2012
8. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-backward error: Automatic detection of tracking failures. In: 2010 20th International Conference on Pattern Recognition (ICPR), pp. 2756–2759, August 2010
9. Kolsch, M., Turk, M.: Fast 2d hand tracking with flocks of features and multi-cue integration. In: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2004) vol. 10, pp. 158. IEEE Computer Society, Washington, DC (2004). http://dl.acm.org/citation.cfm?id=1032641.1033046
10. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Cehovin, L., Nebehay, G., Fernandez, G., Vojir, T., Gatt, A., Khajenezhad, A., Salahledin, A., Soltani-Farani, A., Zarezade, A., Petrosino, A., Milton, A., Bozorgtabar, B., Li, B., Chan, C.S., Heng, C., Ward, D., Kearney, D., Monekosso, D., Karaimer, H., Rabiee, H., Zhu, J., Gao, J., Xiao, J., Zhang, J., Xing, J., Huang, K., Lebeda, K., Cao, L., Maresca, M., Lim, M.K., El Helw, M., Felsberg, M., Remagnino, P., Bowden, R., Goecke, R., Stolkin, R., Lim, S., Maher, S., Poullot, S., Wong, S., Satoh, S., Chen, W., Hu, W., Zhang, X., Li, Y., Niu, Z.: The visual object tracking vot2013 challenge results. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 98–111, December 2013
11. Lebeda, K., Hadfield, S., Matas, J., Bowden, R.: Long-term tracking through failure cases. In: Proceeedings of the IEEE workshop on Visual Object Tracking Challenge at ICCV 2013. IEEE, Sydney, December 2, 2013
12. Lim, M.K., Chan, C.S., Monekosso, D.N., Remagnino, P.: Swatrack: A swarm intelligence-based abrupt motion tracker. In: MVA. pp. 37–40 (2013)
13. Maresca, M.E., Petrosino, A.: MATRIOSKA: A Multi-level Approach to Fast Tracking by Learning. In: Petrosino, A. (ed.) ICIAP 2013, Part II. LNCS, vol. 8157, pp. 419–428. Springer, Heidelberg (2013)
14. Nebehay, G., Pflugfelder, R.: Consensus-based matching and tracking of keypoints for object tracking. In: Winter Conference on Applications of Computer Vision. IEEE, March 2014
15. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Comput. Vision 77(1–3), 125–141 (2008). http://dx.doi.org/10.1007/s11263-007-0075-7
16. Salaheldin, A., Maher, S., El Helw, M.: Robust real-time tracking with diverse ensembles and random projections. In: 2013 IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 112–120, December 2013
17. Sevilla-Lara, L.: Distribution fields for tracking. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012, pp. 1910–1917. IEEE Computer Society, Washington, DC (2012)
18. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: An experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 99(PrePrints), 1 (2013)
19. Čehovin, L., Kristan, M., Leonardis, A.: Robust visual tracking using an adaptive coupled-layer visual model. IEEE Computer Society, April 2013

20. Vojíř, T., Matas, J.: Robustifying the flock of trackers. In: Wendel, A., Sternig, S., Godec, M. (eds.) CVWW 2011: Proceedings of the 16th Computer Vision Winter Workshop, pp. 91–97. Graz University of Technology, Inffeldgasse 16/II, Graz (2011)
21. Vojíř, T., Matas, J.: The Enhanced Flock of Trackers. In: Cipolla, R., Battiato, S., Farinella, G.M. (eds.) Registration and Recognition in Images and Video. SCI, vol. 532, pp. 111–138. Springer, Heidelberg (2014)
22. Wu, C., Zhu, J., Zhang, J., Chen, C., Cai, D.: A Convolutional Treelets Binary Feature Approach to Fast Keypoint Recognition. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 368–382. Springer, Heidelberg (2012)
23. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2411–2418, June 2013
24. Wu, Y., Shen, B., Ling, H.: Online robust image alignment via iterative convex optimization. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1808–1814, June 2012
25. Xiao, J., Stolkin, R., Leonardis, A.: An enhanced adaptive coupled-layer lgtracker++. In: Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops, ICCVW 2013, pp. 137–144. IEEE Computer Society, Washington, DC (2013). http://dx.doi.org/10.1109/ICCVW.2013.24
26. Zhang, T., Ghanem, B., Liu, S., Ahuja, N.: Low-Rank Sparse Learning for Robust Visual Tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 470–484. Springer, Heidelberg (2012)

A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration

Yang Li and Jianke Zhu(✉)

College of Computer Science, Zhejiang University Zhejiang, Hangzhou, China
jkzhu@zju.edu.cn

Abstract. Although the correlation filter-based trackers achieve the competitive results both on accuracy and robustness, there is still a need to improve the overall tracking capability. In this paper, we presented a very appealing tracker based on the correlation filter framework. To tackle the problem of the fixed template size in kernel correlation filter tracker, we suggest an effective scale adaptive scheme. Moreover, the powerful features including HoG and color-naming are integrated together to further boost the overall tracking performance. The extensive empirical evaluations on the benchmark videos and VOT 2014 dataset demonstrate that the proposed tracker is very promising for the various challenging scenarios. Our method successfully tracked the targets in about 72% videos and outperformed the state-of-the-art trackers on the benchmark dataset with 51 sequences.

Keywords: Visual Tracking · Correlation Filter · Kernel Learning

1 Introduction

Visual tracking is one of the fundamental research problem in computer vision community for its various applications in video surveillance, robotics, human computer interaction and driverless vehicle. Although great progress has been made in the past decade, the model-free tracking is still a tough problem due to illumination changes, geometric deformations, partial occlusions, fast motions and background clutters.

Recently, correlation filter is introduced into visual community, which has already been applied in many applications [2] [10] [13] [27]. As described in Convolution Theorem, the correlation in time domain corresponds to an element-wise multiplication in Fourier domain. Thus, the intrinsic idea of correlation filter is that the correlation can be calculated in Fourier domain in order to avoid the time-consuming convolution operation. Meanwhile, the correlation filter is treated as similarity measure between the two signals in signal processing, which gives a reliable distance metric and explains the reason of the promising performance achieved by the previous approaches. Bolme et al. [7] and Henriques et al. [13] introduce the correlation filter into the tracking application. Although achieved the appealing results both in accuracy and robustness, these correlation

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 254–265, 2015.
DOI: 10.1007/978-3-319-16181-5_18

filter-based trackers employ the template with the fixed size, which is not able to handle the scale changes of a target.

In this paper, we propose a novel scale adaptive kernelized correlation filter tracker with multiple feature integration. The proposed approach overcomes the limitations of the conventional correlation filter trackers by a multiple scales searching strategy. To solve the scale change issue in object tracking, we sample the target with different scales, and resize the samples into a fixed size to compare with the leant model at each frame. Meanwhile, we adopt a multiple feature integration scheme, which employs the raw pixel, Histogram of Gradient [9] and color-naming [32] to further enhance the proposed tracker for dealing with the more challenge scenarios. Our experimental evaluation demonstrates that the proposed scale adaptive and multiple feature integration method achieves a significant performance gain (over 10%) comparing the state-of-the-art approach. Moreover, our method successfully tracks the targets in almost 72% sequences in the benchmark [33] with 51 videos in total.

The main contributions of this paper can be summarized as follows. Firstly, we extend the correlation filter-based tracker with the capability of handling scale changes, which obtains an impressive performance gain in accuracy. Secondly, we conduct the extensive experiments to compare the previous studies of the correlation filter-based trackers [14] [4] [12] with our proposed method that includes multiple features integration, scale adaptive scheme and a full system. These experiments reveals the underline clues on the importance of the different components for a modern tracking-by-detection tracker. Finally, the proposed tracker achieved a very appealing performance both in accuracy and robustness against the state-of-the-art trackers.

2 Related Works

Tracking-by-detection trackers [11] [1] [16] [34] are very popular due to its high performance and efficiency. As these methods usually employ the binary classifier to distinguish the tracked object from the background, which are usually denoted as the discriminative methods. Struck [11] is one the most representative discriminative trackers, which employs the structured Support Vector Machine(SVM) to directly link the target's location space with the training samples. It achieves the appealing result in the recent benchmark [33]. TLD [16] exploits a set of structural constraints with a sampling strategy using boosting classifier. The re-detection function makes the TLD method more robust in the challenge videos. Inspired by the compressive sensing techniques, Zhang et al. [34] train a Naive Bayes classifier with the compressive features projected from the original space. MIL [1] explores the idea of a bag of positive samples with a boosting variant algorithm to construct the tracker. Meanwhile, generative model-based trackers [22] [15] [21][29] [3] [30] [24] aim to build the metric model to search the most similar patches for the tracked object. SCM [36] combines the discriminative classifier and generative model to achieve the high accuracy and robustness. However, it involves with the heavy computational cost, which hinders its capability on real-time applications. Additionally, some trackers [5] [35] employ the

structure information in the scene to enhance the tracking performance while others [31] exploits the deep learning techniques in the object tracking task.

Our proposed approach is closely related to the correlation filter-based trackers [14] [4] [12] [7] [6], which adopt the correlation filter in traditional signal processing technique into the tracking applications. CSK [12] is proposed to explore the structure of the circulant patch to enhance the classifier by the augmentation of negative samples, which employs the kernel correlation filter to achieve the high efficiency. Based on CSK [12], KCF [14] adopts the HoG feature [9] instead of raw pixel to improve both the accuracy and robustness of the tracker. To further boost the performance of CSK tracker, Danelljan et al. [4] adopt the color-naming feature into the object tracking task, which is a powerful feature for the color objects [17] [19] [18]. Meanwhile, MOSSE [7] formulates the problem in the view of learning a filter .

3 The Tracker

In this section, we firstly review the kernelized correlation filter (KCF) tracker [14], and then introduce the powerful features used in our approach. Moreover, a scale adaptive scheme is presented to improve the correlation filter-based trackers.

3.1 The KCF Tracker

Our approach is built on KCF tracker [14], which achieves very impressive results on Visual Tracker Benchmark [33]. Although the idea of KCF tracker is very simple, it achieves the fastest and highest performance among the recent top-performing trackers. The key of KCF tracker is that the augmentation of negative samples are employed to enhance the discriminative ability of the track-by-detector scheme while exploring the structure of the circulant matrix for the high efficiency. In the following, we briefly review the main idea of KCF tracker [14].

In KCF [14], Henriques et al. assume that the cyclic shifts version of base sample is able to approximate the dense samples over the base sample. Suppose that we have a one-dimensional data $\mathbf{x} = [x_1, x_2, ..., x_n]$, a cyclic shift of $\mathbf{x}$ is $\mathbf{Px} = [x_n, x_1, x_2, ..., x_{n-1}]$. The experiments show that such assumption is held reasonably in most of cases. Therefore, all the cyclic shift visual samples, $\{\mathbf{P^u x} | u = 0...n-1\}$, are concatenated to form the data matrix $\mathbf{X} = C(\mathbf{x})$. As the data matrix is purely generated by the cyclic shifts of $\mathbf{x}$, it is called *circulant matrix*. It has an intriguing property [28] that all the circulant matrices can be expressed as below:

$$\mathbf{X} = \mathbf{F}^H \mathbf{diag}(\mathbf{Fx})\mathbf{F} \tag{1}$$

where $\mathbf{F}$ is known as the DFT matrix, which transforms the data into Fourier domain, and $\mathbf{F}^H$ is the Hermitian transpose of $\mathbf{F}$. The decomposition of circulant matrix can be employed to simplify the solution of linear regression. The

objective function of linear ridge regression can be formulated as follows:

$$\min_{\mathbf{w}} \sum_{i}^{n} (f(\mathbf{x}_i) - y_i)^2 + \lambda ||\mathbf{w}|| \tag{2}$$

where the function f can be written as the linear combination of basis samples: $f(\mathbf{x}) = \mathbf{w}^{\mathbf{T}}\mathbf{x}$. The ridge regression has the close-form solution, $\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. Substituted by Eqn.1, we have the solution $\hat{\mathbf{w}}^* = \frac{\hat{\mathbf{x}}^* \odot \hat{\mathbf{y}}}{\hat{\mathbf{x}}^* \odot \hat{\mathbf{x}} + \lambda}$, where $\hat{\mathbf{x}} = \mathbf{F}\mathbf{x}$ denotes the DFT of $\mathbf{x}$, and $\hat{\mathbf{x}}^*$ denotes the complex-conjugate of $\hat{\mathbf{x}}$. Compared with the prevalent method, this solution saves the computational cost of both extracting patches explicitly and solving a general regression problem [14]. In the case of no-linear regression, kernel trick, $f(\mathbf{z}) = \mathbf{w}^T\mathbf{z} = \sum_{i=1}^{n} \alpha_i \mathcal{K}(\mathbf{z}, \mathbf{x}_i)$, is applied to allow more powerful classifier. For the most commonly used kernel functions, the circulant matrix trick can also be used [14]. The dual space coefficients $\boldsymbol{\alpha}$ can be learnt as below

$$\hat{\boldsymbol{\alpha}}^* = \frac{\hat{\mathbf{y}}}{\hat{\mathbf{k}}^{\mathbf{xx}} + \lambda} \tag{3}$$

where $\mathbf{k}^{\mathbf{xx}}$ is defined as *kernel correlation* in [14]. Similar to the linear case, the dual coefficients are learnt in Fourier domain. The inference is valid for the case that kernel function treats each dimension of the data equally [14]. In this paper, we adopt the Gaussian kernel which can be applied the circulant matrix trick as below:

$$\mathbf{k}^{\mathbf{xx}'} = \exp(-\frac{1}{\sigma^2}(||\mathbf{x}||^2 + ||\mathbf{x}'||^2) - 2\mathbf{F}^{-1}(\hat{\mathbf{x}} \odot \hat{\mathbf{x}}'^*)) \tag{4}$$

As the algorithm only requires dot-product and DFT/IDFT, the computational cost is in $O(n \log n)$ time. The training label $\mathbf{y}$ is a Gaussian function, which decays smoothly from the value of one for the centered target to zero for other shifts. As zero means the negative sample, we need to enlarge the original target bounding box to enclose the negative samples. In this paper, we employ the window with the size of 2.5 times larger than its original target box for training. Although the cyclic shift lost lots of information on the original frame, the classifier obtains the dense samples to fit the model more precisely.

The circulant matrix trick can also be applied in detection to speed up the whole process. The patch $\mathbf{z}$ at the same location in the next frame is treated as the base sample to compute the response in Fourier domain,

$$\hat{\mathbf{f}}(\mathbf{z}) = (\hat{\mathbf{k}}^{\widetilde{\mathbf{x}}\mathbf{z}})^* \odot \hat{\boldsymbol{\alpha}} \tag{5}$$

where $\widetilde{\mathbf{x}}$ denotes the data to be learnt in the model. When we transform $\hat{\mathbf{f}}(\mathbf{z})$ back into the spatial domain, the translation with respect to the maximum response is considered as the movement of the tracked target. The motion model implied that the searching range is the window size for the base patch. Although the whole model follows the tracking-by-detection scheme, there are only two samples in the process, both at the same position sampled in the last frame and current frame. Intuitively, it is more like a similarity metric in Fourier domain. In addition, Bolme et al. [7] give another interpretation on the whole process. For the more detailed formulation, please refer to [14] [7].

3.2 Multiple Feature Integration

Since the kernel correlation function only needs to calculate the dot-product and vector norm, multiple channels can be applied for the image features. Suppose the multiple channels of the data representation are concatenated into a vector $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_C]$. Eqn. 4 can be rewritten as follows:

$$\mathbf{k}^{\mathbf{x}\mathbf{x}'} = \exp\left(-\frac{1}{\sigma^2}(||\mathbf{x}||^2 + ||\mathbf{x}'||^2) - 2\mathbf{F}^{-1}\left(\sum_{\mathbf{C}} \hat{\mathbf{x}}_{\mathbf{c}} \odot \hat{\mathbf{x}}_{\mathbf{c}}'^{*}\right)\right) \tag{6}$$

which allows us to use the more strong features rather than the raw greyscale pixels. Moreover, we can employ various powerful features to exploit the advantages of feature fusion. There are three types of features used in our proposed tracker. Besides the raw greyscale pixel of the original image, we adopt two commonly used features in visual tasks.

Histogram of Gradient (HoG) is one of the most popular visual features in vision community, since it is very effective in practical applications and can be computed very efficiently. The feature extracts the gradient information from a cell, which is a range of pixels. HoG counts the discrete orientation to form the histogram. As in [9], we employ the 31 gradient orientation bins variant in our method.

Color-naming or color attributes, is a perspective space, which is the linguistic color label assigned by human to describe the color. Being better than the RGB space, the distance in color label space is more similar to human sense. As achieved the promising results in other visual tasks such as object recognition, object detection and action recognition [17] [19] [18], we employ the mapping method described in [32] to transform the RGB space into the color names space, which is an 11 dimensional color representation. Color names provide the perception of object color, which usually contains the important information on the target.

The two features are complementary to each other. HoG puts emphasis on the image gradient while color naming focuses on the color information. In section 4.2, we will testify the efficacy of these features separately. Although the idea is quite straightforward, the performance gain is very promising. Note that the feature sizes do not consist with each other at first and alignment should be applied for the features data for the correlation filter.

3.3 Multiple Scale Kernelized Correlation Filter

As described in Section 3.1, the whole process is straightforward. Moreover, KCF is unable to deal with the scale changes in videos. To this end, we propose a scale adaptive method to enable the naive correlation filter tracker to deal with the scale variations.

In Section 3.1, the searching strategy is implied in the kernel correlation filter. We employ the bilinear interpolation to enlarge the image representation space from the countable integer space into the uncountable float space. We fix the template size as $\mathbf{s_T} = (s_x, s_y)$, and define a scaling pool $\mathbf{S} = \{t_1, t_2, ..., t_k\}$. Suppose that the target window size is $\mathbf{s_t}$ in the original image space. For the current frame, we sample k sizes in $\{t_i\mathbf{s_t}|t_i \in \mathbf{S}\}$ to find the proper target. Note that the dot-product in kernel correlation function needs the data with the fixed size. In this paper, we employ bilinear-interpolation to resize the samples into the fixed template size $\mathbf{s_T}$, and the final response is calculated by

$$\arg\max \mathbf{F}^{-\mathbf{1}}\hat{\mathbf{f}}(\mathbf{z}^{t_i}) \tag{7}$$

where $\mathbf{z}^{t_i}$ is the sample patch with the size of $t_i\mathbf{s_t}$, which is resized to $\mathbf{s_T}$. Since the response function obtains a vector, the max operation is employed to find its maximum scalar. As the target movement is implied in the response map, the final displacement needs to be tuned by t to get the real movement bias.

Note that all the templates are registered to the same size. Thus, the update procedure is straightforward. There are two sets of coefficients should be updated. One is the dual space coefficients $\boldsymbol{\alpha}$, and another is the base data template $\widetilde{\mathbf{x}}$. As in [14], we linearly combine the new filter with the old one as below:

$$\bar{\mathbf{T}} = \theta\mathbf{T}_{new} + (1-\theta)\bar{\mathbf{T}} \tag{8}$$

where $\mathbf{T} = [\boldsymbol{\alpha}^T, \widetilde{\mathbf{x}}^T]^T$ is the template to be updated. With the scale adaptive scheme, the proposed tracker is able to deal with the size changes. The overall algorithm is summarized into Algorithm 1.

Algorithm 1. Overall algorithm of SAMF

Require:
The template for the tracked target, $\widetilde{\mathbf{x}}$;
The dual space coefficient, $\boldsymbol{\alpha}$;
The newly arrived observation, $\mathbf{y}$;
The last frame position, $\mathbf{p}_{old}$;

Ensure:
The updated template for the tracked target, $\widetilde{\mathbf{x}}$;
The updated dual space coefficient, $\boldsymbol{\alpha}$;
The new position, $\mathbf{p}_{new}$;

1: **for** every t_i in **S do**
2: Sample the new patch $\mathbf{z}^{t_i}$ based on size $t_i\mathbf{s_t}$ and resize it to $\mathbf{s_T}$ with multiple features.
3: calculate the response $\hat{\mathbf{f}}(\mathbf{z}^{t_i})$ with Equation 5 and 6.
4: **end for**
5: Get final position $\mathbf{p}_{new}$ and size $t_i\mathbf{s_t}$ according to Equation 7
6: Get $\widetilde{\mathbf{x}}_{new}$ based on new position $\mathbf{p}_{new}$ and size $t_i\mathbf{s_t}$, and calculate $\boldsymbol{\alpha}_{new}$ with Equation 3.
7: Use Equation 8 to update $\widetilde{\mathbf{x}}$ and $\boldsymbol{\alpha}$ with $\widetilde{\mathbf{x}}_{new}$ and $\boldsymbol{\alpha}_{new}$.
8: **return** updated $\widetilde{\mathbf{x}}$ and $\boldsymbol{\alpha}$;

4 Experiments

We conduct three experiments to evaluate the efficacy of our proposed tracker. Firstly, we implemented three trackers with various settings, including Multiple Features tracker (MF), Scale Adaptive tracker (SA) and the proposed Scale Adaptive with Multiple Features tracker (SAMF). We compare them with other correlation filter-based trackers. Secondly, we evaluate our proposed tracker against the state-of-the-art trackers to show the effectiveness of our proposed SAMF tracker. Additionally, we report the detailed evaluation on VOT 2014 dataset.

4.1 Experimental Setup and Methodology

We implemented the proposed tracker by native Matlab without optimization. All the experiments are conducted on an Intel i5-760 CPU (2.80 GHz) PC with 16 GB memory. Our proposed SAMF tracker runs at about 7 fps. The σ used in Gaussian function is set to 0.5. The cell size of HoG is 4×4 and the orientation bin number of HoG is 9. The learning rate θ is set to 0.01. We use the scaling pool $\mathbf{S} = \{0.985, 0.99, 0.995, 1.0, 1.005, 1.01, 1.015\}$. All parameters are same for all following experiments.

In all the experiments, two evaluation criteria are used. The first one is mean center location error (CLE). CLE is the difference between the center of tracked results and the ground truth, where the smaller value means the more accurate result. The second criteria is the Pascal VOC overlap ratio (VOR) [8]. It is defined as $VOR = \frac{Area(B_T \cap B_G)}{Area(B_T \cup B_G)}$, where B_T is the tracking bounding box, and B_G is the ground truth bounding box. The larger value means the more accurate result.

To make comprehensive evaluation on the proposed approach, we employ the whole 51 video sequence in the benchmark [33] for the first two experiments. Moreover, we run the proposed tracker on VOT 2014 dataset containing 25 sequences. In VOT 2014 challenge, the accuracy is measured by the VOR score. The robustness indicates the failing time for a tracker on the sequence.

4.2 Experiment 1: Comparison between Correlation Filter-based Trackers

To evaluate the performance gain of our proposed scale adaptive scheme with multiple features, we run six variants of trackers on the benchmark [33], including SAMF, MF, SA, KCF, CN and CSK. All of these trackers takes advantage of the circulant matrix or kernel correlation filter. Table 1 summarizes the difference for these trackers. Figure 2 shows the CEL curves and VOR curves for those trackers. Although their ideas are very similar, the tracking performances are quite different. This indicates that the visual features and search strategy are essentially important to the visual tracking tasks. CSK only employs the raw pixel, whose rank is the lowest one among the compared trackers. CN adopts both

Table 1. The difference among six trackers

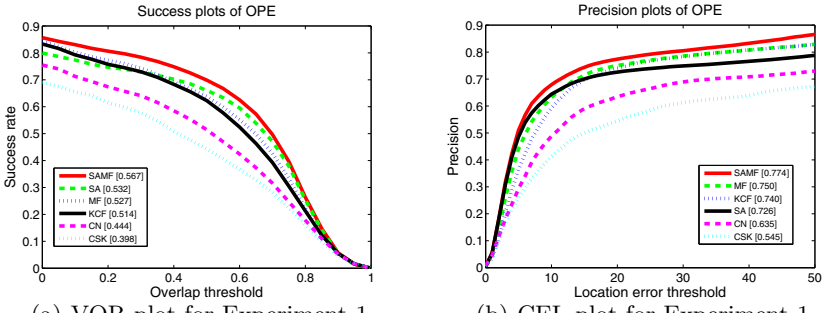
name	Features	Scale adaptive
SAMF	raw pixel, HoG, Color label	Yes
SA	HoG	Yes
MF	raw pixel, HoG, Color label	No
KCF [14]	HoG	No
CN [4]	raw pixel, Color label	No
CSK [12]	raw pixel	No

color names and raw pixel as features, and achieves a few improvement upon CSK. MF outperforms the KCF by augmenting the features space with color information and raw pixel. As shown in VOR curve, SA obtains a large improvement in accuracy shows. However, the robustness is decayed in the CEL curve. This demonstrates that expanding the search range will lead to the problem of local maximum. By taking advantage of the fusion features and the proposed scale adaptive scheme, SAMF tracker achieved the best performance in both VOR and CEL metrics.

The results from our experiment shows that our proposed tracker is very promising both in robustness and accuracy. The experiment also suggests that the feature and search strategy play very important role in visual tracking. Comparing to KCF, the VOR performance gains of SA and MF are 3.8% and 2.7% respectively while the SAMF gets a 10.6% improvement upon KCF. This indicates that the SAMF is not just the simple combination of the MF and SA, which can effectively capture the color information while accurately estimating the size of object.

4.3 Experiment 2: Comparison with the State-of-art Trackers

Table 2 illustrates the overall performance for the six trackers compared with the top two trackers reported in benchmark [33]. In the experiments, we observe



(a) VOR plot for Experiment 1 (b) CEL plot for Experiment 1

Fig. 1. The benchmark overall plot of the six kernel correlation filter based trackers

that the mean VOR score will be below 50% when the tracker loses the target in the sequence. Therefore, we define the successfully tracked sequence for a given tracker when the mean overlap of the whole sequence is above 0.5. The total number of the successfully tracked sequences can be viewed as a comprehensive metric of the tracker. The trackers with HoG feature achieved the very appealing performance compared against SCM and Struck in all the methods. SAMF achieves the best performance in terms of both mean CLE and mean VOR. Impressively, our approach achieves 57.4% in mean VOR overall, which is 10% improvement over the KCF tracker. In addition, the proposed tracker successfully tracked 37 of 51 sequences in the benchmark. This demonstrates that 72.5% of the sequences in the benchmark can be tracked, which is a big improvement for the visual object trackers.

Figure 2 shows the detailed report of SAMF compared with the top rank trackers, KCF [14], SCM [36], Struck [11], CN [4], TLD [16], ASLA [15], CXT [5], VTS [20], DFT [29], CPF [26], LSK [23], LOT [25] and VTD [21] in the benchmark. SAMF ranks the first with a large margin comparing to other trackers. Although the SAMF is not specially designed for occlusions, deformations and out-of-plane rotations, surprisingly, the proposed tracker obtains very appealing performances on these challenging video sequences. These promising results

Table 2. Overall comprehensive evaluation

	SAMF	SA	MF	KCF	CN	CSK	SCM	Struck
mean CEL	30.09	39.91	34.55	35.49	64.68	88.78	54.13	50.57
mean VOR	0.574	0.539	0.533	0.519	0.448	0.401	0.505	0.478
Passed Num.	37	32	32	31	23	18	28	28

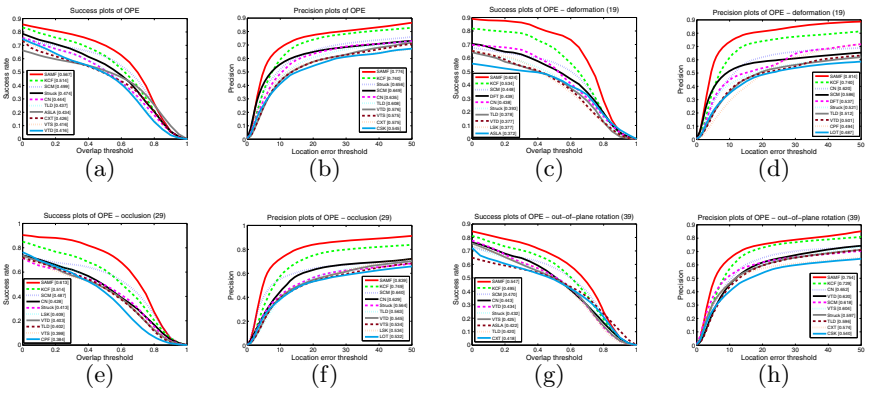


Fig. 2. The plot curves for the proposed tracker compared with 9 state-of-art trackers in the benchmark. (a)-(h) indicate the VOR and CEL of overall, deformation, occlusion and out-of-plane rotation, respectively.

Table 3. The results of VOT 2014

	accuracy						robustness					
	SAMF	KCF	NCC	SAMFn	KCFn	NCCn	SAMF	KCF	NCC	SAMFn	KCFn	NCCn
ball	0.772	0.702	0.740	0.738	0.640	0.633	1	1	30	0.47	1	26.1
basketball	0.748	0.574	0.573	0.640	0.562	0.577	0	2	30	0	2	30.7
bicycle	0.613	0.454	0.717	0.659	0.516	0.678	0	1	9	0.13	0.4	9.93
bolt	0.562	0.522	0.206	0.555	0.510	0.447	2	3	33	1.93	2.6	32.7
car	0.508	0.421	0.708	0.521	0.402	0.646	0	0	6	0.07	0	6.07
david	0.817	0.746	0.691	0.763	0.691	0.623	0	0	16	0	0	14.9
diving	0.245	0.233	0.269	0.209	0.226	0.233	4	5	8	4.4	4.8	6.87
drunk	0.568	0.434	0.364	0.542	0.481	0.423	0	0	4	0	0.53	4.07
fernando	0.394	0.402	0.575	0.393	0.393	0.331	1	1	15	1	1.13	13.3
fish1	0.495	0.438	0.564	0.472	0.445	0.541	3	3	16	2.73	3.27	16.5
fish2	0.296	0.299	0.265	0.294	0.257	0.189	5	4	14	4.80	5.47	12.4
gymnastics	0.536	0.528	0.663	0.467	0.489	0.402	2	3	8	2.47	2.2	7.4
hand1	0.544	0.389	0.515	0.417	0.408	0.378	3	6	13	5.33	4.8	14.8
hand2	0.462	0.438	0.275	0.400	0.443	0.230	5	8	15	7.07	7.87	16.8
jogging	0.819	0.760	0.795	0.674	0.655	0.696	1	1	3	0.93	1.07	3.33
motocross	0.400	0.372	0.326	0.351	0.349	0.208	4	5	9	3.4	4	9.07
polarbear	0.708	0.662	0.750	0.672	0.649	0.620	0	0	3	0	0	2.6
skating	0.452	0.488	0.675	0.526	0.530	0.563	0	0	26	0.07	0.4	26.7
sphere	0.879	0.713	0.643	0.796	0.664	0.674	0	0	1	0	0	2.27
sunshade	0.758	0.761	0.775	0.684	0.718	0.723	0	0	5	0	0	5.33
surfing	0.800	0.797	0.889	0.728	0.738	0.793	0	0	0	0	0	0
torus	0.840	0.757	0.507	0.752	0.687	0.376	0	0	17	0.07	0.27	15.9
trellis	0.814	0.546	0.600	0.732	0.506	0.525	0	0	29	0	0	27.5
tunnel	0.545	0.318	0.719	0.494	0.292	0.639	0	0	10	0	0	9.33
woman	0.758	0.755	0.745	0.734	0.687	0.611	1	2	23	1	2.07	21.5
Mean	0.613	0.540	0.582	0.569	0.518	0.510	1.28	1.80	13.72	1.43	1.75	13.44

suggest that the effective features and proper search strategy are more effective than the complicated models for deformations and occlusions.

4.4 Experiment 3: VOT 2014

Finally, we evaluate our proposed tracker on VOT 2014 dataset. The results are summarized into Table 3. Compared against KCF [14] and the baseline NCC tracker provided by the VOT organizer [1], SAMF achieves the higher performance both in accuracy and robustness. NCC performs quite well in accuracy but poor in the robustness. This is because NCC obtains more ground truth labels when it fails to track the target. Benefited from the correlation filter, KCF achieves an appealing score in robustness, however, it ranks at the last place in the accuracy due to the template with the fixed size. The proposed SAMF achieves the best results on both the accuracy and robustness. It can be seen that our proposed SAMF tracker performs especially well in case of robustness meanwhile it maintains the highest accuracy compared with other two trackers. This consists with the experimental results illustrated in Section 4.3.

[1] http://votchallenge.net/vot2014/index.html

5 Conclusions

This paper presented a very effective tracker based on the framework of correlation filter. We proposed the scale adaptive scheme to deal with the problem of the fixed template size in the conventional kernel correlation filter tracker. Moreover, the powerful features including HoG and color naming are fused together to further boost the overall performance for the proposed tracker. The extensive empirical evaluations on the benchmark videos and VOT 2014 dataset demonstrate that the proposed method is very promising for the various challenging scenarios. Our method successfully tracked the targets in about 72% videos and outperformed the state-of-the-art trackers on the benchmark dataset with 51 sequences.

Acknowledgments. The work was supported by National Natural Science Foundation of China under Grants (61103105 and 91120302).

References

1. Babenko, B., Yang, M.H., Belongie, S.: Robust object tracking with online multiple instance learning. TPAMI **33**(8), 1619–1632 (2011)
2. Boddeti, V.N., Kanade, T., Kumar, B.V.: Correlation filters for object alignment. In: CVPR (2013)
3. Chen, D., Yuan, Z., Wu, Y., Zhang, G., Zheng, N.: Constructing adaptive complex cells for robust object tracking. In: ICCV (2013)
4. Danelljan, M., Khan, F.S., Felsberg, M., van de Weijer., J.: Adaptive color attributes for real-time visual tracking. In: CVPR (2014)
5. Dinh, T.B., Vo, N., erard Medioni, G.: Context tracker: Exploring supporters and distracters in unconstrained environments. In: CVPR (2011)
6. D.S.Bolme, B.A.Draper, J.R.Beveridge: Average of synthetic exact filters. In: CVPR (2009)
7. D.S.Bolme, J.R.Beveridge, B.A.Draper, Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: CVPR (2010)
8. Everingham, M., Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classese(voc) challenge. IJCV **88**(2), 303–338 (2010)
9. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. TPAMI (2010)
10. Galoogahi, H.K., Sim, T., Lucey, S.: Multi-channel correlation filters. In: ICCV (2013)
11. Hare, S., Saffari, A., Torr, P.H.S.: Struck: Structured output tracking with kernels. In: ICCV (2011)
12. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: Exploiting the Circulant Structure of Tracking-by-Detection with Kernels. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 702–715. Springer, Heidelberg (2012)
13. Henriques, J.F., Carreira, J., Caseiro, R., Batista, J.: Beyond hard negative mining: Efficient detector learning via block-circulant decomposition. In: ICCV (2013)
14. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. TPAMI (2014)

15. Jia, X., Lu, H., Yang, M.H.: Visual tracking via adaptive structural local sparse appearance model. In: CVPR, pp. 1822–1829. Providence, June 2012
16. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-learning-detection. In: PAMI (2011)
17. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A., Lopez, A., Felsberg, M.: Coloring action recognition in still images. IJCV 105(3), 205–221 (2013)
18. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A., Vanrell, M., Lopez, A.: Color attributes for object detection. In: CVPR (2012)
19. Khan, F.S., van de Weijer, J., Vanrell, M.: Modulating shape features by color attention for object recognition. IJCV 98(1), 49–64 (2012)
20. Kwon, J., Lee, K.M.: Tracking by sampling trackers. In: ICCV (2011)
21. Kwon, J., Lee, K.M.: Visual tracking decomposition. In: CVPR (2010)
22. Lim, J., Ross, D., Lin, R.S., Yang, M.H.: Incremental learning for visual tracking. In: NIPS (2004)
23. Liu, B., Huang, J., Yang, L., Kulikowsk, C.: Robust tracking using local sparse appearance model and k-selection. In: CVPR (2011)
24. Mei, X., Ling, H.: Robust visual tracking using l1 minimization. In: ICCV (2009)
25. Oron, S., Bar-Hillel, A., Levi, D., Avidan, S.: Locally orderless tracking. In: CVPR (2012)
26. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-Based Probabilistic Tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
27. Revaud, J., Douze, M., Cordelia, S., Jgou, H.: Event retrieval in large video collections with circulant temporal encoding. In: CVPR (2013)
28. Gray, R.M.: Toeplitz and circulant matrices: A review. Now Publishers **77**(1–3), 125–141 (2006)
29. Sevilla-Lara, L., Learned-Miller, E.: Distribution fields for tracking. In: CVPR (2012)
30. Wang, D., Lu, H., Yang, M.H.: Least soft-thresold squares tracking. In: CVPR. Portland, June 2013
31. Wang, N., Yeung, D.Y.: Learning a deep compact image representation for visual tracking. In: NIPS (2013)
32. van de Weijer, J., Schmid, C., Verbeek, J.J., Larlus, D.: Learning color names for real-world applications. TIP 18(7), 1512–1524 (2009)
33. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR (2013)
34. Zhang, K., Zhang, L., Yang, M.-H.: Real-Time Compressive Tracking. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 864–877. Springer, Heidelberg (2012)
35. Zhang, L., van der Maaten, L.: Structure preserving object tracking. In: CVPR (2013)
36. Zhong, W., Lu, H., Yang, M.H.: Robust object tracking via sparsity-based collaborative model. In: CVPR, pp. 1838–1845, Providence, June 2012

W10 - Computer Vision + ONTology
Applied Cross-Disciplinary Technologies

Uncertainty Modeling Framework for Constraint-Based Elementary Scenario Detection in Vision Systems

Carlos Fernando Crispim-Junior(✉) and Francois Bremond

INRIA Sophia Antipolis, Valbonne, France
carlos-fernando.crispim_junior@inria.fr

Abstract. Event detection has advanced significantly in the past decades relying on pixel- and feature-level representations of video-clips. Although effective those representations have difficulty on incorporating scene semantics. Ontology and description-based approaches can explicitly embed scene semantics, but their deterministic nature is susceptible to noise from underlying components of vision systems. We propose a probabilistic framework to handle uncertainty on a constraint-based ontology framework for event detection. This work focuses on elementary event (scenario) uncertainty and proposes probabilistic constraints to quantify the spatial relationship between person and contextual objects. The uncertainty modeling framework is demonstrated on the detection of activities of daily living of participants of an Alzheimer's disease study, monitored by a vision system using a RGB-D sensor (Kinect®, Microsoft©) as input. Two evaluations were carried out: the first, a 3-fold cross-validation focusing on elementary scenario detection (n:10 participants); and the second devoted for complex scenario detection (semi-probabilistic approach, n:45). Results showed the uncertainty modeling improves the detection of elementary scenarios in recall (*e.g.,* In zone phone: 84 to 100 %) and precision indices (*e.g.,* In zone Reading: 54.5 to 85.7%), and the recall of Complex scenarios.

Keywords: Uncertainty Modeling · Ontology · Event Detection · Activities of Daily Living · Older People

1 Introduction

Event detection has been significantly advancing since the past decade within the field of Computer vision giving birth to applications on a variety of domains like safety and security (*e.g.,* crime monitoring [9]), medical diagnosis and health monitoring [23][5], and even as part of a new paradigm of human-machine interface in gaming and entertainment (Microsoft© Kinect®). Event detection methods in computer vision may be categorized in (adapted from Lavee *et al.* [11]): classification methods, probabilistic graphical models (PGM), and semantic models; which are themselves based on at least one of the following data

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 269–282, 2015.
DOI: 10.1007/978-3-319-16181-5_19

abstraction level: pixel-based, feature-based, or event-based. Artificial Neural Networks, Support-Vector Machines (SVM), and Independent Subspace Analysis (ISA) are examples of classification methods. For instance, Le *et al.*[12] have presented an extension of the ISA algorithm for event detection, where the algorithm learned invariant spatio-temporal features from unlabeled video data. Wang *et al.* [21] have introduced new descriptors for dense trajectory estimation as input for non-linear SVMs. Common examples of PGMs approaches are Bayesian Network (BN), Conditional Random Fields, and Hidden Markov Models (HMM). BNs have been evaluated at the detection of person interactions (e.g., shaking hands) [16], left luggage [13], and traffic monitoring [9]. Kitani *et al.* [8] has proposed a Hidden Variable Markov Model approach for event forecasting based on people trajectories and scene features. Despite the advances, PGMs have difficulty at modeling the temporal dynamics of an event. Izadinia and Shah [7] have proposed to detect complex events from by a graph representation of joint the relationship among elementary events and a discriminative model for complex event detection.

Even though the two previous classes of methods have considerably increased the performance of event detection in benchmark data sets, as they rely on pixel-based and feature-based abstractions they have limitations in incorporating the semantic and hierarchical nature of complex events. Semantic (or Description-based) approaches use descriptive language and logical operators to build event representations using domain expert knowledge. The hierarchical nature of these models allow the explicit incorporation of event and scene semantic with much less data than Classification and PGM methods.

Ceusters *et al.* [3] proposes the use of Ontological Realism to provide semantic knowledge to high-level events detected by a multi-layer hierarchical and dynamical graphical model in a semi-supervised fashion (human in the loop). Zaidenberg *et al.* [22] have evaluated a constraint-based ontology language for group behavior modeling and detection in airport, subways, and shopping center scenes. Cao *et al.* [2] have proposed an ontology for event context modeling associated to a rule-based engine for event detection in multimedia monitoring system. Similarly, Zouba *et al.* [23] have evaluated a video monitoring system at the identification of activities of daily living of older people using a hierarchical constraint-based approach. Oltramari and Lebiere [15] presents a semantic infra-structure for a cognitive system devoted for event detection in surveillance videos.

Although Semantic models advantage at incorporating domain expert knowledge, the deterministic nature of their constraints makes them susceptible to noise from underlying components - *e.g.,* people detection and tracking components in a pipeline of computer vision system - as they lack a convenient mechanism to handle uncertainty. Probabilistic reasoning has been proposed to overcome these limitations. Ryoo and Aggarwal [17] [18] have proposed hallucination concept to handle uncertainty from low-level components in a context-free grammar approach for complex event detection. Tran and Davis [19] have proposed Markov logic networks (MLNs) for event detection in parking lots. Kwak *et al.* [10] have proposed the detection of complex event by the combination

of primitive events using constraint flows. Brendel et al [1] propose probabilistic event logic to extend an interval-based framework for event detection; by adopting a learned weight to penalize the violation of logic formulas.

We present a uncertainty modeling framework to extend the generic constraint-based ontology language proposed by Vu *et al.* [20] by assessing the probability of constraint satisfaction given the available evidence. By combining both frameworks we allow domain expert to provide event models following a deterministic process, while probabilistic reasoning is performed in second plan to cope with the uncertainty in constraint satisfaction. In this paper we focus on handling uncertainty of elementary events.

2 Uncertainty Modeling Framework

Uncertainty may come from different levels of the event modeling task; from failures on the low-level components which provided input-data for the event detection task (*e.g.,* sudden change in person estimated dimension) to the model expressiveness at capturing the real-world event. For instance, constraint violation may be due to person-to-person differences in performing an event (event intra-class variation). In both cases it may be desirable that the event model be still detected even with a smaller probability.

We propose here a framework to handle uncertainty on elementary events. The framework may be decomposed on: event modeling, uncertainty modeling, and inference. In event modeling step domain experts use the constraint-based video event ontology proposed in [20] to devise event models based on attributes of tracked physical objects (*e.g.,* a person) and scene semantics (*contextual objects*). In uncertainty modeling step we learn the conditional probability distributions about the constraints using annotation on the events and the event models provided by domain experts. The inference step is performed by the temporal algorithm of Vu *et al.* [20] adapted to also compute event probability. The probability computation sub-step infers how likely a model is given the available evidence based on pre-learned conditional probabilities about the evaluated constraints.

2.1 Video Event Ontology

The constraint-based framework is composed of a temporal scenario (event) recognition algorithm and a video event ontology for event modeling. The video event ontology is based on natural terminology to allow end users (*e.g.,* medical experts) to easily add and change event models of a system. The models take into account *a priori* knowledge of the experimental scene, and attributes of objects (herein called Physical Objects, *e.g.,* a person, a car, etc.) detected and tracked by the vision components. *A priori* knowledge consists of the decomposition of a 3D projection of the scene floor plan into a set of spatial zones which carry semantic information about the monitored scene (*e.g.,* zones like “TV”, “armchair”, “desk”, “coffee machine”). The temporal algorithm is responsible for

the inference task, where it takes as input low-level data from underlying vision components, and evaluates whether these objects (or their properties) satisfy the constraints defined in the modeled events. An event model is composed of (up to) five parts [20]:

- **Physical Objects** refer to real-world objects involved in the detection of the modeled event. Examples of physical object types are: mobile objects (*e.g.,* person, or vehicle in another application), contextual objects (equipment) and contextual zones (chair zone).
- **Components** refer to sub-events of which the model is composed.
- **Constraints** are conditions that the physical objects and/or the components should hold. These constraints could be logical, spatial and temporal.
- **Alert** define the level of importance of the event model, and
- **Action** is an optional clause which works in association with the Alert type describes a specific course of action which should be performed in case the event model is detected,(*e.g.,* send a SMS to a caregiver responsible to check a patient over a possible falling down).

The physical object types depend on the domain of application. Two disjoint default types are presented, Mobile and Contextual Objects, with one extensions each, respectively, Person and Contextual Zone. Mobile is a generic class which defines the basic set of attributes for any moving object detected in the scene (*e.g.,* 3D position, width, height, depth). Person is an extension of Mobile class whose attributes are body posture and appearance signature(s). Contextual Object (CO) type refer to *a priori* knowledge of the scene. Contextual zone is an extension of CO commonly used to define a set of vertices in the ground plane which corresponds to a region with semantic information (*e.g.,* eating table, tv, desk) for an event model. Contextual objects may be defined at the deployment of the system by the domain experts or by launching an object detection algorithm for scene description at system installation, and specific times where object displacement is identified. Physical object types can be expanded accordingly to describe all types of objects in the scene.

Constraints define conditions that physical object properties and/or components must satisfy. They can be non-temporal, such as spatial (person->position *in* a contextual zone; or displacement(person1) >1 m) and appearance constraints (person1->AppearanceSignature = person2->ApperanceSignature); or temporal to capture specific duration patterns or time ordering between a model sub-events (components). Temporal relation are defined following Allen's interval algebra (*e.g., before, and, meet, overlaps*). Fig. 1 describes the model *Person changing from zone*1 *to zone* 2; which is defined in terms of a temporal relationship between two sub-events: *e.g., c*1, *Person in zone* 1 before *c*2, *Person in zone* 2.

The ontology hierarchically categorizes event models according to their complexity as (in ascending order):

- **Primitive State** models property(ies) and/or relationship among physical object(s) constant on a time interval (person posture, or person inside a contextual zone).

```
CompositeEvent(Person changing from zone1 to zone 2,
    PhysicalObjects( (per:Person), (z1: Zone), (z2: Zone) )
    Components (
    (c1: PrimitiveState Person_in_zone_1 (p1,z1)
    (c2: PrimitiveState Person_in_zone_2 (p1,z1)
    )
    Constraints( (c1 before c2) )
    Alert( NOTURGENT )
)
```

Fig. 1. Person changing from zone 1 to zone 2

- **Composite State** refers to a composition of two or more primitive states.
- **Primitive Event** models a change in a value of physical object property (*e.g.,* person changes from sitting to standing posture), and
- **Composite Event** refers to the composition of two previous event models which should hold a temporal relationship (person changes from sitting to standing posture before person in corridor zone).

2.2 Uncertainty Modeling for Elementary Scenarios

For uncertainty modeling purposes we divided the constraint-based ontology event models into two categories: elementary and composite scenarios. The term scenario is used to differentiate the modeling and inference tasks. Elementary Scenario have a direct correspondence to the primitive state type of the ontology, and the Composite Scenario represents all other ontology event types (Primitive Event, Composite States and Composite Events). This simplification is performed since these ontology event categories were devised to help domain experts at devising models in a modular fashion and then reduce model complexity and increase its re-usability. But, none difference exists for the inference algorithm while processing these event categories besides to the hierarchy depth of the sub-events they define a relationship for.

The uncertainty modeling framework is based on the following concepts:

- **Elementary Scenario**(ES) is composed of physical objects and constraints. This scenario constraints are only related to instantaneous values (*e.g.,* current frame) of physical object(s) attribute(s).
- **Composite Scenario**(CS) is composed of physical objects, sub-scenarios (components) and constraints; where the latter generally refer to composition and/or temporal relationships among model sub-scenarios.
- **Constraint** is a condition that physical object(s) or sub-scenarios must satisfy, and refer to the constraint types presented on the constraint-based ontology section.
- **Attributes** correspond to the properties (characteristics) of real world objects measured by the underlying components of the event detection task (*e.g., vision system*).

- **Observation** corresponds to the amount of evidence on a constraint or a scenario model.
- **Instance** refers to an individual detection of a given scenario.

Fig. 2 presents a description for the elementary scenario *Person in zone Tea*. This scenario is based on the physical objects *Person* and the semantic zone *zoneTea*. For instance, *zoneTea* would be polygon drawn on the floor - close or around the table where the kitchen tools to prepare tea are commonly placed - *a priori* defined by a domain expert during system installation or automatically detected by the system. The model has two constraints: the logic constraint that the target zone is *zoneTea*; and a spatial constraint called *In* which verifies whether the person position lies inside the given zone. Fig. 3 illustrates an example of a scene where semantic zones were manually drawn on the floor plane where contextual objects are located.

```
ElementaryScenario(Person_in_zone_Tea,
    PhysicalObjects( (per:Person), (zT: Zone) )
    Constraints(
       (per->Position In zT->Vertices)
       (zT->name = "zoneTea")
       (displacement(per->Position) < stopConstant)
    )
)
```

Fig. 2. Elementary Scenario Person in zone Tea

2.3 Computation of Elementary Scenario Uncertainty

The uncertainty of an Elementary Scenario is formalized as function of the framework confidence on the satisfaction of the Elementary Scenario constraints. Equation 1 presents an formalization of Elementary Scenario Uncertainty using Bayes Rule.

$$P(E_i|C_i) = \frac{P(C_i|E_i) * P(E_i)}{P(C_i)} \tag{1}$$

where,

- $P(E_i|C_i)$: Conditional Probability of Event E_i given its observed constraints C_i;
- $P(C_i|E_i)$: Probability of constraints which intervene on E_i at the current frame; and
- $P(E_i)$: Prior Probability of Event.

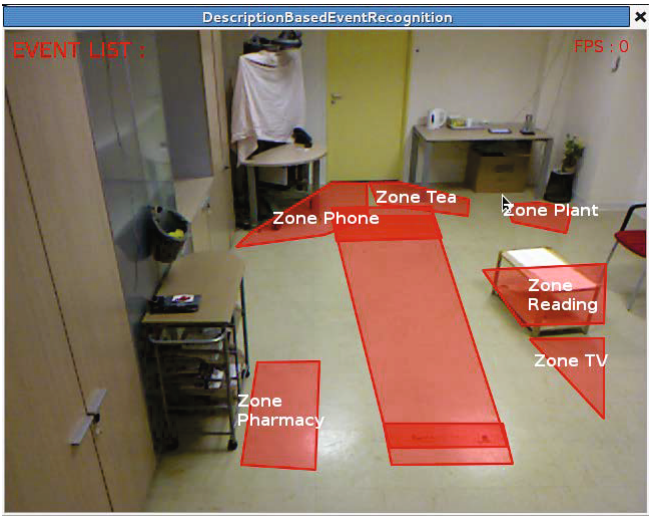


Fig. 3. Scene semantic zones

The conditional probability of event E_i given its set of observed constraints C_i is given by the multiplication of the individual conditional probabilities of its constraints. We assumed all constraints contribute equally to the event model detection and are conditionally independent (see Equation 2).

$$P(C_i|E_i) = \prod_{c_{i,j} \in C_i}^{N_j} P(c_{i,j}|E_i) \tag{2}$$

where $C_{i,j}$:

– Conditional probability of Constraint j of given event i.

To avoid computing $P(C_i)$ which can become costly as the number of constraints increase, we opted to use the non-normalized probability of $P(E_i|C_i)$ as described in Equation 3.

$$\tilde{P}(E_i|C_i) = P(E_i) \prod_{c_{i,j} \in C_i}^{N_j} P(c_{i,j}|E_i) \tag{3}$$

In its final form the proposed formula for elementary scenario uncertainty (Equation 3) addresses small violations of constraints from noise coming from underlying components and due to event intra-class variations.

2.4 Probabilistic Constraints

The uncertainty of a scenario model or its conditional probability given the evidence is addressed by associating each of its constraints to a Probability Density

Function (PDF) responsible for quantifying how likely the constraint would be satisfied given the available evidence. The use of PDFs provide a modular and flexible way to model and change the uncertainty process that governs the conditional probability distribution of a constraint given the available evidence - e.g., by modeling the variation of the low level data the constraint is conditioned on during the targeted event execution - and allowing us to avoid the fully specification of the set of assignments of a conditional probability table. Moreover, different constraints may use different PDFs according to the low-level data, and the PDF may be easily changed without any other changes to the event model.

Besides to selecting the fitting PDF to a given constraint it is also important to how we evaluate the constraint goal in a probabilistic fashion. In the case of the spatial operator In its deterministic version is susceptible to different sources of uncertainty: firstly, from the estimated position of the person which may be influenced by noise from low-level computer vision components; and secondly, from the semantic zone *zoneTea* - *a priori* defined by an expert - which may not accommodate the complete floor surface where people may stand to prepare tea. Its probabilistic counter-part should quantify how likely is the person position to be inside the zone of interest given these sources of noise. We here propose two probabilistic alternatives to the deterministic constraint In: the Center *In* and the Border *In*.

- The Center *In* is fully based on a PDF with respect to the relative distance between the centroid of the person - projected onto the floor - and the central position of the given semantic zone.
- The Border *In* is a hybrid implementation which provides maximum probability (100 %) when the person is anywhere inside the semantic zone, and a probability proportional to the distance of the person to the closest zone edge otherwise.

To model the conditional probability distribution of the distances between the person position and the semantic zone we have used Equation 4. Briefly, this equation converts the observed distance among objects into the corresponding value in an uniform Gaussian distribution using expected parameters pre-learned per semantic object. The corresponding value is then applied to an exponential function to obtain the probability of the constraint given the evidence, *e.g.*, a specific low-level data value for elementary scenario. The resulting PDF provides a probability curve with maximum value around the mean parameter and a monotonically decreasing behavior is observed as the observed value distances from the mean.

$$P(C_{i,j}) = \exp(\frac{1}{2} * (\frac{observed\_value - \bar{x}}{s})^2) \tag{4}$$

where, $\bar{x}$: learned mean of constraint value, and s: standard deviation of $\bar{x}$

2.5 Learning Constraint Conditional Probabilities

The conditional probability distribution of the elementary constraints were obtained by a learning step based on the event models provided by domain

experts - using the constraint-based ontology - and annotated RGB-D recordings of the targeted events. The learning step was performed as follows: firstly, an event detection process was performed using the deterministic event models. Each time the deterministic *In* was evaluated the relative distance used by the probabilistic counterparts was stored independent of whether the current constraint is satisfied. Secondly, using the event annotation we collect the distance values frequently assumed by the In variants when elementary scenario annotation is present for the given RGB-D recording. Thirdly and finally, we computed statistics about the the collected values of the attribute the constraint was conditioned on. By performing the learning step using event models combined with event annotation (both provided by domain experts) we aim at capturing the Conditional Probability Distribution (CPD) of the constraints according to the event model semantics and maybe reduce the semantic gap between the event model and the real-world event.

Elementary Scenarios are assumed to be equally probable as their evidence is mainly related to a single time unit (e.g., a frame). The Temporal aspect of scenario models such as instance filtering is currently performed by a threshold method which removes low-probability events. The influence of previous instances probabilities into the evaluated time unit will be evaluated in the future in conjunction with uncertainty modeling at Composite Scenario level (Composite Event).

3 Evaluation

The proposed framework has been evaluated at modeling the uncertainty of activities of daily living of participants of a clinical protocol for Alzheimer's disease study. Two evaluations were performed, firstly on the detection of elementary scenarios, and secondly on the detection of complex events by using uncertainty framework for elementary scenarios as basis for the deterministic complex event models. The latter evaluation intends to assess the improvement brought to the detection of high-level scenario by low-level uncertainty modeling. For both evaluations contextual objects were defined *a priori* by domain experts and mostly refer to static furniture in the scene.

Concerning the learning step necessary to obtain the parameters for the constraint conditional probabilities, in the first evaluation the parameters were computed following the rules of the 3-fold cross-validation procedure. For the second evaluation, the 10 videos involved in the 3-fold cross-validation procedure were used for the learning procedure, and the complex detection performance was evaluated on a set of recordings of 45 participants new to the system, which were only annotated in terms of Composite Events.

3.1 Data Set

Participants aged 65 years and over were recruited by the Memory Center of Nice Hospital. Inclusion criteria of the Alzheimer Disease (AD) group are: diagnosis

of AD according to NINCDS-ADRDA criteria and a Mini-Mental State Exam (MMSE) score above 15. AD participants who have significant motor disturbances (per the Unified Parkinson's Disease Rating Scale) are excluded. Control participants are healthy in the sense of behavioral and cognitive disturbances. Experimental recordings used a RGB-D camera (Kinect®, Microsoft©).

The clinical protocol is divided into three tasks: directed tasks, semi-directed tasks, and discussion with the clinician task. The directed tasks (10 minutes) are divided on two sub-tasks: physical directed- and vocal directed-tasks. In the semi-directed task (15 minutes) the participants are asked to undertake a set of Instrumental Activities of Daily Living in a Hospital observation room furnished with home appliances [6]. The participants enter the room alone with a list of activities to perform and are advised to leave the room only feeling all the required tasks are accomplished.

For this framework evaluation we have focused only on the semi-directed task. The list of semi-directed activities is composed as follows:

- Read 1 article and answer three questions,
- Turn on the TV,
- Establish the account balance,
- Pay the phone bill (check writing),
- Answer the phone,
- Call the psychologist to confirm the appointment afterwards,
- Find on a bus map the line that takes you to the train station,
- Prepare the drug box for tomorrow according to the prescription,
- Water the plant,
- Prepare a hot tea.

3.2 RGB-D Monitoring System

The framework for uncertainty modeling was evaluated using a RGB-D sensor-based monitoring system, built on the event detection framework proposed by Vu *et al.* [20], and later evaluated on the detection of daily living activities of older people by Crispim-Junior *et al.* [5] using a 2D-RGB camera as the input sensor.

The evaluation monitoring system can be composed into three main steps: people detection, people tracking, and event detection. People detection step is performed by a depth-based algorithm proposed in Nghiem *et al.* [14], since we have replaced the 2D-RGB camera by a RGB-D sensor. The depth-based algorithm performs as follows: first, background subtraction is employed on the depth image provided by the RGB-D camera to identify moving regions. Then, region pixels are clustered in objects based on their depth and neighborhood information. Finally, head and shoulder detectors are employed to detect people amongst other types of detected objects.

The set of people detected by the previous algorithm is then evaluated by a multi-feature tracking algorithm proposed in Chau *et al.* [4], which employs as

features the 2D size, the 3D displacement, the color histogram, and the dominant color to discriminate among tracked objects.

Event detection step has as input the set of tracked people generated in the previous step and *a priori* knowledge of the scene provided by a domain expert. This step was evaluated for two different components for comparison purposes: the proposed framework for uncertainty modeling, and the deterministic event modeling framework proposed by Vu *et al.* [20] and evaluated by Crispim-Junior *et al.* [5]. Both components frameworks used the same underlying components.

3.3 Performance Measurement

The framework performance on event detection is evaluated using the indices of Recall (Rec.) and Precision (Prec.) described in Equations 5 and 6, respectively in comparison to ground-truth events annotated by domain experts.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

where TP: True Positive rate, FP: False Positive rate and FN: False Negative rate.

4 Results and Discussion

Table 1 presents the performance of the uncertainty modeling framework on elementary scenario (primitive state) detection in a 3-fold cross-validation scheme. The cross-validation scheme used 10 RGB-D recordings of participants of the clinical protocol data set. "Deterministic" stands for the deterministic constraint-based approach. Results are reported as the average performance on the frameworks on the validation sets.

Table 1. Framework Performance on Elementary Scenario Detection on a 3-fold-cross-validation scheme

	Deterministic		Border In		Center In	
IADL	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
In zone Pharmacy	100.0	71.4	100.0	100.0	100	83.3
In zone Phone	84.0	95.45	92.0	92.0	100.0	100.0
In zone Plant	100.0	81.8	100.0	34.6	100.0	81.8
In zone Tea	93.3	77.7	100.0	36.6	93.3	73.7
In zone Read	75.0	54.5	100.0	38.1	75.0	85.7

$N:$ 10 *participants*; 15 *min. each*; *Total* : 150 *min.*

The proposed probabilistic constraints outperformed the deterministic approach on the recall index and on precision index in a few cases such as "In

zone reading" and "In zone Pharmacy" with *Center In* constraint. *Border In* constraint presented the highest recall, but the lowest average precision.

Table 2 presents the results of the framework on Composite Event Detection. Here an hybrid strategy is adopted where the uncertainty modeling is used on elementary scenarios and the deterministic constraint-based framework is used on composite event modeling.

Table 2. Framework Performance on Composite Event Detection Level

	Deterministic		Border In		Center In	
IADL	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.
Talk on Phone	88.76	89.77	89.88	70.79	88.76	85.86
Preparing Tea/Coffee	81.42	73.07	95.71	40.36	92.85	55.08
Using Pharmacy Basket	87.75	97.72	89.79	95.65	89.79	97.77
Watering plant	78.57	84.61	100.0	23.14	100.0	28.86

$N:$ 45 *participants*; 15 *min. each*; $Total:$ 675*min.*

The results on complex event detection showed *Center In* and *Border In* had similar performance on recall index outperforming the deterministic approach. *Center In* outperformed *Border In* in the precision index for this test but was still worse than the deterministic approach in most cases. The worse performance in precision index may be attributed to other model constraints which did not have their uncertainty addressed. Based on the results presented we select *Center In* constraint as the probabilistic alternative for the deterministic *In*.

5 Conclusions

We have presented a uncertainty modeling framework to handle uncertainty from low-level data in constraints of elementary scenarios (low-level events). The framework improves the detection performance of elementary scenarios in recall and precision and of composite scenarios in recall.

Further work will extend the framework to model composite scenarios and the uncertainty related to composite and temporal relations among its sub-components. Moreover, we will also investigate alternatives to allow small deviations from the scenario constraint without the need of performing a supervised learning step.

References

1. Brendel, W., Fern, A., Todorovic, S.: Probabilistic event logic for interval-based event recognition. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3329–3336, June 2011
2. Cao, Y., Tao, L., Xu, G.: An event-driven context model in elderly health monitoring. In: Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 120–124 (2009)

3. Ceusters, W., Corso, J.J., Fu, Y., Petropoulos, M., Krovi, V.: Introducing ontological realism for semi-supervised detection and annotation of operationally significant activity in surveillance videos. In: Proceedings of the 5th International Conference on Semantic Technologies for Intelligence, Defense and Security (STIDS) (2010). http://www.cse.buffalo.edu/jcorso/pubs/stids2010_istare_withresponses.pdf
4. Chau, D.P., Bremond, F., Thonnat, M.: A multi-feature tracking algorithm enabling adaptation to context (2011)
5. Crispim-Junior, C., Bathrinarayanan, V., Fosty, B., Konig, A., Romdhane, R., Thonnat, M., Bremond, F.: Evaluation of a monitoring system for event recognition of older people. In: 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 165–170, August 2013
6. Folstein, M.F., Robins, L.N., Helzer, J.E.: The mini-mental state examination. Archives of General Psychiatry 40(7), 812 (1983). http://dx.doi.org/10.1001/archpsyc.1983.01790060110016
7. Izadinia, Hamid, Shah, Mubarak: Recognizing Complex Events Using Large Margin Joint Low-Level Event Model. In: Fitzgibbon, Andrew, Lazebnik, Svetlana, Perona, Pietro, Sato, Yoichi, Schmid, Cordelia (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 430–444. Springer, Heidelberg (2012)
8. Kitani, Kris M., Ziebart, Brian D., Bagnell, James Andrew, Hebert, Martial: Activity Forecasting. In: Fitzgibbon, Andrew, Lazebnik, Svetlana, Perona, Pietro, Sato, Yoichi, Schmid, Cordelia (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 201–214. Springer, Heidelberg (2012)
9. Kumar, P., Ranganath, S., Weimin, H., Sengupta, K.: Framework for real-time behavior interpretation from traffic video. IEEE Transactions on Intelligent Transportation Systems **6**(1), 43–53 (2005)
10. Kwak, S., Han, B., Han, J.H.: Scenario-based video event recognition by constraint flow. In: CVPR, pp. 3345–3352. IEEE (2011)
11. Lavee, G., Rivlin, E., Rudzsky, M.: Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **39**(5), 489–504 (2009)
12. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, pp. 3361–3368. IEEE Computer Society, Washington, DC (2011). http://dx.doi.org/10.1109/CVPR.2011.5995496
13. Lv, F., Song, X., Wu, B., Kumar, V., Nevatia, S.R.: Left luggage detection using bayesian inference. In: PETS (2006)
14. Nghiem, A.T., Auvinet, E., Meunier, J.: Head detection using kinect camera and its application to fall detection. In: 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), pp. 164–169. July 2012
15. Oltramari, A., Lebiere, C.: Using ontologies in a cognitivegrounded system: Automatic action recognition in video surveillance. In: Proceedings of STIDS 2012 (7th International Conference on "Semantic Technology for Intelligence, Defense, and Security (2013)
16. Park, S., Aggarwal, J.K.: A hierarchical bayesian network for event recognition of human actions and interactions. Multimedia Syst. **10**(2), 164–179 (2004)
17. Ryoo, M.S., Aggarwal, J.K.: Recognition of composite human activities through context-free grammar based representation. In: CVPR (2), pp. 1709–1718. IEEE Computer Society (2006)

18. Ryoo, M.S., Aggarwal, J.K.: Semantic representation and recognition of continued and recursive human activities. International Journal of Computer Vision **82**(1), 1–24 (2009)
19. Tran, Son D., Davis, Larry S.: Event Modeling and Recognition Using Markov Logic Networks. In: Forsyth, David, Torr, Philip, Zisserman, Andrew (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 610–623. Springer, Heidelberg (2008)
20. Vu, V.T., Bremond, F., Thonnat, M.: Automatic video interpretation: A novel algorithm for temporal scenario recognition. In: Proc. 8th Int. Joint Conf. Artif. Intell., pp. 9–15 (2003)
21. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176, June 2011
22. Zaidenberg, S., Boulay, B., Brmond, F.: A generic framework for video understanding applied to group behavior recognition. CoRR abs/1206.5065 (2012)
23. Zouba, N., Bremond, F., Thonnat, M.: An activity monitoring system for real elderly at home: Validation study. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 278–285, August 2010

Mixing Low-Level and Semantic Features for Image Interpretation

A Framework and a Simple Case Study

Ivan Donadello[1,2(✉)] and Luciano Serafini[1]

[1] Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
[2] Department of Information and Communication Technology, University of Trento, Via Sommarive 14, 38123 Trento, Italy
{donadello,serafini}@fbk.eu

Abstract. Semantic Content-Based Image Retrieval (SCBIR) allows users to retrieve images via complex expressions of some ontological language describing a domain of interest. SCBIR adds some flexibility to the state-of-the-art methods for image retrieval, which support query either by keywords or by image examples. The price for this additional flexibility is the generation of a semantically rich description of the image content reflecting the ontology constraints. Generating these semantic interpretations is an open research problem. This paper contributes to this research line by proposing an approach for SCBIR based on the somehow natural idea that the interpretation of a picture is an (onto) logical model of an ontology that describes the domain of the picture. We implement this idea in an unsupervised method that jointly exploits the ontological constraints and the low-level features of the image. The preliminary evaluation, presented in the paper, shows promising results.

Keywords: Computer vision · Ontologies · Semantic image interpretation

1 Introduction

In recent years internet has seen a terrific increase of digital images. Thus the need of searching for images on the basis of human understandable descriptions, as in the case of textual documents, is emerging. For this reason, sites as YouTube, Facebook, Flickr, Grooveshark allow the tagging of the media and support searching by keywords and by examples. Tags associated to media constitute a simple human understandable representation of the media content. Tagging activity is very stressful and often is not well done by users. For this reason, methods for automatically generate a description of the image content, as in textual document understanding, become a real necessity. There are many approaches to image understanding which try to generate a high level description of an image by analysing low-level information (or features), such as colours, texture and contours, thus providing such a high level description in terms of

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 283–298, 2015.
DOI: 10.1007/978-3-319-16181-5_20

semantic concepts. This would allow a person to search, for instance, for an image containing "a man is riding an horse". The difficulty to find the correspondence between the low-level features and the human concepts is the main problem in content-based image retrieval. It is the so-called *semantic gap* [17]. It's widely recognised that, to understand the content of an image, contextual information (aka background knowledge) is necessary [21]. Background knowledge, relevant to the context of an image, can be expressed in terms of logical languages in an ontology [6]. Ontologies can play two main roles in image processing. First, they allow to express a set of constraints on the possible interpretations of an image and the satisfaction of such constraints can be checked via logical reasoning. Second, the terminology introduced by the ontology can be used as formal language to describe the image content. This will enable semantic image retrieval using queries expressed in the language introduced by the ontology. These two roles can be obtained by designing ontologies that formalize human understandable concepts (aka object types) and relations that can be found in the set of considered pictures (e.g., rides, part-of, nearby, is-talking-to, etc.). Furthermore, the background knowledge encoded in ontologies provides constraints on types of objects and relations, e.g. a vehicle has at least two wheels or horses can be ridden by men. The advantage of having the tags as concepts coming from a background knowledge allows to reason over the image. For example the tag "horse" enables to infer the presence of an animal.

In the present work we adopt the natural idea, envisaged in [19,23], that the interpretation of an image, in the context of an ontology, is a (partial) model of the ontology, which expresses the state of affairs of the world in the precise moment in which the picture has been taken. We propose to formalize the notion of image interpretation, w.r.t. an ontology, as *a segmented image, whose segments are associated with a set of objects of a partial model of the ontology.* To cope with the fact that a picture reports only partial information on the state of affairs we use the notion of partial model of a logical theory [30]; to cope with the possibility of having multiple alternative interpretations of a picture we introduce the notion of *most plausible partial model* of an image. The most plausible partial model for a picture is a partial model that maximizes a given scoring function, which depends from the low-level features of the image.

To have a preliminary evaluation of the above idea, we implemented this framework for a specific and limited case. We developed a fully unsupervised method to generate image interpretations able to infer the presence of complex objects from the parts present in the picture, thus inferring the relative "part-whole" structure. The method jointly exploits the constraints on the part-whole relation given by the ontology, and the low-level features of the objects available in the image. This work should be considered preliminary. Nevertheless, the evaluation shows promising results.

The paper is organized as follows. In Section 2, we present an overview on semantic image interpretation (SII). Section 3 describes our formal framework for SII. Section 4 shows how we adapt our general framework to the specific task of interpreting part-whole relation. Finally Section 5 describes the preliminary evaluation.

2 Related Work

The pure logical approach to image interpretation considers the information coming from a knowledge base for generating a semantic interpretation of an image. It is the most popular and satisfactory method. The first work that faced the problem in a logical approach is described in [23]. The authors propose a framework, based on first-order logic (FOL), for the depiction and interpretation of images. They address the image interpretation problem as finding the set of logical models of a knowledge base under the closed world assumption (CWA). The framework is presented with the example of interpreting hand drawn geographical maps, but it can be applied to other domains. The uncertainty is treated adding assertions on the specific case. A possible drawback is that an interpretation based on a total segmentation of the image using the CWA is unreasonable. This critique was described in [26] where the authors further explore the notion of logic-based approach to image interpretation. They introduce the notion of partial model for finding an image interpretation. Moreover, they propose a DL language with a calculus system for computing such a partial model. Uncertainty is not addressed. The growing interest in DL led to the first DL framework for computer vision [18]. In this work the authors investigate reasoning about spatial information in order to understand objects in a scene. The output are simple assertions on the objects and uncertainty is not handled. Following the DL-based approach, the authors of [20] explore a framework for the general high-level scene understanding task. The main interest of the work is in the conceptual structure for describing the basic components of a scene: the aggregates. An aggregate is a set of parts that compose a concept in a scene with some constraints. For example, an aggregate can be the concept of laying a table, its parts are physical objects as the table cover, actions as the transport of a dish and temporal constraints: the tablecloth has to be put before the dishes. Thus, the task of scene interpretation is the instantiation of aggregates driven by the evidence. The output of the framework is a partial model and uncertainty is not handled. This work has been extended in [19], where the authors propose a DL framework for knowledge-based high-level scene understanding. The framework remarks the necessity of a partial model and, finally, it introduces the notion of the most plausible partial model. Indeed, more interpretations can arise, so the construction of a partial model has to be guided for selecting the most probable one using a probabilistic approach. Uncertainty is not addressed. Another approach for selecting the most plausible partial model, or explanation, for a multimedia is given in [22]. Here the authors propose a DL framework for the multimedia interpretation based on abduction. The abductive reasoning [13] infers a possible explanation from a set of facts, or evidence. In this work, the evidence coming from the media analysis is the input for the abduction process that computes a plausible high-level interpretation (a partial model) of a knowledge base. The preferred explanation for the media is the one that contains more evidence and less hypotheses. This method requires a set of DL rules for defining what is abducible and uncertainty is not handled. A recent method for performing abduction, for scene understanding problem, is given by

the algebraic erosion over the concept lattice of a background knowledge [4]. A survey on logical approaches to multimedia interpretation can be found in [9].

The above-mentioned works assume that the information coming from the low-level image analysis is certain and without errors. But it is possible that this information, such as the labels or the spatial relations between regions, can be incomplete, vague and contradictory. We can have regions without labels, or more weighted labels or even contradictory labels. Fuzzy DL [31] is an appropriate formalism to reason in presence of imprecision. Fuzzy DL can reduces the semantic gap as in [14] where the authors propose a fuzzy DL ontology of spatial relations. The goal is to recognize objects exploiting the spatial information extracted from the image. A fuzzy DL framework for handling the vagueness and the inconsistency of the semantic features is proposed in [7]. The presented system enriches the image with new labels taken from an ontology.

Alternative approaches rely on Gestalt theory, attribute grammars and machine learning techniques. In [32] a generic framework for scene understanding that integrates domain knowledge with Gestalt theory [28] is proposed. The framework exploits the Gestalt laws of grouping such as similarity, closure and continuty with domain knowledge to perform the semantic segmentation of images. The work described in [12] uses an attribute graph grammar and a top-down/bottom-up inference algorithm for bulding the parse tree of manmade scenes such as buildings, hallways, kitchens ect. The algorithm maximizes a Bayesian posterior probability. In [29] the authors train a recursive neural network for parsing natural scene images. They recover the intrinsic structure of the natural scene by individualizing objects and capturing part-whole and proximity relations among them. The work in [3] detects structured objects, building façades, using a hierarchical approach based on layers. Every layer detects and classifies structures in the image for the next layer that computes higher level semantic structures. Every layer selects the best interpretation of the image using an ad hoc similarity distance between graphs. Uncertainty is addressed using this similarity distance. This method is generalized in [2] using a kernel function for the graph similarity. The above methods perform the parsing of the scene starting from low-level information of the image, but the structures they build lack of a formal semantics as the logic approaches provide.

Probabilistic approaches are alternatives to fuzzy DL for handling the vagueness but also for driving the construction of the most plausible model. A well-known formalism that combines FOL knowledge bases and probabilistic graphical models in a unique representation is given by Markov Logic Networks [24]. Another significant approach is given by combining FOL with kernel machines [8].

3 Problem Formulation

We start by introducing some assumptions and definitions which constitute the basic elements of the proposed framework.

Background knowledge. We suppose that background knowledge is contained in a knowledge base expressed in a logic of the family of Description Logics (DLs) [5]. In the following we briefly introduce DL formalism. Given three disjoint sets of symbols $\Sigma = \Sigma_C \uplus \Sigma_R \uplus \Sigma_I$, denoting concepts, relations (or roles) and individuals respectively, a $\mathcal{SHIQ}$ concept is defined by the following grammar:

$$C, D := A \mid \neg C \mid C \sqcap D \mid C \sqcup D \mid \exists R.C \mid \forall R.C \mid (\geq n)R.C \mid (\leq n)R.C$$

where $A \in \Sigma_C$, and $R \in \Sigma_R$. Furthermore, we suppose that Σ_R is closed under inverse role, i.e., if $R \in \Sigma_R$ then R^- (the inverse of R) is in Σ_R. Axioms are expressions of the following forms:

Axioms of the T-box	**Axioms of the A-box**
$C \sqsubseteq D$, concept inclusion axiom	$C(a)$, object class assertion
$R \sqsubseteq S$, role inclusion axiom	$R(a, b)$, role assertion

An interpretation $\mathcal{I}$ of the signature Σ is a pair $\langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$, where $\Delta^{\mathcal{I}}$ is a non empty set called the interpretation domain of $\mathcal{I}$. The symbol $\cdot^{\mathcal{I}}$ is a function from Σ to the subsets, the relations and the elements of $\Delta^{\mathcal{I}}$ satisfying the following constraints: $\cdot^{\mathcal{I}} : \Sigma_C \longrightarrow 2^{\Delta^{\mathcal{I}}}$, concept names are interpreted as subsets of the domain; $\cdot^{\mathcal{I}} : \Sigma_R \longrightarrow 2^{\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}}$, role names are interpreted as binary relations; and $\cdot^{\mathcal{I}} : \Sigma_I \longrightarrow \Delta^{\mathcal{I}}$, individual names are interpreted as elements of the domain. The function $\cdot^{\mathcal{I}}$ can be extended to all the concept expressions as follows:

$(\neg C)^{\mathcal{I}} = \Delta^{\mathcal{I}} \setminus C^{\mathcal{I}}$	$(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$	$(C \sqcup D)^{\mathcal{I}} = C^{\mathcal{I}} \cup D^{\mathcal{I}}$
$(\exists R.C)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \text{for some } (d, d') \in R^{\mathcal{I}},\ d' \in C^{\mathcal{I}}\}$		
$(\forall R.C)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \text{for all } (d, d') \in R^{\mathcal{I}},\ d' \in C^{\mathcal{I}}\}$		
$((\geq n)R.C)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \#(\{d' \in C^{\mathcal{I}} \mid (d, d') \in R^{\mathcal{I}}\}) \geq n\}$		
$((\leq n)R.C)^{\mathcal{I}} = \{d \in \Delta^{\mathcal{I}} \mid \#(\{d' \in C^{\mathcal{I}} \mid (d, d') \in R^{\mathcal{I}}\}) \leq n\}$		

where $\#(A)$ is the cardinality of the set A. A knowledge base $\mathcal{KB}$ is a set of axioms. $\mathcal{I}$ is a *model* of a knowledge base $\mathcal{KB}$ if it satisfies all the axioms in $\mathcal{KB}$, i.e. $\mathcal{I} \models \phi$ for all $\phi \in \mathcal{KB}$, where the satisfiability relation is defined as follows:

Axioms of the T-box	**Axioms of the A-box**
$\mathcal{I} \models C \sqsubseteq D$, iff $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$	$\mathcal{I} \models C(a)$, iff $a^{\mathcal{I}} \in C^{\mathcal{I}}$
$\mathcal{I} \models R \sqsubseteq S$, iff $R^{\mathcal{I}} \subseteq S^{\mathcal{I}}$	$\mathcal{I} \models R(a, b)$, iff $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in R^{\mathcal{I}}$

An interpretation that satisfies $\mathcal{KB}$, namely a model of $\mathcal{KB}$, is a *complete* representation (at a certain level of abstraction) of a possible state of affairs of the real world. The knowledge base, by means of its axioms, imposes constraints on possible states. The states of affairs corresponding to interpretations that do not satisfy $\mathcal{KB}$ are considered impossible. So, for instance, the axiom $\mathsf{House} \sqsubseteq \exists \mathsf{hasPart}.\mathsf{Door}$ imposes that the state of affairs where a house has no door will never be the case.

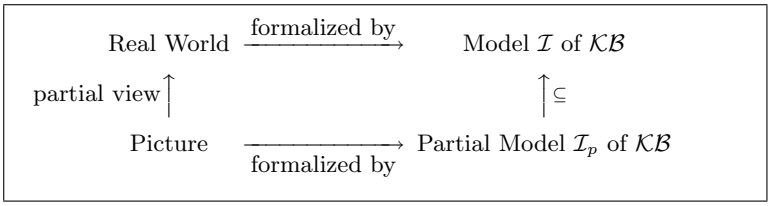


Fig. 1. The world is formalized by a model of $\mathcal{KB}$ and the partial view of the world contained in the picture is formalized by a partial model

Partial models. An image is a partial view of the world. Therefore, a formal representation of the content of an image should be a *partial view* of a model of $\mathcal{KB}$. This view can be considered as an interpretation of the language of $\mathcal{KB}$, but it does not necessarily satisfy all the axioms of $\mathcal{KB}$. The intuition is represented in Figure 1. For example, in a picture we can see a car with only two wheels, the others could be not visible due to the perspective of the view. The claim that a car has four wheels is not satisfied in the picture but it is satisfied in the real world supposing to be in a normal situation. Thus, if we formalize the world as a model of our knowledge base $\mathcal{KB}$ we formalize the picture with the notion of *partial model* $\mathcal{I}_p$. A *partial model* for a knowledge base $\mathcal{KB}$ is an interpretation $\mathcal{I}_p = \langle \Delta^{\mathcal{I}_p}, \cdot^{\mathcal{I}_p} \rangle$ of the knowledge base, such that there is a model $\mathcal{I} = \langle \Delta^{\mathcal{I}}, \cdot^{\mathcal{I}} \rangle$ of $\mathcal{KB}$, called *the completion of* $\mathcal{I}_p$ such that:

1. $\Delta^{\mathcal{I}_p} \subseteq \Delta^{\mathcal{I}}$
2. $a^{\mathcal{I}_p} = a^{\mathcal{I}}$ for all $a \in \Sigma_I$
3. $A^{\mathcal{I}_p} = A^{\mathcal{I}} \cap \Delta^{\mathcal{I}_p}$ for all $A \in \Sigma_C$
4. $R^{\mathcal{I}_p} = R^{\mathcal{I}} \cap \Delta^{\mathcal{I}_p} \times \Delta^{\mathcal{I}_p}$ for all $R \in \Sigma_R$.

Labelled picture. Our starting point is a segmented picture where every segment is associated with a set of labels paired with a confidence level. Labels are symbols taken from the alphabet of a knowledge base which is used to describe the real world from which the picture is taken. Given the current states of image processing software this seems a realistic assumption. We assume therefore that an image is divided into regions where every region has a set of weighted labels. Labels are taken from the signature Σ of the knowledge base. An example of labels and weights of a region is $\{(\mathsf{Duck}, 0.8), (\mathsf{DonaldDuck}, 0.7), (\mathsf{isArguingWith}, 0.4)\}$. We now provide a formal definition of labelled segment with the notion of *patch*.

A *labelled picture* $\mathcal{P}$ is a finite set of *labelled patches* $\mathcal{P} = \{p_1, \ldots, p_n\}$. A *labelled patch* p is a pair $p = \langle P, L \rangle$ where:

- P is a set of adjacent pixels $(i, j) \in \mathbb{N}^2$ of the labelled image $\mathcal{P}$. The pair (i, j) is the coordinates of the pixel in the image.
- L is a set of weighted labels of the patch and it is defined as $L \subseteq \Sigma \times \mathbb{R}$.

The function $Labels : \mathcal{P} \to \Sigma$ returns the set of labels (without weights). Namely for every $p = \langle P, \{\langle l_1, w_1 \rangle, \ldots, \langle l_n, w_n \rangle\} \rangle$, $Labels(p) = \{l_1, l_2, ...l_n\}$.

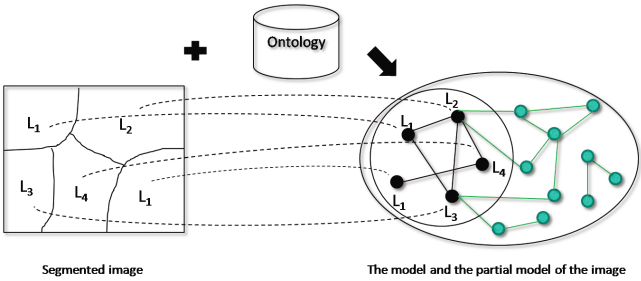


Fig. 2. Alignment between a labelled picture and its semantic interpretation

Problem definition. Following the intuition about partial models we define the semantic image interpretation as computing a partial model $\mathcal{I}_p = \langle \Delta^{\mathcal{I}_p}, \cdot^{\mathcal{I}_p} \rangle$ of the knowledge base. Thus, the solution is to find a method for creating the individuals (the nodes) of $\Delta^{\mathcal{I}_p}$, typing them and linking together (the arcs) according to $\cdot^{\mathcal{I}_p}$, in order to create the structured information representing the semantic content of the image. Having this graph describing the image content is not enough. We need also the information about the segmentation, e.g. in an information retrieval system it could be also necessary returning the single patches. So, we need a link between the individuals of our partial model and their corresponding segments, see Fig. 2. This consideration leads to the following formal definition of the semantic interpretation task.

Definition 1 (Semantic interpretation of a labelled image). *Given a knowledge base $\mathcal{KB}$ with signature Σ and a labelled picture $\mathcal{P}$, a* semantic interpretation of a labelled image *is a couple $(\mathcal{I}_p, cf)$ where:*

- $\mathcal{I}_p = \langle \Delta^{\mathcal{I}_p}, \cdot^{\mathcal{I}_p} \rangle \subseteq \mathcal{I}$ *is a partial model for $\mathcal{KB}$;*
- $cf : \mathcal{P} \rightarrow \Delta^{\mathcal{I}_p}$ *is called* conceptualization function *from the set of patches $\mathcal{P}$ to individuals, that is:*

$$\begin{aligned} cf(p) = i \in \Delta^{\mathcal{I}_p} : \exists l \in Labels(p) : \\ i = l^{\mathcal{I}_p}, \ \textit{with } l \in \Sigma_I, \\ i \in l^{\mathcal{I}_p}, \ \textit{with } l \in \Sigma_C, \\ \exists j \in \Delta^{\mathcal{I}} : (i,j) \in l^{\mathcal{I}}, \ \textit{with } l \in \Sigma_R \ . \end{aligned} \tag{1}$$

Preference relation between (partial) models. In general there are many possible explanations of the content of a picture. Formally this means that there are many partial models. On the other hand the interpretation of a picture should be unique, we have therefore to select one among a set of possible partial models. To face this problem, we introduce a scoring function $\mathcal{S}$ that assigns a score to a partial model based on its adherence to the image content, the highest the adherence the highest the score. Our problem turns to construct a partial model $\mathcal{I}_p^*$ that maximizes $\mathcal{S}$. In symbols:

$$\mathcal{I}_p^* = \underset{\mathcal{I}_p \in \mathbb{M}_p}{\operatorname{argmax}} \mathcal{S}(\mathcal{I}_p) \tag{2}$$

where $\mathbb{M}_p$ is the set of all possible partial models. This function can not be addressed in a purely logical manner but in a statistical framework that mixes low-level features with the logical constraints between concepts (the axioms of the knowledge base). There will be the necessity of a dataset for learning the correlation between objects and relations.

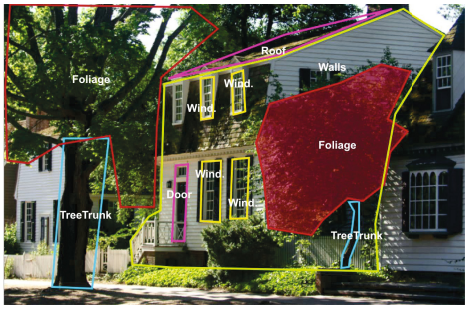
Issues in constructing an image interpretation. To construct the partial model $\mathcal{I}_p$ we have to determine its elements $\Delta^{\mathcal{I}_p}$, their types, their relations, and to search for a completion $\mathcal{I} \supseteq \mathcal{I}_p$ which satisfies all the axioms of $\mathcal{KB}$. There are several problems to face. Decide which are the elements of $\Delta^{\mathcal{I}}$ and $\Delta^{\mathcal{I}_p}$ that correspond to the picture patches, for example two regions labelled with car can be assigned to the same individual due to occlusions in the image. There can be also elements in $\Delta^{\mathcal{I}_p}$ which correspond to the composition of a set of patches. For instance, an individual of type House corresponds to the region obtained by joining the regions labelled with Window, Door, Roof, and Wall.

We also have to decide which are the types of the elements of $\Delta^{\mathcal{I}_p}$, this can be done using the labels contained in the corresponding patch as well as the axioms in the ontology. In general labels are not unique and weights need to be taken into consideration.

Another problem is to decide which are the relations between the elements of $\Delta^{\mathcal{I}_p}$. This can be achieved mixing visual and semantic features. For instance, by clustering with respect to the position of the patches, we can instantiate new individuals and linking them according to the part-whole relation. These inferences strongly depends on the type of relation we are considering.

4 Recognizing Complex Objects from their Parts

In this section we apply our framework to a specific subtask of semantic image interpretation: inferring the presence of complex objects from the presence of their parts. We consider the simplified scenario of a segmented image where patches can be labelled with at most one (non weighted) label. The background knowledge (and constraints) about part-whole relation is described by a simple ontology. Preference relation between partial models is inspired by a general principle of the mereology: the parts of the same object are topologically close in the space. Thus, we will prefer models where close parts in the image are considered parts of the same complex object. But we have to consider that sometimes close parts are not always parts of the same complex object. Therefore, to compute this preference, we need to take into account low-level features, such as the topological distance between patches, as well as semantic features, in order to prefer models that group together parts close in the space belonging to the same object. To compute the best partial model (i.e., the best grouping of parts in wholes) we use clustering techniques.



$$
\begin{aligned}
\text{House} &\sqsubseteq \exists \text{hasParts.Door} \\
&\sqcap \exists \text{hasParts.Window} \\
&\sqcap \exists \text{hasParts.Roof} \\
&\sqcap \exists \text{hasParts.Chimney} \\
&\sqcap \exists \text{hasParts.Walls} \\
\text{Tree} &\sqsubseteq \exists \text{hasParts.Foliage} \\
&\sqcap \exists \text{hasParts.TreeTrunk} \\
\text{Car} &\sqsubseteq \text{StreetVehicle} \\
\text{Motorbike} &\sqsubseteq \text{StreetVehicle} \\
\text{isPartOf} &\equiv \text{hasParts}^{-}
\end{aligned}
$$

Fig. 3. The image of our running example. Every segment has one label among `Foliage`, `TreeTrunk`, `Window`, `Walls`, `Door`, and `Roof`. The labels are taken from a simple ontology $\mathcal{O}$. The right part shows an excerpt of it.

We explain our method via a running example. Suppose that we start from the labelled image $\mathcal{P}$ of Figure 3. The set of patches of $\mathcal{P}$ and their labels are highlighted by the segments in the figure, e.g. a patch of the image is $p = (P, (\texttt{window}, 1))$. We have manually built a simple ontology $\mathcal{O}$ containing part-whole axioms about houses and vehicles, as well as some concept inclusion axioms. An excerpt of $\mathcal{O}$ is shown on the right side of Figure 3. Despite the simplicity of this example, and the manual construction of $\mathcal{O}$, we believe that this can be highly automatized and scaled to a larger domain since there are several knowledge bases describing objects from a mereological and taxonomical point of view, e.g. Wordnet [10].

Partial Model Initialization. According to the approach described in Section 3, building a semantic image interpretation means to construct a partial model $\mathcal{I}_p$ and the conceptualization function cf. To construct $\mathcal{I}_p$ we have to create the set of individuals $\Delta^{\mathcal{I}_p}$ corresponding to the patches of the picture, assign them the correct concepts, and find relations between them. Finally, we have to check if $\mathcal{I}_p$ is a partial model for $\mathcal{O}$, i.e., if there is a completion of $\mathcal{I}_p$ that is a model for $\mathcal{O}$. This last task can be easily solved by the inference services provided by DL reasoners, such as Racer [11] or Pellet [27]. Reasoners perform the completion of an ABox: they search for a model satisfying the ontology and the statements in the ABox. Moreover, they are able to infer new knowledge from the ABox exploiting the axioms in the ontology. From this consideration it follows that the main steps for the semantic interpretation of $\mathcal{P}$ are:

- for every patch $p \in \mathcal{P}$ create a new individual $\mathtt{i}_p$ in the ABox of $\mathcal{O}$;
- typing $\mathtt{i}_p$ according to $Labels(p)$;
- starting the reasoner for a possible completion of the ABox.

In the specific, given a patch p we instantiate a statement as $\texttt{Concept}(\mathtt{i_p})$ in the ABox of $\mathcal{O}$, where $\mathtt{i}_p$ is a new individual and $\texttt{Concept} \in Labels(p)$.

This procedure links together two levels: the concrete level, i.e. the labelled image showing a part of the reality, and an abstract level, i.e. the mathematical entity called partial model. The procedure not only creates the partial domain $\Delta^{\mathcal{I}_p}$ but also the conceptualization function cf. In the running example the partial domain $\Delta^{\mathcal{I}_p}$ is composed by the individuals `foliage1`, `foliage2`, `treeTrunk3`, `treeTrunk4`, `window5`, `window6`, `window7`, `window8`, `walls9`, `door10`, `roof11`. Furthermore, the typing of these individuals brings to the following ABox assertions: `Foliage(foliage1)`, `Foliage(foliage2)`, `TreeTrunk(treeTrunk3)`, `TreeTrunk(treeTrunk4)`, `Window(window5)`, `Window(window6)`, `Window(window7)`, `Window(window8)`, `Walls(walls9)`, `Door(door10)`, `Roof(roof11)`. Now, if we run a reasoner on $\mathcal{O}$ with the ABox it does not raise any inconsistency, this means that there exists a model extending the ABox, thus the latter is a partial logical model of $\mathcal{O}$.

Clustering Parts for Discovering New Complex Objects. The obtained partial model is not so informative, it is necessary to fill it with part-whole relations between individuals. This means to guide the construction of a semantic interpretation of $\mathcal{P}$ towards the most plausible partial model. Such a partial model is obtained according to a general principle, the most plausible model is the one relating together parts of the same object. The idea is to group together the several parts of an object and then inferring a new individual corresponding to that object. We clustered together the several parts of the same object, so different clusters mean different objects. Then, with abductive reasoning, we provide the best explanation for every cluster, that is, the whole object underlying the presence of some parts in the cluster. This approach takes into account geometrical features of the patches and semantic features in a clustering algorithm. Indeed, we need both kind of features because some objects can be close in the Euclidean space but far from a semantic point of view and we do not want to group them together. For example, an house and a tree could be close in the picture, but they are distant in the semantics so they cannot belong to the same cluster. Moreover, two objects can have the same parts but they do not share them. For example, two different houses have as parts some windows, but they do not share them. This is the case where objects can be near in the semantics but distant in the space.

The idea is to define a joint input space for a clustering algorithm. Such a space has to embed low-level with semantic features and its elements are associated to every patch. These elements are vectors representing the joint features of the patch, specifically:

- the (x, y) coordinates of the centroids;
- the semantic distance between the concept expressed by the patch respect to the concepts expressed by other patches.

There are many methods for calculating the semantic distance between concepts, our method is based on the part-whole relations between concepts [16]. Given a patch $p \in \mathcal{P}$ let `L` its label (the concept it expresses), (x_p, y_p) the coordinates of

its centroid, $\{\mathtt{L}_i\}_{i=1}^{n} \subseteq \Sigma_C$ the set of concepts expressed by the other patches, $d_{PW}(\mathtt{L}_j, \mathtt{L}_k)$ the semantic distance according part-whole relation between concepts $\mathtt{L}_j$, $\mathtt{L}_k$, the input space function $\mathcal{IS}_{PW}$ associating patches to their features according to part-whole relation is:

$$\begin{aligned} \mathcal{IS}_{PW} &: \mathcal{P} \rightarrow \mathbb{R}^{n+2} \\ &: p \mapsto \langle x_p, y_p, d_{PW}(\mathtt{L}, \mathtt{L}_1), ..., d_{PW}(\mathtt{L}, \mathtt{L}_n) \rangle \end{aligned} \tag{3}$$

Thus, our input space is the image of $\mathcal{IS}_{PW}$ over $\mathcal{P}$. In our example, an element of the input space associated to a patch p labelled with `door` has the form:

$$p \mapsto \langle x_p, y_p, d_{PW}(\mathtt{Door}, \mathtt{Walls}), d_{PW}(\mathtt{Door}, \mathtt{Foliage}), d_{PW}(\mathtt{Door}, \mathtt{Roof}), \dots \rangle$$

With such an input space we aim at clustering together patches both close in the Euclidean space and in the semantics. In this manner we guide the construction of the partial model towards the most plausible one, i.e. the one that groups parts belonging to the same object in the image. After the clustering we have a set of clusters $\mathcal{CL} = \{cl_1, ..., cl_m\}$. In our running example the clustering algorithm (see Section 5 for details) individualized 2 clusters:

$$\begin{aligned} cl_1 &= \{\mathtt{foliage1}, \mathtt{foliage2}, \mathtt{treeTrunk3}, \mathtt{treeTrunk4}\} \\ cl_2 &= \{\mathtt{window5}, \mathtt{window6}, \mathtt{window7}, \mathtt{window8}, \mathtt{walls9}, \mathtt{door10}, \mathtt{roof11}\}. \end{aligned}$$

For the sake of presentation clarity the clusters contain the individuals corresponding to the patches and not the elements of the input space. The first cluster should group only one foliage and a trunk, the reason is these parts are too close in the Euclidean space and the unsupervised learning (as clustering) is not able to distinguish between them, see Section 5 for details.

Inferring New Individuals from Clusters. The construction of the partial model follows from the set of clusters containing parts belonging to the same object. Indeed, we need to create a new individual in the ABox corresponding to this object and typing it. Technically, we have to compute the least common concept containing the types in the cluster. More generally, we have to find the best explanation underlying a certain cluster. The reasoning that gives an explanation to some evidence is called abductive reasoning. We present a method for typing the most likely object given a cluster of its parts and an ontology. The idea is to find, for every cluster, the ontology concept whose existential concept restrictions maximize the concepts expressed by the cluster elements. This procedure is a further step towards the construction of the partial model that mostly adheres to the image.

This idea needs the following formalism to be expressed. Let us consider the axioms of $\mathcal{O}$ with the form $\mathtt{A} \sqsubseteq \prod_i \exists \mathtt{R}.\mathtt{B}_i$, where $\mathtt{B}_i \subseteq \Sigma_C$ and $\mathtt{R} \in \Sigma_R$. We call $\mathtt{B}_i$ the set of types of the existential restrictions through $\mathtt{R}$. Consequently, let $CF_R : \Sigma_C \rightarrow 2^{\Sigma_C}$, where $\mathtt{R} \in \Sigma_R$, the function that assigns to every concept $\mathtt{A} \in \Sigma_C$ the set of types of its existential restriction through $\mathtt{R}$. For example,

in our ontology $CF_{\texttt{hasParts}}(\texttt{House}) = \{\texttt{Door}, \texttt{Window}, \texttt{Roof}, \texttt{Chimney}, \texttt{Walls}\}$ and $CF_{\texttt{hasParts}}(\texttt{Tree}) = \{\texttt{Foliage}, \texttt{TreeTrunk}\}$. Our approach is to compare the clusters with our ontology, thus we need to extract the concepts expressed by the parts in the clusters and a similarity measure between set of concepts. Given a cluster cl, the function CE extracts the concepts it expresses: $CE : \mathcal{CL} \rightarrow \Sigma_C$. In our running example, $CE(cl_1) = \{\texttt{Foliage}, \texttt{TreeTrunk}\}$. With this formalization it is simple to compare a cluster cl with each concept $\texttt{A}$ by defining a simple kernel set K based on the intersection between sets:

$$K(CE(cl), \texttt{A}) = \frac{\mid CE(cl) \cap CF_{\texttt{hasParts}}(\texttt{A}) \mid}{|CF_{\texttt{hasParts}}(\texttt{A})|}. \tag{4}$$

The abduction step now reduces to:

- perform the kernel set similarity between a given cluster and all the concepts $\texttt{A} \in \Sigma_C$, with $CF_{\texttt{hasParts}}(\texttt{A}) \neq \emptyset$;
- choose the concept that scores best;
- instantiate a new individual, in the ABox of $\mathcal{O}$, with that concept as type.

Thus, given cluster cl, $\texttt{A} \in \Sigma_C$ such that $CF_{\texttt{hasParts}}(\texttt{A}) \neq \emptyset$, we formalize the abductive step as instantiating a new individual $\texttt{newInd} \in \texttt{M}^{\mathcal{I}_p}$ in $\Delta^{\mathcal{I}_p}$, such that:

$$\texttt{M} = \underset{\texttt{A} \in \Sigma_C}{\operatorname{argmax}} K(ce(cl), \texttt{A}). \tag{5}$$

This new individual represents the whole object that best explains the several parts/patches in the cluster. Moreover, the presence of this individual in $\Delta^{\mathcal{I}_p}$ improves the plausibility of the partial model. After its creation we instantiate the `hasParts` relations with the individuals corresponding to its parts. In our running example, the two new individuals after the abductive step are of type `Tree` and `House` for cl_1 and cl_2 respectively.

Remarks. Some considerations are needed. Sometimes, there is not enough semantic information (labels) to discriminate two objects, e.g. can we distinguish a car from a motorbike knowing only the concepts of `Bodywork` and `Wheel`? In this case the kernel could be the same. Objects in the real world are categorized according to a taxonomy (`isA` relation) and a general principle exists: the more general a concept is the less attributes it has. That is, more general concepts have less types of existential restrictions and thus they have a bigger kernel. For example, given the concepts of `Bodywork` and `Wheel`, the kernel with best score will not return the concepts of `Car` or `Motorbike`, but the more general one of `StreetVehicle`.

We have seen that clustering together semantic and low-level features allows to discover objects far in space and semantics, close in space but far in the semantics and vice-versa. But what about objects close in the space and in the semantics? For example, a wheel of a car could be close to the bodywork of a motorbike and the clustering algorithm clusters together the two objects. This is a still open problem, a possible solution will be to exploit further low-level

features. We have a partial solution. After the abduction process of creating new individuals in the ABox we start the reasoner in order to: (i) infer knowledge about the new individuals and (ii) to check the consistency of $\mathcal{O}$ with the new assertions in the ABox. This second step allows us to discard wrong clusters. For example, if there is an axiom where cars have only one bodywork and there is a cluster with two of them with some wheels there will have inconsistency. Thus, that cluster will be discarded with the generation of a new one.

5 Experimental Results

We evaluated the task of discovering part-whole relations by defining a gold standard: given the single parts we want to discover the whole object underlying such parts. This evaluation has been achieved by constructing a small dataset of 15 labelled images where every image has been labelled using the tool LabelMe [25]; labels are taken from an ontology $\mathcal{O}$ similar to the one described in Section 4. We concentrated on two image domains: houses with trees and street vehicles, but the method is general and can be easily extended to whatever domain. We obtained our ground truth labelling the single parts composing an object, such as foliages and tree trunks, and the object itself, the tree. Moreover, we also linked the singles parts to the corresponding object according to the part-whole relation. Parts are linked together using only one level of part-whole relation, i.e. we do not have chains of parts connected by the relation.

The next step was to compare the ground truth with the output of our framework: a partial model of $\mathcal{O}$, i.e. a predicted ABox $\mathcal{A}_P$ consistent with the axioms of $\mathcal{O}$. As described in the Section 4, $\mathcal{A}_P$ contains the individuals corresponding to the parts and to the whole objects, this process has been carried out using clustering techniques. Specifically, the experiments were conducted using the Java-ML library [1] with a clustering technique based on Kohonen's Self-Organizing Maps [15]. Such a technique was the one with better performance.

$\mathcal{A}_P$ is a set of assumptions over $\mathcal{O}$, so the goal is to compare such statements with the ground truth. Thus, we converted every labelled image into an ABox $\mathcal{A}_{GT}$ with the corresponding part-whole relations instantiated. In both the ABoxes we used the same identifiers for the individual names of the single parts, while the whole objects have different individual names. This is obvious because our goal is to predict the whole objects, so we cannot use the corresponding name of the ground truth. The idea is to compare the two ABoxes by individualizing groups of parts corresponding to the same object, i.e. in `partOf` relation with it. We are not interested in the name of such an object but only on its parts. Thus, for both the ABoxes we extracted pairs of individuals corresponding to parts of the same object. For $\mathcal{A}_P$ the set of these pairs is called *positive prediction* (P), the pairs coming from $\mathcal{A}_{GT}$ are the ground truth (T) and their intersection are the true positives (TP). Table 1 shows the performance of our framework, for every image in the ground truth, in terms of precision, recall and F-measure. The mean of these metrics are, respectively, 0.89, 0.87 and 0.84.

The results show a high F-measure, and for the 46.7% of the images we generate a fully correct interpretation. Nonetheless, there are problematic cases.

Table 1. Evaluation of the framework in terms of precision, recall and F-measure

Domain	Image	$\|P\|$	$\|T\|$	$\|TP\|$	Precision	Recall	F-measure
Street Vehicles	1	18	18	18	1.00	1.00	1.00
Street Vehicles	2	42	36	26	0.62	0.72	0.67
Street Vehicles	3	14	22	14	1.00	0.64	0.78
Street Vehicles	4	8	8	8	1.00	1.00	1.00
Street Vehicles	12	32	32	32	1.00	1.00	1.00
Street Vehicles	13	4	4	4	1.00	1.00	1.00
Street Vehicles	14	4	12	4	1.00	0.33	0.50
Street Vehicles	15	12	12	12	1.00	1.00	1.00
Houses, Trees	5	242	122	122	0.50	1.00	0.67
Houses, Trees	6	62	62	62	1.00	1.00	1.00
Houses, Trees	7	56	24	24	0.43	1.00	0.60
Houses, Trees	8	54	46	46	0.85	1.00	0.92
Houses, Trees	9	40	110	40	1.00	0.36	0.53
Houses, Trees	10	68	60	60	0.88	1.00	0.94
Houses, Trees	11	12	12	12	1.00	1.00	1.00

This is due to the fact that the clustering algorithm cannot correctly group the parts of an object. In the cases of low precision (e.g. image 7) the algorithm generates less clusters w.r.t. the ground truth; in the cases of low recall (e.g. image 14) the algorithm generates more clusters w.r.t. the ground truth.

6 Conclusions

In this work we addressed the semantic image interpretation as a procedure to extract structured information from images using an ontology. A possible use of such a structure is semantically querying images about their content. The novelty of this work is a fully formalization of the problem in terms of partial logical model of the ontology based on a simple intuition: as an image is a partial view of the world it has to be formalized as a partial model. Moreover, we stated that a partial model should adhere, as much as possible, to the image, so we need a heuristic to guide its construction towards the most plausible partial model. We applied the framework to a specific subtask: the extraction of part-whole relations between objects in an image. The heuristic guiding the construction of the partial model was based on a simple principle: the parts of an object are close in the space. We implemented this idea with a clustering technique that exploits both low-level and semantic features of the image. The method was tested on a built dataset obtaining, in average, good results.

As future work we aim to find a more efficient method for discriminating objects near in the space and in the semantics. In order to better evaluate the soundness of our framework we want to extend the experiments to a larger dataset. Furthermore, we want to generalize our method to patches with more weighted labels, exploring, for example, fuzzy DL approaches. An important open problem is finding heuristics guiding the construction of plausible partial models for other relations. This can be address, for example, using supervised learning techniques o probabilistic graphical models.

References

1. Abeel, T., Van de Peer, Y., Saeys, Y.: Javaml: A machine learning library. J. Mach. Learn. Res. **10**, 931–934 (2009). http://dl.acm.org/citation.cfm?id=1577069.1577103
2. Antanas, L., Frasconi, P., Costa, F., Tuytelaars, T., Raedt, L.D.: A relational kernel-based framework for hierarchical image understanding. In: Gimel'farb, G.L., Hancock, E.R., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K. (eds.) SSPR/SPR. LNCS, vol. 7626, pp. 171–180. Springer, Heidelberg (2012)
3. Antanas, L., van Otterlo, M., Mogrovejo, J.O., Tuytelaars, T., Raedt, L.D.: A relational distance-based framework for hierarchical image understanding. In: Carmona, P.L., Sánchez, J.S., Fred, A.L.N. (eds.) ICPRAM (2), pp. 206–218. SciTePress (2012)
4. Atif, J., Hudelot, C., Bloch, I.: Explanatory reasoning for image understanding using formal concept analysis and description logics. IEEE Transactions on Systems, Man, and Cybernetics: Systems **44**(5), 552–570 (2014)
5. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.): The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press, New York (2003)
6. Bannour, H., Hudelot, C.: Towards ontologies for image interpretation and annotation. In: Martinez, J.M. (ed.) 9th International Workshop on Content-Based Multimedia Indexing, CBMI 2011, June 13–15, Madrid, Spain, pp. 211–216. IEEE (2011)
7. Dasiopoulou, S., Kompatsiaris, I., Strintzis, M.G.: Applying fuzzy dls in the extraction of image semantics. J. Data Semantics **14**, 105–132 (2009)
8. Diligenti, M., Gori, M., Maggini, M., Rigutini, L.: Bridging logic and kernel machines. Machine Learning **86**(1), 57–88 (2012)
9. Espinosa, S., Kaya, A., Möller, R.: Logical formalization of multimedia interpretation. In: Paliouras, G., Spyropoulos, C.D., Tsatsaronis, G. (eds.) Multimedia Information Extraction. LNCS, vol. 6050, pp. 110–133. Springer, Heidelberg (2011). http://dx.doi.org/10.1007/978-3-642-20795-2_5
10. Fellbaum, C. (ed.): WordNet: an electronic lexical database. MIT Press (1998)
11. Haarslev, V., Hidde, K., Möller, R., Wessel, M.: The racerpro knowledge representation and reasoning system. Semantic Web Journal **3**(3), 267–277 (2012)
12. Han, F., Zhu, S.C.: Bottom-up/top-down image parsing by attribute graph grammar. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1778–1785 (October 2005)
13. Hobbs, J.R., Stickel, M.E., Appelt, D.E., Martin, P.: Interpretation as abduction. Artif. Intell. **63**(1–2), 69–142 (1993). http://dx.doi.org/10.1016/0004-3702(93)90015-4
14. Hudelot, C., Atif, J., Bloch, I.: Fuzzy spatial relation ontology for image interpretation. Fuzzy Sets and Systems 159(15), 1929–1951 (2008); from Knowledge Representation to Information Processing and Management Selected papers from the French Fuzzy Days (LFA 2006). http://www.sciencedirect.com/science/article/pii/S0165011408001012
15. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990)
16. Liu, H., Bao, H., Xu, D.: Concept vector for semantic similarity and relatedness based on wordnet structure. J. Syst. Softw. **85**(2), 370–381 (2012). http://dx.doi.org/10.1016/j.jss.2011.08.029

17. Liu, Y., Zhang, D., Lu, G., Ma, W.Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition **40**(1), 262–282 (2007). http://dx.doi.org/10.1016/j.patcog.2006.04.045
18. Moller, R., Neumann, B., Wessel, M.: Towards computer vision with description logics: some recent progress. In: Proceedings of the Integration of Speech and Image Understanding, pp. 101–115 (1999)
19. Neumann, B., Mller, R.: On scene interpretation with description logics. Image and Vision Computing **26**(1), 82–101 (2008) cognitive Vision-Special Issue. http://www.sciencedirect.com/science/article/pii/S0262885607001394
20. Neumann, B., Weiss, T.: Navigating through logic-based scene models for high-level scene interpretations. In: Crowley, J.L., Piater, J.H., Vincze, M., Paletta, L. (eds.) ICVS 2003. LNCS, vol. 2626, pp. 212–222. Springer, Heidelberg (2003). http://dl.acm.org/citation.cfm?id=1765473.1765497
21. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognitive Sciences **11**(12), 520–527 (2007)
22. Peraldi, I.S.E., Kaya, A., Möller, R.: Formalizing multimedia interpretation based on abduction over description logic aboxes. In: Grau, B.C., Horrocks, I., Motik, B., Sattler, U. (eds.) Description Logics. CEUR Workshop Proceedings, vol. 477. CEUR-WS.org. (2009)
23. Reiter, R., Mackworth, A.K.: A logical framework for depiction and image interpretation. Artificial Intelligence **41**(2), 125–155 (1989)
24. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning **62**(1–2), 107–136 (2006)
25. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. Int. J. Comput. Vision **77**(1–3), 157–173 (2008). http://dx.doi.org/10.1007/s11263-007-0090-8
26. Schroder, C., Neumann, B.: On the logics of image interpretation: model-construction in a formal knowledge-representation framework. In: Proceedings. of the Int. Conf. on Image Processing, vol. 1, pp. 785–788 (September 1996)
27. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical owl-dl reasoner. Web Semant. **5**(2), 51–53 (2007). http://dx.doi.org/10.1016/j.websem.2007.03.004
28. Smith, B., von Ehrenfels, C., Verlag, P.: Foundations of Gestalt theory. Philosophia Verlag Munich, Germany (1988)
29. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 129–136 (2011)
30. Staruch, B., Staruch, B.: First order theories for partial models. Studia Logica **80**(1), 105–120 (2005)
31. Straccia, U.: Reasoning within fuzzy description logics. J. Artif. Intell. Res. (JAIR) **14**, 137–166 (2001)
32. Zlatoff, N., Tellez, B., Baskurt, A.: Image understanding and scene models: a generic framework integrating domain knowledge and gestalt theory. In: International Conference on Image Processing, ICIP 2004, vol. 4, pp. 2355–2358 (October 2004)

Events Detection Using a Video-Surveillance Ontology and a Rule-Based Approach

Mohammed Yassine Kazi Tani[1], Adel Lablack[2(✉)], Abdelghani Ghomari[1], and Ioan Marius Bilasco[2]

[1] RIIR Laboratory, University of Es-Sénia, Oran, Algeria
yassine.kazi@gmail.com, ghomari65@yahoo.fr
[2] Laboratoire D'Informatique Fondamentale de Lille, Université de Lille 1, Lille, France
{adel.lablack,marius.bilasco}@lifl.fr

Abstract. In this paper, we propose the use of a Video-surveillance Ontology and a rule-based approach to detect an event. The scene is described using the concepts presented in the ontology. Then, the blobs are extracted from the video stream and are represented using the bounding boxes that enclose them. Finally, a set of rules have been proposed and have been applied to videos selected from PETS 2012 challenge that contain multiple objects events (e.g. Group walking, Group splitting, etc.).

Keywords: Ontology · Video surveillance · Blobs · Rules

1 Introduction

Nowadays, a growing amount of videos are available. This large amount of data that needs to be stored and indexed should be processed using efficient content based methods. Some of the existing works in video indexing use low-level features like color or motion for indexing video clips [5,7]. Other approaches have their indexing system based on high-level features such as human interpretation using meta-data and keywords [15,20]. These latter systems suffer from the exhaustive manual operations, and the semantic inconsistencies caused by different subjective interpretations made by people.

The semantic gap that exists between the low-level and the high-level features for an event could be solved by combining both levels using an ontology [8]. The use of ontologies for prior knowledge representation and scene understanding of video data is popular in many applications [12,21,22]. Gruber [9] defines the ontology as the representation of the semantic terms and their relationships. It consists of the representation of the concepts, their properties, and the relationship between concepts expressed in linguistic terms. The most important property is the derivation of an implicit knowledge through automated inference. It provides a formal framework to define domain knowledge [2].

We propose to use the concepts of a video surveillance ontology to derive rules that allows events detection from video sequences. The Ontology Web Language

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 299–308, 2015.
DOI: 10.1007/978-3-319-16181-5_21

(OWL) [17] has been used to represent our ontology and the Semantic Web Rule Language (SWRL) [13] to generate the inference rules.

The remainder of this paper is organized as follows. Section 2 reviews some related work in the field of video processing using ontologies. In Section 3, we describe the architecture of our ontology of the video surveillance domain. We describe the methodology used to derive the rules based on the video surveillance domain ontology in Section 4 using the PETS 2012 dataset as a case of study. Finally, we give concluding remarks and potential future work in Section 5.

2 Related Work and Background

Several works based on an ontology have been proposed to overcome the semantic gap between low-level and high-level features. Bagdanov et al. [1] present a system to solve the semantic gap between the high-level concepts and the low-level descriptors using a multimedia ontology. It contains visual prototypes that represent each cluster and act as a bridge between the domain ontology and the video structure ontology. Dasiopoulou et al. [8] have used color homogeneity as descriptor. The visual objects have been included in the ontology and the semantic concepts have been derived from color clustering and reasoning. Bertini et al. [3] have used both generic and domain specific descriptors to identify visual prototypes that represent elements of visual concepts. New instances of visual concepts are then added to the ontology through an updating mechanism of the existing concepts. Finally, the prototypes are used to classify the events and the objects that are observed in video sequences.

In video surveillance applications, some specific events like abnormal events have to be detected from streams provided generally by stationary cameras. An ontology can be used to support the indexing process. Xue et al. [21] proposed an ontology-based surveillance video archive and retrieval system. Lee et al. [10] implement a framework called Video Ontology System (VOS) to classify and index video surveillance streams. Snidaro et al. [18] have used a set of rules in SWRL language for event detection in video surveillance domain. In order to overcome the problem of the manual rules creation by human experts, Bertini et al. [4] proposed an adaptation of the First Order Inductive Learner technique (FOIL) for Semantic Web Rule Language (SWRL) named FOILS.

Most of the previous works in the surveillance domain have used the ontology tool and demonstrate its efficiency to help and manage the indexing and retrieval process. They consider events such as abandoned object, stolen object, a person who is walking from right to left, an airplane that is flying, etc. SanMiguel et al. [11] have proposed an ontology for representing the prior knowledge related to a video event analysis. It is composed of two types of knowledge related to the application domain and the analysis system. Domain knowledge involves all the high level semantic concepts (objects, events, context, etc.) while system knowledge involves the abilities of the analysis system (algorithms, reactions to events, etc.). However, this ontology determines only the best visual analysis framework (or processing scheme) without any inference reasoning for objects tracking and events detection or analysis.

In this paper, we propose to use an ontology based-approach to detect single/multiple objects events through a set of SWRL rules. It allows the transition from the blobs extracted using visual analysis module to the detection of an event.

3 The Architecture of the Ontology

The ontology approach is an effective way to support various processes for events detection in video surveillance domain. The scene is described using the concepts presented in the ontology and a video analysis module extracts the blobs from the streams using some low level property such as color, position, size, etc. The ontology considers these blobs as an input through the bounding boxes that enclose them and instantiate their features for creating the different DataType Property in the ontology. Then, the reasoner of our ontology classifies, in the first step, the different bounding boxes in their respective semantic meaning (Group_Of_Person/ Person) using a set of SWRL rules [13] and associates, in a second step, this video sequence, using another set of SWRL rules, to the appropriate video event class regarding the behavior of its objects.

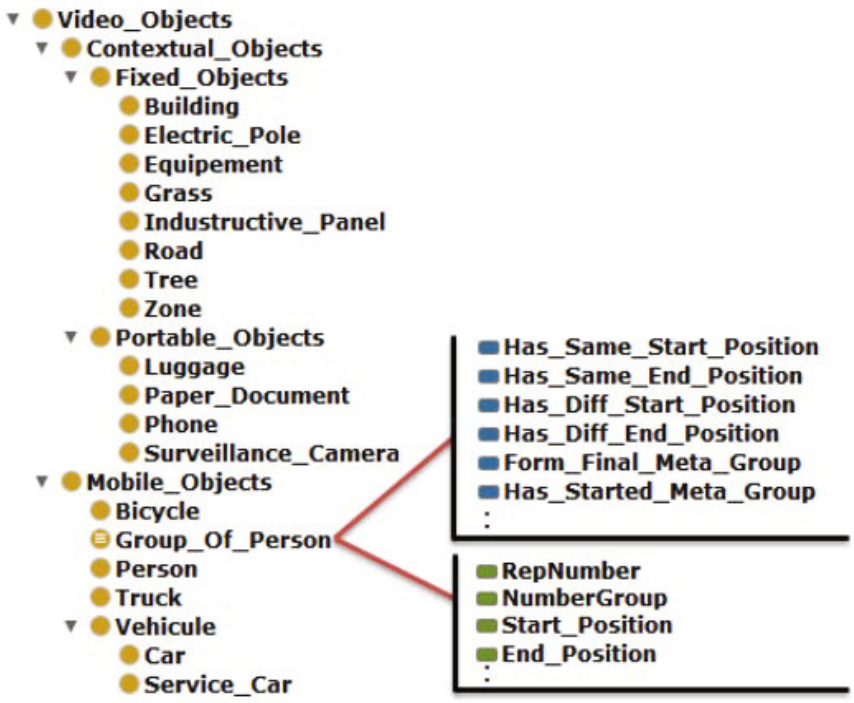


Fig. 1. Video_Objects class hierarchy sample

In order to have an efficient representation of the video surveillance domain, we preserved the same organization proposed by SanMiguel et al. [11] and complete it by adding new concepts. We organize our ontology in four categories, ranging from high-level concepts to low-level features : Video Events (gather all events that can happen in the video surveillance domain), Video Objects (represents a set of objects that can appear in a video sequence), Video Sequences (all the video sequences that could be indexed by our Ontology) and Bounding boxes (all the bounding boxes that enclose the blobs detected by the video analysis module in a video sequence with their low level features). The Figure 1 depicts a sample of the Video_Objects class hierarchy.

4 The Rule Based Approach

In this section, we propose to use the PETS 2012 dataset as a case of study to depicts our rule based approach that allows to handle a video surveillance ontology for events detection in video streams.

4.1 PETS 2012 Dataset

A set of events selected from PETS 2012 challenge [6] are used to experiment the efficiency of the proposed rules. This dataset contains different crowd activities and the task is to provide a probabilistic estimation of some events and to identify the start and the end of the events as well as transitions between them.

4.2 Scene Representation

In order to determine the best configuration of the processing schemes for detecting the events, we describe the scene in terms of concepts of our ontology. The Figure 2 shows an ideal and very precise segmentation of two scenes extracted from PETS 2012 challenge. Although some automatic techniques might be use for segmentation, we have started from a manual segmentation of the scene as the scene contains static elements that will not change over time (building, grass, electric pole, road, trees, car parks, restrictive roads). These elements have a strong semantic meaning, that can enhance the reasoning process and interpret the events resulting from other (volatile) elements (service car) that are subject to movements within the scene setting. For instance, special attention should be raised if moving objects are present in the Restrictive Road and deep analysis should be run to see if the moving objects are pedestrian or cars. Changes in appearances of studied objects can also be relevant in extracting meaningful events (a tree going reddish, might be a strong feature in detecting an abnormal event). Although, we are more focusing on movement reasoning, both kinds of changes (movement and appearance) result in the presence of regions yielding similar characteristics in terms of appearance and/or motion commonly called blobs.

Fig. 2. Scene Representation from PETS 2012 challenge camera views

4.3 Blobs Extraction

We propose an event detection approach based on blob regions. Blobs have proven to be a better feature cue than points, corners or edges as they usually have a larger coverage area and total occlusion of the subject is more unlikely to happen. So, we should identify all the major blobs in the scene. A major blob is defined as a blob that shows potential area size to be considered [16,19]. This is an essential step towards determining potential person/group.

In order to collect these blobs, several algorithms could be used. A background subtraction algorithm will classify the pixels of the input image into foreground and background. Then the blobs are extracted by groping together the foreground pixels belonging to a single connected component. We can also use optical flow by extracting the characteristics of each pixel in each motion image. These flows are then grouped into blobs that have coherent motion and are modeled by a mixture of multivariate Gaussians. The optical flow is useful to characterize each moving pixel according to certain features of the flow vector.

The Figure 3 highlights the bounding boxes that enclose the detected blobs in different situations like Group walking, Group running, Group Splitting, etc.

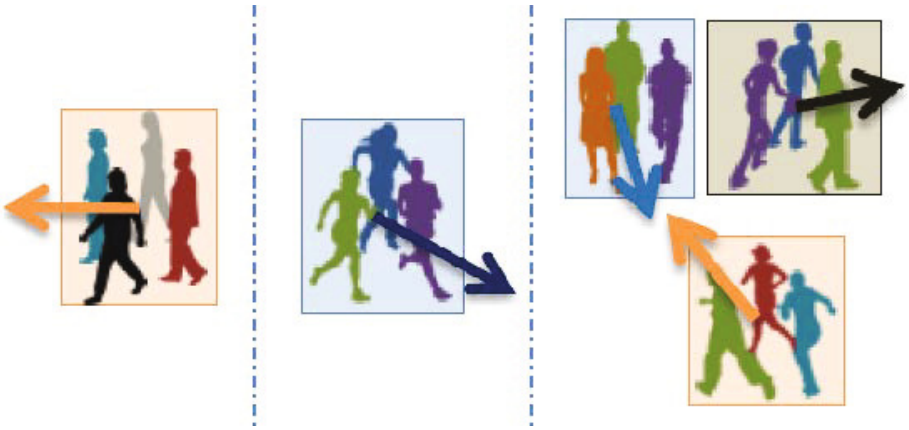


Fig. 3. Events from PETS 2012 challenge: Group walking, Group running, Group merging and Group splitting

A pre-processing stage is often applied to select the major blobs. It is done by applying some anthropomorphic assumptions and morphological operations. The following morphological operations are performed:

- Closing: Morphological closing smoothness sections of contours, fuse together narrow breaks and long gulfs.
- Fill holes: A flood-fill operation is performed to close up the remaining small holes.
- Removal of motion at boundary: Pixels of the motion region that are located along the boundary are eliminated to avoid ambiguity of the region belonging to a possible moving object.

At this stage each blob can represent either an entire object, an object sub-part or can be generated by noise. It is identified by a label and the surrounding bounding box. These bounding boxes are then used as input for the rule stage. The aim of this rules is to ensure the identification of semantically significant objects by analysing detected blobs over consecutives frames.

By comparing the bounding boxes found in two consecutive frames, our rule based approach is able to assess for each blob of the previous frame if it has been found or if undergoes a split or takes part in a merger. It consists in establishing the associations between the objects found in the previous frame and the blobs just extracted and grouped within the bounding boxes. We describe now our strategy according to the blobs that have been detected in the current frame:

- Straightforward tracking: this is the simplest case and it corresponds to two blobs without neighboring ones which are detected approximately in the same position in two successive frames and there are no splits nor merges (blob size is preserved or slightly varies). The concept of approximately in the same position is implemented trough the definition of a threshold on a distance measurement between the blobs.

- Splitting: a split is detected when a blob breaks in two distinct ones. We validate every split as soon as it occurs, creating two new objects. However, the original object identity is resumed if this fragmentation of the object into two blobs is temporary which may be due, for example, to an error during the detection phase.

- Merging: we detect a merging event when two objects having close past trajectories and detected up to frame at time t-1 merge their bounding boxes in the frame at time t. If these conditions are satisfied, the algorithm creates a new object joining the trajectories of the two previous ones

Some events that could happen may introduce a confusion in this process such as:

- Disappearance: an object detected in a frame at time t-1 is classified as lost in the current frame if no blob is present in the neighbourhood of the expected object position at time t. If an object is lost in proximity of an image border, the algorithm assumes that the object has left the scene, else waits for the appearance of the object in proximity of the place where it disappeared. Still, we should ensure that no other blob belonging to another semantically significant object was/is around, and takes the place of the previous.

- Occlusion: it is distinguished from merging/splitting events on the basis of the direction of the past trajectories. When an occlusion occurs, we wait to analyze the scene for a specific number of frames to find the correct association between the objects found before and after the occlusion.

4.4 The Rules Construction

Different events from the PETS 2012 challenge could be used to depict the efficiency of the proposed approach such as:

- Group running and walking events: it consists to estimate if the people forming a group are walking or running. These events can be identified using the motion magnitude in each image. High magnitude event means running while a low magnitude means walking event. The detection is done either by defining an experimental threshold or using a classifier with feature such as the average speed of movement.
- Group formation and splitting events: it consists in the detection and the analysis of the position, the orientation and the speed of the groups.

We have used the Rule plugin of Protégé [14] to write the inference rules of our engine in SWRL language. Our rules are divided into 3 categories:

- Distance rules: it consists on checking the distance between the detected bounding boxes in the current frame. The bounding boxes that are close to each other are grouped into a major bounding box.
- Tracking rules: it consists on tracking the major bounding boxes generated by the previous category over the frames to detect the start/end position.
- Event rules: it consists in analyzing the behaviour of the groups identified in the previous category in order to detect the appropriate event.

The left side of the rule (before the arrow) is checked by the inference engine and the reasoner infer or not the right side. The Figure 4 depicts the construction of a distance rule. It checks if two bounding boxes could be grouped into a major bounding box.

BB(?BBx), BB(?BBy), Frame(?F1), MBB(?MBB1), MBB(?MBB2), BB_Detected_In_Frame(?BBx, ?F1), BB_Detected_In_Frame(?BBy, ?F1), BB_Bottom_Left_Point_Y(?BBx, ?h), BB_Bottom_Right_Point_Y(?BBy, ?d), BB_Number(?BBx, ?n4), BB_Number(?BBy, ?n5), BB_Top_Left_Point_X(?BBx, ?a), BB_Top_Left_Point_X(?BBy, ?f), BB_Top_Left_Point_Y(?BBx, ?e), BB_Top_Left_Point_Y(?BBy, ?l), BB_Top_Right_Point_X(?BBx, ?j), BB_Top_Right_Point_X(?BBy, ?b), BB_Top_Right_Point_Y(?BBy, ?c), MBB_ID(?MBB1, ?n1), MBB_ID(?MBB2, ?n1), Number_BB_In_Frame(?F1, 2), Number_Frame(?F1, ?n1), Number_MBB(?MBB1, ?n2), Number_MBB(?MBB2, ?n3), add(?x2, ?b, 20), greaterThan(?a, ?b), greaterThan(?h, ?d), greaterThan(?n3, ?n2), greaterThanOrEqual(?b, ?x1), greaterThanOrEqual(?e, ?c), lessThanOrEqual(?a, ?x2), lessThanOrEqual(?e, ?d), subtract(?x1, ?a, 20), subtract(?z1, ?j, ?f), subtract(?z2, ?h, ?l) -> BB_Represent_MBB(?BBx, ?MBB1), BB_Represent_MBB(?BBy, ?MBB1), MBB_Detected_In_Frame(?MBB1, ?F1), MBB_H(?MBB1, ?z1), MBB_Top_Left_Point_X(?MBB1, ?f), MBB_Top_Left_Point_Y(?MBB1, ?l), MBB_W(?MBB1, ?z2)

Fig. 4. A rule for grouping two bounding boxes into a major bounding box

This rule presented above is constructed as follow: (i) The reasonner checks in the current frame if the positions of the two bounding box ($BB1$, $BB2$) are close in the X and Y axis. The Bounding boxes are then tested as $BBx \rightarrow BB2$ and as $BBy \rightarrow BB1$ using the following conditions: (i) $Top\_Right\_Point\_YBB1 \leqslant Top\_Left\_Point\_YofBB2 \leqslant Bottom\_Right\_Point\_YofBB1$, (ii) $Top\_Left\_Point \_XofBB2 \leqslant Top\_Right\_Point\_XofBB1 + 20$ and $Top\_Right\_Point\_XofBB1 \geqslant Top\_Left\_Point\_XofBB2 + 20$. In this case, the reasoner will infer that both bounding boxes belong to the same Major Bounding Box and updated it.

A large set of rules is proposed to model all the situations that could happen in the scene according to the events handled by our ontology. The output of each category could be used as input for another one. Indeed, an event is detected using a rule that took as input the information inferred by a tracking rule that has been applied to major bounding boxes identified using distance rules.

The inherent difficulty of writing down rules in SWRL or equivalent language is the fact that the events are spanning over various time intervals. Various time windows can be applied to the same event detection. A split event can occur in a very short time-frame, if the groups are evolving at high speed or it could take a long time-frame if the groups are evolving at low speed. However, we are using a fixed time-window in order to simplify writing rules.

5 Conclusion

Video Surveillance systems become popular in our daily life to ensure security and safety and allows to study human behavior. In this paper, we have presented our rule based approach that allows to handle a video surveillance ontology to detect single or multiple objects events.

In our future work, we will extend our ontology to model new concepts and improve our SWRL rules for handling different events that can occur in video surveillance domain.

References

1. Bagdanov, A.D., Bertini, M., Del Bimbo, A., Serra, G., Torniai, C.: Semantic annotation and retrieval of video events using multimedia ontologies. In: International Conference on Semantic Computing (ICSC), pp. 713–720 (2007)
2. Ballan, L., Bertini, M., Del Bimbo, A., Serra, G.: Semantic annotation of soccer videos byvisual instance clustering and spatial/temporal reasoning in ontologies. Multimedia Tools and Applications **2**, 313–337 (2010)
3. Bertini, M., Del Bimbo, A., Torniai, C.G.C., Cucchiara, R.: Dynamic pictorial ontologies for video digital libraries annotation. In: 1st ACM Workshop on The Many Faces of Multimedia Semantics, pp. 47–56 (2007)
4. Bertini, M., Del Bimbo, A., Serra, G.: Learning ontology rules for semantic video annotation. In: 2nd ACM Workshop on Multimedia Semantics (2008)
5. Del Bimbo, A., Pala, P., Vicario, E.: Spatial arrangement of color flows for video retrieval. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 413–416 (2001)
6. PETS 2012 challenge (2012). http://www.cvg.rdg.ac.uk/pets2012/a.html
7. Chupeau, B., Forest, R.: An evaluation of the effectiveness of color attributes for video indexing. In: SPIE Storage and Retrieval for Media Databases, pp. 470–481 (2001)
8. Dasiopoulou, S., Mezaris, V., Kompatsiaris, I., Papastathis, V.-K., Strintzis, M.G.: Knowledge assisted semantic video object detection. IEEE Transactions on Circuits and Systems for Video Technology **10**, 1210–1224 (2005)

9. Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, (5–6):907–928, November-December 1995
10. Lee, J., Abualkibash, M.H., Ramalingam, P.K.: Ontology-based shot indexing for videosurveillance system. In: Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering, pp. 237–242 (2008)
11. Miguel, J.C.S., Sanchez, J.M.M., García-Martín, A.: An ontology for event detection and its application in surveillance video. In: 6th IEEE International Conference Advanced Video and Signal based Surveillance (AVSS), pp. 220–225 (2009)
12. Noyet, N.F., McGuinness, D.L.: Ontology development 101: A guide to creating your first ontology. Technical report (2001)
13. O'Connor, M.F., Knublauch, H., Tu, S., Grosof, B.N., Dean, M., Grosso, W., Musen, M.A.: Supporting rule system interoperability on the semantic web with SWRL. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 974–986. Springer, Heidelberg (2005)
14. Protégé. The protégé project (2012). http://protege.stanford.edu
15. Sanchez, J.M., Binefa, X., Vitria, J., Radeva, P.: Linking visual cues and semantic terms under specific digital video domains. Journal of Visual Languages and Computing **11**(3), 253–271 (2000)
16. See, J., Wei, L.S., Hanmandlu, M.: Human motion detection using fuzzy rule-base classification of moving blob regions. In: International Conference on Robotics, Vision, Information and Signal Processing (ROVISP) (2005)
17. Smith, M.K., Welty, C., McGuinness, D.L.: Owl web ontology language guide. In: W3C Recommendation (2004). http://www.w3.org/TR/2004/REC-owl-guide-20040210/
18. Snidaro, L., Belluz, M., Foresti, G.L.: Representing and recognizing complex events in surveillance applications. In: 4th IEEE International Conference Advanced Video and Signal based Surveillance (AVSS), pp. 493–498 (2007)
19. Di Stefano, L., Mola, M., Neri, G., Varani, E.: A rule-based tracking system for video surveillance applications. In: International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES) (2002)
20. Wu, Y., Zhuang, Y., Pan, Y.: Content-based video retrieval integrating human perception. In: SPIE Storage and Retrieval for Media Databases, pp. 562–570 (2001)
21. Xue, M., Zheng, S., Zhang, C.: Ontology-based surveillance video archive and retrieval system. In: 5th International Conference on Advanced Computational Intelligence (ICACI) (2012)
22. Yusuf, J.C.M., Su' ud, M.M., Boursier, P., Alam, M.: Extensive overview of an ontology-based architecture for accessing multi-format information for disaster management. In: International Conference on Information Retrieval and Knowledge Management (CAMP), pp. 294–299 (2012)

Semantic-Analysis Object Recognition: Automatic Training Set Generation Using Textual Tags

Sami Abduljalil Abdulhak[1(✉)], Walter Riviera[1], Nicola Zeni[2], Matteo Cristani[1], Roberta Ferrario[2], and Marco Cristani[1]

[1] Department of Computer Science, Cá Vignal 2, Verona, Italy
{sami.naji,walter.riviera,matteo.cristani,marco.cristani}@univr.it
[2] Laboratory for Applied Ontology, Consiglio Nazionale Delle Ricerche (CNR), Via Alla Cascata 56/c, Trento, Italy
{nicola.zeni,roberta.ferrario}@loa.istc.cnr.it

Abstract. Training sets of images for object recognition are the pillars on which classifiers base their performances. We have built a framework to support the entire process of image and textual retrieval from search engines, which, giving an input keyword, performs a statistical and a semantic analysis and automatically builds a training set. We have focused our attention on textual information and we have explored, with several experiments, three different approaches to automatically discriminate between positive and negative images: keyword position, tag frequency and semantic analysis. We present the best results for each approach.

Keywords: Training set · Semantic · Ontology · Semantic similarity · Image retrieval · Textual tags · Flickr · Object recognition

1 Introduction

The process of automatically building a training set of images for object recognition given a class name is a recent challenge originated from the Semantic Robot Vision Challenge [1]. The idea is to mine on-line repositories of images and use them to support image classifiers in object recognition tasks [2]. Given this strategy, the goal is to exploit search engines and retrieve images that can be used to feed a training set for a specific class.

The problem falls under the topic of Image Retrieval (IR): given a certain query in a form of a keyword or an image, the system should present images related to the query. Two main strategies have been deployed to tackle such problem: content-based image retrieval (CBIR) [3] and tag/keyword-based image retrieval (TBIR)[4].

CBIR leverages on the concept of visual similarity between the querying image and the retrieved ones using elementary visual features such as color and shape, through a matching of their properties, while TBIR tries to overcome the

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 309–322, 2015.
DOI: 10.1007/978-3-319-16181-5_22

limitations presented by the CBIR system through the exploitation of the textual information conveyed with images, applying document retrieval techniques to boost the retrieval performances. Nevertheless TBIR performances are influenced by the availability and quality of the textual information users supply with images. In fact, while manually annotating images, users often misuse tags or provide incomplete textual descriptions of the image content [5–7].

The use of the textual information conveyed with images in the process of image retrieval or image classification is not a novel strategy, there have been several works that explore how the textual information can be used, among them [8–11]. Recent approaches explore the use of tags completion either by mining extra textual information obtained from Internet or by using content image analysis to fill the gap[6,12].

In the present work we propose a framework that helps to automate the entire process of training data set construction. The main idea is to use textual information that comes along with images on the web to fully automate the training set generation. To achieve this, we assume that the user annotation process is not always reliable since users are not experts and may annotate images with different purposes. Even though users upload images in a social context where other users can use collaborative tagging to annotate images, tags are not validated and so the subjectivity elements are not removed. Moreover, since users are non expert, they tend to use ambiguous and inappropriate tags to describe images content. The main idea is to explore how statistical and semantic analysis of textual information can help to fully automate the training set construction. In particular, we employ statistical and semantic analysis to filter the textual information, pruning noisy tags and retaining only those that are highly correlated with the content of an image, thus discriminating positive from negative images[1]. We use statistical measures such as frequency and tags distribution, as well as WordNet and semantic distances between tags to evaluate their correlation and explore their contribution in the discriminative process. Our starting assumption is that, by incrementally injecting semantic techniques into the analysis of textual annotation, performances rise and, to validate such assumption, a set of experiments are presented.

The rest of the paper is structured as follows: Section 2 describes the challenges of the image retrieval task and provides an overview of works in the area. The method we propose is introduced in Section 3. Sections 4 discusses the experimental setup and evaluation method, while the evaluation results are presented in Section 5. Finally, conclusions and directions for future work are presented in Section 6.

[1] We consider as positive those images in which the prominence of the object presented in the image indicates that the image fully represents it. On the contrary, we consider as negative those images where the target object is absent or only partially present/visible, as indicated in the list in section 4.

2 Related Work

Annotation is a widely used technique to characterize objects portrayed in images by adding textual tags. The textual tags associated with images have been shown to be useful, improving the access to photo repositories both using temporal [13] and geographical information [14]. One of the popular online tag-based photo sharing repositories is Flickr, allowing users to freely assign one or more chosen keywords for an image for personal organization or retrieval purposes. In other words, it allows users to perform tagging, that is the act of adding words to images, describing the semantics of the visual contents. Users are thus implicitly encouraged to add more keywords, creating relatively large amounts of rich descriptions of objects presented in images. However, the textual tags associated with images are often noisy and unreliable, posing a number of difficulties when dealing with IR.

A number of approaches have been proposed to measure the reliability of the textual tags accompanying images [15–17]. In [17], the authors present a Flickr distance to measure the correlation between different concepts obtained from Flickr. Given a pair of concepts (e.g., car-dog), the algorithm tries to compute the semantic distance between them using square root of Jensen-Shennon divergence. The authors rely on the scores by considering the higher score distance as an indication of high relatedness of a pair of concepts. Related researches have been also focused on investigating which objects people observe most in an image, which they annotate or tag first, and what influence them in choosing words to describe objects depicted in images.

Spain and Perona [15] study the idea of *"importance"* of objects in an image and conclude that important objects are most likely to be tagged first by humans when asked to describe the contents of an image. The authors develop a statistical model validating the notion of dominant object in an image, demonstrating that one can foresee a set of prominent keywords based on the visual cues through regression. A work that is closely related to ours is presented by Hwang and Grauman [18]. They introduce an unsupervised learning method for IR that uncovers the implicit information about the object importance in an image, exploiting a list of keyword tags provided by humans. The proposed method is able to disclose the relationship between human tendencies in tagging images (e.g., words order in the tag list) and the relative importance of objects in an image.

Traditional techniques rely on features extracted from visual contents with visual category models learnt directly from image repositories that require no manual supervision [8–11]. The intuition behind the approach proposed in [9] is to learn object categories from just a few training images in an incremental manner, using a generative probabilistic model. Similarly, Li-Jia Li and Fei-Fei Li [10] propose an incremental learning framework, capable of automatically collecting large image datasets. The authors build a database from a sample of seed images and use the database to filter out newly crawling images by eliminating irrelevant examples.

Fergus et al. [11] introduce a method able to learn object categories by their name, exploiting the raw images automatically downloaded from the Google

image search engine. The introduced approach is able to incorporate spatial information in translation and scale invariant style, possessing the ability to tackle the high intra-category variability and isolate irrelevant images produced by the search engine.

Vijayanarasimhan and Grauman [19] propose an unsupervised approach to learn visual categories by their names using a collection of images pooled from keyword-based search engines. The main goal underneath the proposed approach is to harvest multiple images, by translating the query names into several languages and crawling the search engines for images using those translated queries. The false positive categories are collected from random sample images found in categories that have different names from the category of interest.

We are working on a challenge that is: given the textual tags provided by humans and associated with images, we want to automatically build a good training set by discriminating images as either related or unrelated to a targeted object.

3 Method

In this paper our goal is to take advantage of the textual tags available with images to automatically select the most representative of an object category for training a classifier, without looking at the nature of the objects therein. To do so, we exploit both semantic analysis and pure statistical approaches. These considerations lead us to focus on three main features:

- **keyword position**, to capture an image as related or unrelated on the basis of a keyword (i.e., object class name) position in a tag list;
- **semantic analysis**, to measure the semantic relatedness by means of semantic distance measures;
- **tag frequency**, to count the frequency of usage of each tag from a list describing the object class.

Figure 1 presents a schematic representation of our framework. A detailed description of the procedure is provided in the subsequent subsections.

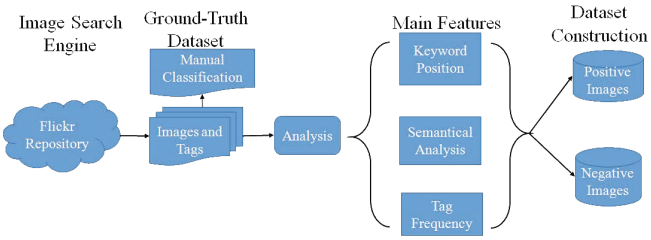


Fig. 1. A schematic representation of our framework.

3.1 Keyword Position

The textual tags given in a tag list and associated with an image describing its content could reasonably help us to derive important and valuable information about the nature of the depicted objects. However, the order in which the textual tags are placed in a tag list is most likely to be influenced by the objects position and size in the visual content [20]. Therefore, it is reasonable to claim that the first textual tags in the list are mostly representing the objects in the center of an image. Taking this keypoint into account, we use this feature to develop 5 different strategies which follow the same algorithmic structure:

Algorithm 1. Keyword Position

Data: a Keyword (i.e.,the object class name) and
$T = \{t_i | \quad \forall \quad Image \quad i \in Keyword, \quad \exists \quad tag - list \quad t_i \quad \}$
Result: A partition of the Images ∈ Keyword in:
Image-$P = \{i_p | i \in$ Images which are usable to build a training dataset$\}$
Image-$N = \{i_n | i \in$ Images which are outliers$\}$

```
Initialization;
foreach i ∈ Images do
    tags ← load t_i;
    clean(tags)tags_n ← extract the first n tags from tags;
    if "keyword" ∈ tags_n then
        Image-P ← i;
    else
        Image-N ← i;
```

Algorithm 1 is designed to demonstrate the systematic workflow of the keyword position feature. Given a tag list comprising a number of textual tags and corresponding to a particular image, the algorithm tries to search for the keyword through the list in the first n positions. The algorithm then labels the image as positive (reliable) if it is related to the class name or negative (outlier) otherwise. It is noteworthy that the clean operation provided in the algorithm is used to remove words with less than three characters, empty strings and non-alphabetic texts. It also splits long sentences into single words, when they are separated by the "_" symbol.

3.2 Semantic Analysis

To define the semantic relatedness or its inverse of the object class characterized by a keyword to the textual tags being used, semantic distance must be measured. Therefore we propose to apply two different standard semantic distance measures: WordNet and Jiang and Conrath [20]. First we adopt the WordNet distance [21]. WordNet is a large-scale lexical database that organizes English terms and their syntactic roles into synsets. Synsets are interlinked by means of conceptual-semantic and a variety of lexical relations. We choose WordNet due to the fact that it is the first attempt to organize a great amount of concepts according to semantic relations and a hierarchy. Since WordNet provides

a lexical relationship between concepts, it is beneficial to semantically measure relatedness of the object class to its related tags by their lexical relationship, such as meronymy (parthood, e.g. bus-wheels) or hypernym (generalization, e.g. bus-vehicle) and so on.

Secondly, we apply the distance measure proposed by Jiang and Conrath in [20]. They formulate their approach in the form of conditional probability of reaching an item of a child synset given an item of one of its parent sysnsets.

We use this feature and run several experiments according to the following algorithmic structure:

Algorithm 2. Semantic Analysis

Data: a Keyword (i.e., the object class name) and
$T = \{t_i | \quad \forall \quad Image \quad i \in Keyword, \quad \exists \quad tag - list \quad t_i \quad \}$
Result: A partition of the Images ∈ Keyword in:
Image-$P = \{i_p | i \in$ Images which are usable to build a training dataset$\}$
Image-$N = \{i_n | i \in$ Images which are outliers$\}$

```
1 Initialization;
2 foreach i ∈ Images do
3     tags ← load t_i;
4     clean(tags);
5     score_i ← sum or mean of the distance values of the tags;
6     if if score_i ≥ a Threshold τ then
7         Image-P ← i;
8     else
9         Image-N ← i;
```

Algorithm 2 is developed to clearly illustrate how we apply the semantic analysis feature to measure the semantic relatedness or its inverse of the object class to its textual tags. As already mentioned above, we adopt two different distance measures: WordNet and Jiang and Conrath. The algorithm takes the object class (represented by a keyword) and each image's tag list, then computes the distance of the keyword to every single textual tag in the tag list, yielding a score for each. If the algorithm finds no semantic distance between the keyword and a textual tag, it discards the tag. The algorithm therefore labels an image as positive (reliable) if its score is equal or above a threshold τ; otherwise it labels it as negative (outlier). The threshold value τ changes with respect to the experiment (see Section 4).

3.3 Tag Frequency

To understand which are the most frequently used tags (words) that describe images related to a certain object class, we compute the frequency values of all the single $tag_{(i,j)}$ as their occurrences probability. The idea is to perform a selection based on the utility of the words used to describe the object depicted

in an image. The frequency value of a single $tag_{(i,j)}$ is computed as follows:

$$Feq(tag_{(i,j)}) = \frac{O - tag_{(i,j)}}{\sum_{i=1}^{N_{images}} length(tag_i)},$$

where $tag_{(i,j)}$ is the j^{th} tag of the tag list associated to image i, and $O - tag_{(i,j)}$ is the total number of a $tag_{(i,j)}$ occurrences. In particular, if a given frequency value of a single $tag_{(i,j)}$ is relatively high, it means that many images of the considered object class require it into their descriptions. In other words, it is natural to think that if we are looking at an image of a "car", we highly expect to observe higher frequency values for tags like *"wheel"* or *"driver"* than *"pizza"* or *"pencil"*.

We use this feature to develop 12 different strategies which follow the same algorithmic structure:

Algorithm 3. Tag Frequency

Data: a Keyword (i.e., the object class name) and
$T = \{t_i | \quad \forall \quad Image \quad i \in Keyword, \quad \exists \quad tag - list \quad t_i \quad \}$
Result: A partition of the Images ∈ Keyword in:
Image-$P = \{i_p | i \in$ Images which are usable to build a training dataset$\}$
Image-$N = \{i_n | i \in$ Images which are outliers$\}$

```
Initialization;
foreach i ∈ Images do
    tags ← load t_i;
    clean(tags);
    score_i ← sum or mean of the frequency values of the tags;
    if if score_i ≥ a Threshold τ then
        Image-P ← i;
    else
        Image-N ← i;
```

Algorithm 3 uses frequency values to determine if a given image is related to the object class. To do this, it combines the frequency values of each $tag_{(i,j)}$ to produce a score. Then, it labels an image i as positive (reliable) if its score is equal or above a threshold τ; otherwise it labels it as negative (outlier). The threshold value τ changes with respect to the experiment (see Section 4).

4 Experiments

We devote this section to demonstrate the systematic workflow of our framework. Firstly, we pool images for a set of 21 object classes taken from the standard Caltech101[2], using Flickr online photo sharing[3]. Each class contains 400 images

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101
[3] https://www.flickr.com/

as well as their corresponding tag lists (tag_i). For simplicity, the number of crawled images has been defined in order to minimize the computational time of downloading images and managing their tags during the experiments. The effective number of classes have been normalized to 16, avoiding the classes that are composed by a bi-gram (i.e., two words). The remaining classes are: *accordian, bonsai, euphonium, face, laptop, menorah, nautilus, pagoda, panda, piano, pyramid, revolver, starfish, sunflower, umbrella, watch.* Since there are 400 images and 400 tag lists per class, the dataset is composed of 6400 images and 6400 tag lists.

To generate the ground-truth for our experiment in a more effective and efficient way, we build a graphical user interface (GUI) that allows us to manually label an image as positive or as negative with respect to the object class. For reliable manual classification, some guidelines are defined and adopted. If the following guidelines are satisfied, then an image is labeled as negative; otherwise as positive:

- an image is completely unrelated with the object specified by the category it belongs to;
- an image contains irrelevant parts of the object, that is, parts that alone are not sufficient to make the category object identifiable;
- an image contains only internal parts of the category object (like a cockpit of an airplane or an engine of a car);
- an image is a drawing or a caricature of the category object.

For each single feature we run several different experiments based on different strategies. Each strategy differs from the others with regard to the method used to compute the threshold. This produces different results in determining if a given tag list is associated to a positive or negative image.

Referring to the algorithms described in the subsection 3.1, 3.2, 3.3, we give a brief explanation of the strategies associated to the threshold which produces the best discrimination results:

Feature 1: Based on experiments performances, we obtain the best result when searching if a keyword is found in the first three positions in the tag list. Surprisingly, this feature does not involve any cleaning mechanism of textual tags in the tag list (it avoids the step number 4 of algorithm 1). However, the feature takes the textual tags as they are provided by Flickr. At this point one may ask why using contaminated textual tags in a tag list is, unexpectedly, producing better results than the cleaned version. The answer lays in the "filtering" mechanism of the textual tags. Cleaning the tag list tag_i implies producing more single words (tag_j) since the tag sentences are split. This increases the probabilities of finding the right match with the keyword, therefore a higher number of tags labeled as positive. This has been confirmed by the number of false positives generated using the other strategies, which is widely higher than the number of false positives produced by the strategy just described. To provide a better understanding of what happens if we do not perform any tag cleaning on the tag list, we present the following example: given the tag list relative to a negative

image of the *panda* class: *"zoo_atlanta", "taishan", "giant_panda"*, the keyword would not be matched since the substring matching is not performed. Therefore, the image is labeled as negative. This results change if we clean the tag list by splitting the sentences into single words. The cleaned tag list becomes: *zoo, atlanta, taishan, giant, panda*. In this case, the keyword would match with the 5^{th} tag and therefore the image is now labeled as positive.

Feature 2: This feature uses two different measures: the standard semantic distance provided by WordNet, and the distance proposed by Jiang and Conrath in [20]. To select the one which produces the best results, we use both metrics to run the 12 strategies. We used these two distances since they are widely adopted in literature. The comparison results are shown in figure 2 .

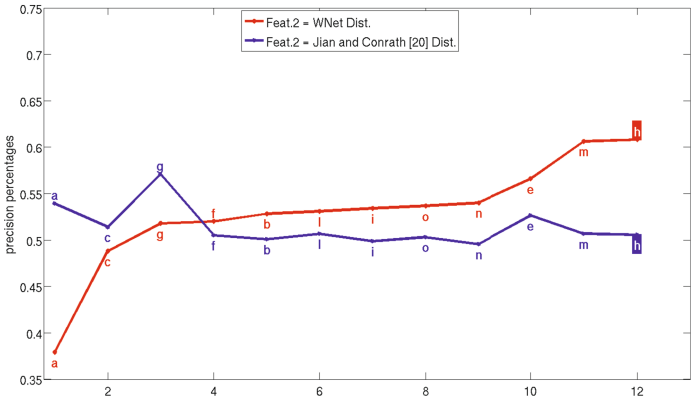


Fig. 2. Summary results obtained by using WordNet and Jiang and Conrath distances in [20] in all the strategies. WordNet distance is outperforming in average in all of the strategies. We compute the precision rate for each strategy $(a, b, \ldots, o)$ as: $\#TruePositive/(\#TruePositive + \#FalsePositive)$.

Using WordNet distance as shown in figure 2, we observe constant increase in the average performances of all strategies. Therefore, in the following description we are mainly referring to the WordNet distance. The strategy based on the WordNet distance, which gives the best results, uses the following criteria to split the images set: defining the $scores_i$ as the mean of the distances between the considered tags and the keyword:

$$scores_i = mean(Distance(tag_{(i,j)} - keyword))$$

$$Feat2(scores_i) = \begin{cases} positive & \text{if } mean(scores_i) \geq \tau \\ negative & \text{otherwise} \end{cases}$$

The best result is obtained using this strategy when the threshold is set to $\tau = median(scores_I)$, where the $scores_I$ is the vector of all the $scores_i$.

Feature 3: The strategy based on the tag frequency feature, which produces the best results, compared with the other strategies, uses the following criteria to split the images set: defines the $scores_i$ as the sum of the frequency values of the considered tags with respect to the keyword:

$$scores_i = \sum_{i=1}^{N_{images}} Fq(tag_{(i,j)})$$

$$Feat3(scores_i) = \begin{cases} positive & \text{if } mean(scores_i) \geq \tau \\ negative & \text{Otherwise} \end{cases}$$

We reach the best results when the threshold is set to $\tau = mean(scores_I)$, where the $scores_I$ is the vector of all the $scores_i$.

5 Performances Evaluations

To assess the reliability of the experimental performances of the features described beforehand, we select n images labeled as positives from all the strategies and from Flickr. Hence, we count the true positives and the false positives that have been generated by the strategies and by Flickr (in this case, the false positives are the ones we manually label as negatives). Since the main goal of this framework is to generate a reliable dataset of images, for this reason, all of our strategies tend to produce more negative than positive labels. This behavior allows to minimize the number of the false positive labels generated during the experiments. Since not all strategies produce the same number of positive labels, to avoid the problem of getting some Null values, we fix $n = min(P - labels)$ of each feature. The selection of the n labels has been done randomly for Flickr, while for our strategies the first n are considered. To ensure the consistency of Flickr performances, we average the results produced after 10 random selections.

Table 1 displays the percentage values of the performances obtained using Flickr and our best strategies. The column $\#P - labels$ contains the different n values used for each class. The column $GT - Positives$ presents the number of true positives within the ground-truth.

To make the performances reported in the table more comparable, we recalculate the precision percentages by fixing $n = 50$ positive labels[4] per class. Also in this case, the selection of the 50 labels has been done randomly for Flickr, while for our strategies it is referred to the first n. In figure 3, we provide the average values of each strategy for all the classes with $n = 50$.

In this last case, an exception is done for the *"euphonium"* category, since it is composed by just 9 positive images also in the ground-truth.

At this point, one may be skeptical about the reliability of our strategies, since we are estimating their performances by considering only 50 images against the 400 downloaded. Therefore, if we observe how the performances change when we consider all the available positive labels shown in table 1, we are more confident

[4] This parameter has been set by considering the lowest common number of labels.

Table 1. Precision results obtained using all the features for 16 classes. Flickr provides the number of correct positive labels from the n images downloaded from the Flickr repository. *Feat* is an abbreviation for feature, where *Feat.1* refers to keyword position, *Feat.2* refers to semantic analysis, and *Feat.3* refers to tag frequency.

Classes	# P- labels	GT-Positives	Flickr	Feat.1	Feat.2	Feat.3
watch	218	386 / 400	94.95	95.87	96.79	96.79
sunflower	178	379 / 400	93.26	97.19	96.63	96.63
bonsai	119	362 / 400	90.76	90.76	92.44	88.24
panda	182	359 / 400	89.56	90.11	32.31	97.25
laptop	171	359 / 400	88.30	92.98	93.57	87.72
pyramid	203	250 / 400	65.02	60.10	64.04	64.04
starfish	170	211 / 400	49.41	60.00	56.47	53.53
piano	50	105 / 400	37.50	58.33	37.50	70.83
umbrella	175	164 / 400	37.14	41.14	41.71	44.00
menorah	148	146 / 400	34.46	33.78	29.73	35.81
accordion	158	118 / 400	31.01	29.75	31.65	28.48
pagoda	167	114 / 400	29.94	32.34	34.13	38.32
face	135	120 / 400	28.15	31.11	25.19	27.41
revolver	127	110 / 400	26.77	38.58	42.52	31.50
nautilus	163	67 / 400	17.79	22.09	25.15	17.79
euphonium	8	9 / 400	0	62.5	0	0

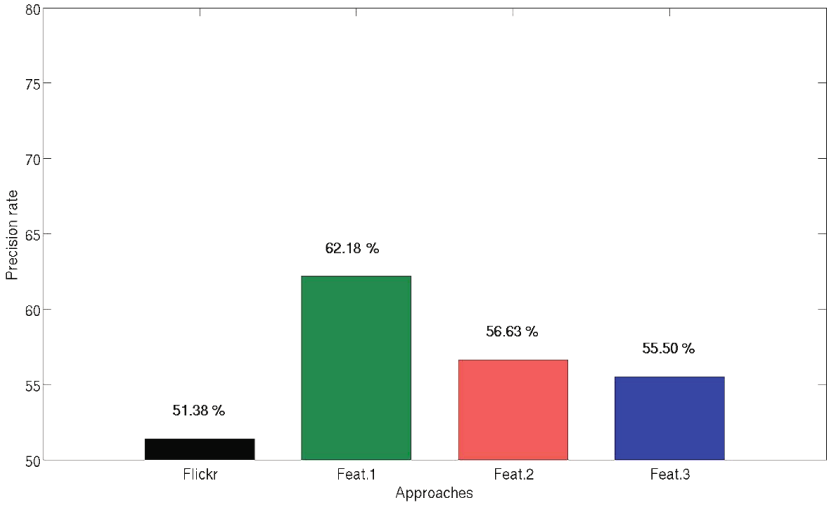


Fig. 3. Summary of results of all the features by fixing $n = 50$. The highest precision is given using *feat.1* (i.e., keyword position).

on our results. Indeed, if we calculate the average of the positive labels considered in the last case, we can observe (see table 2) that the performances remain

constant when setting $n \neq 50$. The overall performance of our strategies still outperforms Flickr. In particular, using keyword position, the average performance obtained is encouragingly good (about 11% higher than Flickr). This information is further enriched since it provides us with a more reliable percentage value than the ones provided by the results of $n = 50$.

Table 2. The average performance of all the features when $n = 50$ and $n \neq 50$

# P- labels	Flickr	Feat.1	Feat.2	Feat.3
$\neq 50$	50.87	61.18	56.62	55.50
$= 50$	50.87	62.18	56.62	55.50

6 Conclusions

We have presented a framework to support the entire process of image and textual retrieval from search engines that, given an input keyword, performs a statistical and a semantic analysis and automatically builds a training set. We have conducted several experiments to validate our assumptions about the analysis of textual information and the evaluation that we have provided on three investigated methods have shown that the position of tags, their order, is relevant. We have investigated the semantic aspects by using semantic distance. Unfortunately, the results achieved show modest benefit for the adopted semantic features. However, the methods suggested are currently under continuous experimentation and need to bee further investigated. In particular, we consider for future work to explore the use of different search engines such as Google[5], ImageNet[6], InstaGram[7] or Pinterest [8] to check if they are interchangeable or can be combined to improve performances. We plan also to extend and investigate other semantic features related to ontological relationships of textual information and combine them with the aim of creating a waterfall model which combines different strategies.

Acknowledgements. This research was supported by the VISCOSO project financed by the Autonomous Province of Trento through the "Team 2011" funding programme.

References

1. Helmer, S., Meger, D., Viswanathan, P., McCann, S., Dockrey, M., Fazli, P., Southey, T., Muja, M., Joya, M., Jim, L., Lowe, D.G., Mackworth, A.K.: Semantic

[5] http://www.google.com
[6] http://www.image-net.org
[7] http://instagram.com/
[8] http://www.pinterest.com

robot vision challenge: current state and future directions. In: IJCAI workshop (2009)
2. Cheng, D.S., Setti, F., Zeni, N., Ferrario, R., Cristani, M.: Semantically-driven automatic creation of training sets for object recognition. Computer Vision and Image Understanding 131, 56–71 (2014)
3. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. **2**(1), 1–19 (2006)
4. Liu, Y., Xu, D., Tsang, I.W., Luo, J.: Textual query of personal photos facilitated by large-scale web data. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 1022–1036 (2011)
5. Heymann, P., Paepcke, A., Garcia-Molina, H.: Tagging human knowledge. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining. WSDM 2010, pp. 51–60. ACM, New York (2010)
6. Lin, Z., Ding, G., Hu, M., Wang, J., Ye, X.: Image tag completion via image-specific and tag-specific linear sparse reconstructions. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1618–1625, June 2013
7. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 716–727 (2013)
8. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 2, pp. 524–531 (2005)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding **106**(1), 59–70 (2007)
10. Li, L.J., Li, F.F.: Optimol: Automatic online picture collection via incremental model learning. International Journal of Computer Vision **88**(2), 147–168 (2010)
11. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005. vol. 2, pp. 1816–1823 (2005)
12. Gilbert, A., Bowden, R.: A picture is worth a thousand tags: automatic web based image tag expansion. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part II. LNCS, vol. 7725, pp. 447–460. Springer, Heidelberg (2013)
13. Dubinko, M., Kumar, R., Magnani, J., Novak, J., Raghavan, P., Tomkins, A.: Visualizing tags over time. In: Proceedings of the 15th International Conference on World Wide Web. WWW 2006, pp. 193–202. ACM, New York (2006)
14. Ahern, S., Naaman, M., Nair, R., Yang, J.: World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In. In Proceedings of the Seventh ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 1–10. ACM Press (2007)
15. Spain, M., Perona, P.: Some objects are more equal than others: measuring and predicting importance. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 523–536. Springer, Heidelberg (2008)
16. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: CHI 2007: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 971–980. ACM Press, New York (2007)
17. Wu, L., Hua, X.S., Yu, N., Ma, W.Y., Li, S.: Flickr distance: A relationship measure for visual concepts. Pattern Analysis and Machine Intelligence, IEEE Transactions on **34**(5), 863–875 (2012)

18. Hwang, S.J., Grauman, K.: Learning the relative importance of objects from tagged images for retrieval and cross-modal search. Int. J. Comput. Vision **100**(2), 134–153 (2012)
19. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8 (June 2008)
20. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. CoRR cmp-lg/9709008 (1997)
21. Fellbaum, C.: WordNet: An Electronic Lexical Database. Language, Speech and Communication. Mit Press (1998)

Characterizing Predicate Arity and Spatial Structure for Inductive Learning of Game Rules

Debidatta Dwibedi(✉) and Amitabha Mukerjee

Indian Institute of Technology, Kanpur, India
debidattadwibedi@gmail.com, amit@iitk.ac.in

Abstract. Where do the predicates in a game ontology come from? We use RGBD vision to learn a) the spatial structure of a board, and b) the number of parameters in a move or transition. These are used to define state-transition predicates for a logical description of each game state. Given a set of videos for a game, we use an improved 3D multi-object tracking to obtain the positions of each piece in games such as 4-peg solitaire or Towers of Hanoi. The spatial positions occupied by pieces over the entire game is clustered, revealing the structure of the board. Each frame is represented as a Semantic Graph with edges encoding spatial relations between pieces. Changes in the graphs between game states reveal the structure of a "move". Knowledge from spatial structure and semantic graphs is mapped to FOL descriptions of the moves and used in an Inductive Logic framework to infer the valid moves and other rules of the game. Discovered predicate structures and induced rules are demonstrated for several games with varying board layouts and move structures.

Keywords: Predicate discovery · Spatial structure discovery · Game rule learning · Semantic graphs · Multi-object tracking · Vision-based ontology discovery · Inductive logic programming · Kinect

1 Introduction

Any formal system is built on a base vocabulary of predicates, functions and constants. These predicates may show much variability while representing the same linguistic terms. In modeling games with moving pieces, predicates such as `move()` or `adjacent()` may vary in argument patterns and semantics owing to differences between games. Thus, in Tic-tac-toe, a `move` involves adding a piece, whereas in Towers of Hanoi or Kalaha, many pieces may be moved at once. Thus, the arity of `move()` varies across games. Similarly, adjacency relations will change depending on the board layout (1-D, 2-D, mixed-vertical, triangle vs grid, etc.). In order for an ontology to be induced for such games, it is crucial that one start with the right predicates. In addition the range of constant values that a variable can take (e.g. the set of valid positions) has to be specified. In this paper, we look at single-person games involving pieces that move, and we

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 323–338, 2015.
DOI: 10.1007/978-3-319-16181-5_23

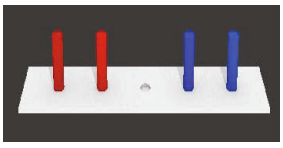
ask if instead of introducing such knowledge implicitly in the background, can we discover such structures by visually observing the game play?

Inductive Logic Programming and allied methods have shown immense advantages in learning domain theories for a wide class of problems [6,18], but the approach is still restricted by an inability to discover a suitable set of predicates, which require grounding in sensorimotor data. Formal systems with polymorphism permit functions with varying arity, but these cannot be handled efficiently is inductive logic situations. Thus, the background input for inductive logic programming invariably involves predicates with fixed arities.

When a child is shown a game of Tic-tac-toe, that each move involves adding a single piece is immediately obvious, whereas in Towers of Hanoi, it is clear that a move may involve several pieces. Similarly, one glance at a chess board tells a learner that it has 8×8 squares, and that the position of any piece can take a value only from these 64 possibilities. This suggests that some aspects of the vocabulary used in the background theory may be inferred by the learner - as opposed to being programmed - thus providing greater flexibility for inducing the domain theory.

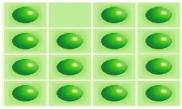
Here, we build on recent work in semantic graph discovery from RGB-D (depth data) images to learn structures of interactions between objects [2,25] to explore the possibility of learning some aspects of predicate structures in games involving moving pieces. Specifically, we attempt to discover a) the arity and structure of base predicates such as `move()`, and b) the underlying spatial structure that provides the set of constants that define admissible values for some fluents like position. In the process, we also construct visual semantic interpreters and generators for these predicates, in terms of the visual routines which result in a discovered cluster.

The approach is demonstrated in three one-person games (or puzzles) involving spatial reconfiguration of pieces : Jumping frogs (1-D); Towers of Hanoi (1-D with vertical) and 4×4 Peg Solitaire (2-D)(Fig. 1). Both Jumping frogs and Peg solitaire have been modeled in simulation using the BlenSor RGBD simulation system[11]; Towers of Hanoi has been tested both on real and simualted data. The datasets and code used is being made available at http://www.cse.iitk.ac.in/users/vision/debidatt/



(a) Jumping frogs puzzle

(b) Towers of Hanoi



(c) 4×4 Peg Solitaire

Fig. 1. Examples of Spatial Reconfiguration games handled. Board spatial layout and predicate structures such as number of pieces involved in moves are inferred from the visual structure. ILP then is able to infer aspects such as that higher disks must be smaller in Towers of Hanoi.

2 Related Work

Inductive logic programming (ILP) attempts to hypothesize the simplest hypothesis explaining a set of (mostly) positive examples using background knowledge [6,18]. More formally, given a set of observed examples E_i (propositions), and the categories c_i they belong to, ILP attempts to find the simplest model H (a first-order-logic theory) s.t. for all training pairs $\langle E_i, c_i \rangle, H \wedge Ei \wedge B \models c_i$, while $\forall c' \neq ci, H \wedge Ei \wedge B \not\models c_i$.

ILP approaches have been used in learning the rules for boardgames like Tic-Tac-Toe and Hexapawn [3], dice-based games [23] and card games [12,16]. In each of these, the backghround knowledge already covers concepts like board representation, adjacency / linearity tests, frame axioms, turns and opponents, piece ownership and spatial predicates. Our objective is to start a bit further back, and try to discover the structure for some these predicates.

However, hypotheses discovered by ILP (Progol) are restricted to essentially single clause hypotheses in the refutation chain, and multi-clause induction is highly inefficient [19,24]. One approach to multi-clause induction is to prioritize the ordering of rules using a set of meta theoretic rules ("top theory") that enables multi-clause refutations [19]. This has been used in learning grammars and also a strategy for the Nim game. Other attempts to extend the paradigm include interleaving induction with abduction models to generate more compact models for modeling event structures [9]. Systems attempting to learn game strategy are better served by using models related to learning planns, which often use a PDDL structure [10]. However, our objective here is at the vision-logic interface, and not in the domain of logic per se, hence we restrict ourselves to Progol for our testing.

2.1 Inducing Domain Theories for Games from Vision

Inducing rules of games using vision as input has been attracting increasing attention in recent years [3,12,13], since they provide a key test for other generalizations that may be possible for real-world problems. In Barbu etal [3], the learned rules are used impressively by a robot to manipulate the pieces onto a wooden frame to actually play the game. They use ILP (Progol) to learn valid moves of the game pieces and winning conditions in six games. The approach proposed by Kaiser [13] is also inductive, requiring a few visual demonstrations to learn rules for games such as Connect4 or Gomoku.

However, these systems needs to be provided with the predicate structure implicitly via background knowledge. Thus [3,4,13] all assume a 2-D grid of known size, and pre-define the set of possible moves and adjacency relations of interest. The priors embedded in the background knowledge thus restrict the generality of such systems. Also, the visual classifiers associated with each predicate are hard-coded and game specific. We show that as part of ths semantic-graph analysis, these visual routines, (and hence the argument structure) can be discovered for predicates like `move()`.

2.2 Representing Scenes with Semantic Graphs

In a series of recent papers, Aksoy and co-workers [1,2] have mapped videos to dynamic graphs with nodes representing objects and edges encoding semantic relations such as contact. Related ideas for learning semantic relations by tracking objects can be found in the semantic segmentation of scenes[7], affordance modeling of objects[15] and manipulation planning[5]. Semantic graphs can model manipulation actions[2][25] in terms of primitives like merging and dividing and used to classify higher-order actions like making a sandwich, cutting a cucumber, pouring liquids, etc. When a piece is moved in a game, manipulations are relatively simpler, since the piece does not deform or merge into others.

A key requirement for our work is that objects must be tracked reliably across visual frames. As in [25], we propose to use Kinect-based RGBD image inputs for the tracking. Contact between pieces is important in some games (e.g. Towers of Hanoi), and this is determined by analyzing four types of relationships between each pair: *touching, overlapping, non-touching and absent.* A matrix encoding all possible relation pairs is created and this is compressed to represent only the change in relation pairs. The dynamic changes in graphs caused by manipulation actions are compared by converting these relations into strings. Thus one may define spatial and temporal similarity measures between different actions, and cluster such actions, resulting in a template for game actions such as `move()`. Other candidates for edges in semantic graphs may be obtained by tracking the hand in 3D videos [20].

In the attempt presented here, part of the structure is being learned via the semantic graph in terms of contact and neighbourhood relations, and this is used to identify the type of primitive predicates that would be used to describe the system. These predicates are added to a sparse human-defined ontology of background knowledge in order to learn rules for games and puzzles from the RGBD videos.

We modify the semantic graph for situations specific to rigid piece motions as in games. We are given a set of game videos as input, but are not told about the spatial structure - whether it is being played on a grid or a line or a triangle or other spatial layout. We also do not know the number of pieces involved in each state-transition and their specific behaviours. In the next section, we see how we do this starting with RGBD videos which enable improved 3D tracking since camera-based depth data is available. For example. clustering all the 3D positions of the pieces enable us to obtain the "cells" that a piece can occupy. Grid layouts are identified using Principal Component Analysis; if the layout is aligned to the dominant eigenvectors, it is a grid. Next, we identify if there is direct contact (as in Towers of Hanoi), if so, contact is used as the edge relation in our semantic graph. Else, we use adjacency relations defined on the board discovered. This initial analysis also tells us the number of changes that occur on different types of moves, and how these can be captured in terms of a "move" or a "transition" predicate.

In our work we analyze the RGBD video of a game. If there are contact situations, we consider contact as a primitive for the Semantic graph analysis; else we

use neighbourhoods on the discovered spatial structure. These relationships are mapped to FOL predicates which are then used in an ILP framework to induce rules for the game.

3 Semantic Graphs of Game Scenes from RGBD Video

In order to generate semantic graphs from images or point clouds, the first task is to robustly segment and track each piece. Challenges include occlusion by the hand or by other objects and altered appearance. Other changes come about due to division or merging (e.g. a tower may be a single merged object in Towers of Hanoi). The above problem is simpler in games because pieces are usually rigid. However, many games have pieces that are identical in colour and shape, throwing up other challenges.

3.1 Game Piece Segmentation

With 3D data, object segmentation can be performed to cluster points close to each together based on Euclidean distance[22]. Algorithm 1 is a modified version where we perform filtering based on the colour in the HSV space before the clusters of points are discovered in the scene by doing Euclidean clustering based on distance. This is done because sometimes game pieces of different colours might be placed on top on another or in contact with each other like in the Towers of Hanoi. So our objective is to extract clusters of points as game pieces. These clusters should either have perceptually different colours or be separated above a particular threshold in space as shown in Fig. 2.

Algorithm 1. Pipeline to extract objects from scene

1. Use a Pass Through filter to focus on the table-top.
2. Use RANSAC to filter out points of the table-top from the cloud.
3. Perform Colour-based filtering of the point cloud in HSV space.
4. Do euclidean clustering of the different colour clouds to give objects that are either separated in space or have perceptually different colours.

Fig. 2. Game pieces found in a scene from the real Towers of Hanoi dataset

3.2 Multi-object Tracking

In the multi-object tracking problem, a label associated with an object needs to be linked with the same object in the next frame and this needs to be done with all objects present in the scene. The problem is challenging owing to all pieces being identical in many games, and further complicated due to occlusion by the player's hand or by other pieces. A model-based detection method cannot be used here since many objects have the same shape and colour.

Aksoy et al.[2], extracted segments from the images using super-paramagnetic clustering in a spin-lattice model[8]. Doing this allowed them to perform robust markerless tracking of the segments. A number of other tracking algorithms[14, 25] attempt to handle objects that may break up (cutting with a knife) or join together (pouring from one glass to another), etc. Since game pieces are usually rigid our tracker can make the assumption that pieces do not break up or merge significantly.

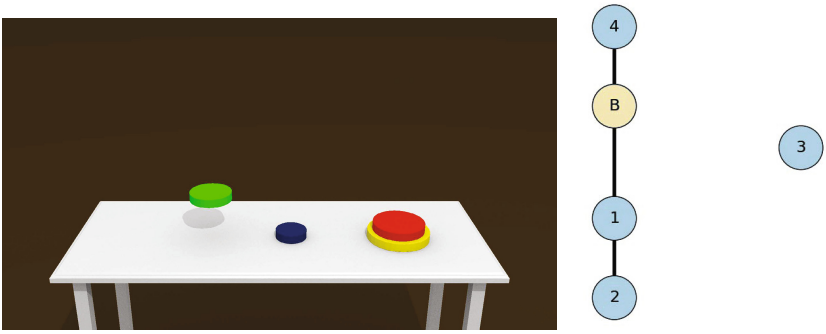
Our proposed method for tracking multiple-objects in a point cloud video is based on the occupancy of voxels by an object in one frame and the next. Multiple object tracking can be reduced to an assignment problem where the objects detected in frame i need to matched with themselves in frame $i + 1$.

The assignment problem is a combinatorial optimization problem. It consists of finding a maximum weight or minimum cost matching in a weighted bipartite graph. In other words, there are two sets $A = \{a_1, a_2, .., a_n\}$ and $B = \{b_1, b_2, ..., b_n\}$. There is a certain cost for matching a a_i with a b_j. The assignment problem is to match each members of set A one member of set B such that the total cost of the assignments is minimized. The Hungarian method is used to solve the label assignment problem in polynomial time.

Using Euclidean distance between the centroids [5] may fail if there are multiple objects moving simultaneously. We use the octree overlap between point clouds that is the amount of overlap between axis-oriented bounding boxes of the objects. The hierarchical octree [17] method reduces complexity by downsampling the point cloud. We build the octree representation of the objects found by segmentation in two consecutive frames. If it moves, there is going to be a spatial overlap between the same object in the two consecutive frames. This overlap will be zero with the other objects present in the scene. We use this overlap in space to track objects by maximizing the sum of all overlaps while assigning labels from one frame to the next. There are two assumptions that make this tracking algorithm work. Our objects of interest are non-planar and rigid. Planar objects may have zero overlap with themselves in the next frame. The action performed by the player is slow enough for the Kinect to record the movement of the objects. If the frame-rate of recording the point clouds is slow there will be no overlap. In our case, however, we recorded gameplay at the usual pace a person plays and there was considerable overlap between the same objects in consecutive frames at normal Kinect recording rates. We also suggest the use of a Kalman filter to improve tracking under full occlusion.

3.3 Semantic Graphs

A semantic graph of the scene encodes the relationships between the objects. Building semantic graphs depends on choosing some primitive relations for the edges, and this often depends on the task one is looking at. An intuitive primitive is to consider contact, e.g. Yang et al.[25], but sometimes an object like a bar, may be privileged [5]. In our situation, the table-top is a special object whose contacts are not listed as predicates. Aksoy et al.[2] encode proximity relationships even if they are not in contact. They also encoded the semantic relationship *overlapping* which meant one segment is included in another.



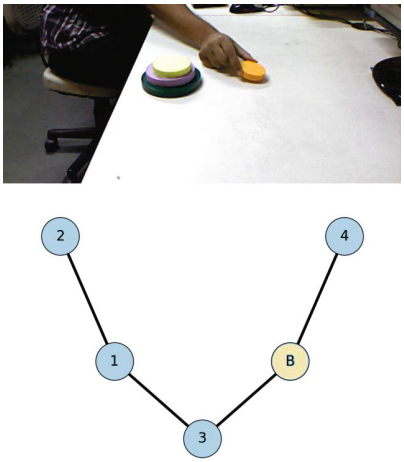
(a) Frame 93 from *Towers of Hanoi* (simulation) dataset (b) Semantic Graph

Fig. 3. Example Semantic Graph

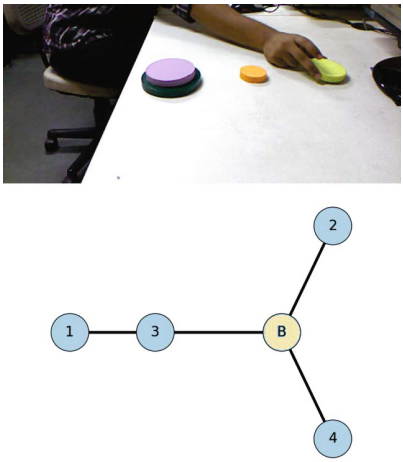
In most board games or puzzles the game state is altered by picking up a piece and placing it somewhere else on the board, but sometimes an intermediate piece or the piece at the target square, if of an opposing colour, may be removed. In games such as the Towers of Hanoi, vertical contact occurs frequently, and this needs to be represented.

In Fig. 3, there are four pieces from largest piece (1, yellow) to smallest (4, blue) with red (2) and green (3) in between. The board is labeled B. Edges reflect contact between pieces. Thus, the graph shows that a stack of 1,2 is on the board, as well as 4, but the green piece (3) is not in contact with anything. Changes in this semantic graph - e.g. 3 being placed on top of 2 - will represent a move action.

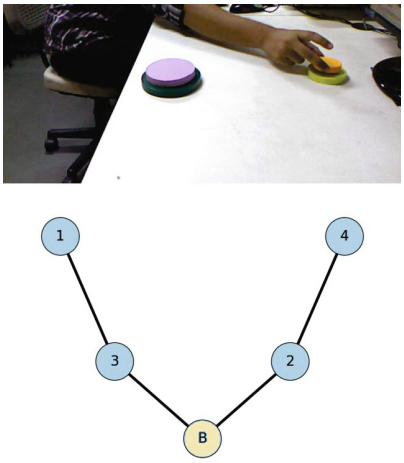
We can now discover the states of the game by looking for configuration changes of the game pieces on the board. Every time a player lifts up a piece, an edge is broken. The moment the player places the piece back on the board or on another piece, a new edge is formed. Hence, game states can easily be discovered from the video by looking for states where the number of edges changes. Each node in the graph also stores meta-information such as the coordinates of its centroid, average colour of the object, number of visible points and the volume occupied by the bounding box of the object in the current frame. After the states



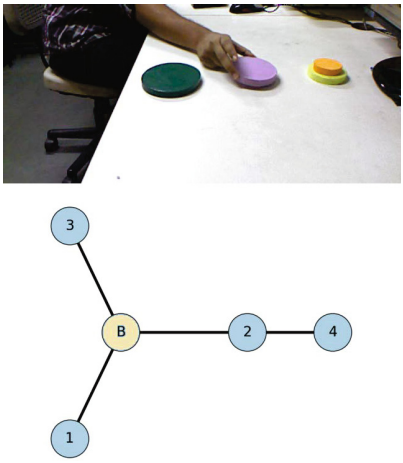
(a) Frame 4 with Semantic Graph



(b) Frame 83 with Semantic Graph



(c) Frame 173 with Semantic Graph



(d) Frame 251 with Semantic Graph

Fig. 4. Automatic detection of game states in the *Towers of Hanoi* Real dataset. Blocks and their labels in the graph:(1,purple),(2,yellow),(3,green),(4,orange). For example, comparing graphs (a) and (b), we find that the move consisted in taking the piece 2 from the stack 3,1,2 to the board.

have been detected, the change from one state to another can be found out by looking for changes in the meta-information. In Fig. 4, some game states from the *Towers of Hanoi* dataset, that were discovered automatically, are shown.

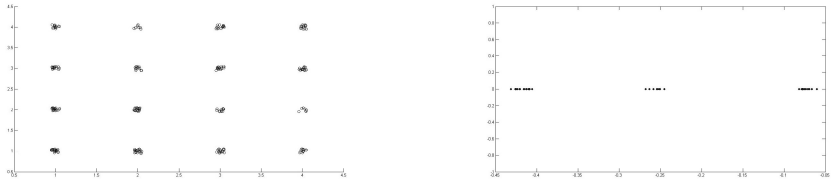
We observe that discovering game states is not a trivial problem. For example in the 4×4 peg solitaire, after a piece has been moved, the intermediate, jumped-over piece is removed. Here the system needs to be told that the intermediate stage does

not constitute a "game state". This could also be learned via a heuristic looking at pauses in the game, but as of now, this has not been implemented.

4 Learning Spatial States

Many logical systems start with an implicit assumption about the board on which the game is being played. But this need not be the case. A human observing a game immediately notes the type of board on which the game is being played. Thus, a game such as a 4×4 peg solitaire will have a 2-D structure in the horizontal plane, whereas the Towers of Hanoi has essentially a 1-D structure with vertical contacts. The distribution of spatial locations of the pieces during an entire game can be used to infer the game board, using the following steps:

1. **Discover intrinsic dimensionality of the game**: The system does not have any idea in the beginning whether the game is 1D or 2D or 3D. After it has discovered the game states by using the methods described in the previous section, it populates a list of the positions of all the game pieces across all the game-state frames. These are data points where game pieces have visited during the game play. By performing Singular Value Decomposition(SVD) on these coordinates the intrinsic dimensionality of the game is known. One-dimensional games have only one significant eigenvalue.
2. **Transform from camera coordinates to board coordinates**: X_b, Y_b, Z_b are coordinates of the object in the frame of the board which will be used to find the clusters. These co-ordinates are obtained by transforming the camera coordinates X_c, Y_c, Z_c by using the cosines of the angles between the axes. $\hat{x}_b, \hat{y}_b$ and $\hat{z}_b$ represent the unit vectors of the axes in the frame of the board. $\hat{z}_b$ is obtained as the average of normals of the points on the board. $\hat{x}_b$ and $\hat{y}_b$ are obtained by SVD mentioned above. The eigenvector corresponding to the largest eigenvalue gives $\hat{x}_b$ if it doesn't coincide with $\hat{z}_b$. Similarly, In 2D games the second significant eigenvector gives $\hat{y}_b$. This can also be found as a cross product of $\hat{z}_b$ and $\hat{x}_b$. The above generalizations don't hold true when the game being played doesn't conform to an usual rectangular grid like triangular peg solitaire.
3. **Discover discrete valid positions of game pieces**: The next step is to look for clusters in the positions occupied by game pieces in the game states. While finding out the optimal number of clusters is an open problem, there are statistical methods to estimate the optimum number of clusters in a dataset like ours. One method will be to look for an elbow or a bend in the sum of squared error(SSE) plot. The locations of the clusters are discovered by performing k-means clustering using the value of k found by using the elbow method. In Fig. 5(a) and Fig. 5(b) there are sixteen clusters and three clusters respectively. Fig. 6 shows the elbow method being used to determine the number of clusters in corresponding to the four holes in one dimension in $4 \times 4 PegSolitaire$.



(a) 16 clusters in *4×4 Peg Solitaire* (b) Three clusters in *Towers of Hanoi*

Fig. 5. Clusters formed in the significant dimension

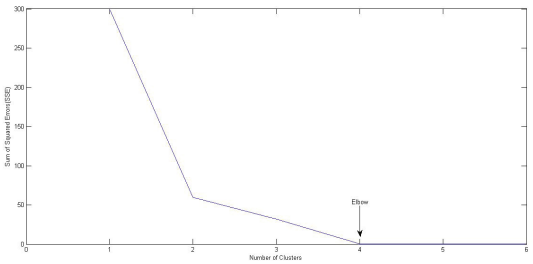


Fig. 6. Elbow method to find number of clusters in one axis of *4×4 Peg Solitaire*

4. **Represent game state**: For each game state, each game piece is assigned to its nearest cluster. Doing so, allows us to generate a general representation of game states of any game. This might leave us with a cluster that is unoccupied which can be represented as empty. We transfer these states to a logic programming system which will be a better domain to induce the rules of games. The first game state(Fig. 1(a)) in Four Frogs will be `[{a},{b},{},{c},{d}]` where a, b, c and d are the labels given to the game pieces. The third hole is unoccupied in the beginning which is represented by the empty set. In Towers of Hanoi, the state shown in Fig. 1(b) will be represented by `[{a,b,c},{d},{}]`. This representation is there to handle games where pieces can be placed one on top another occupying the same discrete cluster on the board. This can be extended to 2D games where a matrix of characters will represent the game state.

4.1 From Semantic Graphs to Horn Clauses

We use meta-information contained in the nodes of the graphs and changes in that from one game state to the next to generate logical clauses that will help us learn the rules. We generate the background knowledge and positive examples (instances seen in video) to come up with hypotheses regarding the rules of the game.

The ontology used to represent games and involves three kind of predicates:

1. *Attributes of game pieces* derived from visual classifiers like size, colour, shape, starting position etc.
2. *Relationships between game pieces* generated from the edges of the semantic graphs like *on*, *contact* etc.
3. *Movement of game pieces* generated from changes in game states and semantic graphs (*move*, *transition*, etc.).

Background Knowledge: We assume that game pieces are objects that will need to be monitored. Attributes of the game pieces like color, shape and size may constrain the possible moves it can make. First, we need to identify the number of pieces. Thus, a 4-piece Towers of Hanoi, may have the following initial declaration: `piece(a). piece(b). piece(c). piece(d).`

In 1-D games, location is described with one variable and in 2-D with two. In the Towers of Hanoi, 3 clusters are discovered on the primary eigenvector. Each cluster is also associated with a number which helps in comparing their position with other clusters. They are declared as follows: `x(l1). x(l2). x(l3). project(l1,1). project(l2,2). project(l3,3).` A set of colours are predefined and associated with a HSV classifier. These are used to declare a colour for each game piece:

```
colour(a,red). colour(b,green). colour(c,yellow). colour(d,blue).
```

Numerical features like size is obtained as the largest dimension of the bounding box of the game piece, rounded off to an integer scale:

```
size(a,1). size(b,3). size(c,9). size(d,10).
```

We do not use shape classifiers in the present analysis since in the games we consider all objects have the same shape. For each numerical feature there is a meta-clause generator that compares their values. For example the clause generated for size is shown below:

```
greatersize(A,B) :- piece(A),piece(B),size(A,NA),size(B,NB),NA>NB.
```

The function *diff* gives us the number of steps a game piece has been moved and in what direction (positive is along the default axis). *absDiff* ignores the direction. In the 4×4 peg-solitaire *diff* and *absDiff* operate on each dimension separately. In the towers of hanoi we also use predicates for *top* and *bottom* in a stack.

```
diff(X1,X2,Diff):- x(X1),x(X2),project(X1,N1),project(X2,N2),
Diff is N1-N2.
abs(X,X) :- X>=0.abs(X,Y) :- X<0, Y is -X.
absDiff(X1,X2,Diff) :- x(X1),x(X2),project(X1,N1),project(X2,N2),
Diff1 is X1-X2, abs(Diff1,Diff).
neighbour(X1,X2) :- absDiff(X1,X2,1).
top(A,[A]).top(A,[B|C]) :- top(A,C).
bottom(A,[A]).bottom(A,[B|C]) :- bottom(A,B).
```

Note that for 2D games, the *diff* is modified *xDiff* and *yDiff* and similarly for *absDiff*.

Given a set of observations we can obtain **Positive examples** of board play. A critical inference has to do with valid **Moves of game pieces**. A move results in a transition from one spatial graph to another, which includes a piece

move along with possible side effects (e.g. removal of the intermediate piece in 4×4 peg solitaire). The relationship *transition* encodes the active piece and the states of clusters that undergo change from one game state to the next. It has the following structure:
`transition(<active pieces>,<initial states>,<final states)>.`
The predicate shown below is from the Towers of Hanoi game and represents a piece d being moved where the set of game pieces at the initial position $l1$ was `[a,b,c,d]` and that at final position $l2$ after the move was `[d]`:
`transition(d,[a,b,c,d],[],[a,b,c],[d]).`
The arity of the transition predicate varies from game to game. In the 4×4 Peg Solitaire, the number of pieces involved in a move are two and the number of positions where there is change from one game state to the next is three. Hence, the *transition* relation example for the move where piece $p1$ in position $l1$ jumps over piece $p2$ in $l2$ to land in $l3$ following which $p2$ is removed looks like this:
`transition(p1,p2,[p1],[p2],[],[],[],[p1]).`

Table 1. Games learnt with their respective modes of data generation

Game	Nature of Dataset
Towers of Hanoi	Animated(generated in Blensor), Real(recorded with a Kinect)
Four Frogs	Animated(generated in Blensor)
4 × 4 Peg Solitaire	Game traces of a simulation

5 Experiments and Results

5.1 Towers of Hanoi

In addition to one real game played, we used the RGBD simulator BlenSor[11] to animate four differently sized blocks with *Towers of Hanoi* puzzle being solved. There are 740 frames of 640 × 480 RGBD images recorded on an artificial Kinect sensor in BlenSor. The real Kinect data with the Towers of Hanoi being by a person has 1200 frames. The ILP system input includes the following:

```
colour(a,yellow).colour(b,red).colour(c,green).colour(d,blue).
size(a,10).size(b,8).size(c,4).size(d,2).
on(d,a).on(d,b).on(d,c).on(c,b).on(c,a).on(b,a).
from(d,[a,b,c,d],[d]).from(c,[a,b,c],[c]).from(d,[d],[c,d]).
```

The rules learnt by PROGOL are:

```
on(A,B) :- greatersize(B,A).
transition(A,B,C,D,E) :- top(A,C), top(A,E).
```

The first rule translates as "No disk may be placed on top of a smaller disk." The second rule says that piece A moves from the top of the stack C and to the top of stack E. The system sees that all six on() relations have occurred over the game, and no negations are given, so it infers that on(A,B) must hold

whenever greatersize(B,A) which is not true but in inferred as time is not taken into account.

5.2 Jumping Frogs puzzle

The animated dataset consists of 560 frames of 640 × 480 RGBD images. There are five cylindrical holes in a row, two red pegs (which can only move right) and two blue pegs (only move left)(Fig. 1(a)). Initially, the red pegs are placed in the two left holes and the blue pegs are placed in the two right holes leaving a hole in between that is empty. The goal of the game is to swap the positions of the red pegs with the blue pegs. PROGOL generalizes the clause *move* and comes up with four rules:

```
move(A,B,C) :- diff(B,C,-2), colour(A,blue).
move(A,B,C) :- diff(B,C,-1), colour(A,blue).
move(A,B,C) :- diff(B,C,1), colour(A,red).
move(A,B,C) :- diff(B,C,2), colour(A,red).
```

We learn that if there is an object that moves right its colour must be red and if there is one which moves left then its colour must be blue. More interestingly, the system discovers that there are two types of moves a piece is able to do that is one step and one jump which implies moving two steps at the same time.

The colours of the pegs were then interchanged. The rules learnt by appending the newer clauses with the older ones are:

```
move(A,B,C) :- diff(B,C,-2), startpos(A,l1).
move(A,B,C) :- diff(B,C,-2), startpos(A,l2).
move(A,B,C) :- diff(B,C,-1), startpos(A,l1).
move(A,B,C) :- diff(B,C,-1), startpos(A,l2).
move(A,B,C) :- diff(B,C,1), startpos(A,l4).
move(A,B,C) :- diff(B,C,1), startpos(A,l5).
move(A,B,C) :- diff(B,C,2), startpos(A,l4).
move(A,B,C) :- diff(B,C,2), startpos(A,l5).
```

Thus the colour dependence is replaced by a clause for the row where the pieces start from. This highlights the fact how the rules learnt by induction learning can undergo radical changes depending on the dataset.

5.3 4 × 4 Peg Solitaire

In the beginning, of this game there are 15 marbles arranged in form of a 4 × 4 grid with one position empty(Fig. 1(c). The marbles can only move by jumping to an empty position and by doing so the piece over which they jumped is removed. The objective is to remove as many pieces as one can, preferably reaching a single piece. We use game traces of a simulation of this game being solved to test how good our system is in inducing the rules in case it has perfect information regarding the game states. The rules learnt by ILP are:

```
move(A,B,C):- xabsdiff(B,C,2). move(A,B,C):- yabsdiff(B,C,2).
transition(A,B,C,D,E,E,E,C):-piece(A),piece(B),top(A,C),
```

```
bottom(A,C),top(B,D),bottom(B,D),empty(E).
```

The two move rules have learned that the moves take place either horizontal or vertical rows of three neighbouring cells. In the transition predicate, the arguments are the pieces involved (here A,B), and the remaining 3+3 arguments are the pieces at the three locations involved, before and after the move. Thus the learned rule says that the state of loc1 and loc2 changes to E, which was the initial state of loc3. The piece at loc3 becomes C which was initially at loc1 (i.e. the piece A is moved to loc3). Thus, the rule infers that A moves from loc1 to loc 3, and that the piece B is removed from the jumped-over position loc2. The three locations are arranged in a horizontal or vertical row of the board.

5.4 Discussion

We observe that in all three cases, the spatial structure can be inferred at the visual level, permitting a set of constants which the position attributes in move() etc can be assigned to. Also the number of pieces and positions affected by move are identified in the vision system. When the resulting game states and transitions are introduced into the ILP system, we find that it is able to derive the right rules, such as identifying that in ToH, the higher disks must be smaller, or that in peg solitaire, adjacency relations (for move) are only row or column-wise. Similarly, in the peg solitaire, the fact that the jumped-over piece (also an argument to move) is removed, is inferred.

6 Conclusion

One of the major challenges in inducing knowledge representations involves discovering the right set of logical primitives to be used. Here we have presented a framework that is able to analyze RGBD videos of game scenes using dynamic semantic graphs, which permit generation of suitable Horn Clause structures. The system uses an improved tracking based on the assumption that game pieces do not change shape or visual attributes (like colour or shape). We then demonstrate its application in learning the rules of game and puzzles. The system can successfully induce the spatial description of boards for 1-D and 2-D games, and also induce vertical contact situations and their ramifications for an otherwise 1-D game such as Towers of Hanoi. The arity of predicates such as "move" varies in these games and is captured via the pre-processing in the Semantic Graph step.

As of now, we have demonstrated this for only three simple games. A number of loose ends remain in the present implementation. As of now, the end states of a game are not being discovered, hence we are not able to generate a Game Description Language(GDL) which will enable the system to start playing these games. In most real situations, the learner often needs to be told about the start and end configurations along with whether it was a winning or losing game, etc. Our system can be enhanced with this start and goal state knowledge to generate the suitable GDL for automatic game playing. Further, the system cannot handle

multi-player games, which require event calculus representations. However, our main focus has been to demonstrate the idea of obtaining descriptors with the correct number of arguments, which would apply equally to event calculus or other planning formalisms.

Also, for any system using vision, improvements are always possible in tracking. Recent research[14][21] on multi-object tracking has shown encouraging results which may be helpful in tracking for games with more game pieces.

However, the main contribution of this work is at the level of the implicit knowledge used in defining logical descriptors. This is a challenging problem for knowledege representation in general that has not been adequately investigated, and this work takes some initial steps in developing vision-based approaches towards discovering this implicit structure.

References

1. Aein, M.J., Aksoy, E.E., Tamosiunaite, M., Papon, J., Ude, A., Worgotter, F.: Toward a library of manipulation actions based on semantic object-action relations. In: IROS-2013, pp. 4555–4562 (2013)
2. Aksoy, E.E., Abramov, A., Dörr, J., Ning, K., Dellen, B., Wörgötter, F.: Learning the semantics of object-action relations by observation. The International Journal of Robotics Research **30**(10), 1229–1249 (2011)
3. Barbu, A., Narayanaswamy, S., Siskind, J.M.: Learning physically-instantiated game play through visual observation. In: 2010 IEEE International Conference on Robotics and Automation (ICRA), pp. 1879–1886. IEEE (2010)
4. Björnsson, Y.: Learning rules of simplified boardgames by observing. In: ECAI, pp. 175–180 (2012)
5. Dantam, N., Essa, I., Stilman, M.: Linguistic transfer of human assembly tasks to robots. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 237–242. IEEE (2012)
6. De Raedt, L.: Inductive logic programming. In: Encyclopedia of machine learning, pp. 529–537. Springer (2010)
7. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 284–298. Springer, Heidelberg (2012)
8. Dellen, B., Erdal Aksoy, E., Wörgötter, F.: Segment tracking via a spatiotemporal linking process including feedback stabilization in an nd lattice model. Sensors 9(11), 9355–9379 (2009)
9. Dubba, K., Bhatt, M., Dylla, F., Hogg, D.C., Cohn, A.G.: Interleaved inductive-abductive reasoning for learning complex event models. In: Muggleton, S.H., Tamaddoni-Nezhad, A., Lisi, F.A. (eds.) ILP 2011. LNCS, vol. 7207, pp. 113–129. Springer, Heidelberg (2012)
10. Edelkamp, S., Kissmann, P.: Symbolic exploration for general game playing in pddl. In: ICAPS-Workshop on Planning in Games. vol. 141, p. 144 (2007)
11. Gschwandtner, M., Kwitt, R., Uhl, A., Pree, W.: BlenSor: blender sensor simulation toolbox. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Wang, S., Kyungnam, K., Benes, B., Moreland, K., Borst, C., DiVerdi, S., Yi-Jen, C., Ming, J. (eds.) ISVC 2011, Part II. LNCS, vol. 6939, pp. 199–208. Springer, Heidelberg (2011)

12. Hazarika, S.M., Bhowmick, A.: Learning rules of a card game from video. Artificial Intelligence Review **38**(1), 55–65 (2012)
13. Kaiser, Ł.: Learning games from videos guided by descriptive complexity. In: Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
14. Koo, S., Lee, D., Kwon, D.S.: Multiple object tracking using an rgb-d camera by hierarchical spatiotemporal data association. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1113–1118. IEEE (2013)
15. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International Journal of Robotics Research **32**(8), 951–970 (2013)
16. Magee, D., Needham, C., Santos, P., Cohn, A., Hogg, D.: Autonomous learning for a cognitive agent using continuous models and inductive logic programming from audio-visual input. In: Proceedings of the AAAI workshop on Anchoring Symbols to Sensor Data, pp. 17–24 (2004)
17. Meagher, D.: Geometric modeling using octree encoding. Computer graphics and image processing **19**(2), 129–147 (1982)
18. Muggleton, S.: Inverse entailment and progol. New generation computing **13**(3–4), 245–286 (1995)
19. Muggleton, S.H., Lin, D., Tamaddoni-Nezhad, A.: MC-TopLog: Complete Multi-clause Learning Guided by a Top Theory. In: Muggleton, S.H., Tamaddoni-Nezhad, A., Lisi, F.A. (eds.) ILP 2011. LNCS, vol. 7207, pp. 238–254. Springer, Heidelberg (2012)
20. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Efficient model-based 3d tracking of hand articulations using kinect. In: BMVC, pp. 1–11 (2011)
21. Papon, J., Kulvicius, T., Aksoy, E.E., Worgotter, F.: Point cloud video object segmentation using a persistent supervoxel world-model. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3712–3718. IEEE (2013)
22. Rusu, R.B.: Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. Ph.D. thesis, Computer Science department, Technische Universitaet Muenchen, Germany, October 2009
23. Santos, P., Colton, S., Magee, D.R.: Predictive and Descriptive Approaches to Learning Game Rules from Vision Data. In: Sichman, J.S., Coelho, H., Rezende, S.O. (eds.) IBERAMIA 2006 and SBIA 2006. LNCS (LNAI), vol. 4140, pp. 349–359. Springer, Heidelberg (2006)
24. Yamamoto, Y.: Research on Logic and Computation in Hypothesis Finding. Ph.D. thesis
25. Yang, Y., Fermuller, C., Aloimonos, Y.: Detection of manipulation action consequences (mac). In: CVPR 2013 (2013)

Perceptual Narratives of Space and Motion for Semantic Interpretation of Visual Data

Jakob Suchan[1(✉)], Mehul Bhatt[1], and Paulo E. Santos[2]

[1] Cognitive Systems, University of Bremen, Bremen, Germany
jsuchan@informatik.uni-bremen.de
[2] Centro Universitario da FEI, São Paulo, Brazil

Abstract. We propose a commonsense theory of *space* and *motion* for the high-level semantic interpretation of dynamic scenes. The theory provides primitives for commonsense representation and reasoning with *qualitative spatial relations, depth profiles*, and *spatio-temporal change*; these may be combined with probabilistic methods for modelling and hypothesising event and object relations. The proposed framework has been implemented as a general activity abstraction and reasoning engine, which we demonstrate by generating declaratively grounded visuo-spatial narratives of perceptual input from vision and depth sensors for a benchmark scenario.

Our long-term goal is to provide general tools (integrating different aspects of space, action, and change) necessary for tasks such as real-time human activity interpretation and dynamic sensor control within the purview of cognitive vision, interaction, and control.

1 Introduction

Systems that monitor and interact with an environment populated by humans and other artefacts require a formal means for representing and reasoning about spatio-temporal, event and action based phenomena that are grounded to real public and private scenarios (e.g., logistical processes, activities of everyday living) of the environment being modelled. A fundamental requirement within such application domains is the need to explicitly represent and reason about dynamic spatial configurations or scenes and, for real world problems, integrated reasoning about space, actions, and change [1]. With these modelling primitives, the ability to perform *predictive* and *explanatory* analyses on the basis of sensory data is crucial for creating a useful intelligent function within such environments.

Commonsense, Space, Change. Qualitative Spatial & Temporal Representation and Reasoning (QSTR) provide a commonsensical interface to abstract and reason about quantitative spatial information [2]. *Qualitative spatial / temporal calculi* are relational-algebraic systems pertaining to one or more aspects of space such as *topology, orientation, direction, size* [3].

The integration of qualitative spatial representation and reasoning techniques within general commonsense reasoning frameworks in AI is an essential next-step

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 339–354, 2015.
DOI: 10.1007/978-3-319-16181-5_24

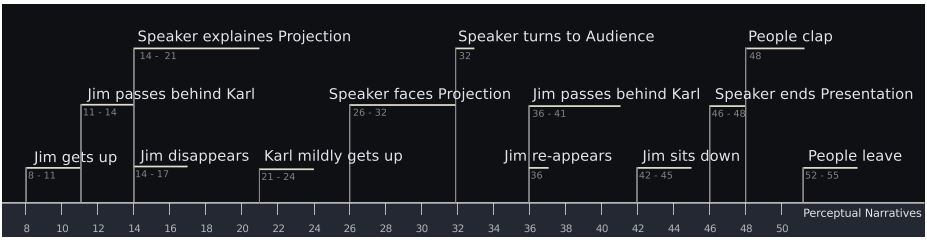


Fig. 1. Semantic Interpretation by Perceptual Narrativisation

for their applicability toward tasks such as spatial planning, spatio-temporal diagnosis and abnormality detection, event recognition and behaviour interpretation [4]. CLP(QS) [5] provides a framework for declarative spatial reasoning.

Perceptual Narratives [6] are declarative models of visual, auditory, haptic and other observations in the real world that are obtained via artificial sensors and / or human input. As an example, consider the *smart meeting cinematography* domain, where *perceptual narratives* as in Fig. 1 are generated based on perceived spatial change interpreted as interactions of humans in the environment. Such narratives explaining the ongoing activities are needed to anticipate changes in the environment, as well as to appropriately influence the real-time control of the camera system.

We suggest that the **semantic interpretation** of activities from video, depth (e.g., time-of-flight devices such as Kinect), and other forms of sensory input requires the representational and inferential mediation of qualitative abstractions of *space, action, and change* [1]. Generation of perceptual narratives, and their access via the declarative interface of logic programming facilitates the integration of the overall framework in bigger projects concerned with cognitive vision, robotics, hybrid-intelligent systems etc.

The particular focus and contributions of this paper are: (*a*) *Space and motion*: declaratively reasoning about qualitative spatial relations (e.g., topology, orientation), and motion in the context of everyday activities involving humans and artefacts (*b*) *Hybridisation*: integrating the qualitative theory with a probabilistic method for hypothesising object relations (*c*) *Semantic characterisation*: as a result of (a) and (b), generation of declarative narratives of perceptual RGB-D data that is obtained directly from people/object tracking algorithms.

2 Related Work

The core emphasis in activity and behaviour recognition has been on supervised learning algorithms requiring preprocessed (e.g., annotated) datasets from sensory streams. Unsupervised methods have received recent attention, with hybrid models integrating machine learning techniques with high-level structured representation and reasoning gaining recent momentum. The literature review below

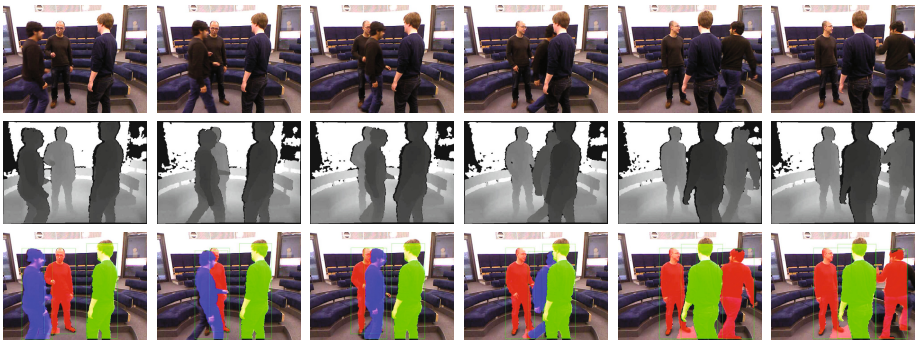


Fig. 2. Activity Sequence: *passing in-between people*, corresponding RGB and Depth profile data

concentrates on proposals concerned with the main aspects of the investigation reported in the present paper, namely, the high-level interpretation of events from the standpoint of Qualitative Spatial & Temporal Representation and Reasoning (QSTR). General reviews of work on activity and behaviour recognition can be found in [7–9].

2.1 Scene Interpretation

Research on scene interpretation has been largely based on probabilistic methods, motivated by the need to deal with sensor noise and image uncertainty [7], leaving aside the representation of general facts about the domain and the interplay between this representation and the actual interpretation of the scenes. Logic-based image interpretation, on the other hand, tackles the problem from the viewpoint of effective representation of general facts about the domain, as well as the generalisation of these facts to problems with infinite variables. Close to the topic of this paper, dos Santos et al. [10] presents a formalism for interpreting events such as *approaching*, *receding*, or *coalescing* from pairs of subsequent images obtained by a mobile robot's stereopair. Fernyhough et al. [11] proposed a technique for generating event models automatically based on qualitative reasoning and a statistical analysis of video input. This line of work has been further developed and has led to a range of related techniques broadly within the umbrella of the field of cognitive vision [12–14]. Dee et al. [14] proposes a method based on unsupervised clustering for building semantic scene models from video data using observed motion. Dubba et al. [12] presents a supervised learning framework to learn event models from large video datasets using inductive logic programming. Tran and Davis [15], and Morariu and Davis [16] present analogous results on the use of spatio-temporal relations within a first-order probabilistic language for the analysis of video sequences obtained in a parking lot. In a similar manner Song et al. [17] present a general framework for recognizing events in RGB-D data using probabilistic first-order logic and use it for tracking kitchen activities. Bohlken et al. [18] present work on a real-time

activity monitoring system defining activity concepts in an ontology which can be automatically transformed into a high-level scene interpretation system.

None of the works related to this paper, however, have considered a qualitative theory about space and motion as the basis to generate probabilistic interpretations of events. The present paper fulfills this gap by extending the qualitative theory proposed in [19] to account for the 3D space, while also combining it with interpretations of events from RGB-D data.

2.2 Cognitive Vision

The field of cognitive vision [20,21] has developed as an approach to enhance classical computer vision systems with cognitive abilities to obtain more robust vision systems, that are able to adapt to unforeseen changes, make sense of perceived data and show goal directed behavior. Vernon [20] defines a cognitive vision system in terms of its capabilities as follows:

> *"A cognitive vision system should be able to engage in purposive goal-directed behavior, it should be able to adapt robustly to unforeseen changes of the visual environment, and it should be able to anticipate the occurrence of objects or events"* Vernon [20]

There are multiple approaches towards the goal of developing a cognitive vision system. A detailed research plan for the development of the field of cognitive vision systems can be found in the technical report of the ECVision (European Research Network for Cognitive Computer Vision Systems) [22]. Among others, a symbolic approach to model knowledge about spatio-temporal phenomena has gained attention [15,23–25]. Cohn et al. [26] present work towards a cognitive vision system built on qualitative spatial and temporal abstractions to ground high-level concepts in visually sensed data.

2.3 QSTR – Qualitative Spatial and Temporal Reasoning

Qualitative Spatial & Temporal Representation and Reasoning (QSTR) [27] abstracts from an exact numerical representation by describing the relations between objects using a finite number of symbols. Qualitative representations use a set of relations that hold between objects to describe a scene. To represent the continuity of spatial change, Freksa [28] introduced the *conceptual neighborhoods*. Relations between two entities are conceptual neighbors if they can be directly transformed from one relation into the other by continuous change of the environment.

In the line of research about qualitative continuous spatial change, Galton [29–31] investigated movement on the basis of an integrated theory of space, time, objects, and position. Muller [32] defined continuous change using 4-dimensional regions in space-time. Hazarika and Cohn [33] build on this work but used an interval based approach to represent spatio-temporal primitives. In [34] Davis discusses the use of transition graphs for reasoning about continuous spatial change and applies them in physical reasoning problems.

Table 1. Spatial Relations and the Corresponding Motion Relations

Σ **Space**	
Topology	
	$discrete(p,q,t), partially\_overlapping(p,q,t),$ $proper\_part(p,q,t), proper\_part\_inverse(p,q,t), equal(p,q,t)$
Extrinsic Orientation (horizontal, vertical, and in depth)	
	$left(p,q,t), overlaps\_left(p,q,t), along\_left(p,q,t), horizontally\_equal(p,q,t),$ $overlaps\_right(p,q,t), along\_right(p,q,t), right(p,q,t)$
	$above(p,q,t), overlaps\_above(p,q,t), along\_above(p,q,t), vertically\_equal(p,q,t),$ $overlaps\_below(p,q,t), along\_below(p,q,t), below(p,q,t)$
	$closer(p,q,t), overlaps\_closer(p,q,t), along\_closer(p,q,t), distance\_equal(p,q,t),$ $overlaps\_further(p,q,t), along\_further(p,q,t), further(p,q,t)$
Σ **Motion**	
Movement	$approaching(p,q,t)$ and $receding(p,q,t)$
Size Motion	$elongating(x,p,t)$ and $shortening(x,p,t)$
Rate of Size Motion	$same\_rate(x,y,t)$, $faster(x,y,t)$,
Presence in the Scene	$appearing(p,t)$ and $disappearing(p,t)$

3 A Theory of Space, and Motion

We present a theory of space and motion to represent spatio-temporal phenomena for activity interpretation. As basic entities of the theory we consider depth profiles (see Fig. 2), which are regions of space, with a depth structure (distance from the sensor). These depth profiles are obtained by the projections of detected individuals in the scene on the image plane, where each point of the projected region has an associated depth value. Based on the depth profile we make different abstractions to encounter different aspects of space, i.e. regions, points (centroid), bounding cuboids, oriented points, lines (object axis) etc. These relations are defined in terms of the following functions on the depth profiles attributes:

depth: $depth\ profile \times time\ point \rightarrow float$, gives an depth profiles average distance from the observer at a time instant;

depth_front: $depth\ profile \times time\ point \rightarrow float$, gives an depth profiles minimal distance from the observer at a time instant;

depth_back: $depth\ profile \times time\ point \rightarrow float$, gives an depth profiles maximal distance from the observer at a time instant;

centroid: $depth\ profile \times time\ point \rightarrow (integer, integer, integer)$, gives the x,y, and z coordinates of the depth profiles centre point

size: $dimension \times depth\ profile \times time\ point \rightarrow integer$, maps a dimension, a depth profile and a time point to the depth profile's size in the given dimension;

dist: $depth\ profile \times bounding\ box \times time\ point \rightarrow float$, maps two depth profiles and a time point to the angular distance separating the depth profiles centroids in that instant.

in_sight: $depth\ profile \times time\ point \rightarrow boolean$, maps a depth profile and a time point to the presence of the depth profile. A depth profile is present at a time point, as long as there is at least one pixel associated with the depth profile.

3.1 Σ Space – Qualitative Spatial Relations

The basic part of our spatial theory consists of spatial relations on pairs of depth profiles, which includes relations on *topology* and *extrinsic orientation* in terms of left, right, above, below relations and depth relations (distance of a depth profile from the Observer).

Topological Relations. We represent the connectedness of pairs of depth profiles by the relations of the region connection calculus [35] for the 2D bounding boxes, omitting the depth. We use the RCC5 [35] subset of the region connection calculus in a ternary version, which contains the relations $discrete(p, q, t)$, $partially\_overlapping(p, q, t)$, $proper\_part(p, q, t)$, $proper\_part\_inverse(p, q, t)$, and $equal(p, q, t)$, where the third argument represents the time point when the relation holds. As the topological relations are defined on the two dimensional image plane, they do not represent the connection of two physical objects but rather the connection of the projection of two physical objects [36]. Due to this fact, the topological relations combined with the depth of the objects can be used to model that one object occludes the other.

Extrinsic Orientation. We represent the extrinsic orientation (relative position) of two depth profiles, with respect to the observer's viewpoint, making distinctions on the *3D position* and the *size* of the depth profiles. To this end, we use the bounding cuboid of the perceived depth profile determined by its *width*, *height*, and *thickness*, given as *depth_front* and *depth_back*. Given that we have 3D objects, we end up with a set of relations that resemble Allen's interval algebra [37] for each dimension, i.e. *horizontal*, *vertical*, and *depth*. However, in terms of depth perception, the interval relations that happen "instantaneously" (namely, *meets*, *starts*, and *finishes*) are irrelevant.

$$closer(p, q, t) \leftrightarrow (depth\_back(p, t) < depth\_front(q, t)); \tag{1a}$$

$$overlaps\_closer(p, q, t) \leftrightarrow (depth\_front(p, t) < depth\_front(q, t)) \wedge (depth\_front(q, t) < depth\_back(p, t)); \tag{1b}$$

$$along\_closer(p, q, t) \leftrightarrow (depth\_front(p, t) < depth\_front(q, t)) \wedge (depth\_front(q, t) < depth\_back(p, t)) \wedge (depth\_back(q, t) < depth\_back(p, t)); \tag{1c}$$

$$depth\_equal(p, q, t) \leftrightarrow (|depth\_front(p, t) - depth\_front(q, t)| < 0) \wedge (|depth\_back(p, t) - depth\_back(q, t)| < 0). \tag{1d}$$

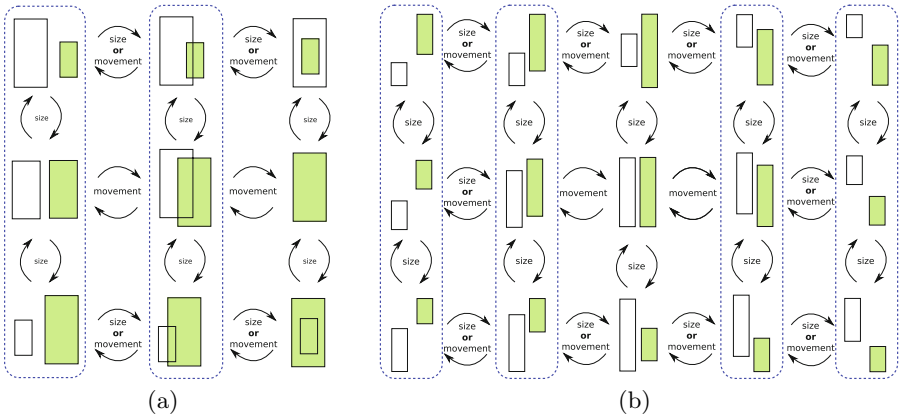


Fig. 3. Continuous Transitions between Spatial Relations on Topology and Extrinsic Orientation: topological and positional changes due to movement and transformation of the projected regions

Additionally we define the relations $further(p, q, t), overlaps\_further(p, q, t)$, and $along\_further(p, q, t)$ as inverse of the relations above. Accordingly to these relations on depth, we define the relations for the horizontal and the vertical dimension as listed in Table 1. To account for small diviations in the depth values, we apply a threshold μ represents the average error in the depth values.

3.2 Σ Motion – Qualitative Spatial Change

Spatial relations holding for perceived depth profiles change as an result of motion of the individuals in the scene (see Fig. 3). To account for this, we define motion relations by making qualitative distinctions of the changes in the depth profiles parameters, i.e. the distance between two depth profiles and its size. In each of the formulae presented below the timepoint t falls within the the open time interval (t1, t2). In this work, such time intervals are assumed to be very small; therefore, the predicates defined below are locally valid with respect to the time point t. We assume that this constraint is respected in this work but do not write it explicitly in the formulae for clarity. Further, we assume that there is a static relation between all relations to represent the case that the distance between two depth profiles stays the same, which is the case where the depth profile does not change in size or relative position.

Relative Movement. The relative movement of pairs of depth profiles is represented in terms of changes in the distance between their *centroids*. We represent these changes in terms of *approaching* and *receding* as defined below.

$$approaching(p, q, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (dist(p, q, t_2) < dist(p, q, t_1)); \tag{2a}$$

$$receding(p, q, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (dist(p, q, t_2) > dist(p, q, t_1)). \tag{2b}$$

Size-Motion. To represent size-motion of a single depth profile, we consider relations on changes in depth profiles *width*, *height* and *thickness* separately. Changes on more than one of these parameters at the same time instant can be represented by combinations of the relations below. In the relations below, the variable x is defined on the set of depth profiles attributes $x \in \{width, height, thickness\}$.

$$elongating(x, p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (size(x, p, t_1) < size(x, p, t_2)); \tag{3a}$$

$$shortening(x, p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge (size(x, p, t_1) > size(x, p, t_2)). \tag{3b}$$

Ordering Relations on the Rate of Size-Motion. We need to define relations that state the rate of relative changes in the *width*, *height*, and *thickness* parameters of a depth profile. The relations introduced to account for these issues are defined below, where variables x and y are defined on the set of depth profile attributes ($x, y \in \{width, height, thickness\}$), and $\Delta(x)$ and $\Delta(y)$ denote the change in these parameters at the time point t which is defined on a short interval $[t_1, t_2]$ as described above.

$same\_rate(x, y, t)$ represents the case when attribute x changes "at the same rate" as y at a time point t (more formally, $\frac{\Delta(x)}{\Delta(y)} = 0$)

$faster(x, y, t)$ represents the case when attribute x changes "faster" than attribute y at a time point t (more formally, $\Delta(x) > \Delta(y)$)

Presence of depth profiles in the scene. The relations *appearing* and *disappearing* represent the events of an depth profile being present in the scene at time t that was not present in the scene at the previous time point, resp. not being present at time t but has been present at the previous time point.

$$appearing(p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge \neg in\_sight(p, t_1) \wedge in\_sight(p, t_2); \tag{4a}$$

$$disappearing(p, t) \leftrightarrow \exists t_1 t_2 (t_1 < t) \wedge (t < t_2) \wedge in\_sight(p, t_1) \wedge \neg in\_sight(p, t_2). \tag{4b}$$

4 Spatial Change between Individuals in the Scene

To describe the observed scene in terms of spatio-temporal phenomena we combine the different aspects of the theory about *space* and *motion* providing a rich vocabulary about qualitative changes in the visual domain. This allows us to describe the ongoing interactions and operations between the physical entities represented by the depth profiles as well as on conceptual objects in the environment.

Individuals and objects in the scene. For the individuals and objects in the scene we assume that they have certain properties, i.e. we assume detected individuals to be rigid and non-opaque. Additionally we define abstract objects to represent the observer and the field of view of the sensing device. These objects are assumed to be non-moveable and for the field of view to have no physical object attached to it.

Visibility with Respect to the Observer. Topological relations of the depth profile's projection on the image plane, can be interpreted as visibility from the observers point of view [36] given, that the represented individuals are rigid and non-opaque. We use this fact to represent that one depth profile is occluded by another depth profile.

$$partially\_occluded(p, q, t) \leftarrow further(p, q, t) \wedge partiall\_overlapping(p, q, t). \tag{5a}$$

$$not\_occluded(p, q, t) \leftarrow discrete(p, q, t) \vee (closer(p, q, t) \wedge partially\_overlapping(p, q, t)). \tag{5b}$$

In the case of a full occlusion, the individual will not be detected any more, so this relation can only be hypothesised in the case of the disappearance of the individual.

Visibility relations changes as a result of motion, either of the individuals in the scene or of the observer. As defined in [38] the space in the environment can be divided into separate regions based on the visibility relations of an object in these regions with respect to an occluding object and the observer. Which results in the three zones, the *Light Zone(LZ)*, the *Twilight Zone(TZ)*, and the *Shadow Zone(SZ)*. To move from one zone to an other the object can only move in a certain way. E.g. to get from the right *Light Zone* to the left *Light Zone*, without passing in front of the occluding object, the object has to pass the right *Twilight Zone*, the *Shadow Zone*, and the left *Twilight Zone*.

Movement Direction with Respect to the Observer. We represent relative movement of a depth profile with respect to the observer by introducing distinct objects for the observer as well as the borders of the cameras field of view.

$$moving\_closer(p, t) \leftarrow approaching(p, observer, t); \tag{6a}$$

$$moving\_further\_away(p, t) \leftarrow receding(p, observer, t); \tag{6b}$$

$$moving\_left(p, t) \leftarrow approaching(p, left\_border, t); \tag{6c}$$

$$moving\_right(p, t) \leftarrow approaching(p, right\_border, t). \tag{6d}$$

In this way we define the relations for: (1). *moving closer*: the depth profile moves towards the observer; (2). *moving further away*: the depth profile moves away from the observer; (3). *moving left / right*: the depth profile approaches the left / right border of the field of view.

5 Human Interactions Grounded in Spatial Change

The abstractions of space and motion described in the previous section reflect changes between individuals in the real world, that are consequences of interactions conducted in the environment (or possible noise). However, in many cases it is not possible to unambiguously map from the changes in the relations to interactions of objects in the world, thus we associate the predicates on spatial change with possible hypotheses on interactions. Towards this, interactions are declaratively defined by there spatio-temporal appearance in the scene, using a

3-layered hierarchical activity model grounded in the spatial change observed in the environment. The activity model consists of the activity, interactions, and operations.

- **Activity** defined by its goal and determined by the specific interaction sequence performed towards this specific goal
- **Interaction** goal driven interactions between individuals in the scene determined by the observed spatial operations involved in the interaction
- **Spatial Operation** elemental parts of an interaction defined by spatial and temporal relations on perceived individuals in the environment

We consider consecutive frames in which the same relation holds for a pair of depth profiles or for a single depth profile as intervals of space and motion, in the sense of Allen's intervals [37]. An Interaction is then defined by spatial operations carried out by individuals involved in the interaction. Spatial operations are the basic elements of an interaction and determine, how an interaction is carried out in the environment, in terms of perceivable change. Operations are defined based on the observed intervals of space and motion using Allen's interval algebra to model temporal relations between these intervals. E.g. the interaction passing behind is declaratively defined in logic programming as depicted in Eq. 7a-d.

$$interaction(passing\_behind, P, Q, I) :- \quad (7a)$$
$$interaction(passing, P, Q, I_1), observation(partially\_occluded, P, Q, I_2),$$
$$discrete\_time(during, I_1, I_2), discrete\_time(equal, I, I_2).$$

$$interaction(passing, P, Q, I) :- \quad (7b)$$
$$operation(changing\_sides, P, Q, I_1), operation(moves, P, I_2),$$
$$discrete\_time(during, I_1, I_2), discrete\_time(equal, I, I_1).$$

$$operation(changing\_sides, P, Q, interval(T2, T3)) :- \quad (7c)$$
$$observation(horizontal(left), P, Q, interval(T1, T2)),$$
$$observation(horizontal(right), P, Q, interval(T3, T4)),$$
$$discrete\_time(meets, interval(T1, T2), interval(T3, T4)).$$

$$operation(moves, P, I) :- observation(moving(\_), P, I). \quad (7d)$$

5.1 Hypotheses on Perceived Spatial Change

Hypotheses on interactions in the real world are generated based on the perceived spatial change represented by the qualitative abstractions of *space* and *motion* and the background knowledge described in the previous section. To make hypotheses on interactions in the environment, one has to take possible noise and faulty observations into account, as well as consistency constraints between concurrent interactions.

- **Uncertainty** due to limitations in the low-level sensing, or to occlusion by other individuals in the scene. E.g. noise, missing observations, and occlusion.

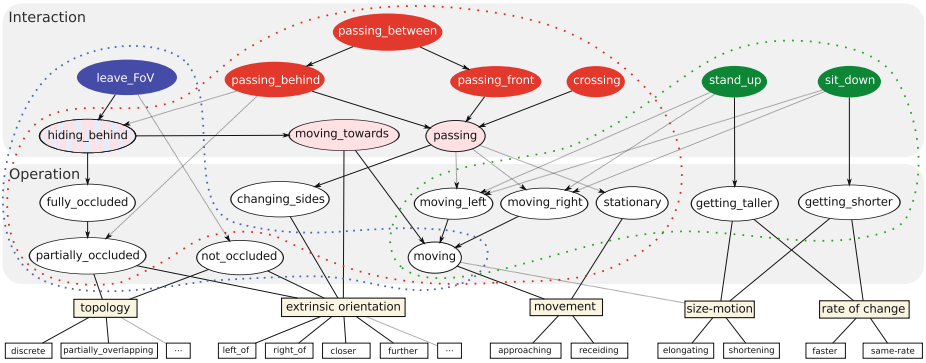


Fig. 4. Interaction taxonomy for the smart meeting domain

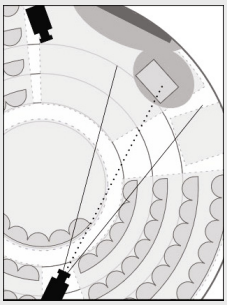
– **Consistency** in terms of concurrently performed interactions and the spatial operations contained in these interactions.

Hypotheses on interactions are arranged in a sequence, in the way, that the interactions and the corresponding spatial operations included in the interactions, best explain the observed spatial change. The probability for a certain interaction is then determined by the probability that the observation reflects the ongoing interactions in the environment, and the evidence that an observation provides for a certain interaction. For the use-case scenario presented in section 6 we use a causal network to evaluate the generated hypotheses given their grounding in observations on spatial change in the environment.

5.2 Perceptual Narratives of Human Activities

Sequences of hypothesized interactions are interpreted as perceptual narratives that describe the interactions performed in the environment with respect to the perceived spatial change. These narratives serve as a basis for reasoning in the sense of *explanation*, *prediction*, and *planning* for spatial control. As the perceptual narratives are grounded in the spatial change observed by the sensors, the narrative does not only reflect the performed interactions, but also states, how these interactions are performed in terms of the involved spatial operations.

Thus the narrative can be used to reason about the activity, the interactions within the activity, and the spatial change reflecting the interactions. And thereby help to explain incomplete or inconsistent observations, to reason about the most likely next steps towards the goal of the activity and thus predict upcoming spatial change, and to plan (spatial) control actions based on the aforementioned reasoning capabilities which is an important ability for dynamic control in smart environments.



Listing 1. Smart Meeting Cinematography
The smart meeting cinematography domain focusses on professional situations such as meetings and seminars. A basic task is to automatically produce dynamic recordings of interactive discussions, debates, presentations involving interacting people who use more than one communication modality such as hand-gestures (e.g., raising one's hand for a question, applause), voice and interruption, electronic apparatus (e.g., pressing of a button), movement (e.g., standing-up) and so forth. The scenario consists of people-tracking, gesture identification closed under a context-specific taxonomy, and also involves real-time dynamic collaborative co-ordination and self-control of pan-tilt-zoom (PTZ) cameras in a *sensing-planning-acting* loop. The long-term vision is to benchmark with respect to the capabilities of human-cinematographers, real-time video editors, surveillance personnel to record and semantically annotate individual and group activity (e.g., for summarisation, story-book format digital media and promo generation).

6 Use-Case: Smart Meeting Cinematography

We demonstrate the applicability of the theory of space and motion in the context of the meeting scenario (Listing 1). In this context, the basic interactions involved in the meeting process in Fig. 4 are considered. For the presented use-case, we assume that the camera is fixed in its position and orientation. Thus the changes observed in the relations are only due to object's motion (or noise in the sensor data).

Tracking and detection of Individuals. The particular hardware setup used in the meeting scenario consists of pan-tilt-zoom (PTZ) cameras, and depth sensors (Kinect), providing RGB-D data consisting of RGB images and corresponding depth information. Open source vision libraries, i.e. OpenCV and OpenNI are then used to detect and track individuals in the scene, which are perceived via their projection on the image plane of the sensor and their depth information. The thereby obtained depth profiles are 2.5 D regions of space, with a depth structure which gives the distance between the sensor and each pixel of the detected individuals.

Interactions in the smart meeting scenario. Interactions as performed in the meeting environment are modeled based on the spatial and temporal appearance of the interactions. For the meeting domain we take the interactions *enter_FoV*, *leave_FoV*, *passing_behind*, *passing_front*, *passing_between*, *crossing*, *stand_up*, and *sit_down* into account. Fig. 4 illustrates the taxonomy of these interactions and how they are defined based on the qualitative abstractions of space and motion. To generate the hypotheses on interactions in the environment we included a

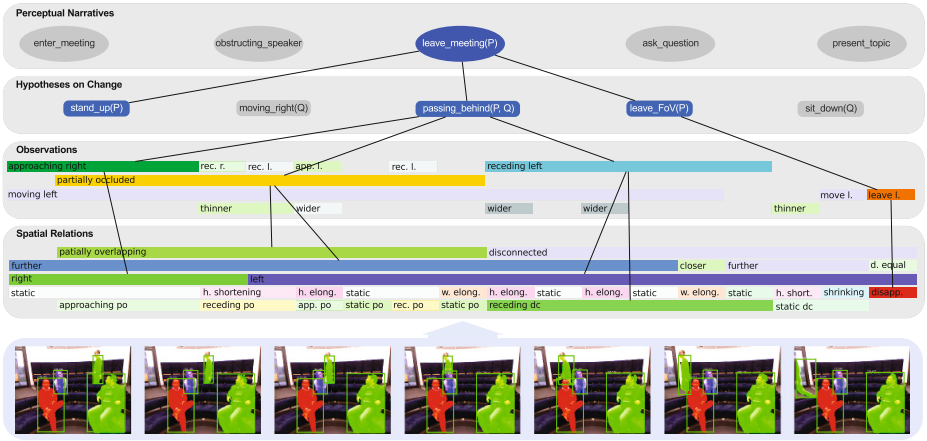


Fig. 5. Perceptual Narratives of Space, and Motion.

simple model of noise occurring in the sensing process and, of the consistency of concurrent interactions.

Resulting perceptual narrative. Using the described theory of space and motion and the interactions defined in the interaction taxonomy, combined with a simple naive bayes method for generating hypotheses, the system is able to generate the following narrative for the exemplary scene in Fig. 5.

$$\left[\begin{array}{rl} I_1 \equiv & interaction(stand\_up(P_4, interval(t_9, t_{13}))). \\ spatial\_operations(I_1) \equiv & getting\_taller(P_4, interval(t_9, t_{13})). \\ I_2 \equiv & interaction(passing\_behind(P_4, P_3, interval(t_{49}, t_{57}))). \\ spatial\_operations(I_2) \equiv & changing\_sides(P_4, P_3, interval(t_{52}, t_{53})) \wedge \\ & partially\_occluded(P_4, P_3, interval(t_{49}, t_{57})) \wedge \\ & moving\_left(P_4, interval(t_{45}, t_{65})) \wedge \\ & stationary(P_3, interval(t_1, t_{66})). \\ & \vdots \\ I_4 \equiv & interaction(leave\_FoV(P_4, interval(t_{66}, t_{66}))). \\ spatial\_operations(I_4) \equiv & moving\_towards(P_4, left\_border, interval(t_{65}, t_{65})) \wedge \\ & hiding\_behind(P_4, left\_border, interval(t_{66}, t_{66})). \end{array}\right] \quad (8)$$

Additionally to the interaction hypotheses, the narrative includes the spatial operations performed as a part of the interaction, and thereby reflect how the interactions are performed in the environment.

7 Conclusion and Outlook

Hypothesised object relations can be seen as building blocks to form complex interactions that are semantically interpreted as activities in the context of the

domain. As an example consider the sequence of observations in the meeting environment depicted in Fig. 5.

> Region P elongates vertically, region P approaches region Q from the right, region P partially overlaps with region Q while P being further away from the observer than Q, region P moves left, region P recedes from region Q at the left, region P gets disconnected from region Q, region P disappears at the left border of the field of view

These observations can be explained by the means of a percaptual narrative in terms of interactions in the real world performed in the meeting situation.

> Person P stands up, passes behind person Q while moving towards the exit and leaves the room.

Toward the generation of (declaratively grounded) perceptual narratives [6] such as the above, we developed and implemented a commonsense theory of qualitative *space* and *motion* for abstracting and reasoning about dynamic scenes. We defined combined relations capturing different spatial modalities in the context of a benchmark domain, namely the smart meeting cinematography scenario of the ROTUNDE initiative [39]. As a proof of concept, we integrated our proposed theory with a basic probabilistic reasoning method to generate hypotheses on interactions performed in the smart meeting scenario based on the combined model of *space* and *motion*. The smart meeting cinematography scenario serves as a challenging benchmark to investigate narrative based high-level cognitive interpretation of everyday interactions. Work is in progress to release certain aspects (pertaining to space, motion, real-time high-level control) emanating from the narrative model via the interface of constraint logic programming (e.g., as a Prolog based library of space–motion). Perceptual narrative based scene interpretation will be used for cognitive camera control consisting of interpreting the observations, to identify important information, and plan control actions based on the spatial requirements and constraints of scene. Work towards this end includes the integration of multiple camera viewpoints, where the system has to reason about perspective changes and visibility based on qualitative spatio-temporal abstractions.

References

1. Bhatt, M.: Reasoning about space, actions and change: a paradigm for applications of spatial reasoning. In: Qualitative Spatial Representation and Reasoning: Trends and Future Directions. IGI Global, USA (2012)
2. Cohn, A.G., Renz, J.: Qualitative spatial reasoning. In van Harmelen, F., Lifschitz, V., Porter, B., (eds.) Handbook of Knowledge Representation. Elsevier (2007)
3. Ligozat, G.: Qualitative Spatial and Temporal Reasoning. Wiley, ISTE (2013)
4. Bhatt, M., Guesgen, H., Wölfl, S., Hazarika, S.: Qualitative spatial and temporal reasoning: Emerging applications, trends, and directions. Spatial Cognition & Computation **11**, 1–14 (2011)

5. Bhatt, M., Lee, J.H., Schultz, C.: CLP(QS): a declarative spatial reasoning framework. In: Egenhofer, M., Giudice, N., Moratz, R., Worboys, M. (eds.) COSIT 2011. LNCS, vol. 6899, pp. 210–230. Springer, Heidelberg (2011)
6. Bhatt, M., Suchan, J., Schultz, C.: Cognitive interpretation of everyday activities - toward perceptual narrative based visuo-spatial scene interpretation. In: Finlayson, M., Fisseni, B., Lwe, B., Meister, J.C., (eds.) Computational Models of Narrative (CMN) (2013)
7. Lavee, G., Rivlin, E., Rudzsky, M.: Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **39**, 489–504 (2009)
8. Gonzlez, J., Moeslund, T.B., Wang, L., (eds.) Special issue on Semantic Understanding of Human Behaviors in Image Sequences. In: Computer Vision and Image Understanding. vol. 116, pp. 305–472. Elsevier (2012)
9. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**, 976–990 (2010)
10. dos Santos, M., de Brito, R.C., Park, H.H., Santos, P.: Logic-based interpretation of geometrically observable changes occurring in dynamic scenes. Applied Intelligence **31**, 161–179 (2009)
11. Fernyhough, J.H., Cohn, A.G., Hogg, D.: Constructing qualitative event models automatically from video input. Image Vision Comput. **18**, 81–103 (2000)
12. Dubba, K.S.R., Cohn, A.G., Hogg, D.C.: Event model learning from complex videos using ILP. In: ECAI, pp. 93–98 (2010)
13. Sridhar, M., Cohn, A.G., Hogg, D.C.: Unsupervised learning of event classes from video. In: AAAI (2010)
14. Dee, H.M., Cohn, A.G., Hogg, D.C.: Building semantic scene models from unconstrained video. Computer Vision and Image Understanding **116**, 446–456 (2012)
15. Tran, S.D., Davis, L.S.: Event modeling and recognition using markov logic networks. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 610–623. Springer, Heidelberg (2008)
16. Morariu, V., Davis, L.: Multi-agent event recognition in structured scenarios. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
17. Song, Y.C., Kautz, H., Allen, J., Swift, M., Li, Y., Luo, J., Zhang, C.: A markov logic framework for recognizing complex events from multimodal data. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. ICMI 2013, pp. 141–148. ACM, New York (2013)
18. Bohlken, W., Neumann, B., Hotz, L., Koopmann, P.: Ontology-based realtime activity monitoring using beam search. In: Crowley, J.L., Draper, B.A., Thonnat, M. (eds.) ICVS 2011. LNCS, vol. 6962, pp. 112–121. Springer, Heidelberg (2011)
19. Santos, P.: Reasoning about depth and motion from an observer's viewpoint. Spatial Cognition and Computation **7**, 133–178 (2007)
20. Vernon, D.: Cognitive vision: The case for embodied perception. Image Vision Comput. **26**, 127–140 (2008)
21. Vernon, D.: The Space of Cognitive Vision. In: Christensen, H.I., Nagel, H.-H. (eds.) Cognitive Vision Systems. LNCS, vol. 3948, pp. 7–24. Springer, Heidelberg (2006)
22. Auer et al.: A research roadmap of cognitive vision. Technical Report v5, ECVISION (2005). www.ecvision.org
23. Sridhar, M., Cohn, A.G., Hogg, D.C.: Learning functional object-categories from a relational spatio-temporal representation. In: ECAI, pp. 606–610 (2008)

24. Dubba, K.S.R., Cohn, A.G., Hogg, D.C.: Event model learning from complex videos using ilp. In: Proc. ECAI. Volume 215 of Frontiers in Artificial Intelligence and Applications, pp. 93–98. IOS Press (2010)
25. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2009, pp. 2012–2019 (2009)
26. Cohn, A.G., Hogg, D.C., Bennett, B., Devin, V., Galata, A., Magee, D.R., Needham, C.J., Santos, P.: Cognitive vision: integrating symbolic qualitative representations with computer vision. In: Christensen, H.I., Nagel, H.-H. (eds.) Cognitive Vision Systems. LNCS, vol. 3948, pp. 221–246. Springer, Heidelberg (2006)
27. Cohn, A., Hazarika, S.: Qualitative spatial representation and reasoning: An overview. Fundam. Inf. **46**, 1–29 (2001)
28. Freksa, C.: Conceptual neighborhood and its role in temporal and spatial reasoning. In: Singh, M., Travé-Massuyès, L. (eds.) Decision Support Systems and Qualitative Reasoning, pp. 181–187. North-Holland, Amsterdam (1991)
29. Galton, A.: Towards an integrated logic of space, time and motion. In: IJCAI, pp. 1550–1557 (1993)
30. Galton, A.: Towards a qualitative theory of movement. In: Frank, A.U., Kuhn, W. (eds.) Spatial Information Theory - A Theoretical Basis for GIS (COSIT'95), pp. 377–396. Springer, Heidelberg (1995)
31. Galton, A.: Qualitative Spatial Change. Oxford University Press (2000)
32. Muller, P.: A qualitative theory of motion based on spatio-temporal primitives. In: Cohn, A.G., Schubert, L.K., Shapiro, S.C., (eds.) Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR'98), pp. 131–143. Morgan Kaufmann, Trento, June 2–5, 1998
33. Hazarika, S.M., Cohn, A.G.: Abducing qualitative spatio-temporal histories from partial observations. In: KR, pp. 14–25 (2002)
34. Davis, E.: Qualitative reasoning and spatio-temporal continuity. In: Hazarika, S.M. (ed.) Qualitative Spatio-Temporal Representation and Reasoning: Trends and Future Directions, pp. 97–146. IGI Global, Hershey (2012)
35. Cohn, A.G., Bennett, B., Gooday, J., Gotts, N.M.: Qualitative spatial representation and reasoning with the region connection calculus. Geoinformatica **1**, 275–316 (1997)
36. Randell, D., Witkowski, M., Shanahan, M.: From images to bodies: Modeling and exploiting spatial occlusion and motion parallax. In: Proc. of IJCAI, Seattle, U.S., pp. 57–63 (2001)
37. Allen, J.F.: Maintaining knowledge about temporal intervals. Commun. ACM **26**, 832–843 (1983)
38. Tassoni, S., Fogliaroni, P., Bhatt, M., Felice, G.D.: Toward a Qualitative 3D Visibility Model. In: 25th International Workshop on Qualitative Reasoning, co-located with the IJCAI-11 Conference, Barcelona, Spain (2011)
39. Bhatt, M., Suchan, J., Freksa, C.: ROTUNDE - A Smart Meeting Cinematography Initiative. In: Bhatt, M., Guesgen, H., Cook, D. (eds.) Proceedings of the AAAI-2013 Workshop on Space, Time, and Ambient Intelligence (STAMI). AAAI Press, Washington (2013)

Multi-Entity Bayesian Networks for Knowledge-Driven Analysis of ICH Content

Giannis Chantas[1], Alexandros Kitsikidis[1], Spiros Nikolopoulos[1], Kosmas Dimitropoulos[1], Stella Douka[2], Ioannis Kompatsiaris[1], and Nikos Grammalidis[1(✉)]

[1] Centre for Research and Technology Hellas, Information Technologies Institute, 6th km Xarilaou-Thermi, Thessaloniki, Greece
ngramm@iti.gr
[2] Department of Physical Education and Sport Science, Aristotle University of Thessaloniki, Thessaloniki, Greece

Abstract. In this paper we introduce Multi-Entity Bayesian Networks (MEBNs) as the means to combine first-order logic with probabilistic inference and facilitate the semantic analysis of Intangible Cultural Heritage (ICH) content. First, we mention the need to capture and maintain ICH manifestations for the safeguarding of cultural treasures. Second, we present the MEBN models and stress their key features that can be used as a powerful tool for the aforementioned cause. Third, we present the methodology followed to build a MEBN model for the analysis of a traditional dance. Finally, we compare the efficiency of our MEBN model with that of a simple Bayesian network and demonstrate its superiority in cases that demand for situation-specific treatment.

Keywords: Semantic analysis · Intangible cultural heritage · Multi-entity bayesian networks

1 Introduction

By the age of six, humans recognize more than 104 semantic concepts [1] and keep learning more throughout their life. Can a computer program learn how to recognize semantic concepts in multimedia content the way a human does? In addressing this question, divergent approaches have been proposed, relying either on the use of explicit knowledge or the abundant availability of data. Advocating the former, [2], [3] are two notable cases where a small number of examples used during learning are able to provide models with sufficient generalization ability. The authors rely on the hypothesis that once a few visual categories have been learned with significant cost, some information may be abstracted from the process to make learning further categories more efficient. Taking a different perspective, the authors of [4] claim that with the availability of overwhelming amounts of data many problems can be solved without the need for complex parametric algorithms. The authors index a large dataset of 79 million images

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 355–369, 2015.
DOI: 10.1007/978-3-319-16181-5_25

and using nearest neighbor matching for image annotation, they claim that given the excessive volume of the indexed images it is reasonable to assume that almost every "unseen" image will be close enough to a ''seen" image. These examples demonstrate the debate around the mechanism of building perceptual models and the discussion on how much of the knowledge should come in an explicit form and how much can be obtained implicitly from the available training samples. Although moving towards the one or the other extreme of the debate may still produce non-trivial recognition models, higher levels of efficiency can only be achieved if explicit and implicit knowledge are effectively combined.

The aim of this work is to verify the aforementioned statement in the specialized domain of Intangible Cultural Heritage (ICH). The term intangible cultural heritage (ICH) (UNESCO, 2013) refers to valuable traditional art forms and creative practices, such as singing, dancing, craftsmanship, etc. In this paper, we advocate the use of Multi-Entity Bayesian Networks (MEBNs) [5] as an efficient scheme to facilitate the analysis of ICH content, mainly due to their ability in combining first-order logic with probability theory. The remaining of this paper is organized as follows: Section 2 describes the particularities of the ICH domain and motivates the use of MEBNs. In Section 4 we describe the most important characteristics of MEBNs and argue about their appropriateness to address the particularities of ICH domain. Section 5 offers some details about the methodology adopted to implement and apply MEBNs for analyzing ICH content. Finally, Section 6 explains the results of our preliminary experimental study, while Section 7 summarizes our concluding remarks.

2 Semantic Analysis in the ICH Domain

The semantic analysis of digital heritage resources is considered a particularly important prerequisite for their preservation. This is even more evident in the domain of ICH. Indeed, given that during the preservation of intangible heritage the significance of heritage artifacts is implied in their context, the scope of digital preservation extends to the preservation of the background knowledge that puts them in proper perspective. For example, Mangalacharan [6] is an invocation dance in Indian Odissi dance form, which is specified in terms of specific and predefined dance actions and it is accompanied by a specific kind of music. The dance actions entail the movement of human body parts and interaction with object and the accompanying music has features that fit to the dance. Moreover, high-level concepts are manifested in the dance that are composed of basic body actions, which are related to the music features. Thus, the preservation of heritage resources requires a solution to the problem of: (a) recognizing media patterns that correspond to elementary domain concepts like objects, postures, actions, audio tempos, etc, and (b) consider these elementary domain concepts as evidence to support a hypothesis stating that the analyzed media item manifests a certain high-level concept.

A major difficulty in representing ICH knowledge is the inherent ambiguity and uncertainty in concepts prevalent in this domain. In meeting this challenge,

explicitly provided logic-based rules need to be combined with a probabilistic inference framework in order to map low-level multimedia features to high-level concepts. Initially, the domain concepts and their relations will have to be expressed in a machine understandable format that should be also capable of encoding different snapshots of the analysis environment (e.g., number of dance steps). Then, low-level multimedia features that may incorporate visual, or other types of signals will have to be analyzed to obtain elementary conceptual information, acting as evidence. Finally, the framework used for probabilistic inference should inherit the logic-based rules encoded in the first step and evaluate the extracted evidence in the context of the domain knowledge. Thus, at the core of semantic multimedia analysis lies the development of a theory that will not only manage to effectively combine logic-based rules with probabilistic inference, but will also offer the necessary flexibility to cope with an un-predictable and dynamically changing environment.

3 Related Work

A number of works have been presented in the literature that aim to represent knowledge in a probabilistic manner. OOBN models [7] have been proposed as an alternative to standard BN for overcoming the inherent inflexible structure of BN. An OOBN object is a collection of domain attributes that extents regular BN nodes, so as to become more flexible to situations that require customization. Probabilistic relational models (PRMs) [8] extend Bayesian networks by introducing the concept of properties, and relations between them. Like MEBNs, PRMs provides a similar mechanism to built situation specific probabilistic models. However, OOBN and PRM expressivity is inferior to MEBN, mainly due to the context limitations used to enforce logical constraints on the model variables.

Ontologies with probabilistic extensions have been also used for the semantic analysis of ICH content. For example, in [6] the authors propose the use of ontology-based mapping for linking cultural heritage content to ICH concepts. More specifically, the ontology used in this framework includes the descriptions of domain concepts that are formally given in terms of the related low-level audio-visual features, appearing in the multimedia content. In this way, a convenient semantic interpretation of the multimedia data is enabled. In another closely related work [9], a semi-automatic ontology construction methodology is proposed for combining bayesian networks with probabilistic inference. The goal of this work is to facilitate the semantic analysis of cultural Indian dances, i.e. detection of specific dance styles and moves in multimedia with cultural content. Note however, that although the ontology is constructed using probabilistic methods (i.e. as a BN of concepts and relations), the BN remains unchanged. This is a serious modeling shortcoming that motivates the use of MEBNs.

Another approach for handling uncertainty relies on Fuzzy DLs which are based on the fuzzy set theory and are capable of performing reasoning under uncertain circumstances. This is in contrast to the probabilistic extension of ontologies where uncertainty handling is based on probabilistic formalisms. One

such example is presented in [10] where the use of fuzzy DLs semantics has been proposed to interpret the output of the classifiers into a semantically consistent interpretation. In this work the authors claim that the use of DLs allows to formally capture the semantics underlying the concepts of interest, while the fuzzy extensions provide the means to model the vagueness encompassed in the extracted classification. The advantage of MEBNs compared to Fuzzy DLs is their flexibility in learning the parameters of the model that can be done based on samples, rather than by specifying ad hoc rules as required in Fuzzy DLs. In the following, we advocate the use of MEBNs as a potential solution to the problem of semantic analysis in the domain of ICH.

4 Multi-entity Bayesian Networks

MEBN logic is a formal system that unifies probability theory and classical first-order logic (FOL). Thus, MEBNs are the outcome of the combination of BN with FOL. From a Bayesian perspective, MEBNs are extended BNs by incorporating FOL. Their main advantage is in combining the capability of BN to model uncertainty with the expressivity of FOL in representing knowledge. The key feature of MEBNs is the ability to build situation specific BNs (SSBN) that are customized according to the snapshot of the environment being modeled in an arbitrary situation. In this way, MEBNs overcome the inflexibilty of BNs to adapt to the volatile environment being model, since they have a fixed structure and conditional probability for each node.

Technically, a MEBN is a collection of MEBN fragments (MFrags). An MFrag includes (among others) *resident* node(s), for each of which a local conditional distribution and a set of parent nodes (if any) are defined. The MFrag of a resident node is called its *home* MFrag. Also, in an MFrag, there are input nodes that are resident nodes in other MFrags. The parents of a resident node can be either resident, input nodes or both. The resident nodes are, in a sense, templates that are used to construct the nodes of the SSBN, i.e., the name of the nodes, the dependencies with other nodes and the conditional probability distribution. The local conditional distribution of a resident node in an MFrag is a function that produces the conditional probabilities of the SSBN nodes produced by the resident node. This function takes as input the structure of SSBN and produces a conditional probability for the related node accordingly.

Another component of an MFrag is the logical variables, placed as arguments in resident and input nodes, and logical constraint nodes, imposing constraints on the logical variables participating in the MFrag. Logical variables and their constraints are the manifestation of the FOL into MEBN modeling. The structure of an SSBN is determined by the logical variables and the admissible by the constraints values to which they can be instantiated, according to the situation of the environment being modeled (e.g., number of nodes a resident node, acting as a template, replicate). In other words, logical variables and their constraints drive the construction of the SSBNs based on the evidence collected by the environment, translated as potential values of the logical variables.

The ultimate goal of modeling with MEBN is inference, which provides us with the ability to analyze the environment being modeled (e.g., a stochastic process or a static snapshot of a closed system), based on evidence. Inference is performed on the SSBN, which results based on evidence. There are two steps in SSBN construction. In the first step, the logical variables are instantiated to values determined by the environment and according to logical constraints. The resulting nodes in the SSBN have a defined set of parent nodes and a conditional distribution. In the second step, a subset of the SSBN random variables (i.e. the *observed* variables) are considered known (observed) and instantiated to their observed (measured) values. Then, Bayesian inference provides the posterior probability distribution of the unknown random variables we want to estimate.

5 Applying MEBN Theory in the Domain of ICH

In order to validate our assumptions in a real world problem, we have used MEBNs as a knowledge representation and analysis tool for recognizing the different styles of a traditional greek dance. There are two main reasons motivating the use of MEBNs for this specific task, namely, uncertainty modeling and situation-specific analysis. Uncertainty in this case is manifested in two cases. In the first, a dancer may unexpectedly deviate from the dance pattern (e.g., skips a dance step). In the second, the step detector may fail to detect a step and/or correctly recognize its features. MEBNs are capable to model both the volatility of the step number and the uncertainty (randomness) aspects of each performance. Also, the situation specific analysis capability is useful due to the fact that, usually, the number of steps is not *a priori* known. SSBN can be proven very beneficial for the dynamic modeling of such situations. In our work, the role of the MEBN is to adapt in each performance and model in a probabilistic Bayesian framework the uncertainty aspects of the dance. Based on that, the ultimate goal is to detect the dance style through probabilistic inference.

The first step in employing MEBNs is to consult the experts in order to elicit and formally encode the domain knowledge. Thus, a methodology for the ontology specification and engineering will have to be employed. Subsequently, the knowledge encoded in this ontology will act as the basis for constructing the corresponding MFrags. Then, the observations extracted from the analysis of sensor signals will be injected to the framework so as to generate the SSBN and perform probabilistic inference. Finally, decisions about the different dance styles can be made based on the posterior probability distribution of the network.

5.1 Ontology Specification and Engineering

Most state-of-the-art methodologies for ontology engineering incorporate the requirements specification activity. The communication tool that is used during this activity is a set of competency questions (CQs) that are posed to the experts. The CQs are answered by the experts in natural language. The terms used in these answers are subsequently analyzed with respect to their frequency and

semantic affinity, so as to extract the terminology (names, adjectives and verbs) that will be formally represented in the ontology by means of concepts, attributes and relations. In accordance with this practice, we have followed the methodology of [11] in order to specify and engineer the ontology presented in Figure 1.

Fig. 1. Tsamiko ontology graph. Some nodes were colored for brevity of demonstration, black: right foot steps, orange: left foot steps. The TsamikoSixStep and TsamikoTenStep are high-foot steps when the dance style is male. Also, grey areas illustrate the 'hasDouble(Right/Left)Step' relation.

The style of Tsamiko dance is characterized as "double" or "single" and as "male" or "female". Thus, there are four different characterizations, and, hence, Tsamiko dance styles: (single, male), (double, male), (single, female) and (double, female). In distinguishing between the different styles, the most important elementary concept has to do with the type of steps. A Tsamiko dance consists of multiple dance cycles, each one consisting of ten distinct steps. Each step is characterized and distinguished from the other steps by its four attributes (i.e. left or right direction, left or right foot, single or double step and foot is in high or

low position) and its place in the sequence (i.e. 1st, 2nd,...,10th). The attributes of some steps depend on the particular style of the dance being performed, while other step attributes remain the same for all styles. More specifically, among the step attributes that do not depend on the dance style, we identify the movement direction (i.e., left or right), as well as the body place (i.e., whether it is the right or the left foot). Particularly, the first six steps have a "right" direction and the rest have a "left" direction. Also, 1st, 3rd, 5th, 8th, 10th are performed with the right foot and the rest with the left foot. Another important attribute that now depends on whether the style is "male" or "female", has to do with the foot lifting movement, which essentially differentiates between a step that is performed with the foot in high position, or in a position close to the floor. More specifically, the foot is high at the 6th and 10th step in a "male" dance while it is always low for "female" style. Also, in a dance of "double" style, the 2nd, 4th and 8th standard step of the dance cycle are characterized by the "double" attribute. On the other hand , all steps have the "single" attribute when the style is "single". Finally, the ontology of Figure 1 reveals also the importance of sequence among the undertaken steps that has to be performed in a rather strict order. Thus, it is evident that the detection of each step along with its attributes is crucial for our analysis framework.

5.2 Sensor Signal Analysis for Elementary Concept Detection

In order to capture the performance of the dancers, we have used markerless motion capture based on depth sensing technology. Microsoft Kinect sensors were employed, which are low-cost real-time depth sensing cameras that can track the volume of a performer and produce skeletal data. Microsoft Kinect SDK [12] has been used as a solution for skeletal tracking and acquisition. It provides the ability to track the 3D positions of 20 predefined skeletal joints of a human body at 30Hz rate. In order to solve occlusion and self-occlusion tracking problems inherent in this type of motion capture and to increase the total area of coverage, several Kinect sensors were placed in an array in front of the dancer (Figure 2 Left). The captured data were combined following a fusion strategy described in [13], leading to an increased robustness of skeletal tracking.

The elementary domain concepts, which are the steps and the way they are executed, were extracted from the analysis of the joint position signals. The analysis consists of two parts: segmentation and feature extraction. Segmentation is performed on two levels of granularity. Initially, the dance periods are detected. Tsamiko dance has a repeating pattern, with the dancer moving on a semi-circle performing several steps to the right direction followed by several steps to the left. This consists of a single period, which is easily detected by analyzing the position of the waist of the dancer. Peak detection of a sub-sampled (to remove noise) waist displacement along the horizontal axis reveals the time instants when the dancer is at the end of the left/right movement (Figure 2 Right). Subsequently, each period is further segmented into steps which we consider an elementary domain concept. The detection of steps is based on the movement of ankles along the horizontal axis relative to the movement of the root of the

body. Once again, local maxima detection is employed for the detection of time instants when the footstep is performed (when the foot touches the ground, the relative horizontal displacement produces local maxima).

After the segmentation, each segment is analyzed to extract the features of each step. The features extracted from each step are: the foot that is moving (left foot or right foot), the direction of movement (left or right), raised foot and double step. Those features are extracted from a rule based analysis of ankle and knee joint position signals of the dancers' legs. The double step is a sequence of two small steps executed sequentially, which we consider as a single step during the segmentation period, since the intermediate steps are small and executed very quickly. The result of this analysis is a sequence of steps together with properties assigned to each step which are used subsequently to infer the dance style.

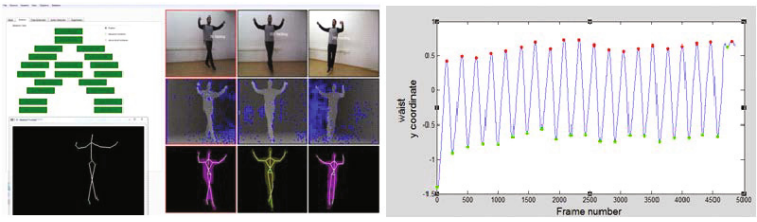


Fig. 2. (Left) Tsamiko performance captured by three depth cameras. Skeletal tracking from each camera can be seen as well as the final fused skeleton tracking result. (Right) Displacement of the waist of the dancer along the horizontal axis. Red and green dots represent the peaks and valleys detected, segmenting the dance into periods.

5.3 MTheory and MFrags

Based on the ontology described in Section 5.1, we have developed the MEBN of Figure 3. In this figure, a MEBN is presented consisting of two MFrags. The TsamikoStepMFrag contains information about the step sequence for different dance styles, along with the style distinguishing characteristics of the steps. The TsamikoStyleMFrag contains the style related MEBN nodes, "genderstyle" and "stepstyle" that can take the values male/female and single/double, respectively. Each MFrag contains input nodes (colored in grey), resident nodes (colored in yellow) and logical nodes (colored in green). The input nodes of the TsamikoStepMFrag are resident nodes in the TsamikoStyleMFrag. The only exception is the node "step", which is used to model a recursive process as described below.

The "step" input node in TsamikoStepMFrag (colored in grey) models the step sequence that comprise a dance cycle. It is very important to understand the concept of recursion in MEBNs, which is manifested in this case by making the input (grey) node "step" the parent of the resident (yellow) node "step"

(both having the same name but different logical variables ($t1$ and $t2$) as arguments). These variables are logical variables that are used to model the aspect of time sequence in the detected steps. For example, if $t2$ is instantiated as $timeStep_i$, then, based on the constraints dictated by the logical (green) nodes, $t1$ is instantiated as $timeStep_{i-1}$. In this way, $t1$ is always the previous time step of $t2$ enforcing the desired recursive process. In our example of Figure 3, the resident (yellow) node "step" has a range set of ten values, TsamikoStep1, ..., TsamikoStep10, each value denoting one of the ten distinct steps in a Tsamiko dance cycle. Thus, with the recursive definition of the node "step" (i.e. both as an input and a resident node) we enable the modeling of a dance step sequence execution. Note that the number of cycles and the starting and ending step are arbitrary.

Besides "step", there are four more resident nodes in the TsamikoStepMFrag as depicted in Figure 3: "hasDirection", "foot", "isFootHigh" and "isDoubleStep", which are essentially the features that declare the execution method of each step. The first node can take either the "leftDirection" or the "rightDirection" value, while the second node can take the "leftFoot" or "rightFoot" values. Both are not directly dependent to the dance style, as shown by the lack of direct arrows between these nodes and the nodes "genderstyle" and "stepstyle". Instead, the impact of "hasDirection" and "foot" to the recognition of the dance style goes through the "step" node that models the execution pattern of the dance. On the other hand, the nodes "isFootHigh" and "isDoubleStep" take boolean values and directly depend on the nodes "genderstyle" and "stepstyle" that determine the dance style. These dependencies are better described in the following paragraph that explains the TsamikoStyleMFrag.

According to the ontology presented in Section 5.1, the style of Tsamiko dance can be characterized as *male* or *female* and as double or single. We have decided to recognize the undertaken style on a per step-basis, meaning that steps of the same sequence can be attributed to different styles. Nevertheless, there is strong correlation between the style detected in $step_i$ and the probability that $step_{i+1}$ will follow the same style. Thus, as in the case of steps, there is an inherent requirement for modeling a recursive relation between the variables determining the style characteristics. To this end, the TsamikoStyleMFrag consists of two variables "genderstyle" and "stylestep" that exist both as resident and input nodes in the same MFrag, modeling the recursive relation between the styles detected for each step. Similar to TsamikoStepMFrag, the TsamikoStyleMFrag consists also of the exact same logical variables $t1$ and $t2$ that are instantiated to time steps and are used to enforce the desired recursive process. Finally, we should note that "genderstyle" and "stepstyle" are also used as input nodes in the TsamikoStepMFrag and act as direct parents of "isFootHigh" and "isDoubleStep".

Figure 4 demonstrates the situation-specific bayesian network that results from the aforementioned MEBN model, for a specific situation where the total performance consists of five steps. This network is used to estimated the posterior probabilities of the unknown variable (yellow nodes) based on the observations collected

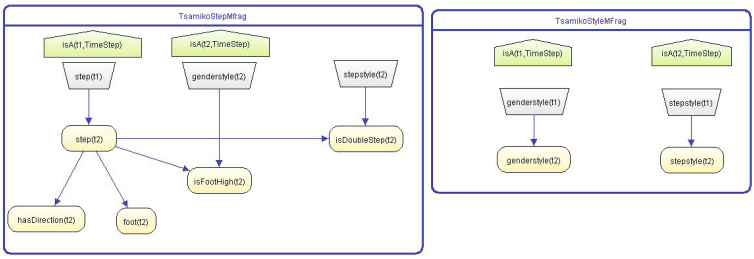


Fig. 3. MEBN developed to facilitate the analysis of greek tsamiko dance

for the known variables (cyan nodes). Finally, bayesian inference is applied to calculate the posterior probabilities of the unknown variables by employing belief propagation [14].

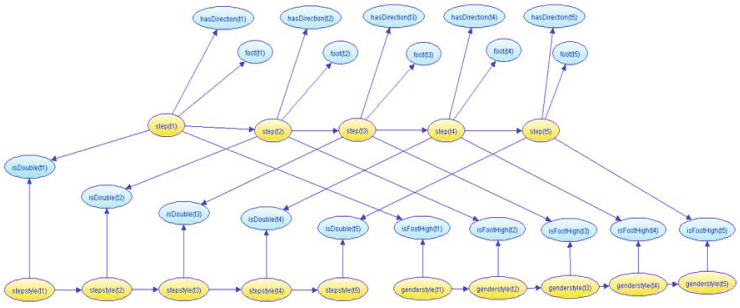


Fig. 4. The situation-specific bayesian network derived from the MEBN model of Figure 3 using five steps

6 Preliminary Experimental Results

The goal of our experimental study is to verify the appropriateness of MEBNs in recognizing the different dance styles based on the undertaken steps. Actually, our interest is not in just classifying a step sequence to one of the existing dance styles, which would constitute a trivial problem. Instead, our goal is to classify each step to one of the existing dance styles and at the same time assess the proficiency level of the performer (i.e. an estimation of how accurately the performer has executed the dance). In the first case we expect that the classification accuracy of the MEBN-based framework will outperform a baseline approach that relies on BNs but does not make any use of the situation specific capability of MEBNs. In the second case, we expect that our MEBN-based analysis framework will rank high the step sequences that have been performed flawlessly and rank low the step sequences that contain one or more execution errors.

Table 1. Characteristics of the different performances used for the tsamiko dance

performance	length	attributes	performance	length	attributes
A1	48	single/female	A9	201	double/male
A2	153	single/female	A10	190	double/female
A3	194	single/male	A11	198	double/female
A4	196	single/female	A12	189	double/female
A5	189	single/female	A13	202	double/male
A6	200	single/male	E1	190	single/male
A7	202	single/male	E2	100	double/female
A8	195	single/female	E3	191	double/male
A9	201	double/male	-	-	-

6.1 Dataset and Evaluation Metrics

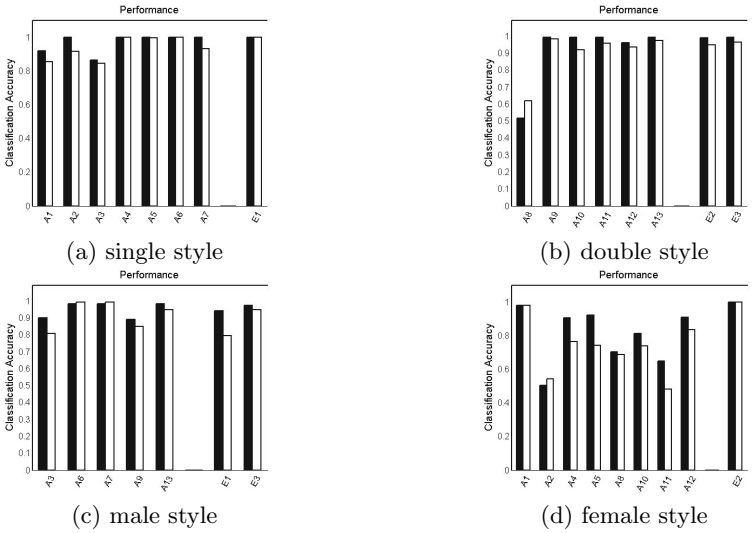
In our experiments we have used 16 recorded performances of Tsamiko dance, with the sequence length ranging from 50 to 200 steps. Out of the 16 recorded performances 3 were executed by professional dancers (E1-E3) while the rest was obtained from apprentice level dancers (A1-A13). All performances were executed with the same musical piece and every performance was annotated with its dance style, as depicted in Table 1.

In order to assess the classification accuracy we apply a threshold on the confidence score extracted for each step. More specifically, when we analyze a step sequence that is annotated as "single" we consider as correctly classified all steps that have caused the posterior probability (i.e. confidence degree) of the "step-style" node to overcome the 0.5 threshold. Similar is the case when analyzing a step sequence annotated with the other style types. Then, we divide this number with the total number of steps so as to calculate the classification accuracy for the entire performance. On the other hand, in order to assess the proficiency level of each performance we use the average of the confidence degrees of all steps in this performance. More specifically, the result of the analysis process for each performnance is a two-dimensional table with its columns corresponding to the different styles, its rows corresponding to the individual steps and its values being the posterior probability of the SSBN random variables modeling the dance style information. Thus, by performing column-wise averaging in this table we obtain four scores (i.e. corresponding to the four dance styles) that are suitable for assessing the proficiency level of the undertaken step sequence (i.e. considering that the closer you get to a 100% score the closer you get to a flawless performance).

6.2 Step Classification Accuracy

In order to verify the benefit of being able to adapt to the situation at hand, we compare our MEBN-based model with a baseline approach that lacks this capability. More specifically, given that one of the auxiliary features offered by the signal analysis module is the total number of steps composing the step sequence, the baseline approach was designed to totally neglect this situation specific information. It is essentially implemented as a straightforward BN with a fixed number

of nodes representing steps (we have used 10 steps which is the standard cycle in a tsamiko dance) and without any provision for the recursive relation between steps and dance cycles. Figure 5 demonstrates the classification accuracy of both frameworks. We can see that the MEBN-based framework outperforms the baseline approach in 28 out of the 32 cases. Given that both frameworks rely on probabilistic inference and both frameworks have been designed based on the same domain knowledge, it is reasonable to attribute the observed improvement in the flexibility of the MEBN-based framework to adapt in the number of steps composing each performance.



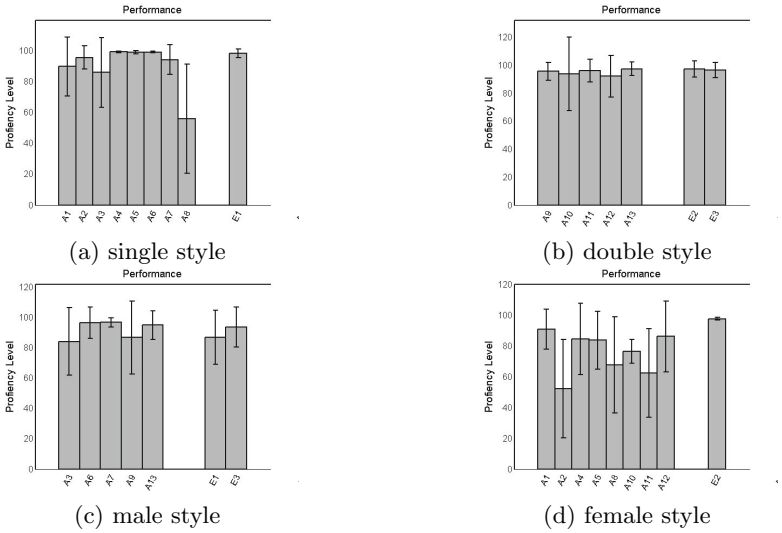
(a) single style (b) double style

(c) male style (d) female style

Fig. 5. The performance results are grouped based on the style. The bar diagrams colored in black correspond to the MEBN-based framework, while the bar diagrams colored in white correspond to the baseline.

6.3 Proficiency Level Assessment

Figure 6 shows the proficiency level results for the 16 performances, grouped based on the dance style and distinguishing between apprentices and experts. Moreover, apart from the average score the standard deviation is also depicted. The obtained results verify our expectation that in the majority of the examined cases our MEBN-based analysis framework is able to distinguish between an apprentice and an expert. This is evident in the case of "female" style where the performance level of the expert is higher than all apprentices. Similar conclusions can be also extracted for the cases of "single" and "double" where despite the fact that the proficiency level of the experts does not supersede all apprentices their superiority is evident in terms of average numbers. Finally, in the case of "male" style we notice that the proficiency level of the experts is lower by approximately 1.5% than the average score of all apprentices. This outcome can



(a) single style (b) double style

(c) male style (d) female style

Fig. 6. Proficiency level results for the 16 performances of our dataset, grouped based on the dance style and distinguishing between apprentices and experts

be attributed to the low performance of our MEBN-based analysis framework in effectively modeling one of the key-features characterizing the "male" style, which is the detection of "isFootHigh". Since the detection of this feature is rather challenging for the signal analysis module, it seems that in this case our MEBN-based model has failed to prevent the propagation of the first stage analysis error to the final outcome. In our future work we plan to examine the score of each step in correlation with the performance of the signal analysis module so as to gain more insights.

7 Conclusions

In this paper we have shown how the theory of MEBNs can be used to combine probabilistic analysis with first-order logic. The proposed framework was employed for the semantic analysis of Tsamiko traditional dances. The purpose of semantic analysis was to recognize the specific style of the tsamiko dance based on the special characteristics of the dance steps. The latter were extracted by a motion analysis module relying on the body movements. Experiments demonstrated that the classification efficiency of the proposed model is significantly better than the standard Bayesian network case. Also, the model was evaluated in terms of the ability to discriminate between expert and apprentice dancers, giving encouraging results. In the future, we plan to augment the MEBN model with information obtained from other than visual based modalities, such as

sound. Precisely, we expect that by exploiting the information from the musical piece accompanying the dance performances, we can improve the accuracy of semantic analysis. However, the task of combining music and body movement information is challenging, since it requires the identification of the musical features that provide useful information (i.e., that have a semantic meaning for the dance) and the detection of their dependency with the dance steps. Finally, we should mention that although the experimental study of this work focuses in just one dance (tsamiko dance) the step-wise approach that we have presented can be easily adopted to analyze other types of dance that require the execution of a pre-determined sequence of steps.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7-ICT-2011-9) under grant agreement no FP7-ICT-600676 "i-Treasures: Intangible Treasures - Capturing the Intangible Cultural Heritage and Learning the Rare Know-How of Living Human Treasures".

References

1. Biederman, I.: Recognition-by-components: A theory of human image understanding. Psychological Review **94**, 115–147 (1987)
2. Li, F.F., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Trans. Pattern Anal. Mach. Intell. **28**(4), 594–611 (2006)
3. Li, J., Wang, J.Z.: Real-time computerized annotation of pictures. IEEE Trans. Pattern Anal. Mach. Intell. **30**(6), 985–1002 (2008)
4. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**, 1958–1970 (2008)
5. Laskey, K.B.: Mebn: A language for first-order bayesian knowledge bases. Artificial Intelligence **172**(2–3), 14–178 (2008)
6. Mallik, A., Chaudhuri, S., Ghosh, H.: Nrityakosha: Preserving the intangible heritage of indian classical dance. ACM Journal on Computing and Cultural Heritage **4**(3), 11 (2011)
7. Koller, D., Pfeffer, A.: Object-oriented bayesian networks. In: Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference. Morgan Kaufmann, San Francisco (1997)
8. Pfeffer, A.: Probabilistic Reasoning for Complex Systems. Stanford University, Stanford (2000)
9. Mallik, A., Pasumarthi, A.P., Chaudhury, S.: Multimedia ontology learning for automatic annotation and video browsing. In: 1st ACM International Conference on Multimedia Information Retrieval (MIR 2008), New York (2008)
10. Dasiopoulou, S., Kompatsiaris, I., Strintzis, M.G.: Applying fuzzy DLs in the extraction of image semantics. In: Spaccapietra, S., Delcambre, L. (eds.) Journal on Data Semantics XIV. LNCS, vol. 5880, pp. 105–132. Springer, Heidelberg (2009)

11. Suárez-Figueroa, M.C., Gómez-Pérez, A., Villazón-Terrazas, B.: How to write and use the ontology requirements specification document. In: Dillon, T., Herrero, P., Meersman, R. (eds.) OTM 2009, Part II. LNCS, vol. 5871, pp. 966–982. Springer, Heidelberg (2009)
12. Microsoft: Kinect for windows — voice, movement and gesture recognition technology (2013). http://www.microsoft.com/en-us/kinectforwindows/
13. Kitsikidis, A., Dimitropoulos, K., Douka, S., Grammalidis, N.: Dance analysis using multiple kinect sensors. In: VISAPP2014 (2014)
14. Jensen, F.V.: Bayesian artificial intelligence: Kevin b. korb, ann e. nicholson, chapman & hall, 354 pages (2004). Pattern Anal. Appl. **7**(2), 221–223 (2004)

$\mathcal{ALC}$(F): A New Description Logic for Spatial Reasoning in Images

Céline Hudelot[1(✉)], Jamal Atif[2], and Isabelle Bloch[3]

[1] Centrale Supelec, Paris, France
celine.hudelot@ecp.fr
[2] PSL, LAMSADE, Université Paris Dauphine, Paris, France
jamal.atif@dauphine.fr
[3] Institut Mines Télécom - Télécom ParisTech - CNRS LTCI, Paris, France
isabelle.bloch@telecom-paristech.fr

Abstract. In image interpretation and computer vision, spatial relations between objects and spatial reasoning are of prime importance for recognition and interpretation tasks. Quantitative representations of spatial knowledge have been proposed in the literature. In the Artificial Intelligence community, logical formalisms such as ontologies have also been proposed for spatial knowledge representation and reasoning, and a challenging and open problem consists in bridging the gap between these ontological representations and the quantitative ones used in image interpretation. In this paper, we propose a new description logic, named $\mathcal{ALC}$(**F**), dedicated to spatial reasoning for image understanding. Our logic relies on the family of description logics equipped with concrete domains, a widely accepted way to integrate quantitative and qualitative qualities of real world objects in the conceptual domain, in which we have integrated mathematical morphological operators as predicates. Merging description logics with mathematical morphology enables us to provide new mechanisms to derive useful concrete representations of spatial concepts and new qualitative and quantitative spatial reasoning tools. It also enables imprecision and uncertainty of spatial knowledge to be taken into account through the fuzzy representation of spatial relations. We illustrate the benefits of our formalism on a model-guided cerebral image interpretation task.

Keywords: Spatial reasoning · Ontology-based image understanding · Description logics

1 Introduction

In image interpretation and computer vision, spatial relations between objects and spatial reasoning are of prime importance for recognition and interpretation tasks [5,6], in particular when the objects are embedded in a complex environment. Indeed, spatial relations allow solving ambiguity between objects having a similar appearance, and they are often more stable than characteristics of the

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 370–384, 2015.
DOI: 10.1007/978-3-319-16181-5_26

objects themselves. This is typically the case of anatomical structures, as illustrated in Figure 1, where some structures, such as the internal grey nuclei (thalamus, putamen, caudate nuclei), may have similar grey levels and similar shapes, and can be therefore easier distinguished for their individual recognition using spatial relations [14,33]. Spatial relations also allow improving object and scene recognition in images such as photographs [18,29], or satellite images [2,22,35].

Spatial reasoning can be defined as the domain of spatial knowledge representation, in particular spatial relations between spatial entities, and of reasoning on these entities and relations. This field has been largely developed in Artificial Intelligence, in particular using qualitative representations based on logical formalisms [1,36]. In image interpretation and computer vision, it is much less developed and is mainly based on quantitative representations [9,23]. Bridging the gap between the qualitative representations and the quantitative ones is a challenging and open issue to make them operational for image interpretation.

Description logics (DL) equipped with concrete domains [27] are a widely accepted way to integrate concrete and quantitative qualities of real world objects with conceptual knowledge and as a consequence to combine qualitative and quantitative reasoning useful for real-world applications. In this paper, we propose a new description logic, named $\mathcal{ALC}(\mathbf{F})$, dedicated to spatial reasoning for image understanding. In this framework, the combination of a description logic with concrete domains and mathematical morphology provides new mechanisms to derive useful concrete representations of concepts and new reasoning tools, as demonstrated in [20,21]. This paper builds upon these works by studying in depth the formal properties of this framework and revisiting the tableau decision algorithm. This framework also enables us to take into account imprecision to model vagueness, inherent to many spatial relations and to gain in robustness in the representations [9]. The rest of this paper is organized as follows. In Section 2, we review some related work and we recall how mathematical morphology can be used to derive fuzzy representations of spatial relations. In Section 3, we briefly present the main concepts of a spatial relation ontology used to represent spatial knowledge. We describe our new logic and its properties in Section 4. The reasoning and inference components are detailed in Section 5, and we illustrate the benefits of this framework for image interpretation tasks in Section 6, with the example of brain structure recognition in 3D images.

2 Spatial Knowledge Representations

As mentioned in Section 1, spatial relations between objects of a scene are of prime importance for semantic scene understanding. Several models for representing spatial relations have been proposed in the literature. These models can be classified according to different viewpoints:

– The nature of the model: quantitative or semi-quantitative models versus qualitative ones. In image interpretation and computer vision, many quantitative or semi-quantitative representations have been proposed. Many of

them assimilate objects to basic entities such as centroid or bounding box [23] and others are based on the notion of histograms [28,30]. On the contrary, in the Artificial Intelligence field, many qualitative and ontological models have been proposed (for instance, see [13] for a review).
- The type of the spatial relations: many authors have stressed the importance of topological relations and have proposed models for them [15,31] but distances and directional relative positions [9,24] are also important, as well as more complex relations such as "between", "surround" or "along" for instance.
- Their ability to model some important characteristics of spatial knowledge and in particular its imprecision [9].

The choice of a representation also depends on the type of question raised and the type of reasoning one wants to perform [10]: (1) which is the region of space where a relation with respect to a reference object is satisfied ? (2) to which degree is a relation between two objects satisfied?

In the following, we briefly present some fuzzy models of spatial relations using mathematical morphology on which we build our logic.

We denote by $\mathcal{S}$ the spatial (image) domain, and by $\mathcal{F}$ the set of fuzzy sets defined over $\mathcal{S}$, defined via their membership functions, associating with each point of space a membership value in $[0, 1]$. The usual partial ordering on fuzzy sets is used, denoted by $\leq_{\mathcal{F}}$, and the associated infimum $\wedge$ and supremum $\vee$. The empty set is denoted by $\emptyset_{\mathcal{F}}$ and the fuzzy set with membership value equal to 1 everywhere by $1_{\mathcal{F}}$. For a t-norm t and its residual implication I, $(\mathcal{F}, \leq_{\mathcal{F}}, \wedge, \vee, \emptyset_{\mathcal{F}}, 1_{\mathcal{F}}, t, I)$ is a residuated lattice of fuzzy sets defined over the image space by $\mathcal{S}$.

As shown in [10] and the references therein, mathematical morphology is a powerful tool to model spatial relations in various settings (sets, fuzzy sets, propositional logics, modal logics...). In the fuzzy set setting, the two main morphological operators, dilation δ and erosion ε, are defined from a t-norm t and its residual implication I as [12]:

$$\forall x \in \mathcal{S}, \delta_\nu(\mu)(x) = \vee_{y \in \mathcal{S}} t(\nu(x - y), \mu(x)), \tag{1}$$

$$\forall x \in \mathcal{S}, \varepsilon_\nu(\mu)(x) = \wedge_{y \in \mathcal{S}} I(\nu(y - x), \mu(x)). \tag{2}$$

The idea for mathematical morphology based spatial reasoning is to define the semantics of a spatial relation by a fuzzy structuring element ν in the spatial domain, and to use morphological operations to compute the region of space where the relation is satisfied with respect to a reference object. For instance, if ν represents the relation "right of", then $\delta_\nu(\mu)(x)$ represents the degree to which x is to the right of the fuzzy set μ (an example is illustrated in Figure 2). This allows answering the first question above. As for the second question, histogram based approaches can be adopted [30], or pattern matching approaches, applied to the previous result and the fuzzy set representing the second object. A review of fuzzy spatial relations can be found in [9].

3 An Ontology of Spatial Relations

The semantic interpretation of images can benefit from representations of useful concepts and the links between them as ontologies. We build on the work of [19] which proposes an ontology of spatial relations with the aim of guiding image interpretation using spatial knowledge. We briefly recall the main concepts of this ontology using description logics (DLs) as a formal language and we rely on the standard notations of DLs (see [3] for an introduction).

One important entity of this ontology, as proposed in [19], is the concept ***SpatialObject*** ($\mathsf{SpatialObject} \sqsubseteq \top$). As mentioned in [26], the nature of spatial relations is twofold: they are concepts with their own properties, but they are also links between concepts and thus an important issue is related to the choice of modeling spatial relations as concepts or as roles in DLs. In [19], a spatial relation is not considered as a role (property) between two spatial objects but as a concept on its own (***SpatialRelation***), enabling to address the two spatial reasoning questions mentioned in Section 2.

- A ***SpatialRelation*** is subsumed by the general concept ***Relation***. It is defined according to a ***ReferenceSystem***:
  $\mathsf{SpatialRelation} \sqsubseteq$
  $\mathsf{Relation} \sqcap \ni \mathsf{type}.\{\mathsf{Spatial}\} \sqcap \exists\ \mathsf{hasReferenceSystem}.\mathsf{ReferenceSystem}$
- The concept ***SpatialRelationWith*** refers to the set of spatial relations which are defined according to at least one or more reference spatial objects RO (hasRO):
  $\mathsf{SpatialRelationWith} \equiv$
  $\mathsf{SpatialRelation} \sqcap \exists\ \mathsf{hasRO}.\mathsf{SpatialObject} \sqcap \geq 1\ \mathsf{hasRO}$
- We define the concept ***SpatiallyRelatedObject*** which refers to the set of spatial objects which have at least one spatial relation (hasSR) with another spatial object. This concept is useful to describe spatial configurations:
  $\mathsf{SpatiallyRelatedObject} \equiv$
  $\mathsf{SpatialObject} \sqcap \exists\ \mathsf{hasSR}.\mathsf{SpatialRelationWith} \sqcap \geq 1\ \mathsf{hasSR}$
- At last, the concept ***DefinedSpatialRelation*** represents the set of spatial relations for which target (hasTargetObject) and reference objects (hasRO) are defined:
  $\mathsf{DefinedSpatialRelation} \equiv$
  $\mathsf{SpatialRelation} \sqcap \exists\ \mathsf{hasRO}.\mathsf{SpatialObject} \sqcap \geq 1\ \mathsf{hasRO} \sqcap$
  $\exists\ \mathsf{hasTargetObject}.\mathsf{SpatialObject} \sqcap = 1\ \mathsf{hasTargetObject}$

4 Proposed Logic for Spatial Reasoning: $\mathcal{ALC}(\mathbf{F})$

In this section, we introduce mathematical morphology as a spatial reasoning tool. In particular, mathematical morphology operators are integrated as predicates of a spatial concrete domain. The main objective is to provide a foundation to reason about qualitative and quantitative spatial relations. The proposed logic is built on $\mathcal{ALCRP}(D)$ [16,17] with the spatial concrete domain $\mathbf{F}$. We name it $\mathcal{ALC}(\mathbf{F})$ in the rest of the paper.

4.1 $\mathcal{ALC}(\mathbf{F})$ - Syntax and Semantics

Definition 1 (Spatial concrete domain). *A spatial concrete domain is a pair* $\mathbf{F} = (\Delta_{\mathbf{F}}, \Phi_{\mathbf{F}})$ *where* $\Delta_{\mathbf{F}} = (\mathcal{F}, \leq_{\mathcal{F}}, \wedge, \vee, \emptyset_{\mathcal{F}}, 1_{\mathcal{F}}, t, I)$ *is a residuated lattice of fuzzy sets defined over the image space* $\mathcal{S}$*,* $\mathcal{S}$ *being typically* $\mathbb{Z}^2$ *or* $\mathbb{Z}^3$ *for 2D or 3D images, with* t *a t-norm (fuzzy intersection) and* I *its residuated implication.* $\Phi_{\mathbf{F}}$ *denotes a set of predicate names on* $\Delta_{\mathbf{F}}$ *which contains:*

- *The unary predicates* $\bot_{\mathcal{S}}$ *and* $\top_{\mathcal{S}}$ *defined by* $\bot_{\mathcal{S}}^{\mathbf{F}} = \emptyset_{\mathcal{F}}$ *and* $\top_{\mathcal{S}}^{\mathbf{F}} = 1_{\mathcal{F}}$*.*
- *The name of the unary predicate* μ_X *defined by* $(\mu_X)^{\mathbf{F}} \in \mathcal{F}$ *(*$(\mu_X)^{\mathbf{F}} : \mathcal{S} \rightarrow [0,1]$*). The predicate associates to a spatial concept* X *a unique fuzzy set in the concrete domain* $\mathbf{F}$*. For each point* $x \in \mathcal{S}$*,* $\mu_X^{\mathbf{F}}(x)$ *represents the degree to which* x *belongs to the spatial representation of the object* X *in the spatial domain (the image in our illustrative example).*
- *The name of the unary predicate* ν_R *defined as* $\nu_R^{\mathbf{F}} \in \mathcal{F}$ *(*$\nu_R^{\mathbf{F}} : \mathcal{S} \rightarrow [0,1]$*). The predicate associates to a spatial relation* R*, the fuzzy structuring element* $\nu_R^{\mathbf{F}}$ *defined on* $\mathcal{S}$ *which represents the fuzzy relation* R *in the spatial domain.*
- *The name of the unary predicate* $\delta_{\nu_R}^{\mu_X}$*, defined by* $(\delta_{\nu_R}^{\mu_X})^{\mathbf{F}} = \delta_{\nu_R^{\mathbf{F}}}(\mu_X^{\mathbf{F}}) \in \mathcal{F}$*, with* δ *a fuzzy dilation defined as in Equation 1.*
- *The name of the unary predicate* $\varepsilon_{\nu_R}^{\mu_X}$*, defined as* $(\varepsilon_{\nu_R}^{\mu_X})^{\mathbf{F}} = \varepsilon_{\nu_R^{\mathbf{F}}}(\mu_X^{\mathbf{F}}) \in \mathcal{F}$*, with* ε *a fuzzy erosion defined as in Equation 2.*
- *The names of two binary predicates* $\sqcap_d, \sqcup_d$*:* $(\mu_{X_1} \sqcap_d \mu_{X_2})^{\mathbf{F}} = \mu_{X_1}^{\mathbf{F}} \wedge \mu_{X_2}^{\mathbf{F}}$ *and* $(\mu_{X_1} \sqcup_d \mu_{X_2})^{\mathbf{F}} = \mu_{X_1}^{\mathbf{F}} \vee \mu_{X_2}^{\mathbf{F}}$*, with* $\wedge$ *and* $\vee$ *the infimum and the supremum of* $\mathcal{F}$*.*
- *The name of a binary predicate* $\setminus_d$*, defined as* $(\mu_{X_1} \setminus_d \mu_{X_2})^{\mathbf{F}} = \mu_{X_1}^{\mathbf{F}} \setminus \mu_{X_2}^{\mathbf{F}}$*, with* $\setminus$ *the difference between fuzzy sets.*
- *The name of an unary predicate* $-$ *which defines the substraction between the membership function of a fuzzy set with a number into* $[0,1]$*.*
- *Names for composite predicates consisting of composition of elementary predicates.*

We now illustrate how these fuzzy concrete domain predicates are used to represent spatial relations. As in [16,17], we assume that each abstract spatial relation concept and each abstract spatial object concept is associated with its fuzzy representation in the concrete domain by the concrete feature hasForFuzzyRepresentation, denoted hasFR (it is a concrete feature because each abstract concept has only one fuzzy spatial representation in the image space).

- $\mathsf{SpatialObject} \equiv \exists\ \mathsf{hasFR}.\top_{\mathcal{S}}$. It defines a SpatialObject as a concept which has a spatial existence in image represented by a spatial fuzzy set.
- In the same way, we have: $\mathsf{SpatialRelation} \equiv \mathsf{Relation} \sqcap \exists\ \mathsf{hasFR}.\top_{\mathcal{S}}$.

Then, the following constructors can be used to define the other concepts of the ontology:

- $\exists\ \mathsf{hasFR}.\mu_X$ restricts the concrete region associated with the object X to the specific spatial fuzzy set defined by the predicate μ_X,

- $\exists$ hasFR.ν_R restricts the concrete region associated with the relation R to the specific fuzzy structuring element defined by the predicate ν_R,
- $\exists$ hasFR.$\delta^{\mu_X}_{\nu_R}$ restricts the concrete region associated with the spatial relation R to a referent object X, denoted R_X, to the spatial fuzzy set obtained by the dilation of $\mu^{\mathbf{F}}_X$ by $\nu^{\mathbf{F}}_R$,
- each concept R_X can then be defined by:
  R_X $\equiv$ SpatialRelation $\sqcap\exists$ hasRO.X $\sqsubseteq$ SpatialRelationWith and R_X $\equiv$ SpatialRelation $\sqcap\exists$ (hasFR,hasRO.hasFR).λ,
  where λ is a binary predicate built with the mathematical fuzzy operators δ and ε. For a relation R which has a referent object X, we write:

$$\text{(hasFR,hasRO.hasFR).}\delta \equiv \text{hasFR. } \delta^{\mu_X}_{\nu_R},$$

- $C \equiv$ SpatialObject$\sqcap$hasSR.R_X denotes the set of spatial objects which have a spatial relation of type R with the referent object X and we have the following axioms:

$$C \sqsubseteq \exists\text{relationTo.X and } C \sqsubseteq \text{SpatiallyRelatedObject.}$$

Examples for Distance Relations. This new formalism can be used to model different types of spatial relations and to derive useful concrete representations of these spatial relations. We illustrate our approach with distance relations. As for other relations, distance relations can be defined using fuzzy structuring elements and fuzzy morphological operators [8]. For instance, the Close_To relation can be defined by the structuring element ν_{Close_To}, which provides a representation of the relation in the spatial domain $\mathcal{S}$. This representation can be learned from examples. We can thus define the abstract spatial relation Close_To as: Close_To $\equiv$ DistanceRelation $\sqcap\exists$hasFR.ν_{Close_To}. Let X $\equiv \exists$ hasFR.μ_X, $\mu^{\mathbf{F}}_X$ being the spatial fuzzy set representing the spatial extent of the object X in the concrete domain (image space). Using the concept-forming predicate operator $\exists f.P$ (see [16]), we can define restrictions for the fuzzy representation of the abstract spatial concept Close_To_X using the dilation operator δ. As a consequence, we have: Close_To_X $\equiv$ DistanceRelation $\sqcap\exists$hasFR.$\delta^{\mu_X}_{\nu_{Close\_To}}$. The value $\delta^{\mu^{\mathbf{F}}_X}_{\nu^{\mathbf{F}}_{Close\_To}}(x)$ represents the degree to which a point x of $\mathcal{S}$ belongs to the fuzzy dilation of the fuzzy spatial representation of X by the fuzzy structuring element $\nu^{\mathbf{F}}_{Close\_To}$. This approach naturally extends to any distance relation expressed as a vague interval.

4.2 Properties

Admissibility of $\mathbf{F} = (\Delta_{\mathbf{F}}, \Phi_{\mathbf{F}})$. A concrete domain $\mathcal{D}$ is called admissible if the set of its predicate names is closed under negation and contains a name $\top_{\mathcal{D}}$ for $\Delta_{\mathcal{D}}$, and the satisfiability problem for finite conjunctions of predicates is decidable [27]. Let us prove that the concrete domain $\mathbf{F} = (\Delta_{\mathbf{F}}, \Phi_{\mathbf{F}})$ is admissible thanks to the algebraic setting of mathematical morphology and fuzzy sets. Indeed, using the classical partial order on fuzzy sets $\leq_{\mathcal{F}}$, $(\mathcal{F}, \leq_{\mathcal{F}})$ is a complete lattice.

1. The name for $\Delta_{\mathbf{F}}$ is $\top_{\mathcal{S}}$.
2. $\Phi_{\mathbf{F}}$ is closed under negation:
   - $\neg\top_{\mathcal{S}} = \bot_{\mathcal{S}}$; $\neg\bot_{\mathcal{S}} = \top_{\mathcal{S}}$;
   - $\forall \mu_X^{\mathbf{F}} \in \mathcal{F}, \neg\mu_X^{\mathbf{F}} \in \mathcal{F}$ (the negation is then a fuzzy complementation and $\mathcal{F}$ is closed under complementation); $\forall \nu_R^{\mathbf{F}} \in \mathcal{F}, \neg\nu_R^{\mathbf{F}} \in \mathcal{F}$;
   - $\forall (\mu, \nu) \in \mathcal{F}^2, \neg\delta_\nu(\mu) = \varepsilon_\nu(\neg\mu)$ and $\neg\varepsilon_\nu(\mu) = \delta_\nu(\neg\mu)$ (duality of erosion and dilation), for dual connectives t and I [11].
3. For decidability of the satisfiability of finite conjunctions of predicates, the same reasoning as in [16] can be applied, leading to the following algorithm:
   - negated predicates can be replaced by other predicates (or disjunctions of predicates), so that only non-negated predicates need to be considered;
   - concrete representations of μ_X and ν_R are computed and considered as variables;
   - relations can be computed between the concrete representations of spatial objects, using classical algorithms of mathematical morphology (here we consider a discrete finite space, and these algorithms are tractable);
   - then it can be directly checked whether a conjunction of predicates is satisfied or not (this is performed in the concrete domain, i.e. a digital finite space, and is therefore tractable).

Let us note that tractability is guaranteed by the fact that the computation of dilations has a low computational complexity. If it is computed using a brute force method, its complexity is in $O(Nn_{se})$ where N is the size of the spatial domain (i.e. number of pixels or voxels) and n_{se} is the size of the support of the structuring element (with $n_{se} \ll N$ in general). Moreover, fast propagation algorithms exist for a number of relations (see e.g. [7] for directions). Additionally, most relations can be computed on sub-sampled images to reduce the computational cost while keeping enough accuracy.

Moreover, several interesting properties for spatial reasoning can be derived from properties of mathematical morphology (for properties of mathematical morphology see [34] and [11,12] for the fuzzy case). We summarize here the most important ones:

1. $\vee$**-commutativity:** $\delta_{\nu_R^{\mathbf{F}}}(\mu_{X_1}^{\mathbf{F}}) \vee \delta_{\nu_R^{\mathbf{F}}}(\mu_{X_2}^{\mathbf{F}}) = \delta_{\nu_R^{\mathbf{F}}}(\mu_{X_1}^{\mathbf{F}} \vee \mu_{X_2}^{\mathbf{F}})$ and $\delta_{\nu_{R1}^{\mathbf{F}}}(\mu_X^{\mathbf{F}}) \vee \delta_{\nu_{R2}^{\mathbf{F}}}(\mu_X^{\mathbf{F}}) = \delta_{\nu_{R1}^{\mathbf{F}} \vee \nu_{R2}^{\mathbf{F}}}(\mu_X^{\mathbf{F}})$ and therefore we have the following rules:
   Rule 1: $\mathsf{R\_X_1 \sqcup R\_X_2 \equiv R\_(X_1 \sqcup X_2)}$.
   Rule 2: $\mathsf{R1\_X \sqcup R2\_X \equiv R12\_X}$,
   where $\mathsf{R12}$ has for representation in the concrete domain $\nu_{R1}^{\mathbf{F}} \vee \nu_{R2}^{\mathbf{F}}$.
2. $\wedge$**-monotony:** $\delta_{\nu_R^{\mathbf{F}}}(\mu_{X_1}^{\mathbf{F}} \wedge \mu_{X_2}^{\mathbf{F}}) \leq_{\mathcal{F}} \delta_{\nu_R^{\mathbf{F}}}(\mu_{X_1}^{\mathbf{F}}) \wedge \delta_{\nu_R^{\mathbf{F}}}(\mu_{X_2}^{\mathbf{F}})$, leading to:
   Rule 3: $\mathsf{R\_(X_1 \sqcap X_2) \sqsubseteq R\_X_1 \sqcap R\_X_2}$.
3. **Increasingness:** $\mu_{X_1}^{\mathbf{F}} \leq_{\mathcal{F}} \mu_{X_2}^{\mathbf{F}} \Rightarrow \forall \nu_R^{\mathbf{F}} \in \mathcal{F}, \delta_{\nu_R^{\mathbf{F}}}(\mu_{X_1}^{\mathbf{F}}) \leq_{\mathcal{F}} \delta_{\nu_R^{\mathbf{F}}}(\mu_{X_2}^{\mathbf{F}})$ and $\nu_{R_1}^{\mathbf{F}} \leq_{\mathcal{F}} \nu_{R_2}^{\mathbf{F}} \Rightarrow \forall \mu_X^{\mathbf{F}} \in \mathcal{F}, \delta_{\nu_{R_1}^{\mathbf{F}}}(\mu_X^{\mathbf{F}}) \leq_{\mathcal{F}} \delta_{\nu_{R_2}^{\mathbf{F}}}(\mu_X^{\mathbf{F}})$ which implies:
   Rule 4: $\mathsf{X_1 \sqsubseteq X_2 \Rightarrow \forall\ R,\ R\_X_1 \sqsubseteq R\_X_2}$.
   Rule 5: $\mathsf{R_1 \sqsubseteq R_2 \Rightarrow \forall\ X,\ R_1\_X \sqsubseteq R_2\_X}$.

4. **Iterativity property:** $\delta_{\nu^{\mathbf{F}}_{R_1}}(\delta_{\nu^{\mathbf{F}}_{R_2}}(\mu^{\mathbf{F}}_X)) = \delta_{\delta_{\nu^{\mathbf{F}}_{R_1}}(\nu^{\mathbf{F}}_{R_2})}(\mu^{\mathbf{F}}_X)$ hence:

   Rule 6: $\mathsf{R_1\_(R_2\_X)} \equiv \mathsf{(R_1\_R_2)\_X}$,
   where $\mathsf{R_1\_R_2}$ is the relation having as fuzzy concrete representation $\delta_{\nu^{\mathbf{F}}_{R_1}}(\nu^{\mathbf{F}}_{R_2})$.

5. **Extensivity:** $\nu^{\mathbf{F}}_R(O) = 1 \iff \forall \mu^{\mathbf{F}}_X \in \mathcal{F}, \mu^{\mathbf{F}}_X \leq_{\mathcal{F}} \delta_{\nu^{\mathbf{F}}_R}(\mu^{\mathbf{F}}_X)$, where O is the origin of $\mathcal{S}$ hence:

   Rule 7: $\mathsf{X} \sqsubseteq \mathsf{R\_X}$ for any relation defined by a dilation with a structuring element containing the origin of $\mathcal{S}$ (with membership value 1).

6. **Duality:** $\varepsilon_{\nu^{\mathbf{F}}_R}(\mu^{\mathbf{F}}_X) = 1 - \delta_{\nu^{\mathbf{F}}_R}(1 - \mu^{\mathbf{F}}_X)$ for dual t and I, which induces relations between some relations. For instance the fuzzy representation of the mereotopological relation $\mathsf{IntB\_X}$ can be written as: $\mu^{\mathbf{F}}_X \setminus (\varepsilon^{\mu_X}_{\nu_0})^{\mathbf{F}} = \mu^{\mathbf{F}}_X \wedge (\delta^{1-\mu_X}_{\nu_0})^{\mathbf{F}} = (\delta^{1-\mu_X}_{\nu_0})^{\mathbf{F}} \setminus (1-\mu_X)^{\mathbf{F}}$, where ν_0 is an elementary structuring element, hence:

   Rule 8: $\mathsf{IntB\_X} \equiv \mathsf{ExtB\_\neg X}$.

These properties provide the basis for inference processes. Other examples use simple operations, such as conjunction and disjunction of relations, in addition to these properties, to derive useful spatial representations of potential areas of target objects, based on knowledge about their relative positions to known reference objects. This will be illustrated in Section 6 on a real example.

5 Reasoning and Inference Method

A knowledge base $\langle \mathcal{T}, \mathcal{A} \rangle$ built with our description logic framework is composed of two components: the terminology $\mathcal{T}$ (i.e. Tbox) and assertions about individuals $\mathcal{A}$ (i.e. Abox). Different kinds of reasoning can be performed using description logics: basic ones, including concept consistency, subsumption, instance checking, relation checking, knowledge base consistency, and non-standard ones [25]. In [3], it has been shown that basic inference services can be reduced to Abox consistency checking. For instance, concept satisfiability, (i.e. C is satisfiable with respect to $\mathcal{T}$ if there exists a model $\mathcal{I}$ of $\mathcal{T}$ such that $C^{\mathcal{I}}$ is not empty) can be reduced to verifying that the Abox $\mathcal{A} = \{a : C\}$ is consistent.

In description logics, this reasoning is often based on tableau algorithms, also known as completion algorithms. A good overview on these algorithms can be found in [4]. The principle of these algorithms is the following: starting from an initial Abox $\mathcal{A}_0$ whose consistency is to be decided, the algorithm iteratively applies completion rules to transform the given Abox into more descendent Aboxes. The algorithm results in a tree of Aboxes (or a forest in the case of Aboxes involving multiple individuals with arbitrary role relationships between them). The algorithm stops either if the produced Abox is complete, i.e. no more rules are applicable, or all leafs in the tree are contradictory (i.e. with clashes). Tableau algorithms often assume that all the concept terms occurring in the Abox are converted in their negation normal form.

In our framework, to combine terminological with quantitative reasoning in the concrete domain, the tableau calculus proposed in [17] is slightly modified.

First, the properties of description logics derived from properties of mathematical morphology can be directly used to expand the knowledge base and to facilitate the consistency checking. For instance, each disjunction of spatial relations is replaced according to the following equivalences: $\mathsf{R\_X_1} \sqcup \mathsf{R\_X_2} \equiv \mathsf{R\_(X_1} \sqcup \mathsf{X_2)}$ and $\mathsf{R1\_X} \sqcup \mathsf{R2\_X} \equiv \mathsf{R12\_X}$.

Moreover, in our framework, we assume as an ontological commitment that each instance of abstract concept is associated with its fuzzy spatial representation in the image space with the feature hasFR. As a consequence, each step of the tableau calculus algorithm enables us to derive spatial constraints on the fuzzy concrete representation using the properties of mathematical morphology. Thus, we consider that an instance of a concept C describing an object having a spatial relation R with the instance of another concept X (i.e. $\mathsf{R\_X}$) is satisfiable if and only if the fuzzy representation in the image domain of the instance of C *fits* with the fuzzy representation of the instance of the relation $\mathsf{R\_X}$ with the function $fit : \mathcal{F} \times \mathcal{F} \rightarrow \{0, 1\}$ which verifies the strict inclusion between fuzzy sets[1]:

$$fit(\mu^{\mathbf{F}}_{X_1}, \mu^{\mathbf{F}}_{X_2}) = 1 \Leftrightarrow \mu^{\mathbf{F}}_{X_1} \leq_{\mathcal{F}} \mu^{\mathbf{F}}_{X_2}.$$

If it does not fit, a clash occurs in the Abox. More precisely, this clash occurs when we have the following assertions in the Abox:

- $s : X, (s, t) : hasSR, t : \mathsf{R\_Y}, (s, \mu_X) : hasFR, (t, \lambda^{\mu_Y}_{\nu_R}) : hasFR$ and, in the spatial domain, $fit(\mu^{\mathbf{F}}_X, (\lambda^{\mu_Y}_{\nu_R})^{\mathbf{F}}) = 0$ where λ is the fuzzy predicate enabling the building of the fuzzy representation of the spatial relation $\mathsf{R\_Y}$.

Other occuring clashes are:

- $a : C \in \mathcal{A},\ a : \neg C \in \mathcal{A}$
- $(a, x) : f \in \mathcal{A},\ (a, y) : f \in \mathcal{A}$ with $x \neq y$

As an example, let us detail some completion rules introduced in our framework for spatial reasoning:

- **Spatial Object Conjunction Rule** $(\mathcal{R}_{\sqcap})$
  - **Premise**: $(a : X \sqcap Y) \in \mathcal{A}$, $(a, \mu) : hasFR$, $(a : X) \notin \mathcal{A}$, $(a : Y) \notin \mathcal{A}$.
  - **Consequence**: $\mathcal{A}' = \mathcal{A} \cup \{a : X, a : Y\}$ and we have the spatial constraint $\mu = \mu_X \sqcap_d \mu_Y$.

  This rule means that if a conjunction is included in $\mathcal{A}$, then each part of the conjunction should be included in $\mathcal{A}$ as well (this is what is meant by "completion"). Here the novelty when using concrete domains is that the constraint $\mu = \mu_X \sqcap_d \mu_Y$ is added as well.
- **Spatial Relation 1** $(\mathcal{R}1_{R_X})$
  - **Premise**: $(a : \mathsf{R\_X}) \in \mathcal{A}$, $((a, \mu) : \mathsf{hasFR}) \in \mathcal{A}$, $\neg\exists r, (r : R) \in \mathcal{A}$, $\neg\exists x, (x : X) \in \mathcal{A}$.

[1] Other functions could be used.

- **Consequence**: $\mathcal{A}' = \mathcal{A} \cup \{r : R, (r, \nu_R) : \mathsf{hasFR}, x : X, (x, \mu_X) : \mathsf{hasFR}, (a, x) : \mathsf{hasRO}\}$ and we have the spatial constraint $\mu = \lambda_{\nu_R}^{\mu_X}$.
  This rule means that from $a : \mathsf{R\_X}$ ($a \in (\mathsf{R\_X})^{\mathcal{I}}$), we can deduce that there must exist an individual r which is an instance of the relation R having the fuzzy representation ν_R (i.e. the relation is well defined in the abstract and the concrete domains) and an individual x which is an instance of X (having the fuzzy representation μ_X), such that $(a, x) \in (\mathsf{hasRO})^{\mathcal{I}}$ and the binary predicate $\lambda_{\nu_R}^{\mu_X}$ holds in the concrete domain. λ is the fuzzy predicate enabling the building of the fuzzy representation of the spatial relation $\mathsf{R\_X}$. Note that this rule is a shortcut of the application of the conjunction rule, the exist restriction rule and the complex role rule of [17] on the assertion $a : \mathsf{R\_X}$ where $\mathsf{R\_X} \equiv \mathsf{SpatialRelation} \sqcap \exists\ (\mathsf{hasFR}, \mathsf{hasRO}.\mathsf{hasFR}).\lambda$.

– **Spatial Relation 2** ($\mathcal{R}2_{R_X}$)
  - **Premise**: $(a : \mathsf{R\_X}) \in \mathcal{A}$, $((a, \mu) : hasFR) \in \mathcal{A}$, $\exists r, (r : R) \in \mathcal{A}$ and $((r, \nu_R) : hasFR) \in \mathcal{A}, \exists x, x : X \in \mathcal{A}$ and $((x, \mu_X) : hasFR) \in \mathcal{A}$.
  - **Consequence**: $\mathcal{A}' = \mathcal{A} \cup \{(a, x) : hasRO\}$ and we have the spatial constraint $\mu = \lambda_{\nu_R}^{\mu_X}$.

– **In Spatial Relation** ($\mathcal{R}_{\exists hasSR}$)
  - **Premise**: $(a : \exists \mathsf{hasSR}.\mathsf{R\_X}) \in \mathcal{A}$, $((a, \mu) : \mathsf{hasFR}) \in \mathcal{A}$, $\neg \exists b, (b : \mathsf{R\_X}) \in \mathcal{A}$ and $((a, b) : hasSR) \in \mathcal{A}$.
  - **Consequence**: $\mathcal{A}' = \mathcal{A} \cup \{b : \mathsf{R\_X}, (a, b) : \mathsf{hasSR}, (b, \lambda_{\nu_R}^{\mu_X}) : \mathsf{hasFR}\}$ and we have the spatial constraint $fit(\mu^{\mathbf{F}}, (\lambda_{\nu_R}^{\mu_X})^{\mathbf{F}}) = 1$.

6 An Illustration in the Domain of Medical Image Interpretation

In this section, we illustrate on a simple but real example how our framework can be used to support terminological and spatial reasoning in a cerebral image interpretation application. In particular, our aim is to segment and recognize anatomical structures progressively by using the spatial information between the different structures. The recognition is performed in 3D magnetic resonance images (MRI) obtained in routine clinical acquisitions. A slice of a typical 3D MRI is shown in Figure 1.

6.1 Modeling and Reasoning

Anatomical knowledge, derived from anatomical textbooks [37] and from existing medical ontologies, such as the FMA [32], is converted in our formalism as follows. We denote respectively LV, RLV and LLV the *Lateral Ventricle*, the *Right Lateral Ventricle* and the *Left Lateral Ventricle*. The other anatomical structures we consider are the *Caudate Nucleus* (denoted CN, RCN, LCN) which are *grey nuclei* (denoted GN) of the brain. We have the following TBox ($\mathcal{T}$) describing anatomical knowledge using our spatial logic $\mathcal{ALC}(\mathbf{F})$:

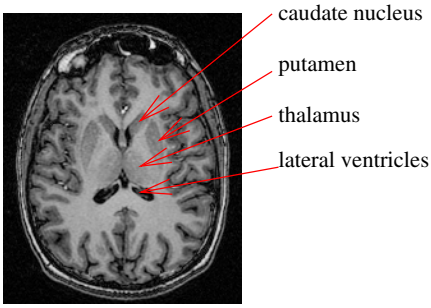


Fig. 1. An example of a slice of a 3D MRI of the brain, with a few anatomical structures indicated

$\mathsf{AnatomicalStructure} \sqsubseteq \mathsf{SpatialObject}$
$\mathsf{GN} \sqsubseteq \mathsf{AnatomicalStructure}$
$\mathsf{RLV} \equiv \mathsf{AnatomicalStructure} \sqcap \exists\ \mathsf{hasFR}.\mu_{RLV}$
$\mathsf{LLV} \equiv \mathsf{AnatomicalStructure} \sqcap \exists\ \mathsf{hasFR}.\mu_{LLV}$
$\mathsf{LV} \equiv \mathsf{RLV} \sqcup \mathsf{LLV}$
$\mathsf{Right\_of} \equiv \mathsf{DirectionalRelation} \sqcap \exists\ \mathsf{hasFR}.\nu_{IN\_DIRECTION\_0}$
$\mathsf{Close\_To} \equiv \mathsf{DistanceRelation} \sqcap \exists\ \mathsf{hasFR}.\nu_{Close\_To}$
$\mathsf{Right\_of\_RLV} \equiv \mathsf{DirectionalRelation} \sqcap \exists\ \mathsf{hasRO.RLV} \sqcap \exists\ \mathsf{hasFR}.\delta^{\mu_{RLV}}_{\nu_{IN\_DIRECTION\_0}}$
$\mathsf{Close\_To\_RLV} \equiv \mathsf{DistanceRelation} \sqcap \exists\ \mathsf{hasRO.RLV} \sqcap \exists\ \mathsf{hasFR}.\delta^{\mu_{RLV}}_{\nu_{Close\_To}}$
$\mathsf{RCN} \equiv \mathsf{GN} \sqcap \exists\ \mathsf{hasSR}.(\mathsf{Right\_of\_RLV} \sqcap \mathsf{Close\_To\_RLV})$
$\mathsf{CN} \equiv \mathsf{GN} \sqcap \exists\ \mathsf{hasSR}.(\mathsf{Close\_To\_LV})$
$\mathsf{CN} \equiv \mathsf{RCN} \sqcup \mathsf{LCN}$

The role forming predicate allows defining explicitly the dilation or erosion as a role (for instance the dilation which leads to the definition of the region to the right of the lateral ventricle):

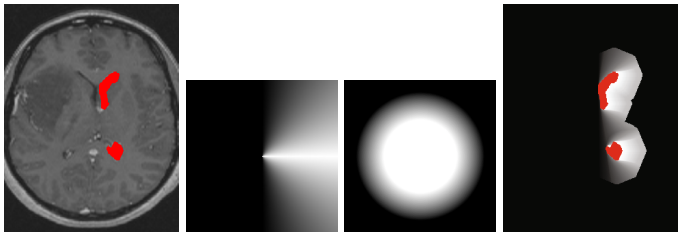
$\mathsf{dilate} \equiv (\mathsf{hasFR},\mathsf{hasRO.hasFR}).\delta$
$\mathsf{Right\_of\_RLV} \equiv \mathsf{Right\_of} \sqcap \exists\ \mathsf{dilate.RLV}$

The situation in Figure 2(a) corresponds to the following Abox $\mathcal{A}$:

c_1: RLV , (c_1,μ_{S_1}): hasFR
r_1: Right_of, $(r_1,\nu_{IN\_DIRECTION\_0})$: hasFR
r_2: Close_To, (r_2,ν_{Close_To}): hasFR

It means that we can observe an instance of the Right Lateral Ventricle (RLV) on Figure 2(a) and that we know its spatial extent in the image domain (μ_{S_1}). Moreover, the spatial relations Right_of and Close_To have been defined in the spatial domain by the learning their parameters from examples of the structuring elements $\nu_{IN\_DIRECTION\_0}$ and ν_{Close_To}.

First Scenario. In a first example, our aim is to find some spatial constraints in the image domain on an instance c_2 of the Right Caudate Nucleus (RCN) given available knowledge, i.e. $\mathcal{K} = (\mathcal{T}, \mathcal{A})$. Our objective is to infer spatial constraints on concrete domains to ensure the satisfiability of RCN given $\mathcal{K}$.



(a) (b) (c) (d)

Fig. 2. (a) The right ventricle corresponding to the image region S_1 is superimposed on one slice of the original image (3D MRI). (b) Fuzzy structuring element representing the semantics of Right_of in the image. (c) Fuzzy structuring element representing the semantics of Close_To in the image. (d) $(\delta^{\mu_{S_1}}_{\nu_{IN\_DIRECTION\_0}})^{\mathbf{F}} \wedge (\delta^{\mu_{S_1}}_{\nu_{Close\_To}})^{\mathbf{F}}$.

Using the basics of description logics reasoning, it means that the Abox enriched with $\{c_2 : \mathsf{RCN}, (c_2, \mu_{S_2}) : \mathsf{hasFR}\}$ is consistent.

First, we replace the concept RCN by its definition in $\mathcal{T}$:

$$\mathcal{A} \cup \{c_2 : \mathsf{GN} \sqcap \exists \mathsf{hasSR}.(\mathsf{Right\_of\_RLV} \sqcap \mathsf{Close\_To\_RLV}), (c_2, \mu_{S_2}) : \mathsf{hasFR}\}.$$

Then, completion rules are used to transform the given Abox into more descendent Aboxes and to derive constraints on the fuzzy representations of concepts in the concrete domain (in our case, the image domain). For instance, the completion rule adds the assertion:

$$c_2 : \mathsf{GN}, c_2 : \exists \mathsf{hasSR}.(\mathsf{Right\_of\_RLV} \sqcap \mathsf{Close\_To\_RLV})$$

and we have an individual name c_3 such that:

$$c_3 : \mathsf{Right\_of\_RLV} \sqcap \mathsf{Close\_To\_RLV}, (c_2, c_3) : \mathsf{hasSR}, (c_3, \mu_{S_3}) : \mathsf{hasFR}$$

In the spatial domain, it means that $\mu^{\mathbf{F}}_{S_2}$ and $\mu^{\mathbf{F}}_{S_3}$ must fit, i.e. $\mathrm{fit}(\mu^{\mathbf{F}}_{S_2}, \mu^{\mathbf{F}}_{S_3}) = 1$.

As c_3 is an instance of a conjunction of spatial objects, its fuzzy spatial representation in the concrete domain is:

$$((\mu_{\mathrm{Right\_of\_RLV}}) \sqcap_d (\mu_{\mathrm{Close\_To\_RLV}}))^{\mathbf{F}}$$

and we add the following assertions in the ABox:

$$c_3 : \mathsf{Right\_of\_RLV}, c_3 : \mathsf{Close\_To\_RLV}$$

The completion rule $\mathcal{R}2_{R_X}$ is applied and we have:

$$\mu_{S_3} = \delta^{\mu_{S_1}}_{\nu_{IN\_DIRECTION\_0}} \sqcap_d \delta^{\mu_{S_1}}_{\nu_{Close\_To}}$$

and the following assertion in the ABox : $(c_3, c_1) : hasRO$.

The set of inferred spatial constraints is:

$$fit(\mu^{\mathbf{F}}_{S_2}, \mu^{\mathbf{F}}_{S_3}) = fit(\mu^{\mathbf{F}}_{S_2}, (\delta^{\mu_{S_1}}_{\nu_{IN\_DIRECTION\_0}} \sqcap_d \delta^{\mu_{S_1}}_{\nu_{Close\_To}})^{\mathbf{F}}) = 1$$

and the following constraint must be verified in the image domain:

$$(\mu_{S_2})^{\mathbf{F}} \leq_{\mathcal{F}} (\delta^{\mu_{S_1}}_{\nu_{IN\_DIRECTION\_0}})^{\mathbf{F}} \wedge (\delta^{\mu_{S_1}}_{\nu_{Close\_To}})^{\mathbf{F}}.$$

The region corresponding to the the right-hand side of the inequality is illustrated in Figure 2(d). No more completion rules can be applied so the concept RCN is satisfiable given $\mathcal{K}$ if and only if this constraint is satisfied.

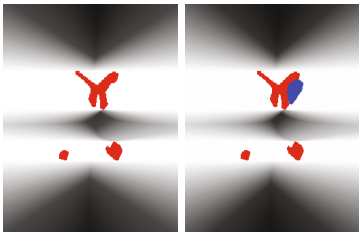
Second scenario. In this second example, illustrating disjunctions of relations, we are interested in all the instances of Caudate Nuclei in the image. A caudate nucleus is a grey nucleus which is either to the right or to the left of the lateral ventricles. This information can be represented by the following axioms:

$$\mathsf{CN} \equiv \mathsf{GN} \sqcap \exists \mathsf{hasSR}.(\mathsf{Right\_of\_LV} \sqcup \mathsf{Left\_of\_LV})$$

Using Rule 2 introduced in Section 4.2, we obtain:

$$\mathsf{Right\_of\_LV} \sqcup \mathsf{Left\_of\_LV} \equiv \mathsf{SpatialRelation} \sqcap \exists \mathsf{hasFR}.\delta^{\mu_{LV}}_{\nu_{RIGHT\_OF} \sqcup_d \nu_{LEFT\_OF}}.$$

As a consequence, the search space for the caudate nuclei is computed by: $\delta_{\nu^{\mathbf{F}}_{RIGHT\_OF} \vee \nu^{\mathbf{F}}_{LEFT\_OF}}(\mu^{\mathbf{F}}_{LV})$, which is equivalent to $\delta_{\nu^{\mathbf{F}}_{RIGHT\_OF}(\mu^{\mathbf{F}}_{LV})} \vee \delta_{\nu^{\mathbf{F}}_{LEFT\_OF}}(\mu^{\mathbf{F}}_{LV})$. The corresponding fuzzy region is represented in Figure 3(a).



(a) (b)

Fig. 3. (a) Fuzzy interpretation of the disjunction of the relations "to the left or to the right of LV". (b) One of the caudate nuclei is displayed.

7 Conclusions

In this paper, we extended the work described in [19] by the proposition of a framework for spatial relationships and spatial reasoning under imprecision based on description logics with fuzzy interpretations in concrete domains and fuzzy mathematical morphology. The resulting framework enables us to integrate qualitative and quantitative information and to derive appropriate representations of concepts and reasoning tools for an operational use in image interpretation. The benefits of our framework for image interpretation has been illustrated

in the domain of medical image interpretation for the progressive segmentation and recognition of brain anatomical structures. Future work aims at formalizing the spatial reasoning in the concrete domain as a constraint satisfaction problem and at further developing the brain imaging example.

References

1. Aiello, M., Pratt-Hartmann, I., Van Benthem, J. (eds.) Handbook of Spatial Logic. Springer (2007)
2. Aksoy, S., Tusk, C., Koperski, K., Marchisio, G.: Scene modeling and image mining with a visual grammar. In: Chen, C. (ed.) Frontiers of Remote Sensing Information Processing, pp. 35–62. World Scientific (2003)
3. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P. (eds.) The Description Logic Handbook: Theory, Implementation, and Applications. Cambridge University Press (2003)
4. Baader, F., Sattler, U.: Tableau algorithms for description logics. Studia Logica **69**, 2001 (2000)
5. Bar, M.: Visual objects in context. Nature Reviews Neuroscience **5**(8), 617–629 (2004)
6. Biederman, I.: Perceiving Real-World Scenes. Science **177**, 77–80 (1972)
7. Bloch, I.: Fuzzy Relative Position between Objects in Image Processing: a Morphological Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(7), 657–664 (1999)
8. Bloch, I.: On Fuzzy Distances and their Use in Image Processing under Imprecision. Pattern Recognition **32**(11), 1873–1895 (1999)
9. Bloch, I.: Fuzzy Spatial Relationships for Image Processing and Interpretation: A Review. Image and Vision Computing **23**(2), 89–110 (2005)
10. Bloch, I.: Spatial Reasoning under Imprecision using Fuzzy Set Theory, Formal Logics and Mathematical Morphology. International Journal of Approximate Reasoning **41**, 77–95 (2006)
11. Bloch, I.: Duality vs. Adjunction for Fuzzy Mathematical Morphology and General Form of Fuzzy Erosions and Dilations. Fuzzy Sets and Systems **160**, 1858–1867 (2009)
12. Bloch, I., Maître, H.: Fuzzy Mathematical Morphologies: A Comparative Study. Pattern Recognition **28**(9), 1341–1387 (1995)
13. Cohn, A.G., Hazarika, S.M.: Qualitative spatial representation and reasoning: An overview. Fundamenta Informaticae **46**(1–2), 1–29 (2001)
14. Colliot, O., Camara, O., Bloch, I.: Integration of Fuzzy Spatial Relations in Deformable Models - Application to Brain MRI Segmentation. Pattern Recognition **39**, 1401–1414 (2006)
15. Freksa, C.: Spatial cognition: An AI perspective. In: de Mántaras, R.L., Saitta, L. (eds.) ECAI, pp. 1122–1128. IOS Press (2004)
16. Haarslev, V., Lutz, C., Moller, R.: Foundations of spatioterminological reasoning with description logics. In: Sixth International Conference on Principles of Knowledge Representation and Reasoning, pp. 112–123. Trento, Italy (1998)
17. Haarslev, V., Lutz, C., Moller, R.: A description logic with concrete domains and a role-forming predicate operator. Journal of Logic and Computation **9**(3), 351–384 (1999)

18. Hernández-Gracidas, C., Sucar, L., Montes-y Gómez, M.: Improving image retrieval by using spatial relations. Multimedia Tools Appl. **62**(2), 479–505 (2013)
19. Hudelot, C., Atif, J., Bloch, I.: Fuzzy Spatial Relation Ontology for Image Interpretation. Fuzzy Sets and Systems **159**, 1929–1951 (2008)
20. Hudelot, C., Atif, J., Bloch, I.: A spatial relation ontology using mathematical morphology and description logics for spatial reasoning. In: ECAI-08 Workshop on Spatial and Temporal Reasoning, pp. 21–25. Patras, Greece, July 2008
21. Hudelot, C., Atif, J., Bloch, I.: Integrating bipolar fuzzy mathematical morphology in description logics for spatial reasoning. In: European Conference on Artificial Intelligence ECAI 2010, pp. 497–502. Lisbon, Portugal, August 2010
22. Inglada, J., Michel, J.: Qualitative Spatial Reasoning for High-Resolution Remote Sensing Image Analysis. IEEE Transactions on Geoscience and Remote Sensing **47**(2), 599–612 (2009)
23. Keller, J.M., Wang, X.: Comparison of spatial relation definitions in computer vision. In: 3rd International Symposium on Uncertainty Modelling and Analysis (ISUMA 1995), p. 679. IEEE Computer Society, Washington, DC (1995)
24. Kuipers, B.J., Levitt, T.S.: Navigation and Mapping in Large-Scale Space. AI Magazine **9**(2), 25–43 (1988)
25. Küsters, R.: Non-standard inferences in description logics. Springer-Verlag New York Inc., New York (2001)
26. Le Ber, F., Napoli, A.: The design of an object-based system for representing and classifying spatial structures and relations. Journal of Universal Computer Science **8**(8), 751–773 (2002)
27. Lutz, C.: Description logics with concrete domains: a survey. Advances in Modal Logics **4**, 265–296 (2003)
28. Matsakis, P., Wendling, L.: A new way to represent the relative position between areal objects. IEEE Transactions on Pattern Analalysis and Machine Intelligence **21**(7), 634–643 (1999)
29. Millet, C., Bloch, I., Hède, P., Moëllic, P.: Using relative spatial relationships to improve individual region recognition. In: 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, pp. 119–126 (2005)
30. Miyajima, K., Ralescu, A.: Spatial organization in 2D segmented images: Representation and recognition of primitive spatial relations. Fuzzy Sets and Systems **65**(2–3), 225–236 (1994)
31. Randell, D., Cui, Z., Cohn, A.: A Spatial Logic based on Regions and Connection. In: Nebel, B., Rich, C., Swartout, W. (eds.) Principles of Knowledge Representation and Reasoning KR'92, pp. 165–176. Kaufmann, San Mateo (1992)
32. Rosse, C., Mejino, J.L.V.: A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy. Journal of Biomedical Informatics **36**, 478–500 (2003)
33. Scherrer, B., Dojat, M., Forbes, F., Garbay, C.: MRF Agent Based Segmentation: Application to MRI Brain Scans. In: Bellazzi, R., Abu-Hanna, A., Hunter, J. (eds.) AIME 2007. LNCS (LNAI), vol. 4594, pp. 13–23. Springer, Heidelberg (2007)
34. Serra, J.: Image Analysis and Mathematical Morphology. Academic Press, New-York (1982)
35. Vanegas, M.C., Bloch, I., Inglada, J.: Detection of aligned objects for high resolution image understanding. In: IEEE IGARSS 2010. Honolulu, Hawai, USA, July 2010
36. Vieu, L.: Spatial Representation and Reasoning in Artificial Intelligence. In: Stock, O. (ed.) Spatial and Temporal Reasoning, pp. 5–41. Kluwer (1997)
37. Waxman, S.G.: Correlative Neuroanatomy, 24th edn. McGraw-Hill, New York (2000)

SceneNet: A Perceptual Ontology for Scene Understanding

Ilan Kadar(✉) and Ohad Ben-Shahar

Ben-Gurion University of the Negev, Beer-Sheva, Israel
ilankad@cs.bgu.ac.il

Abstract. Scene recognition systems which attempt to deal with a large number of scene categories currently lack proper knowledge about the perceptual ontology of scene categories and would enjoy significant advantage from a perceptually meaningful scene representation. In this work we perform a large-scale human study to create "SceneNet", an online ontology database for scene understanding that organizes scene categories according to their perceptual relationships. This perceptual ontology suggests that perceptual relationships do not always conform the semantic structure between categories, and it entails a lower dimensional perceptual space with "perceptually meaningful" Euclidean distance, where each embedded category is represented by a single prototype. Using the SceneNet ontology and database we derive a computational scheme for learning non-linear mapping of scene images into the perceptual space, where each scene image is closest to its category prototype than to any other prototype by a large margin. Then, we demonstrate how this approach facilitates improvements in large-scale scene categorization over state-of-the-art methods and existing semantic ontologies, and how it reveals novel perceptual findings about the discriminative power of visual attributes and the typicality of scenes.

Keywords: Scene understanding · Scene gist recognition · Scene categories · Perceptual relations · Perceptual space

1 Introduction

Scene recognition is a challenging problem in high-level computational vision, especially when the number of categories is large. While humans are able to learn and process hundreds of scene categories, the performance of existing scene recognition approaches drops dramatically as the number of categories increases [1]. In this paper we address two important limitations in the development of scene recognition systems which deal with a large number of categories: (a) the lack of proper knowledge about the ontology of scene categories; and (b) the absence of meaningful scene representation. To address both points, we introduce a new ontology database called "SceneNet" [2], a comprehensive ontology of scene categories that was derived directly from human vision via a large-scale human study. The SceneNet ontology organizes scene categories according to their *perceptual*

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 385–400, 2015.
DOI: 10.1007/978-3-319-16181-5_27

relationships and provides lower dimensional scene representation with "perceptually meaningful" (Euclidean) distance measure, all of which facilitate large-scale scene understanding operations.

While the concept of SceneNet is general, in this paper we report of SceneNet-100, the current version which consists of 100 scene categories from the SUN database [1], with the eventual goal of targeting all of its 908 categories. as we demonstrate later, in addition to significantly better computational results on various large-scale scene understanding operations, the SceneNet database provides important insights into human scene representation and organization and may serve as a key element in better understanding of this important perceptual capacity.

While traditional scene recognition approaches rarely consider the possibility of ontological organization of scene categories and indeed treat each category separately and independently [1,3–5], learning and using ontologies of categories is not new and has been explored in the context of object recognition in different forms in the past [6–17]. For example, several approaches have been developed for learning ontologies based on image features [6–10] to speed up classification for a small cost of categorization performance. However, by construction these approaches depend on the classifier and the selected features. The use of ontologies was recently promoted by exploiting WordNet [18] as a semantic relationships database for object recognition [12,14–17,19]. For example, researchers have shown the benefits of using WordNet for organizing images [16], reducing computational complexity [12], improving classification and search engine results [14,17], and learning similarity functions for better image retrieval [19]. Indeed, semantic relationships can be extracted quite conveniently from WordNet. Still, as we will show later in Sec. 2, *semantic* relationships between categories do not necessarily agree with their *perceptual* relationships.

Arguing that a proper knowledge of the ontology of scene categories should be based on perceptual criteria and inferred from human vision, our contributions and course of action are summarized as follows:

- We perform a large-scale human study to create the SceneNet-100 database, a publically available online ontology for scene understanding that organizes scene categories according to their perceptual relationships.
- We embed scene categories along with their perceptual relationships in a lower dimensional *perceptual space* with "perceptually meaningful" Euclidean distance, where each category is represented by a single prototype.
- We extend the large margin taxonomy embedding algorithm [20] to kernels for learning a non-linear mapping of scene images into the perceptual space, where each scene image is closest to its category prototype than to any other prototype by a large margin.
- We show how our approach leads to significant improvements in large-scale scene categorization over state-of-the-art methods and existing semantic ontologies.
- We exploit the proposed SceneNet database for novel perceptual findings about the discriminative power of visual attributes and the typicality of scenes.

2 SceneNet: An Online Database for Scene Understanding

Establishing a comprehensive ontology of real-world scenes is critical for further research in scene understanding. In this section we describe the construction of our large-scale *perceptual* ontology derived directly from human vision. To this end, we first perform a large-scale human study to determine the perceptual relationships between scene categories using a large set of scene categories that approximates the richness of the real world. Next, we embed the scene categories in a lower dimensional *perceptual space* which represents the perceptual relationships between classes in a meaningful and usable manner.

2.1 The Scene Categorization Pair-Matching Task

In order to measure the perceptual relationships between scene categories, we develop a crowd-source version of the "category discrimination task" recently proposed by Kadar and Ben-Shahar [21]. In particular, we presented workers on Amazon Mechanical Turk (AMT) with a *Scene Categorization Pair-Matching Game*, where participants viewed a series of *briefly* presented pairs of scenes and were asked to judge whether the two scenes belong to the same category or not (i.e., same/different forced choice task). Given the short exposure times (see below), this seems a rather challenging task. Still, evidence for parallel processing in high level categorization of natural images has already been reported, showing that humans are as fast in dual scene presentations as they are for single scene presentations [22].

The dataset for our "game" consisted of 100 scene categories borrowed from the SUN database [1]. The selection of scene categories was carefully done to focus on categories that represent minimal semantic confusion and are maximally diverse and representative of the space of scenes. Scene images were reduced to monochrome and adapted in size to 256×256 pixels.

Each trial of the experiment began with a fixation cross, followed by the simultaneous presentation of two images from our dataset for one of 3 different presentation times (PTs): 50, 100, 200 ms (Note that PTs shorter than 50 ms were excluded for inability to ensure small relative error in their value when executed on unknown computer platform and display device via AMT). The longest PT was introduced as control (i.e., "catch trials") to verify participants' awareness. High error rates in this PT would indicate unreliable participant (see below). By design, 50% of trials in our experiment constitutes a pair of images from the *same* category while the other 50% used images from *different* categories. Chance level performance was therefore 50%. After presentation for the selected PT, the two images were then masked by a pair of masks, each selected at random from a pool of eight random masks having $1/f$ amplitude spectrum. Participants pressed *Same* if they judged the two images to match in category or *Different* if not.

In the beginning of each experiment participants were shown the instructions while the system randomly selected 4 different categories out of the total 100.

Participants then completed a category familiarity procedure using 24 images (6 from each category) so that they could get acquainted with the scene category labels. Then they ran 5 practice trials so they could become familiar with the experimental procedure and task. The experiment itself followed all these steps and consists of 50 trials. Including category familiarity and practice phases, the entire experiment lasted around 5 minutes for each participant.

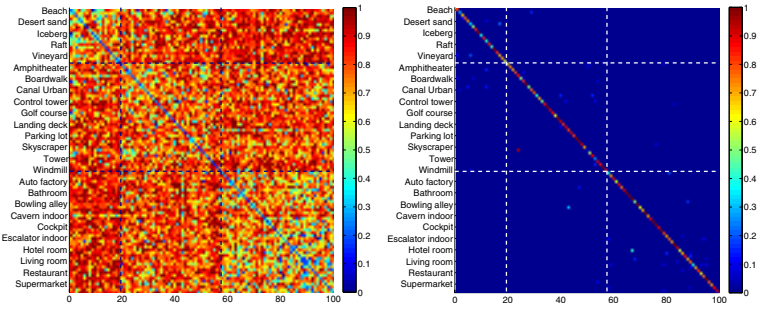
A total of **3262** workers from AMT (with better than 96% approval rate and at least 500 approved HITs) performed the game to provide a large pool of **163,100** trials. Workers were compensated with \$0.5 per HIT, plus \$0.1 bonus to high-scoring participants ($> 85\%$ average discrimination accuracy in the experiment itself) to motivate them to do their best.

The primary difficulty of using a large, non-expert work-force is ensuring that the collected data is reliable. We first analyzed participants response in the catch trials with PT=200ms to confirm participants awareness. To exclude unreliable participants, we set a threshold = 0.75% on average discrimination accuracy (i.e., at the mid point between chance level and perfect discrimination) in PT=200ms trials (and only these trials). Once unreliable participants were filtered out, we were left with **2229** reliable participants over all PTs, whose response data was then used for the analysis and construction of our database.

2.2 Building Perceptual Ontology

Having all (reliable) subjects' response in the same/different discrimination task, we then explored the perceptual relationships between all pairs of scene categories in our dataset by analyzing discrimination accuracy over all trials and PTs. In particular, we calculated subjects' probability to respond *Different* for each pair of categories. Since this probability is expected to increase when such judgment is easier, and since the latter case is expected when scenes become more "perceptually different", this probability is termed as the "perceptual distance" (PD) between pair of visual scene categories [21]. But what are the benefits that such information may provide? We compare the matrix of perceptual distance (PD) with SUN's human confusion matrix [1]. A visual depiction of the two matrices is shown in Fig. 1. Both are organized according to the main semantic classes of the SUN's manually defined ontology [1] (Natural, Manmade Outdoor, and Indoor categories). Fig. 2 further illustrates the perceptual distance with several examples.

Several conspicuous initial observations can be made upon inspection of Fig. 1. First, while the vast majority of entries in the SUN confusion matrix are zeros, the entries in the PD matrix varies between all pairs of scene categories in our dataset to obtain a more informative matrix that can be used for building ontology. Second, given the unlimited presentation time, the confusions in the SUN confusion matrix are likely to be semantic-based rather than perceptual-based (e.g., SUN workers confused between Canal-Urban and Canal-Natural while perceptually they are far apart with PD=0.77; at the same time SUN workers did not confuse Beach and Desert-Sand while perceptually they share



(a) Perceptual Distance Matrix (b) SUN Matrix [1]

Fig. 1. (a) A visualization of the *perceptual* distance between all pairs of scene categories in our dataset. The elements in the diagonal represent the perceptual distance *within* the category while all the other elements represent the perceptual distance *between* their corresponding categories. (b) A visualization of the SUN's "good workers" classification confusions between all pairs of scene categories in our dataset. In both cases, scene categories are organized to Natural (top left), Manmade Outdoor (center) and Manmade Indoor (bottom right), separated by black dashed lines. To avoid clutter only a subset of the scene category labels are presented.

similar perceptual properties, PD=0.42). Indeed, quite often the perceptual relationships are strongly inconsistent with their semantic counterparts. For example, as illustrated in Fig. 2, the "Baseball Field" category is perceptually more related to natural scene categories (e.g., "Desert-Sand", "Beach" and "Field Cultivated") than to most of the manmade categories (e.g., "Castle","Doorway Outdoor" and "Pagoda"), although semantically the opposite holds [18]. Similarly, the "Harbor" category is perceptually more related to several natural scene categories (e.g., "Lake", "Islet") than to many manmade scene categories (e.g., "Street", "Corridor") while semantically the opposite holds again [18] (see Fig. 2). It is this *new* information on scene categories that we wish to exploit for better scene understanding representation and operations.

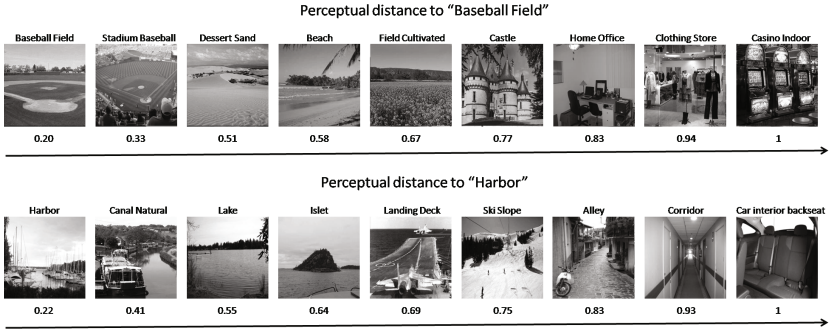


Fig. 2. Our perceptual distance metric for two scene categories examples "Baseball Field" and "Harbor". The other scene categories are labeled with their perceptual distance to the two examples.

2.3 Embedding Categories in Perceptual Space

Our next step is to embed the scene categories along with their perceptual relationships into a possibly lower dimensional perceptual space $\mathbf{R}^c$ such that their Euclidean distances are "perceptually meaningful". One way to carry such embedding is *Multidimensional Scaling (MDS)* – a technique from statistical inference and data visualization to embed a set of objects in Euclidean space while preserving their "distance" as much as possible [23]. In our case these "distances" are the perceptual distances obtained from human vision and although the dimension c of $\mathbf{R}^c$ can be lower, in our analysis we select $c = 58$ in order to preserve the perceptual distances as much as possible. This choice was mandated by the projection of the PD matrix onto the cone of positive semi-definite matrices by forcing negative eigenvalues to zero.

Let $P = [p_1, ..., p_c] \in \mathbf{R}^{c \times c}$ be a matrix whose columns consist of sought-after scene category prototypes, where p_α is the prototype that represents scene category α. We aim to embed the category prototypes such that the distance $||p_\alpha - p_\beta||_2$ reflects the perceptual distance specified in $PD(\alpha, \beta)$ (i.e., the perceptual distance between categories α and β). More formally, our problem becomes

$$P_{SceNet} = \arg\min P \sum_{\alpha,\beta=1}^{c} \left(||p_\alpha - p_\beta||^2 - PD(\alpha, \beta)\right)^2 \tag{1}$$

and it can be solved with metric multi-dimensional scaling [23]. Fig. 3 illustrates the embedding of all the scene categories in our dataset into the first two dimensions of the perceptual space. Interestingly, even with just two dimensions visualized, the results reveal that perceptual relationships do not necessarily conform to their semantic counterparts (e.g., see "Baseball Field", "Gulf Course","Green House Indoor","Phone Booth","Market Outdoor", "Shop Front"). As we demonstrate later (see section 4), the use of this perceptual ontology and space provides significant improvements in scene recognition over the SUN semantic ontology [1], suggesting that the use of the perceptual space over the semantic one should be prioritized in general. At the same time, in agreement with the Spatial Envelope model [24], the first (and most dominant) perceptual dimension appears to be related to the *Naturalness* and *Openness* attributes of the scene. While this can be observed intuitively from the visualization of the obtained perceptual space (see Fig. 3), these findings invite further analysis of the discriminative power of visual attributes (see Sec. 5.1).

Indeed, what are the benefits that such a large-scale perceptual organization may provide over previous perceptual studies that were at much smaller scale with only 8 categories [21,24,25]? We argue (and later demonstrate in Sec. 4) that the perceptual space just described is already highly useful in facilitating significant improvements in large-scale scene recognition applications over state-of-the-art methods and existing semantic ontologies. At the same time, since its larger scale set of categories provides much better representation for richness of the real world, the perceptual structure offered by SceneNet could also provide important insights into human scene organization and representation.

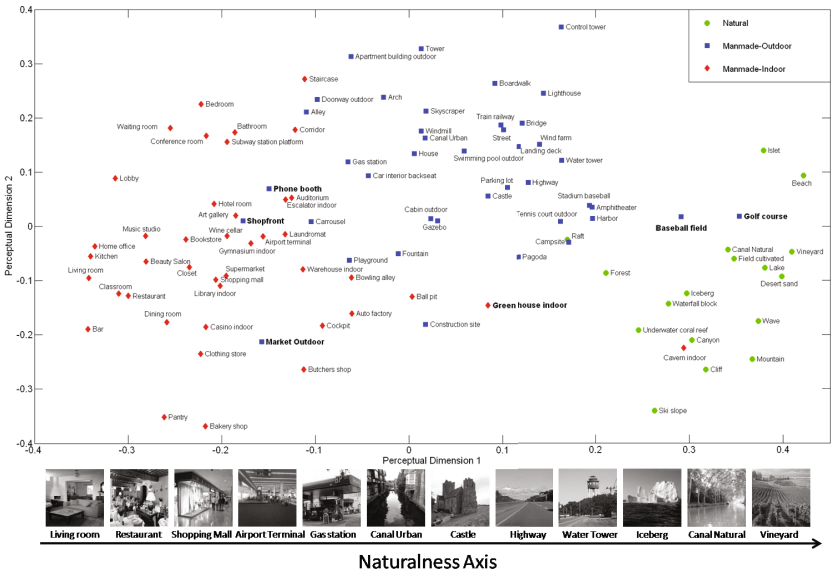


Fig. 3. Visualization of the first two dimensions of the perceptual space. Note that Natural, Manmade outdoor, and Manmade indoor scene categories are colored green, blue, red, respectively. Several categories that are referenced in the text are shown in bold face for faster localization.

In what follows we demonstrate this by exploiting our new perceptual ontology for novel findings about visual attributes and in particular about their discriminative power. For that, we combine SceneNet with the *SUN attribute database* which was recently proposed by Patterson and Hays [26].

3 Large-Scale Scene Recognition with SceneNet

With the SceneNet Database established via experimental analysis as above, we turn to discuss how it may be exploited for large-scale scene recognition. To do so we extend the document taxonomy embedding by Weinberger and Chapelle [20] to allow non-linear ontology embedding via kernels.

3.1 Scene Mapping with Regression

Let a scene image represented as feature vector $x_i \in \mathbf{R}^d$. Once we found a suitable embedding P of the scene category prototypes into $\mathbf{R}^c$, out next step is to find an appropriate linear mapping $W \in \mathbf{R}^{c \times d}$ that maps each input image x_i with category label y_i as close as possible to its category prototype p_{y_i} in the perceptual space. We can find such a linear transformation $z_i = Wx_i$ by setting

$$W = \arg\min W \sum_{i=1}^{n} ||p_{y_i} - Wx_i||^2 + \lambda ||W||^2 \tag{2}$$

where n is the number of input images and λ determines the depth of regularization on W, which is necessary to prevent potential overfitting due to the high number of features. The minimization in Eq. 2 is an instance of *linear ridge regression* whose closed-form solution is

$$W = PJX^T(XX^T + \lambda I_d)^{-1} \tag{3}$$

where $I_d \in \mathbf{R}^{d\times d}$ is the identity matrix, $X = [x_1, ..., x_n]$, and $J \in \{0,1\}^{c\times n}$, with $J_{\alpha i} = 1$ if and only if $y_i = \alpha$.

The above formulation can be easily extended to *kernel ridge regression* [27] to use kernels in the following way

$$W = PJ\kappa(x)(K + \lambda I_n)^{-1} \tag{4}$$

where $K \in \mathbf{R}^{n\times n}$ with elements $K_{ij} = \phi(x_i)^T\phi(x_j)$, $\kappa(x) \in \mathbf{R}^n$ with elements $\kappa_i = \phi(x_i)^T\phi(x)$, and ϕ is the feature mapping function.

In order to categorize a new input x_k, we first map it into the perceptual space

$$z_k = Wx_k = PJ\kappa(x_k)(K + \lambda I_n)^{-1}\ . \tag{5}$$

Then, we estimate its label $\hat{y_k}$ as the category with the closet prototype p_α, i.e., via direct nearest neighbor

$$\hat{y_k} = \arg\min \alpha ||p_\alpha - z_k||^2 \tag{6}$$

3.2 Large Margin Scene Mapping

So far we have learned the scene category prototypes P based on the SceneNet-100 ontology (i.e., directly from human vision and independent of the input data X) and found a mapping W that maps each input scene closest to the prototype of its category in the perceptual space. Still, a better and more robust generalization would allow for the correct prototype p_i to lie much closer to z_i than any other prototype p_α *by a large margin*. Moreover, it would be also preferable if perceptually dissimilar prototypes would be further separated by a *larger* margin than those that are more perceptually related. In the following we briefly describe the large margin formulation [20] to learn P and W *jointly* for better generalization.

Following Eq. 4, let us define the following matrix A:

$$A = J\kappa(x)(K + \lambda I_n)^{-1}\ . \tag{7}$$

Eqs. 4 and 7 suggest that $W = PA$ and that A is completely independent of P and can be computed directly from the input data X. Scene category prototype p_α and query z_i can now be rephrased as follows:

$$p_\alpha = Pe_\alpha \quad and \quad z_i = Px'_i \tag{8}$$

where $x'_i = Ax_i$ and $e_\alpha = [0, ..., 1, ..., 0]^T$ (i.e., vector with all zeros and a single 1 in the α^{th} position). This allow us to reduce the problem to a single optimization

problem to determine P while enforcing large margin constraints with respect to the perceptual relationships between scene categories (i.e., $PD_{y_i\alpha}$) and using regularization with weight $\mu \in [0,1]$ to ensure that P will be as similar as possible to P_{SceNet} (cf. Eq. 1). We hence define the following constrained optimization

$$\begin{aligned} &\arg\min P(1-\mu)\sum_{i,\alpha}\xi_{i\alpha} + \mu||P - P_{SceNet}||^2 \quad \textbf{subject to} \\ &(1) \quad ||P(e_{y_i} - x_i')||^2 + PD_{y_i\alpha} \leq ||P(e_\alpha - x_i')||^2 + \xi_{i\alpha} \\ &(2) \quad \xi_{i\alpha} \geq 0 \end{aligned} \tag{9}$$

where $PD_{y_i\alpha}$ now represents the (large) margin we wish to enforce on the correct classification while the slack variables $\xi_{i\alpha}$ absorb the amount of violation of prototype $p_{\alpha \neq y_i}$ into the margin of the correct prototype p_{y_i} [20].

As is later demonstrated in Sec. 4, the use of regularization term in the objective function is necessary to prevent overfitting due to the high number of features, since while the training input data might differ from the test data, the perceptual ontology remains the same. While the constraints in Eq. 9 are quadratic with respect to P and the optimization is therefore not convex, we can make Eq. 9 convex by defining $Q = P^T P$ and rewriting all distances in terms of Q while requiring that Q is positive semi-definite. With

$$||P(e_\alpha - x_i')||^2 = (e_\alpha - x_i')^T Q (e_\alpha - x_i') = ||e_\alpha - x_i'||_Q^2 \tag{10}$$

we therefore rewrite the final convex optimization problem as follows:

$$\begin{aligned} &\arg\min Q(1-\mu)\sum_{i,\alpha}\xi_{i\alpha} + \mu||Q - Q_{SceNet}||^2 \quad \textbf{subject to} \\ &(1) \quad ||(e_{yi} - x_i'||_Q^2 + PD_{y_i\alpha} \leq ||e_\alpha - x_i'||_Q^2 + \xi_{i\alpha} \\ &(2) \quad \epsilon_{i\alpha} \geq 0 \\ &(3) \quad Q \geq 0 \end{aligned} \tag{11}$$

where $Q_{SceNet} = P_{SceNet}^T P_{SceNet}$. This optimization is a particular instance of semi-definite program (SDP) [28] that can be solved very efficiently with special purpose sub-gradient solvers [29]. Once the optimal solution Q^* is found, one can obtain P with svd $Q^* = P^T P$ and the mapping W from $W = PA$.

4 Large-Scale Scene Categorization

While the goal of this paper and the SceneNet ontology and database is not necessarily limited to improved scene categorization, in this section we demonstrate how the use of the SceneNet-100 ontology embedding facilitates significant improvements in this central and popular task. To do so, we compared our approach, abbreviated here as *SceNet-Ontem*, to the one-vs-all Support Vector Machines (*SVM 1/all*) using publicly available code [1] with the descriptor and kernel defined as above. Additionally, we show that this improvement

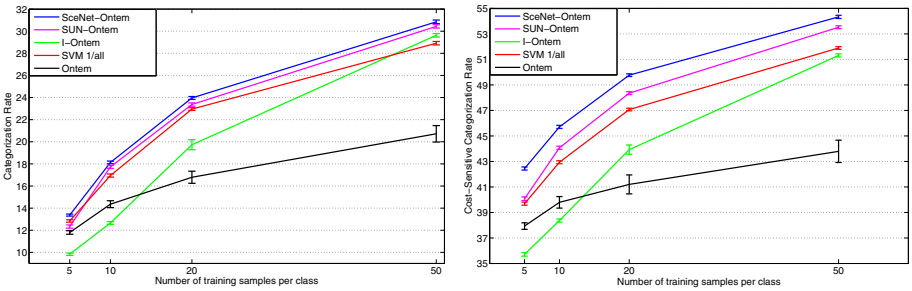


Fig. 4. Scene Categorization: Performance of all discussed algorithms (*SceNet-Ontem,SUN-Ontem,SVM 1/all, I-Ontem*, and *Ontem*) as the number of training examples is increased. **Left**: The standard categorization rate that treats each misclassification equally. **Right**: The cost sensitive categorization rate that treats each misclassification according to the perceptual distances between scene categories. Error bars represent standard error of the means.

results from the very specific ontology represented by SceneNet-100 which was inferred experimentally and reflects human perception. To do so, we also compared *SceNet-Ontem* to *I-Ontem*, an instance of our ontology embedding where the SceneNet ontology is *ignored* and P is set to be the identity matrix $I \in \mathbb{R}^{c \times c}$ such that all category prototypes are placed in constant distance from each other in the perceptual space. Furthermore, we also compared *SceNet-Ontem* to *SUN-Ontem*, an instance of our ontology embedding where the manually defined sematic ontology from SUN is used [1]. Finally, we tested another control classifier, dubbed here as *Ontem*, where the regularization term in the SDP (which is used to enforce similarity between P and P_{SceNet}) is completely ignored by setting $\mu = 0$.

In all cases we randomly split each category to disjoint training and testing sets, with $n_{training} = 5, 10, 20, 50$ and $n_{test} = 50$. The same sets where then used with the five algorithms (i.e., *SceNet-Ontem,SUN-Ontem ,SVM 1/all, I-Ontem,Ontem*) and repeated 20 times (to control for the random selection of samples). The GIST descriptor that was proposed specifically for scene recognition tasks [24] was used with an RBF kernel using the code available online [1]. We set the regularization weights to $\lambda = 1$ for the kernel ridge regression and $\mu = 0.9$ for the SDP. Preliminary experiments have shown that regularization was important but the exact settings of the λ and μ had no crucial impact, except for the need for μ to be closer to one than to zero to insure that P will be similar enough to P_{SceNet}. We evaluated the performance of two measures of categorization accuracy (each measure treats the misclassification differently): (1) the *conventional* categorization rate that weighs each misclassification equally; (2) the *cost sensitive* categorization rate that weighs each misclassification according to the perceptual distance between scene categories. The latter measure is inspired by the observation that quite often the implications of confusing certain classes is less critical than others. For example, it is easy to conceive an application where it is significantly worse to misclassify a *coastal* scene as a *kitchen* rather than a *lake*.

A comparison of the five algorithms and two measures of performance is provided in Fig. 4. As the results show, the use of the SceneNet ontology embedding yields significant improvement over *SVM 1/all* in all training set sizes. The graphs also show that the ontology used cannot be arbitrary but rather it must reflect the true relations between scene categories. Indeed, when all scene categories have constant distance from each other (as in *I-Ontem*), or when P is not required to be similar to P_{SceNet} (as in *Ontem*), performance drops significantly. Finally, while using semantic ontology (cf. *SUN-Ontem*) may improve performance compared to *SVM 1/all*, the use of the SceneNet ontology yields significantly better performance in all training set sizes.

5 Perceptual Insights

Apart from significantly better computational results, the SceneNet database could also provide important findings in human scene organization and representation. In what follows we demonstrate this by exploiting our new perceptual ontology for novel findings about the discriminative power visual attributes and the typicality of scenes.

5.1 Discriminative Power of Visual Attributes

In their recent attempt to enable deeper understanding of scenes, Patterson and Hays [26] proposed the SUN attribute database that spans over 700 categories and 14,000 images with 102 discriminative attributes related to materials, surface properties, lighting, functions/affordance, and spatial envelope properties. While they reported that scene category can be predicted only from scene attributes, using SceneNet we now attempt to determine which among these attributes are the most discriminative, or more generally, to obtain insights about the discriminative power of all attributes. Specifically, we argue that the more discriminative attributes account for most of the distance between scene categories in our perceptual space while less discriminative attributes have only minor effect on the perceptual distance between scene categories. In other words, exploring the interaction between these two databases may reveal this new information very explicitly.

Following the information in the SUN attribute database, let a scene image be represented as attribute feature vector $a \in \mathbf{R}^{102}$. For each pair of scenes a_i and a_j from two distinct scene categories α and β from SceneNet-100, we calculate the vector $d_{ij} \in \mathbf{R}^{102}$ of their absolute pairwise differences. Since d_{ij} reflects the attributes that distinguish scenes a_i and a_j, we refer to it as the attribute-distance vector between scenes a_i and a_j. Next, we trained a support vector regression (ϵ SVR) to map each attribute-distance vector d_{ij} to the perceptual distance value between their corresponding categories α and β (i.e., $PD_{\alpha\beta}$). With the trained support vector regression we could now predict the discriminative power of each visual attribute separately with the input $e_z = [0, ..., 1, ..., 0]^T$ (i.e., vector with all zeros and a single 1 in the z^{th} position for attribute z). Fig. 5

plots these results for all visual attributes in the SUN attribute database, sorted by discriminative power.

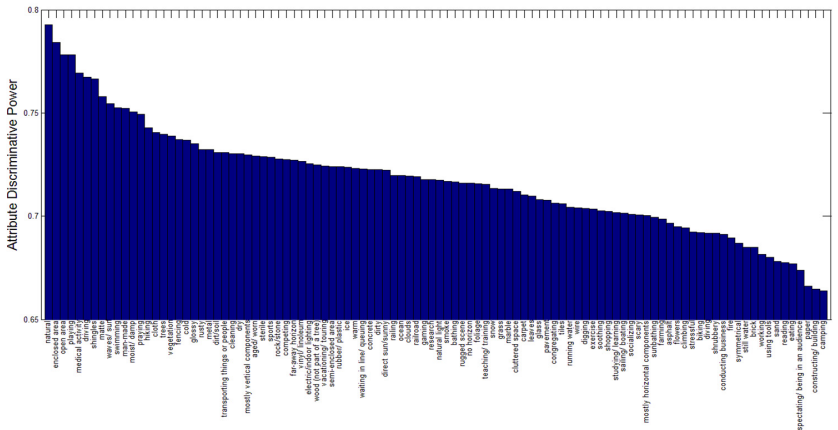


Fig. 5. The discriminative power of the visual attributes in the SUN attribute database. Consistent with the spatial envelope model [24], the most discriminative attributes are "Natural" and "Openness" (i.e., enclosed area, open area). Here, however, with a number of scene categories and attributes that is an order of magnitude larger than that of [24], we provide a rigorous perceptual basis to support and validate this claim.

Consistent with the spatial envelope model [24], the most discriminative attributes are "Natural" and "Openness" (i.e., enclosed area, open area). Here, however, with a number of scene categories and attributes that is an order of magnitude larger than [24], we provide a rigorous perceptual basis to support and validate this claim. More significantly, we provide a full evaluation of the discriminative power for the most comprehensive list of visual attributes available to-date, which enables deeper understanding of visual attributes and their relations to human perception, and could possibly facilitate better attribute-based scene representation for scene recognition.

5.2 Typicality of Scenes

One of the most robust findings in categorization is that category membership is graded and that humans seem to consistently identify typical and atypical exemplars of a category [30,31]. More importantly, there is a large body of work supporting the influence of typicality on categorization (see [32] for a detailed review). For example, it has been found that observers response time is faster for queries involving typical exemplars (e.g., "is a robin a bird") than for atypical exemplars (e.g., "is a chicken a bird") [33], and that learning of category representations is faster when typical rather than atypical exemplars are presented earlier in the sequence [34]. Arguing that a proper scene representation should take these findings into account and be consistent with them, we use our perceptual space scene representation to obtain a new perceptual typicality measure that correlates highly with the typicality ranking of humans.

The perceptual space scene representation (as opposed to discriminative methods such as SVM) has the advantage of representing a soft decision about the degree to which an image belongs to a category. We measure the image typicality by computing the distance between scene images and their categorical prototypes in perceptual space. Examples of the most typical and atypical images by our approach are shown in Fig 6.

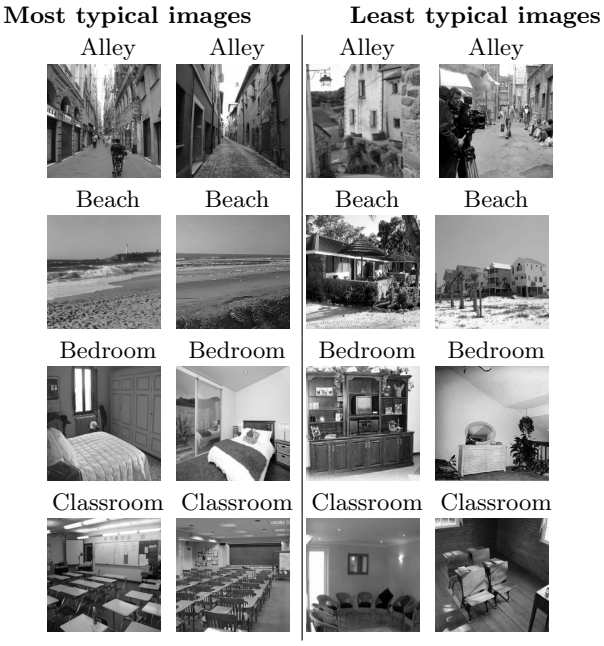


Fig. 6. Examples of the most typical and atypical images by our approach.

Next, we conducted a psychophysical experiment to compare the typicality measure based on the SceneNet perceptual space scene representation (SceneNet typicality measure) with the typicality ranking of humans. In particular, we presented workers on Amazon Mechanical Turk (AMT) with the *Image Typicality Task*, where participants were given the name of a scene category from the SUN-100 database, a short definition of the scene category, and two images. Workers were asked to select which of the two images best described the name and definition (one of the two images was drawn from the 10 most typical images by our approach and the other was drawn from the 10 most atypical images by our approach). A total of 42 workers from AMT (with better than 97% approval rate, at least 5000 approved HITs, and located in the United States) performed the task to provide a large pool of 1000 trials. Workers were compensated with $0.02 per HIT. An sample trial from the Image Typicality Task is shown in Fig 7.

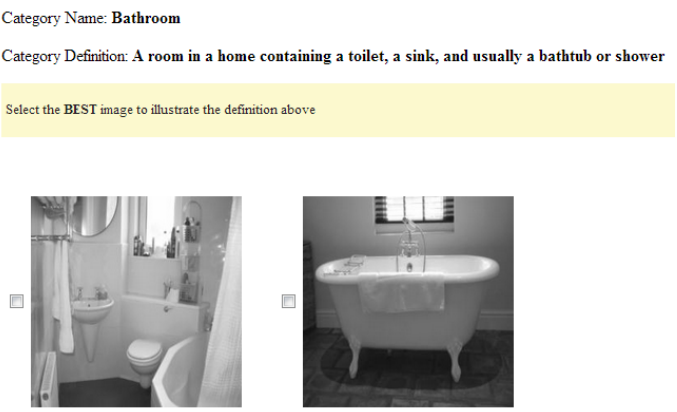


Fig. 7. An example of a trial in the Image Typicality Task. In each trial, participants were given the name of a scene category from the SUN-100 database, a short definition of the scene category, and two images. Workers were asked to select which of the two images best described the name and definition (one of the two images was drawn from the 10 most typical images by our approach and the other one was drawn from the 10 most atypical images by our approach)

Having all worker responses in the Image Typicality Task, we then assessed the degree of agreement between the SceneNet typicality measure and the human subjects' typicality ranking. Our analysis revealed that scenes that humans rate as more typical examples of their category are more likely to be close to their categorical prototype in the perceptual space. Indeed, participants selected an image from the most typical scene images by our approach in 84.38% of the trials, indicating that the SceneNet perceptual space scene representation and the SceneNet typicality measure are perceptually plausible.

6 Conclusion

In this paper we argue that in order to advance the field of scene understanding a proper knowledge of the *perceptual* ontology of scene categories is required. We have proposed such an ontology and provided SceneNet-100, an ontological database of 100 scene categories that was derived directly from human vision through a large-scale human study. The SceneNet ontology and database organizes scene categories according to their *perceptual* relationships and provides a lower dimensional scene representation with "perceptually meaningful" Euclidean distances. We show that the use of SceneNet facilitates significant improvements in large-scale scene categorization and provides important insights into human scene representation and organization for the benefit of future exploration of scene understanding.

References

1. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large scale scene recognition from abbey to zoo. In: CVPR (2010)
2. SceneNet: An Online Perceptual Ontology Database for Scene Understanding. (2013) Anonymous URL. Concealed for blind review
3. Fei-Fei, L., Perona, P.: A bayesian hierarchy model for learning natural scene categories. In: CVPR (2005)
4. Bosch, A., Zisserman, A., Muñoz, X.: Scene Classification Via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178 (2006)
6. Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: CVPR (2008)
7. Bart, E., Porteous, I., Perona, P., Welling, M.: Unsupervised learning of visual taxonomies. In: CVPR (2008)
8. Ahuja, N., Todorovic, S.: Learning the taxonomy and models of categories present in arbitrary images. In: ICCV (2007)
9. Marszałek, M., Schmid, C.: Constructing Category Hierarchies for Visual Recognition. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 479–491. Springer, Heidelberg (2008)
10. Sivic, J., Russell, B., Zisserman, A., Freeman, W., Efros, A.: Unsupervised discovery of visual object class hierarchies. In: CVPR (2008)
11. Li, L., Wang, C., Lim, Y., Blei, D., Fei-Fei, L.: Building and using a semantivisual image hierarchy. In: CVPR (2010)
12. Marszalek, M., Schmid, C.: Semantic hierarchies for visual object recognition. In: CVPR (2007)
13. Torralba, A., Fergus, R., W.T., F.: 80 million tiny images: a large dataset for non-parametric object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1958–1970 (2008)
14. Fergus, R., Bernal, H., Weiss, Y., Torralba, A.: Semantic Label Sharing for Learning with Many Categories. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 762–775. Springer, Heidelberg (2010)
15. Deselaers, T., Ferrari, V.: Visual and semantic similarity in imagenet. In: CVPR, pp. 1777–1784 (2011)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
17. Verma, N., Mahajan, D., Sellamanickam, S., Nair, V.: Learning hierarchical similarity metrics. In: CVPR (2012)
18. Miller, G.: Wordnet: A lexical database for english. In: Communications of the ACM (1995)
19. Deng, J., Berg, A., Fei-Fei, L.: Hierarchical semantic indexing for large scale image retrieval. In: CVPR (2011)
20. Weinberger, K., Chapelle, O.: Large margin taxonomy embedding for document categorization. In: NIPS, pp. 1737–1744 (2008)
21. Kadar, I., Ben-Shahar, O.: Small sample scene categorization from perceptual relations. In: CVPR, pp. 2711–2718 (2012)
22. Rousselet, G.A., Fabre-Thorpe, M., Thorpe, S.J.: Parallel processing in high-level categorization of natural images. Nature Neuroscience **5**(7), 629–630 (2002)

23. Torgerson, W.S.: Multidimensional scaling: theory and method. Psychometrika **17**(6), 401–419 (1952)
24. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision **42**(3), 145–175 (2001)
25. Greene, M., Oliva, A.: Forest before the trees: the precedence of global features in visual perception. Cognit. Sci. **58**, 137–179 (2009)
26. Patterson, G., Hays, J.: SUN attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR (2012)
27. Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: ICML, p. 515521 (1998)
28. Boyd, S., Vandenberghe, L. (eds.): Convex Optimization. Cambridge University Press (2004)
29. Weinberger, K., Saul, L.: Fast solvers and efficient implementations for distance metric learning. In: ICML, pp. 1160–1167 (2008)
30. Vogel, J., Schiele, B.: Semantic typicality measure for natural scene categorization. In: Annual Pattern Recognition Symposium (2004)
31. Ehinger, K., Xiao, J., Torralba, A., Oliva, A.: Estimating scene typicality from human ratings and image features. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 2562–2567 (2011)
32. Murphy, G.L. (ed.): The big book of concepts. MIT Press (2002)
33. Rosch, E.: Cognitive representations of semantic categories. J. Exp. Psych. (1975)
34. Mervis, C., Pani, J.: Acquisition of basic object categories. Cognit. Sci. **12** (1980)

W11 - Visual Perception of Affordances and Functional Visual Primitives for Scene Analysis

Affordances in Video Surveillance

Agheleh Yaghoobi[1], Hamed Rezazadegan-Tavakoli[2(✉)], and Juha Röning[3]

[1] Electronics Laboratory, Dehradun, India
agheleh.yaghoobi@ee.oulu.fi
[2] Center for Machine Vision Research, Oulu, Finland
hamed.rezazadegan@ee.oulu.fi
[3] Computer Science and Engineering Department, University of Oulu, Oulu, Finland
juha.roning@ee.oulu.fi

Abstract. This paper articulates the concept of affordances use as the building block of an automated video surveillance system which learns and evolves over time. It grounds its arguments on the basis of a visual attention hardware and affordances.

Keywords: Surveillance · Attention modeling · Affordances

1 The Problem Scope

Video surveillance is an old demand influencing computer vision. Despite the recent impressive progress, there is still a long way to achieve a fully automated system and many of the prerequisites in this area require careful attention and are somehow a challenge, e.g., background subtraction [2], anomaly detection [3], and etc.

Traditionally successful commercial systems (e.g. SISTORE CX series from Siemens [9]) perform a centralized scene analysis in which violation of a series of predefined rules, which are usually imposed by an operator, trigger an alert. While it seems to be a long way to achieve having automated surveillance system which evolves and learns over time, the affordances theory [4] and visual attention modeling [1,12] somehow promise to pave the way towards such an ultimate aspiration.

In this context, an automated framework is constrained by limited computational resources, volatile conditions of the environment (e.g. amount of crowd), the running site (e.g. a university campus or a factory), understanding the relation between the entities (i.e. scene understanding), and etc. Notwithstanding the difficulties, probably, one can still achieve a degree of automation by utilizing new concepts adopted from cognitive studies. Thus, the research question is: Can one utilize affordances to advance the surveillance to the next level?

2 Is Affordances a Solution?

The answer is not a straightforward affirmative phrase, neither a negative response. Although it is not the sole solution, it can be an important part of the answer

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 403–405, 2015.
DOI: 10.1007/978-3-319-16181-5_28

by facilitating efficient scene processing. It possibly provides the necessities of an ontological-based surveillance platform which evolves over time. The following elaborates how one can implement such a system.

2.1 A Rough Sketch of the Design

A practical approach for implementing such a platform consists of a combination of hardware and software units. Contrarily to the commercially available systems, the camera unit shall be updated to carry out part of the processing. Thus, an array of way more intelligent cameras will feed images and extra information, called meta-tags, to a central system. Meta-tags provide complementary information about the elements of the scene, e.g., it is a moving element, how contrasting the item is compared to its surrounding, and etc. These information are extracted using bottom-up visual attention models or salience modeling techniques, e.g. [6,8], which are embedded in hardware. Afterwards. the video frame and the meta-tags are efficiently encoded [7] to be transferred to a central processing unit.

In the central unit, each video frame is processed using a series of contextual priors [10] and atomic or compositional rules imposed by predefined affordances. Atomic rules are defined as properties of an element independent of its behavior, e.g., *move-ability* defines if an element is able to move or not. On the other hand, a compositional rule consists of several atomic affordances at the same time, e.g., *aggressive movement* can be identified by existence of fast movement towards the site which requires identification of a moving element with particular movement pattern and characteristic.

Contextual priors are also important. In essence, they define the operation environment of the system, more accurately each camera unit. In such a system, two kind of priors exists, 1) excitery priors and 2) inhibitory priors. The first defines the existence of an element and its properties such as probable location. For instance, if a camera shall expect human presence in its field of view or vehicles and where should look for them. Contrarily, the inhibitory priors ban occurrence or existence of an element. In a nutshell, contextual priors ease the building process of an ontological tree [11] that helps understanding the environment.

The ontology evolves either via user interaction or recognition modules which identify the presence of elements and their interactions with the advent of affordances. While a user can alter both domain and conceptualization of the system, the recognition modules only affect changes in the domain (i.e. field of view of each camera). The automatic ontology enrichment and evolving is possible via graphical models in which a mapping between the ontology and an appropriate Bayesian network is derived, e.g. [5].

In the end, the outcome will be a set of hardware and software that vigilantly performs video surveillance, easily adapts to environment, and enhances over time. Also, a series of new affordances defined in the context of the entities of assigned task are expected. Eventually, a system, which integrates bottom-up

visual attention techniques, contextual priors and affordances, will be implemented to derive ontological scene understanding in a less general scenario. In summary, the broad vision is a step toward convergence of cognitive sciences, electrical engineering and computer vision.

References

1. Borji, A., Rezazadegan-Tavakoli, H., Sihite, D.N., Itti, L.: Analysis of scores, datasets, and models in visual saliency prediction. In: ICCV (2013)
2. Bouwmans, T.: Recent advanced statistical background modeling for foreground detection: a systematic survey. Recent Patents on Computer Science **4**(3), 147–176 (2011)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Computing Surveys **41**(3) (2009)
4. Gibson, J.J.: The theory of affordances. In: Perceiving, Acting, and Knowing (1977)
5. Ishak, M.B., Leray, P., Amor, N.B.: A two-way approach for probabilistic graphical models structure learning and ontology enrichment. In: KEOD, pp. 189–194. SciTePress (2011)
6. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell **20**(11), 1254–1259 (1998)
7. Ma, T., Hempel, M., Peng, D., Sharif, H.: A survey of energy-efficient compression and communication techniques for multimedia in resource constrained systems. Commun. Surveys Tuts **15**(3), 963–972 (2013)
8. Mancas, M., Riche, N., Leroy, J., Gosselin, B.: Abnormal motion selection in crowds using bottom-up saliency. In: ICIP, pp. 229–232 (2011)
9. Siemens: SISTORE CX, Configuration Manual. Siemens Building Technologies AG
10. Torralba, A.: Contextual priming for object detection. Int. J. Comput. Vision **53**(2), 169–191 (2003)
11. Town, C.P.: Ontology based Visual Information Processing. Ph.D. thesis, University of Cambridge (2004)
12. Tsotsos, J.K.: A Computational Prespective on Visual Attention. The MIT Press, Cambridge (2011)

Affordance-Based Object Recognition Using Interactions Obtained from a Utility Maximization Principle

Tobias Kluth(✉), David Nakath, Thomas Reineking, Christoph Zetzsche, and Kerstin Schill

Cognitive Neuroinformatics, University of Bremen, Enrique-Schmidt-Straße 5, 28359 Bremen, Germany
tkluth@math.uni-bremen.de

Abstract. The interaction of biological agents within the real world is based on their abilities and the affordances of the environment. By contrast, the classical view of perception considers only sensory features, as do most object recognition models. Only a few models make use of the information provided by the integration of sensory information as well as possible or executed actions. Neither the relations shaping such an integration nor the methods for using this integrated information in appropriate representations are yet entirely clear. We propose a probabilistic model integrating the two information sources in one system. The recognition process is equipped with an utility maximization principle to obtain optimal interactions with the environment

Keywords: Affordance · Sensorimotor object recognition · Information gain

1 Introduction

The ability of humans to reliably recognize objects in the environment is still a largely unsolved problem for artificial systems. Usually, object recognition is understood as a classification problem where a fixed mapping from feature vectors to object classes is learned. This is in line with the classical view of perception as the input from world to mind and of action as the output from mind to world [6], which implies a dissociation between the capacities for perception and action. However, there is strong evidence that object recognition cannot be understood independently of the interaction of an agent with its environment [8]. "Active perception" approaches [1,2] take this partially into account, but actions are not merely means for acquiring new information, they also provide evidence themselves for the recognition [5]. What an agent perceives is thus also determined by what it does or what it is able to do [8].

Research in the direction of affordances by Gibson [3] also provides evidence that affordances are key ingredients of the perceptual process. A variety of studies regarding object recognition show that the visual information of a manipulable

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 406–412, 2015.
DOI: 10.1007/978-3-319-16181-5_29

object causes an activation of representations of actions which can typically be executed on the object [4]. The advantageous interplay between sensory and action information, which was also recognized by Neisser [7], should be considered in the recognition process.

The strong interrelation between motor actions and sensory perceptions is basis for the concept of a sensorimotor representation [8,10]. Similarly to the affordance point of view the processing stages for sensory and motor information are not separated. The approach including the actions in the representation gives the opportunity to choose the next action such that a specific objective is pursued. Generally, the problem of action selection can be solved in numerous ways, but as information gathering should be one major purpose of motor actions it is appropriate to consider an information-theoretic utility function. Prior research in this area often found that the principle of *information gain* is well suited to select an appropriate next action.

In this paper, we propose a system for object recognition which incorporates both the information gain principle from sensorimotor systems and the theoretical concept of affordances. Building upon the investigations in [11], we developed a sensomotoric probabilistic reasoning system for affordance-based object recognition. The design of our architecture is motivated by two main goals: i) to provide a clear relation to Bayesian inference approaches, and ii) to enable a comparison between the classic sensory approach and the sensorimotor, affordance-oriented approach within one common probabilistic framework.

2 Object Recognition System

The system described in the following is a generic architecture (see Fig. 1). The recognition loop starts out with a particular pose of an object which is perceived by a sensor. The sensor passes its raw data to the sensory processing module. After processing, the sensory data becomes part of a new sensorimotor feature, which is then fed into the probabilistic reasoning module. The processed sensory data are also used to obtain a set of possible interactions, i.e., the affordances offered by the sensory data related to the abilities of the agent. The Bayesian inference module calculates the new posterior distribution based on a previously-learned sensorimotor representation. This representation contains the learned perceptual consequences of an interaction in a given state for every object class. The posterior distribution constitutes the current belief of the system. This belief is used by the information gain strategy to choose an optimal next action from the set of possible interactions. The selected interaction then also becomes part of the sensorimotor feature and is subsequently executed by the agent. The whole process results in a new state, which in turn delivers new raw sensory data to enter the next cycle of the recognition loop.

More formally speaking, the system depends on an *agent*, which can be controlled such that it perceives information about a specific aspect of the world. In Fig. 1, the two arrows pointing from the states to the sensory processing module correspond to the mapping $A : U \times X \rightarrow R$, where U is the space of all

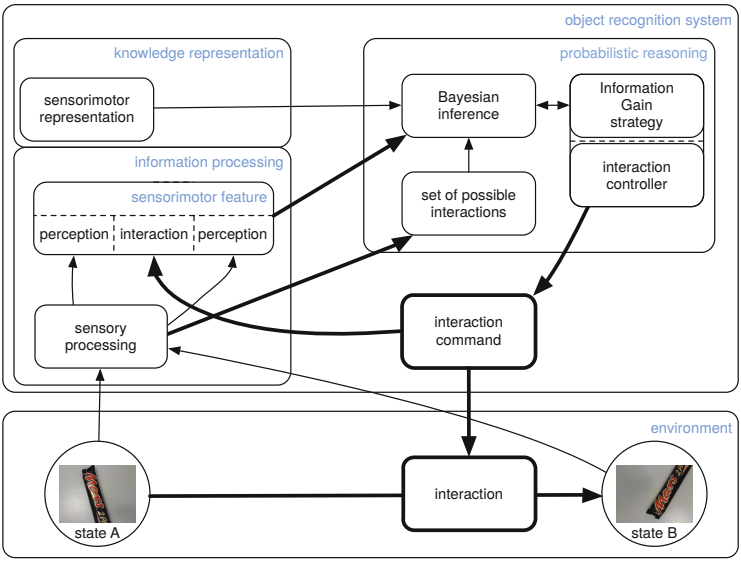


Fig. 1. Architecture of the object recognition system

interactions which are currently possible, X is the state space, and R is the raw sensor data space.

The system has no explicit knowledge about the actual state, and the currently possible interactions U. The possible interactions are of course dependent on the state but nevertheless both information must be obtained from the sensor data. The sensoric dependency on the state is formalized by the mapping $U : X \to \mathcal{P}(\Omega_U)$, where Ω_U is the set of all possible interactions and $\mathcal{P}$ denotes the power set. Note that U comprises the link from the state to the sensory processing module and the following link to the set of possible interactions in Fig. 1, i.e., the perceived affordances. Assuming that the output of the function U is given, we write U instead of $U(x),\ x \in X$, for convenience. Considering the state-agnostic behavior, the influence of the agent can be formally redefined to $A_x : U \to R$ with $A_x(u) := A(x, u) = r,\ x \in X,\ u \in U(x),\ r \in R$. The only time-dependent variables are the state x and the interaction u.

The raw sensor data $r \in R$ is fed into the *sensory processing* (SP) which mainly extracts the relevant features belonging to a feature space F, i.e., $SP : R \to F$. Subsequently, the quantization operation $Q_S : F \to S$ maps the features to a specific feature class in the finite space S. The possible interactions are mapped with $Q_M : \Omega_U \to M$ to the finite set of interactions M to yield a manageable product space of sensory information and actions. The results of these quantizations then become part of a sensorimotor feature (SMF). The single quantizations are represented in Fig. 1 by the arrows from the sensory processing module and the interaction command to the sensorimotor feature which is defined as the triple

$$SMF_i := (s_{i-1}, m_{i-1}, s_i), \tag{1}$$

where $m_{i-1} := Q_M(u_{i-1})$ is the interaction between the sensor information s_{i-1} and s_i at time step t_{i-1} and t_i. The whole chain of operations to obtain the sensor information at a time step t_i can be described by $s_i := (Q_S \circ SP \circ A_x)(u_{i-1})$.

The *knowledge representation* is comprised of the learned sensorimotor representation (SMR), which is a full joint probability distribution of SMFs and the classes represented by the discrete random variable Y. Every possible SMF is generated on a set of known objects in a training phase. This means that, from every possible state x, the sensory consequence of every possible action u is perceived, resulting in

$$SMR := P(SMF_i, Y) = P(S_{i-1}, M_{i-1}, S_i, Y). \tag{2}$$

The *probabilistic reasoning* module consists of a Bayesian inference approach accompanied by an information gain strategy. They rely on bottom-up sensory data and top-down information from the knowledge representation. The information gain strategy can choose an optimal next interaction for the agent, thus improving the input of the following Bayesian inference step.

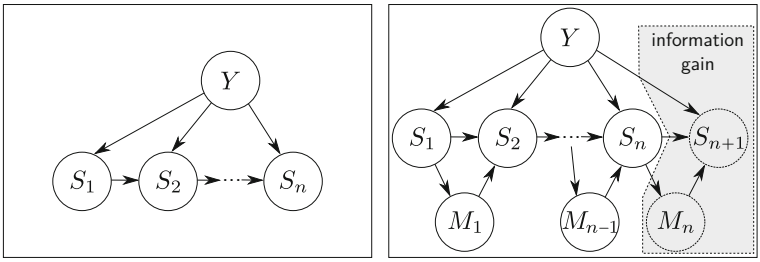
3 Model Implementation and Outlook

Based on the schematic outline presented above, we applied our system to object recognition using a robotic arm interacting with objects in 3D space. We used a discrete set of interactions M of a robotic arm with an object which comprise the relative position/pose of the visual sensor to the object ($\Omega_U = M$, $Q_M = Id$).

In the learning phase, features are extracted from the training data (images from every reachable state). GIST-features [9] are used to describe the sensory input, i.e., defining SP. The quantization Q_S is then learned by performing a k-means clustering on the extracted features. In order to build the individual SMFs, features are extracted and the results are assigned to clusters with the aid of the previously defined mapping Q_S. These labels are combined with the corresponding interactions in a set of SMFs. Finally, all generated SMFs are stored in a Laplace-smoothed SMR.

The probabilistic reasoning is comprised of a Bayesian inference module in the form of a dynamic Bayesian network (BN) and a corresponding information gain strategy. Two of these probabilistic reasoning modules were implemented to examine the difference between *sensor networks*, which only take into account sensory information (which also implies that no information gain strategy is used), and *affordance-based networks*, which integrate sensory perceptions and interactions. The object recognition in the sense of computer vision then takes place by classification which is performed by choosing the class with the maximum posterior probability.

The representative of the *sensor networks* is Bayesian network 1 (BN1) (see Fig. 2a), which resembles an extended naive Bayes model that additionally allows



(a) BN1 **(b)** BN2

Fig. 2. In Bayesian network BN1 (a) sensory information S_n is processed only to obtain the object class Y. Bayesian network BN2 (b) is equipped with the information gain startegy which takes also the action M_n into account.

for statistical dependencies between the preceding and the current sensor information, s_{i-1} and s_i, resulting in

$$P(y|s_{1:n}) \propto P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, y), \tag{3}$$

where $s_{1:n}$ is a short notation for the n-tuple $(s_1, \ldots, s_n)$.

Bayesian network 2 (BN2) (see Fig. 2b) uses the full information of the SMF and therefore belongs to the *affordance-based networks*. The assumption that the current sensory input s_i depends on the action m_{i-1} integrates sensory perceptions and actions in the recognition process and permits the application of the information gain strategy to choose the next optimal interaction. Additionally, it is assumed that the action m_{i-1} statistically depends on the preceding sensory input s_{i-1}. The inference can then be conducted by

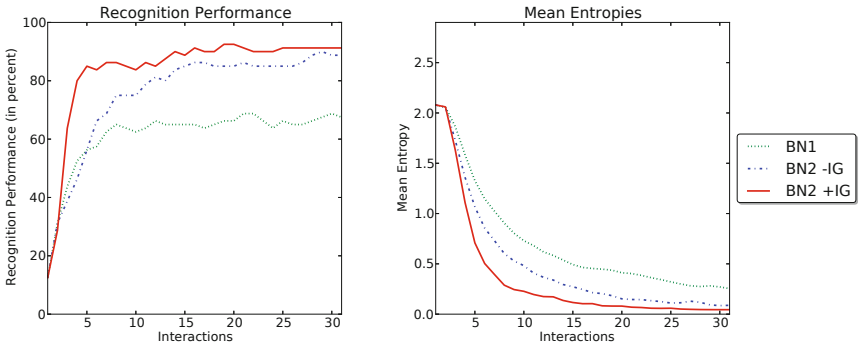
$$P(y|s_{1:n}, m_{1:n-1}) \propto P(y)P(s_1|y)\prod_{i=2}^{n} P(s_i|s_{i-1}, m_{i-1}, y)P(m_{i-1}|s_{i-1}). \tag{4}$$

The strategy for action selection should satisfy two main properties: (i) The strategy should adapt itself to the current belief state of the system and (ii) the strategy should not make decisions in an heuristic fashion but tightly integrated into the axiomatic framework used for reasoning. The information gain strategy presented in this paper complies with both of these properties.

The information gain IG of a possible next action m_n is defined as the difference between the current entropy and the conditional entropy,

$$IG(m_n) := H(Y|s_{1:n}, m_{1:n-1}) - H(Y|S_{n+1}, m_n, s_{1:n}, m_{1:n-1}). \tag{5}$$

This is equivalent to the mutual information of Y and (S_{n+1}, m_n) for an arbitrary m_n. As the current entropy $H(Y|s_{1:n}, m_{1:n-1})$ is independent of the next action



(a) Recognition performance **(b)** Mean entropies

Fig. 3. Results of the robotic arm evaluation (8 object classes, 10 objects per class, 30 discrete viewpoints). BN 1 and 2 -IG executed random movements while BN2 +IG executed information-gain-guided movements.

m_n the most promising action m^* can be calculated by minimizing the expected entropy with respect to S_{n+1},

$$m_n^* = \arg\min_{m_n}(\underset{S_{n+1}}{E}[H(Y|s_{1:n}, S_{n+1}, m_{1:n})]). \tag{6}$$

Because the sensory input s_{n+1} is not known prior to executing m_n, the expected value over all possible sensory inputs s_{n+1} is taken into account. The selected action $m^* \in M$ is integrated into the next sensorimotor feature. The inverse mapping of Q_M can then be used to obtain a top-down interaction command $u \in U$, which is then executed by the agent.

Preliminary results are shown in Fig. 3. In the future, we plan to conduct a more extensive evaluation of our approach (using different sensory features) by comparing it to established object recognition approaches on a larger data set. Furthermore we want to extend our approach by a saliency feature detector.

Acknowledgments. This work was supported by DFG, SFB/TR8 Spatial Cognition, project A5-[ActionSpace], and DLR projects "EnEx" and "KaNaRiA".

References

1. Aloimonos, J., Weiss, I., Bandyopadhyay, A.: Active vision. International Journal of Computer Vision **1**(4), 333–356 (1988)
2. Bajcsy, R.: Active perception. Proceedings of the IEEE **76**(8), 966–1005 (1988)
3. Gibson, J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1992)
4. Grèzes, J., Decety, J.: Does visual perception of object afford action? Evidence from a Neuroimaging study. Neuropsychologia **40**(2), 212–222 (2002)

5. Helbig, H.B., Graf, M., Kiefer, M.: The role of action representations in visual object recognition. Experimental Brain Research **174**(2), 221–228 (2006)
6. Hurley, S.L.: Consciousness in action. Harvard University Press (2002)
7. Neisser, U.: Cognition and reality: Principles and implications of cognitive psychology. WH Freeman/Times Books/Henry Holt & Co. (1976)
8. Noë, A.: Action in Perception. MIT Press (2004)
9. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in Brain Research **155**, 23–36 (2006)
10. O'Regan, J.K., Noë, A.: A sensorimotor account of vision and visual consciousness. Behavioral and Brain Sciences **24**(5), 939–972 (2001)
11. Schill, K., Umkehrer, E., Beinlich, S., Krieger, G., Zetzsche, C.: Scene analysis with saccadic eye movements: top-down and bottom-up modeling. Journal of Electronic Imaging **10**(1), 152–160 (2001)

Detecting Fine-Grained Affordances with an Anthropomorphic Agent Model

Viktor Seib(✉), Nicolai Wojke, Malte Knauf, and Dietrich Paulus

Active Vision Group (AGAS), University of Koblenz-Landau,
Universitätsstr. 1, 56070 Koblenz, Germany
{vseib,nwojke,mknauf,paulus}@uni-koblenz.de
http://agas.uni-koblenz.de

Abstract. In this paper we propose an approach to distinguish affordances on a fine-grained scale. We define an anthropomorphic agent model and parameterized affordance models. The agent model is transformed according to affordance parameters to detect affordances in the input data. We present first results on distinguishing two closely related affordances derived from *sitting*. The promising results support our concept of fine-grained affordance detection.

Keywords: Affordances · Fine-grained affordances · Visual affordance detection · Object classification

1 Introduction

We address the task of detecting affordances on a fine-grained scale in a home environment. Affordances as defined by Gibson [3], [4] inherit the concept of direct perception and the complementary nature of an agent and its environment. Whether or not direct perception can be used in computer vision is still an open debate as discussed e.g. by Şahin et al. [6] and Chemero et al. [2].

In the presented approach we exploit the complementary nature of an agent and its environment. We propose to model the agent as an anthropomorphic body and define a set of parameterized affordance models. A home or office environment for humans must reflect human body characteristics. A system equipped with these models is thus able to detect affordances in the environment.

We present first results on two closely related affordances: *sitting without backrest* and *sitting with backrest* which stem e.g. from the objects stools and chairs, respectively. Traditionally, these two affordances would be both *sitting*. Our results suggest that objects used by humans in a home environment provide distinct affordances on a fine-grained scale.

The remaining of this paper is structured as follows. A brief overview on related work is given in Sect. 2 and a detailed explanation of our method is provided in Sect. 3. Section 4 presents and Sect. 5 discusses the results that we obtained from various test objects. Finally, Sect. 6 concludes this paper and gives an outlook to our ongoing work.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 413–419, 2015.
DOI: 10.1007/978-3-319-16181-5_30

2 Related Work

There have been many approaches to apply ideas coming from the theory of affordances to robotics. We shortly review some approaches exploiting only visual hints for affordance detection. Hinkle and Olson [5] propose a method that uses physical simulation to extract an object descriptor. The simulation consists of spheres falling onto an object from above. A feature vector is extracted from each object depending on where the spheres come to rest. Subsequently, objects are classified as cup-like, table-like or sitable.

A method for office furniture recognition is presented by Wünstel and Moratz [7]. Object classes are modeled explicitly in a graph structure, where nodes represent the object's parts and edges the spatial distances of those parts. Affordances are used to derive the spatial arrangement of the object's components.

Bar-Aviv and Rivlin [1] use an embodied agent to classify objects. The object in question is moved to a virtual simulation environment where the compatibility of different agent poses with the object is tested. The object is assigned the label of the hypothesis with the highest score.

Similar as Wünstel and Moratz [7] we use a plane segmentation approach in our method. However, we encode the spatial information needed for affordance detection in an anthropomorphic agent model rather than creating explicit object models. Contrary to Bar-Aviv and Rivlin [1] who also use an embodied agent, our method operates directly on the data. We do not segmented and move the objects to a simulation environment where they are tested to belong to different classes. In our case, segmentation is a direct consequence of the detected affordances.

3 Model Definitions for Fine-Grained Affordance Detection

In this section we describe our method of detecting fine-grained affordances with an anthropomorphic agent model. Our approach is based solely on visual data.

3.1 Agent and Affordance Models

Our anthropomorphic body model is defined as a directed acyclic graph $\mathcal{H}$ (Fig. 1). In this graph, nodes represent joints in a human body and edges represent parameterized spatial relations between these joints. The nodes contain information on how the joints can be revolved without harming the human. The environment $\mathcal{E}$ is a set of features. So far, we limit the features to arbitrarily oriented planes that are segmented from the input data.

A fine-grained affordance is a property of an affordance that specializes the relation of an agent and its environment. We take the *sitting* affordance as an example. The affordance *sitting* is a generalization of more precise relations that an agent and its environment form. In this paper, we demonstrate our ideas by distinguishing between the fine-grained affordances *sitting without backrest* and *sitting with backrest*.

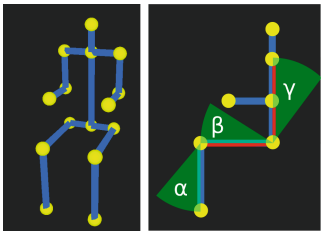


Fig. 1. The anthropomorphic agent model: nodes are depicted in yellow, edges in blue. A perspective view of the model in a sitting pose is shown on the left. This pose serves as the initial body pose for affordances derived from *sitting*. Control areas (red) that must be supported by features as well as relevant joint limits (green) for the *sitting with backrest* affordance are displayed on the right.

Fig. 2. The top row shows example objects from our evaluation. Chairs and stools served for the two fine-grained affordances. The bottom row presents affordance detection results: the *sitting with backrest* affordance is shown in green, whereas the *sitting without backrest* affordance is shown in blue.

3.2 Detecting Affordances

The algorithm used for affordance detection is outlined in Alg. 1. It operates on single scene views from an RGB-D camera. The affordance models f_1 and f_2 denominate the *sitting without backrest* and *sitting with backrest* affordances, respectively. First, plane segmentation on the input point cloud $\mathcal{P}$ is performed. Then, all horizontal planes from the abstract view of the environment $\mathcal{E}$ are tested to comply with the agent model $\mathcal{H}$ and the affordance model f_1 as described in Sec. 3.3. Every plane that affords sitting for the given agent is added to the set S of sitable planes. Then, for each sitable plane s vertical planes in close proximity are found. Each of the vertical planes is again tested to comply with the agent and the affordance models. If the sitable plane s and the vertical plane v together afford f_2 for the given agent, both planes are added to the output point cloud $\mathcal{P}_2$. Otherwise, the sitable plane s is added to the output point cloud $\mathcal{P}_1$ which contains points for the affordance f_1. Thus, the algorithm additionally provides a segmentation of the found affordances. Please note that in Fig. 2 the bounding box around s was extended to the ground plane and all points inside this bounding box were added to $\mathcal{P}_2$ and $\mathcal{P}_1$ for visualization purposes.

3.3 Checking Model Parameters

In Alg. 1 model checking is carried out in two cases. First, to assure that a plane p is sitable and second to assure that a plane v can support the agent's back while it is seated on p.

By varying the angle parameters α and β in the sitting affordance with the constraint that the agent's feet always touch the floor a valid range for the height of the sitting plane is found. Similarly, for the plane v the angle γ is varied to check whether the sitting agent can make use of it.

The dimensions for both planes are directly derived from the agent model. They are given by the body width, the length of the thigh and the height of the back, respectively. Since the size of the planes does not have to match the model proportions exactly to allow sitting or back support, the size is considered valid if it is between the D_{min} and D_{max} percentage parameters of the affordance. For example, for a model width of 0.4 m and $D_{min} = 0.7$ and $D_{max} = 1.3$, the allowed plane sizes would be between 0.28 m and 0.52 m.

Algorithm 1 Fine-grained Affordance Detection.

Require: Point cloud $\mathcal{P}$, Affordance models f_1, f_2, Agent model $\mathcal{H}$
Ensure: Point cloud with segmented affordances $\mathcal{P}_1$ and $\mathcal{P}_2$
$\mathcal{E} \leftarrow segmentPlanes(\mathcal{P})$
$S \leftarrow \emptyset$
for all horizontal planes $p \in \mathcal{E}$ **do**
  if $supportsModels(p, \mathcal{H}, f_1)$ **then**
    $S \leftarrow S \cup p$
  end if
end for
for all $s \in S$ **do**
  $V \leftarrow$ vertical planes $\in \mathcal{E}$ close to s
  if $supportsModels(v, \mathcal{H}, f_2)$ **and** v is biggest plane $\in V$ that supports the models **then**
    $\mathcal{P}_2 \leftarrow \mathcal{P}_2 \cup v$
    $\mathcal{P}_2 \leftarrow \mathcal{P}_2 \cup s$
  else
    $\mathcal{P}_1 \leftarrow \mathcal{P}_1 \cup s$
  end if
end for

4 Experiments and Results

For our experiments we acquired data from 17 different chairs and 3 stools to represent fine-grained affordances. From these data, we extracted 247 different views of the chairs and 47 different views of the stools. Example views of these objects are shown in Fig. 2. Additionally, negative data (i.e. data without the two affordances) from 9 different furniture objects was obtained and 109 views of these objects extracted. Negative data includes objects like a bed, desks, tables, dressers and a heating element. The whole evaluation dataset contains 403 scene views with 294 positive and 109 negative data examples.

The influence of the five parameters (the angle parameters α, β, γ and the size range parameters D_{min}, D_{max}) was tested with 59 different parameter sets. In the first round the parameters were varied systematically over a wide range to obtain 35 different configurations for evaluation. For the second round we inspected the best parameters from the first round and created 24 additional configurations close to the best configurations from the first round. As Hinkle and Olson [5] we included the F-measure, a harmonic mean between precision and recall, in our evaluation. Precision, recall and F-measure for the second round of experiments are shown in Fig. 3. Best results for both fine-grained affordances are shown in Tab. 1, while the best parameter values are presented in Tab 2.

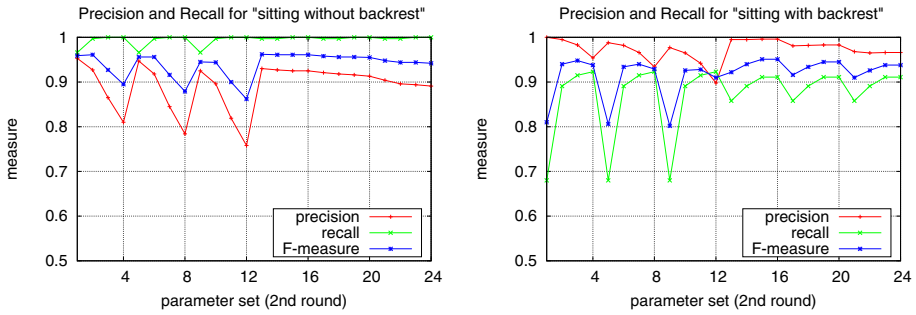


Fig. 3. Illustration of the precision and recall for both fine-grained affordances over all tested parameter sets in the second round of experiments.

Table 1. Best F-measure values for both affordances. In each line also the result of the other affordance is shown.

	sit. w/o backrest	sit. with backrest
best w/o backr.	0.962	0.922
best with backr.	0.961	0.951

Table 2. Affordance model parameters that result in highest F-measure values for the detection of both fine-grained affordances

α, β	γ	D_{min}	D_{max}
30°	35°-40°	0.5	1.4-1.6

5 Discussion

For the *sitting without backrest* affordance (in our test cases derived from the stool objects) the quality of the results was best for α and β between 20° and 40°. As is shown in Fig. 1 these parameters change the angles in the agent's legs. With the constraint that the agent's feet always touch the ground for comfortable sitting, α and β directly influence the allowed heights of the sitting planes. We observed a drop of performance for values higher than 40°. This is due to numerous planes in the datasets that are of low height, but otherwise would allow sitting. Also, if D_{min} is chosen to be only little restrictive (below 0.5) too many small planes and clutter are considered "big enough" for sitting, resulting in a drop of precision. On the other hand, D_{max} has only a moderate effect.

The *sitting with backrest* affordance is additionally influenced by the parameter γ for the inclination of the backrest. For γ, higher values than 40° cause many false positives. The additional effect of D_{min} and D_{max} include the valid dimensions for the size of the backrest that is compared with the agent's back. Again, D_{min} has more significant effects on the results, while D_{max} does not seem to have any effect at all for values higher than 1.6.

The employed parameters influence the results in many different ways. However, as shown in Tab. 1 and Tab. 2 parameters exist that allow high detection rates for both fine-grained affordances while at the same time limiting the number of false negative detections. These first results strongly support our approach of fine-grained affordance detection.

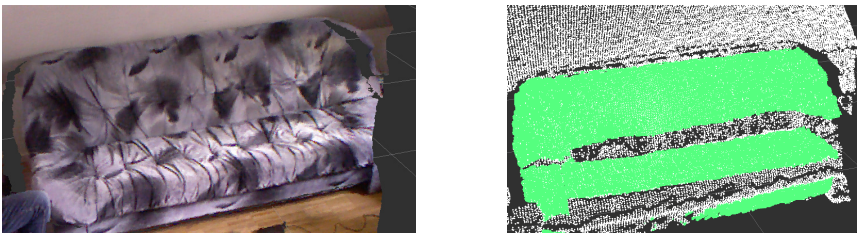


Fig. 4. Detection example of the fine-grained affordance *sitting with backrest* on a sofa with parameters $\alpha = \beta = 30°$, $\gamma = 35°$, $D_{min} = 0.5$ and $D_{max} = 5$. The original image (left) and the detection result in green (right) are shown.

The presented approach of fine-grained affordance detection originally stems from an algorithm to acquire hints to whether or not a stool or chair is present in the input data. Thus, our approach is tailored to this use case. However, the presented method needs to be further generalized to include the detection of fine-grained affordances present on other sitting furniture like sofas. To this point, to detect affordances on sofas, the model parameters need to be altered: D_{max} has to be set to higher values to support wider planes (Fig. 4).

6 Conclusion and Outlook

In this paper we presented an approach to detect affordances on a fine-grained scale by applying an anthropomorphic agent model and affordance models. In its current state our system is able to differentiate between two fine-grained affordances. The high values of the F-measure of 0.956 supports our approach of fine-grained affordance detection.

We continue our work in the two following aspects. First, our current algorithm is feature-centered as we initially detect features (planes) to create an abstract environment representation. However, we expect significant improvement if the agent model is directly fitted into the data (agent-centered approach). This would not only decrease the influence of the plane size parameters, but also allow detecting fine-grained affordances on mixed objects (e.g. a stool without backrest standing close to a wall that can support an agent's back while seated).

Second, we plan to evaluate our approach on a larger test set and include more fine-grained affordances that can be detected with a sitting pose of the agent (e.g. *sitting with armrest* and *sitting in front of a table*). An open question is also how an anthropomorphic agent model can be exploited to detect more fine-grained affordances from different body poses than sitting. As an example for a lying body pose the fine-grained affordances *lying flat* and *lying with pillow* can be distinguished. Fine-grained affordances without a body pose, but with similar actions include knobs attached to drawers and doors that can be *pulled open* or *pulled open while rotating* (about the hinge). We are currently looking for more examples for both cases (with and without body poses) to generalize and formalize our approach of fine-grained affordances.

References

1. Bar-Aviv, E., Rivlin, E.: Functional 3d object classification using simulation of embodied agent. In: BMVC, pp. 307–316 (2006)
2. Chemero, A., Turvey, M.T.: Gibsonian affordances for roboticists. Adaptive Behavior **15**(4), 473–480 (2007)
3. Gibson, J.J.: The concept of affordances. Perceiving, Acting, and Knowing, 67–82 (1977)
4. Gibson, J.J.: The ecological approach to visual perception. Routledge (1986)
5. Hinkle, L., Olson, E.: Predicting object functionality using physical simulations. In: Proc. of IROS 2013, pp. 2784–2790. IEEE (2013)
6. Şahin, E., Çakmak, M., Doğar, M.R., Uğur, E., Üçoluk, G.: To afford or not to afford: A new formalization of affordances toward affordance-based robot control. Adaptive Behavior **15**(4), 447–472 (2007)
7. Wünstel, M., Moratz, R.: Automatic object recognition within an office environment. In: CRV, vol. 4, pp. 104–109 (2004)

A Bio-Inspired Robot with Visual Perception of Affordances

Oscar Chang(✉)

Universidad Central de Venezuela, Caracas, Venezuela
ogchang@gmail.com

Abstract. We present a visual robot whose associated neural controller develops a realistic perception of affordances. The controller uses known insect brain principles; particularly the time stabilized sparse code communication between the Antennal Lobe and the Mushroom Body. The robot perceives the world through a webcam and canny border openCV routines. Self-controlled neural agents process this massive raw data and produce a time stabilized sparse version, where implicit time-space information is encoded. Preprocessed information is relayed to a population of neural agents specialized in cognitive activities and trained under self-critical isolated conditions. Isolation induces an emergent behavior which makes possible the invariant visual recognition of objects. This later capacity is assembled into cognitive strings which incorporate time-elapse learning resources activation. By using this assembled capacity during an extended learning period the robot finally achieves perception of affordances. The system has been tested in real time with real world elements.

Keywords: Affordance perception · Robotic vision · Cooperative neural agents · Deep learning

1 Introduction

Affordance is a quality of an object or environment that allows (or suggests) an individual to perform an action [3]. The term is used in various fields including AI, cognition, perceptual psychology, industrial design, HCI, etc. Perceiving affordances has been related to infants development [10] and has opened vigorous research movement in AI [8] , artificial vision [5] and robotics [6]. Affordance demands the recognition of a class of objects (or environments) with no clear-cut differences, with many diffuse characteristics, with arbitrary boundaries, sizes and designations [3]. It may also trigger in the individual a complex response, such as moving toward the object and sitting on it. To initiate or not a real action will depend in the mediation of others agents. The execution of this excitation-response agreement, trivial for living creatures, combines difficult problems such as tracking and recognizing a moving object [1], the growth of cognitive abilities [4] and the formation of agents societies [9]. In this work we present a visual driven robot whose neural controller support expansible invariant object recognition [2]. In order to extend the robots learning period this

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 420–426, 2015.
DOI: 10.1007/978-3-319-16181-5_31

paper incorporates a self-controlled grow algorithm in which the robot's available learning resources are distributed over an extended period of time. The combination of a long educational experience and the gradual release of learning capabilities finally develop in the robot a credible visual perception of affordances. The controller has an implicit biological structure where cooperative neural agents mimic two key insect brain elements: the antennal lobe (AL) and the Mushroom Body (MB).

2 Previous Works

In previous works by Chang the following operative tools were established [2],[1]:

1) A neural artificial vision system which relies on the computer models of the AL and MB of insects.
2) A flow of circulating information defined by time stabilized sparse code.
3) An expansible learning capacity based upon isolated tunable agents (ITAs).

In this paper we incorporate two new elements: 1) An operative unit called "cognitive string", formed by several ITAs ruled by a common time-released learning mechanism. 2) A "selective reward system" in which short-term learning events are aimed at specific cognitive string.

3 The Robot and Its Multi-agent Neural Controller

The used robot has one moving eye and two final effectors (servomotors) which handle the physical flags P and C. The robot watches the world through a two axis moving webcam and takes as visual input different classes of untailored 2D and 3D images (Fig 1). After training it develops affordance perception for some specific object classes which activate the final effectors.

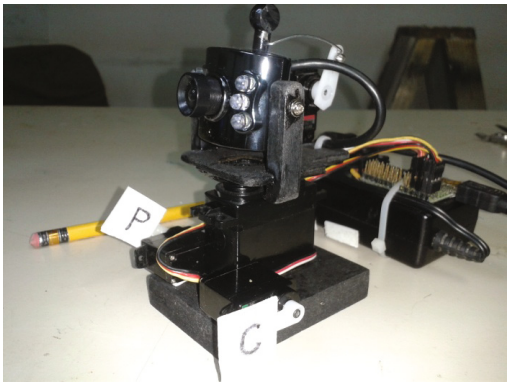


Fig. 1. Visual affordances perceiving robot. The robot observes the world through a two axis moving webcam. Some images afford "painting" and activate the effector P. Some others afford "cutting" and activate the effector C.

The used neural controller utilizes two key insect brain agents: the antennal lobe (AL) and the Mushroom Body (MB) (Fig 2). In the modeled AL primary receptors are pixels in a 100x100 moving region of interest (ROI) image, captured with a webcam and simplified with canny edge detection (a). These pixels feed an ANN pre trained as a crosshair reticle tracker (b). This ANN participates in a close loop feedback system (c) and becomes a generic tracking agent, producing a continuous flow of space-time related unstable code (d). An averaging agent (e) stabilizes this flow and passes a sparse version to the equivalent MB, formed by a set of isolated tunable agents (ITAs) composed by small ANNs (f) specialized in learning, recognition and memory formation. Through OR like operators ITAs' output are grouped into cognitive strings (g) which finally activate the physical effectors (h).

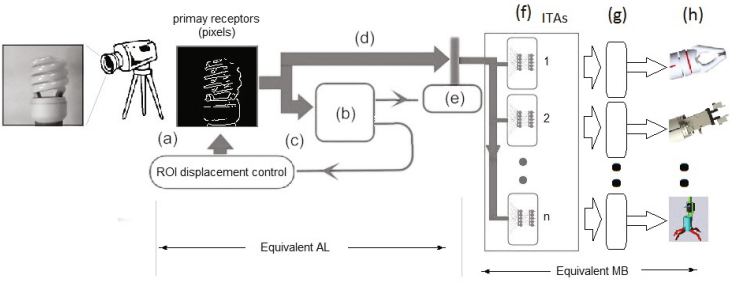


Fig. 2. The neural controller. A 100x100 canny image feeds a modeled Antennal Lobe (AL) which generates time stabilized sparse code. This resource is passed to an artificial Mushroom Body (MB) where isolated tunable agents (ITAs) carry out cognitive duties. ITAs are assembled into cognitive strings which finally activate the effectors.

4 The Artificial AL and MB

In insects the AL converts crude sensors data to a special form of space-temporal code essential for object recognition and relayed to the cognitive elements in the MB [7]. In our AL a backpro trained ANN operates in a closed loop mechanism where images from a video stream control image position [2]. This loop generates a flow of space-time related data which is subsequently stabilized and sparsed. The resulting *time stabilized sparse code* (TSSC) is relayed to cognitive agents in the simulated MB. The insects' MB serves as a large screen where objects can be much more easily discriminated [7]. In our MB cognitive agents called ITAs (Isolated Tunable Agents) are built with trainable three layers ANNs formed by 2500 inputs, 10 hidden and 5 outputs neurons. As in biology ITAs use as input neurons the TSSC coming from the AL (2500 signals).

5 Isolated Learning

When learning to recognize an object each ITA behaves as an auto-critical individual who uses the following learning rules:

Rule 1 Look toward the outside world. See the object for a while and use backpro to:

```
1a) learn to fire with the object.
1b) partially forget what you have learned somewhere else.
```

Rule 2 Look inside yourself. See your own noise source for a while. Use backpro to:

```
2a) learn not to fire with noise.
2b) partially forget what you have learned somewhere else.
```

Using these rules a reward R is defined as a short term learning experience during which one selected ITA receives 100 consecutive backpro cycles watching the chosen object followed by 100 cycles watching white noise. Targets are properly set so that ITA's central output neuron learns to fire with the object and not to fire with noise. At 50 frames/sec a reward lasts 4 seconds and about 5 rewards are needed to memorize one object. Rewards shall not exceed a maximum number or the affected ITA will be degraded (overexposure). When trained under the above principles an ITA shows an emergent capacity to discriminate the learned object from many others, while absorbing a finite quantity (roughly 20%) of visual variances and white noise.

6 The Time-Released Learning Resources

Our next goal is to expand the number of ITAs dedicated to the learning of one object so that class recognition is attained. To this end ITAs are assembled into cognitive strings S^1, S^2,... S^n formed by m by consecutive ITAs numbered from 1 to m. To avoid overexposure a self-controlled time-released mechanism operates in each string distributing the received rewards as: The firs active ITA is the number 1. At any given time only the active ITA in the string receives rewards. Every active ITA i, which receives 15 (or so) consecutive rewards freezes its weight information and passes the active condition to the i+1 ITA. Once trained and for recognition purposes the ITAs' outputs in the same string are "ored" together. A selective reward R^i is now defined as a reward that only affects the active ITA in the cognitive string S^i. This selective norm make possible to dedicate a whole string to the invariant recognition of one object thus expanding object recognition into class recognition.

7 Results

7.1 Experiment 1: The Emergence of Affordance Perception

In this experiment the robot develops perception of affordance for two classes of objects: class P represented by brushes which afford "painting" and class C represented by scissors which afford "cutting". These classes were chosen because physical samples of them were readily available and because they both represent difficult to recognize items, very sensitive to rotational translation. Two cognitive strings S^P and S^C in the MB are selected to develop affordances for painting (brushes) and cutting (scissors) respectively. Once trained the robot demonstrates its perceptions by activating its final effectors P and C. Each string comprises 20 ITAs which cover the rotational image variances of one full rotation per object. For training a human places objects (scissor or brushes) in the robot field of view and sends selective rewards R^P or R^S aimed at the respective strings. Since each trained ITA absorbs about 20 degrees of object rotation, 20 of them cover a full turn. In figure 3 (upper right) two trained ITAs process original images turned into canny images and TSSC. Time stabilized sparse features have been created in the ITAs' hidden layer (weights of one hidden neuron are shown). Using a Pentium Core i5 the learning time is about 8 minutes. Once trained the robot visually scans the shown landscape (left) and after three minutes correctly perceives the eight existing affordances. Some image zooming is tolerated and look alike objects such as pliers and relays are rejected.

Fig. 3. Affordance perception for multiple object visualization

7.2 Experiment 2: A Non-easily Distractible Eye

To test the consistency of its perception of affordance the above trained robot is set to explore the whole Caltech 101_Object Categories data set. After examining the 9146 images in 5 hours the robot reports the 9 mistaken, look-alike elements shown in figure 4. It also recognizes 31 out of 39 true affordances in the "scissor" category.

Fig. 4. Searching for affordances in the 101_ObjectCategories Caltech dataset

8 Discussion

The robot shows a robust affordance perception capacity. For the whole Caltech dataset the error is limited to 0.098 %. More ITAs per cognitive string can be used as to cover full tilting and zooming for each desired affordance. A credible perception of affordance appears only after a prolonged learning period, which in turn requires a space-time distribution of learning resources prepacked in cognitive strings. The used learning and feed forward mechanisms have a natural parallel structure so high speed operation could be expected when using parallel computing.

According to the false positives found in the Caltech 101 the neural controller might be taken as a global shape descriptor. For an external observer, however, this condition may be indistinguishable from true affordance perception. The proposed method could be considered a form of deep learning in which the features available to the ANN for learning are time stabilized space-time relations created by the generic tracking agent in its dynamic pursue of image stability.

9 Conclusions

We have developed and tested a robotic vision system capable of showing clear relations between affordances and perception-action under broad visual conditions. The proposed neural controller uses cooperative neural agents organized as the artificial versions of the AL and MB of living insects. In the proposed MB basic cognitive agents called ITAs, sensitive to short term learning experiences, are assembled into operative modules called cognitive strings. Inside the strings orderly activated ITAs store time stabilized sparse features of selected objects. The combination of a prolonged educational experience, time-elapse release of learning resources and the time stabilized sparse feature extraction finally develops in the robot a credible form of visual affordance perception.

In extended images search the neural visual controller shows a good rejection of false affordances. This may be relevant for constructing efficient, no easily distractible robots. In principle the proposed techniques can expanded to higher pixel resolution and many affordance perceptions.

References

1. Chang, O.: Recognizing a moving object by using neural nets and ocular micro tremor. Revista de la Facultad de Ingenieria Universidad Central de Venezuela **28**, 49–55 (2013). http://ucv.academia.edu/OscarChang/Papers
2. Chang, O.: Reliable object recognition by using cooperative neural agents. In: 2014 International Joint Conference on Neural Networks (IJCNN), Beijing, pp. 2571–2578, July 2014. http://ucv.academia.edu/OscarChang/Papers
3. Gibson, J.: The Ecological Approach To Visual Perception. Taylor & Francis (2013). http://books.google.co.ve/books?id=yv_9hU_26KEC
4. Goertzel, B., Bugaj, S.V.: Stages of cognitive development in uncertain-logic-based AI systems. In: Proceedings of the 2007 Conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006, pp. 174–194. IOS Press, Amsterdam (2007). http://dl.acm.org/citation.cfm?id=1565455.1565468
5. Hermans, T., Rehg, J.M., Bobick, A.: Affordance prediction via learned object attributes. In: IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration (May 2011)
6. Horton, T.E., Chakraborty, A., St. Amant, R.: Affordances for robots: a brief survey. AVANT. Pismo Awangardy Filozoficzno-Naukowej **2**, 70–84 (2012)
7. Huerta, R.: Learning pattern recognition and decision making in the insect brain. In: AIP Conference Proceedings, vol. 1510, pp. 101–119 (2013)
8. Mateas, M.: Expressive AI: A semiotic analysis of machinic affordances. In: Proceedings of the 3rd Conference on Computational Semiotics and New Media, University of Teesside, UK (2003)
9. Minsky, M.: The Society of Mind. A Touchstone Book, Simon & Schuster (1986). http://books.google.co.ve/books?id=veVOAAAAMAAJ
10. Saccani, R., Valentini, N.C., Pereira, K.R., Müller, A.B., Gabbard, C.: Associations of biological factors and affordances in the home with infant motor development. Pediatrics International **55**(2), 197–203 (2013)

Integrating Object Affordances with Artificial Visual Attention

Jan Tünnermann(✉), Christian Born, and Bärbel Mertsching

GET Lab, University of Paderborn, Pohlweg 47–49, 33098 Paderborn, Germany
{tuennermann,born,mertsching}@get.uni-paderborn.de

Abstract. Affordances, e.g., grasping possibilities, play a role in the guidance of human attention. We report experiments on the integration of affordance estimation with artificial visual attention in a prototypical model. Furthermore, Growing Neural Gas is discussed as a potential framework for future attention models that deeply integrate affordance, saliency and further attentional mechanisms.

Keywords: Attention · Saliency · Affordance

1 Introduction

With the transition of robots from specialized automata performing predefined tasks to general autonomous agents, the requirements to perceive, reason about, and interact with their environment have drastically increased. A recent development in robotics is to model aspects of environmental psychology, which deal with the interaction between humans and their surroundings. A popular concept is the *affordance of objects*, introduced by J. J. Gibson in 1977 [8]. In this holistic view, objects possess certain affordances, i.e., objects or their parts can afford certain actions. A common example is a mug, whose handle affords grasping.

This idea has been transferred to technical systems, not only to enhance grasping actions, but also to benefit object recognition and semantic scene perception (see e.g., [5,17,24]). In many cases, objects are better defined by actions the object supports, than by visual attributes. Coming back to the example of a mug, even though colors and shapes may differ widely, mugs in general afford grasping (possibly by some kind of handle), containing liquid and drinking [4]. Therefore, recent research integrates affordance estimation with object recognition [4,9] and the semantic interpretation of scenes and objects [25,26].

Artificial visual attention is a concept inspired by cognitive psychology. The main idea is to filter relevant from irrelevant information very early in processing, and distribute processing resources accordingly. Attention can be guided bottom-up by saliency (local feature contrasts) [11] or in a top-down manner by incorporating knowledge, task demands [12] or the "gist of the scene" [13].

Findings from psychology suggest that affordances influence human visual attention. This has been shown in reductions of reaction times when affordances

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 427–437, 2015.
DOI: 10.1007/978-3-319-16181-5_32

were used to guide attention towards target locations [7] and effects on event related signals in electrophysiological and brain imaging research [10].

This design paper is an update of the report we presented at the First Workshop on Affordances: Affordances in Vision for Cognitive Robotics [23]. The remainder of the paper is organized as follows: Section 2 contains a compressed report of the experiments conducted in [23]. In section 3 we discuss Growing Neural Gas as a framework for artificial attention that we believe has the potential to integrate bottom-up saliency, affordance-based attention and top-down mechanisms in a consistent architecture and improve on several disadvantages of current region-based attention systems. Section 4 concludes the paper.

2 Change Detection Experiments on Saliency and Affordance in Human Attention

In a previous study [21], we employed a "single-shot" change detection task with natural images (see e.g., [19]) to measure the participants' distribution of attention towards salient or affording objects. The phenomenon of change blindness due to short scene interruptions renders the detection of changes in objects difficult. An observer's performance depends on the allocation of attention towards the objects [15]. For the evaluation of psychologically inspired computer vision systems, the change blindness paradigm has the great advantage that images with natural scenes can be used, whereas many other psychophysical tasks require the use of highly artificial synthetic stimuli. We found that human observers performed better in reporting the changes that were made to objects selected by an affordance-based model than when those selected by the saliency model were changed.

The single-shot paradigm contains a single change from the original to the altered image which are shown only briefly (usually between 100 and 500 ms) and a blank screen is shown between the two images. The presentation usually lasts for less than a second and participants respond afterwards, when the image is already gone. Thus, there is only a single binary hit-or-miss measurement per change. Furthermore, the same images cannot be repeated and therefore the amount of trials is limited to the number of available images. Their creation is quite an effort, due to editing in the changes (object removals in our case). Because of the limited number of trials and the binomially distributed response, a large number of subjects is required (40 – 80) to obtain reliable results.

Hence, in the first experiment described here, we tested the so called "flicker paradigm" (see e.g., [15]): the presentation is similar as described above, but it is repeated until the change is reported. Therefore, a more informative measure, namely the time it takes the subject to detect the change, can be obtained. This not only reduces the number of participants required, but may also allow to relate the degree of affordance and saliency to the response time. Therefore, the objective of this first experiment is, using the stimulus material from [21], to investigate whether the effect that affordances are more important than saliency in change detection can be replicated using the flicker paradigm. Furthermore, a

first insight in the influence of the saliency and affordance values on the response time is provided.

2.1 Experiment 1

Participants: Twelve volunteers (average age of 26.82, $SD = 3.6$) participated in this experiment. All had normal or corrected-to-normal vision and not seen the images before.

Stimuli: The stimulus material reported in [21] was used. This consisted of 28 natural scenes, mostly pictures of office environments that contained a number of objects in the reachable action space and some in background areas which would not be reachable by the observer of the scene. For every image, two changed versions were created by locally altering the image (locally blending in an identical image in which the object had been removed at that location). In one altered image of the same scene, an object selected by the saliency model by Itti et al. [11] had been removed, in the other altered image, one object selected by an affordance-based prediction (density of grasping possibilities per image area; this corresponds to the affordance stream described in see section 2.2) had been removed. Refer to [21] for more details regarding the stimulus generation and the actual pictures.

Design and Procedure: The experiment was conducted on a 12.1" touch-screen laptop[1]. Participants were presented with every original image paired with one of the possible changes. The number of times for which each changed image appeared with the affordance or saliency change was balanced over all subjects.

In contrast to the task used in [21], the images cycled back and forth between the original and changed image until the change was reported by touching the screen at the location of the change. If no response was made within one minute, the current trial was aborted and the next trial started. The timing of the image sequence was: "1000 ms initial blank"–"300 ms original image"–"300 ms blank"–"100 ms changed image"–"300 ms blank". For every trial, the response time was recorded.

Results and Discussion: Figure 1 shows the average response time to saliency- and affordance-based changes. Affordance-based changes are reported significantly faster, $t(11) = -5.03, p < 0.001$, confirming our earlier results [21] from the single shot hit-or-miss task. This provides further evidence for the importance of affordances in the deployment of attention. The one minute time limit was reached only four times in the 336 changes presented over all subjects.

[1] Note that the change blindness effect is very robust and does not require highly accurate timing that can be only established with CRT monitors or specialized equipment, which is the case for many other psychophysical paradigms.

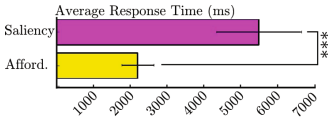


Fig. 1. Average response times for affordance- and saliency-based changes (error bars show the SEM; *** $p < 0.001$).

In [23] we show response time distributions over the different changes and images. No long response times are found when saliency and affordance of a change were high. Especially for the saliency-based changes, long response times occur where the affordance is close to zero. Therefore, an attention model that combines saliency and affordance could show a better performance than models based on each individual component.

2.2 Experiment 2

This second experiment is based on predictions from a prototypical model that combines affordance and saliency estimation. It is intended to investigate whether predictions based on combined saliency and affordance better reflect human attention than the individual components.

A Combined Model of Saliency and Affordance: The model is outline in figure 2. The left image of a stereo image pair is segmented into homogeneously colored regions ① (see e.g., [3]). These regions can be considered proto-objects at pre-attentional stages.

In the saliency stream, the regions are used to generate feature magnitude maps for *color*, *orientation*, *eccentricity*, *symmetry* and *size* (s2). The feature *color* is obtained as the average color of all pixels of a region. *Orientation*, *eccentricity* and *symmetry* are calculated based on 2D central moments of the spatial distribution of a region's pixels. *Size* is the number of pixels in a region.

As a next step in this stream, saliency maps are calculated for each feature dimension individually (s3). This is done by applying a voting style procedure, where each region collects votes from its neighbors, regarding the dissimilarities in every feature dimension (details for the feature and saliency computations can be found in [1]).

In the affordance stream, the left image is used to generate *2D Texlets* which are small local texture patches (a1). Using stereo disparities (a2), the *2D Texlets* are transformed into *3D Texlets* (a3). Small groups of neighboring *Texlets* are created by applying a position-based k-means clustering. Planes are fitted through the *groups* to form *Surflings* (a4), which are further grouped (when close to each other and similarly oriented) to generate *Surfaces* (a5) [14].

Grasping hypotheses in 3D space are generated by fitting a simulated gripper (see figure 3a) to elements of the scene considering the surfaces generated in the process outlined above (a6). Details of this process can be found in [14]. The result, which we make use of in this study, is the estimated contact points of the gripper on the surfaces. Note that in the present study the simulated gripper performs simple two-fingered grasping.

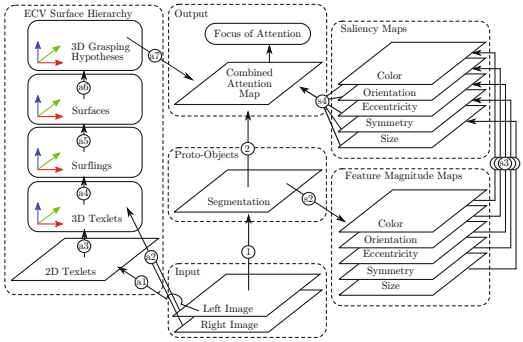


Fig. 2. Structure of the proposed model. The type and flow of data is described in the main text referring to this figure.

The combined attention map is then obtained by integrating the individual saliency maps (s4) [1] and the contribution from the affordance stream using the regions from the initial segmentation. While the saliency contribution is already in region-form, the grasping hypotheses (contact points) have to be projected into 2D first. All points which fall into a certain region (a7)+(2) are summed and normalized by the region size. Because the contact points can often be found on the edges of objects (in their 2D projection), instead of considering a single point in this process, each back-projected point is expanded to 5×5 points in a square region surrounding the initial location, with their contribution decreasing with distance from the original location. This can be seen in figure 3b. In this first attempt to combine saliency and affordance in a technical model, we combine both linearly with equal weights. More advanced strategies to combine different feature channels in attention models are discussed in [11].

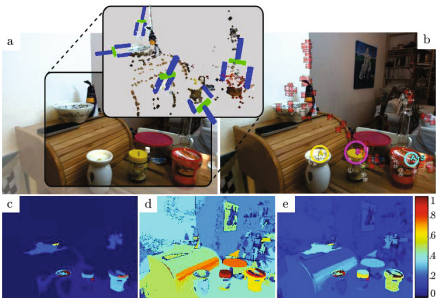


Fig. 3. a: A test scene. Inset: exemplary grasps fitted to a sparse 3D representation. **b:** Grasp points projected into 2D. White patterns represent grasps towards reachable locations, red patterns indicate locations out of reach. "A" affordance, "S" saliency, and "A+S" combined selection. **c–e:** Underlying affordance (c), saliency (d), and combined (e) maps.

Participants: Thirty volunteers (average age of 28.46, $SD = 5.86$) participated.

Stimuli, design and procedure: Stimulus material, experimental design and procedure mostly correspond to the description of the first experiment in section 2.1. The only differences were the use of a new image set (see figure 4a) with an

additional third possible change based on the combined prediction. Furthermore, the saliency-based prediction was obtained with the region-based saliency model [1], which constitutes the saliency channel in the combined model, in contrast to the first experiment where the model by Itti and colleagues [11] was used.

Due to the fact that three predictions (affordance, saliency, combined) are required for each image, and the images focus mainly on the action space where saliency and affordance are both expected to be relatively high, sometimes the same object was selected by two or all three predictions. In such a case, the scene was slightly rearranged by unsystematically shifting objects or the camera, and the scene was rerecorded, until three distinct predictions were obtained.

Fig. 4. a: Change locations marked in nine exemplary images (all 29 images are shown in [23]). **b:** Average response times for the changes based on affordance, saliency (region-based), and combined predictions (error bars show the SEM).

Results and Discussion: Figure 4b shows the average response times to changes based on the (region-based) saliency, affordance, and combined predictions. According to an one-way repeated measures ANOVA, no effect of prediction type was found, $F(2, 29) = 0.63$, $p = 0.54$. This is in contrast to the result of our first experiment, where the responses to affordance-based changes were significantly faster. Furthermore, the saliency conditions (from experiment 1 and experiment 2), as well as the affordance conditions (from each experiment), differ significantly, $t(40) = 5.82$, $p < 0.001$ (affordance), $t(40) = 4.3$, $p < 0.001$ (saliency) according to Holm-Bonferroni corrected two-tailed t-tests.

The long response times in the second experiment indicate that the task was more difficult than in the first experiment. Moreover, the scenes were arranged to contain a large number of affording objects in the action space and thus also the saliency-based selections were mainly such foreground objects, whereas the stimulus material used in the first experiment contained saliency-based changes which were frequently in the background. Inspection of the distribution of the individual changes' average response times (refer to [23]), hints that saliency-based changes may benefit from increasing affordance, while the same seems not to be the case for affordance-based changes.

Notably, the one minute limit was reached twelve times for affordance- and nine times for saliency-based changes, and only once in the combined condition.

Hence, whether and how saliency and affordance enhance each other remains unclear. In the prototypical model, affordance and saliency have been processed

based on different representations (ECV [14] vs. region-based), normalized in different ways, and integrated eventually. Due to this process it becomes difficult to assess the relative contributions of both channels and relate them to response times (we attempted this to some degree in [23]). Furthermore, for future technical applications, calculating and maintaining two separate representations is rather unpractical. Therefore, in the remainder of the paper Growing Neural Gas as a potential structure for a fully integrated representation and attention model is discussed.

3 Growing Neural Gas: An Architecture for Combining Saliency, Top-Down Attention and Affordances?

In the long term, a fully integrated architecture for artificial attention is desirable. In such a framework, feedback loops, which are known to be highly important in biological vision, can be established: results from higher levels of the architecture, such as affordances or top-down information can be used to influence the generation or propagation of the scene representation from lower levels. Furthermore, the integration of different dimensions, such as various saliency dimensions, specific and gist-based top-down influences and object affordances benefits from a common consistent representation which can preserve the relative strength of each dimension's contribution (a prerequisite for more advanced combination strategies as suggested in [11]).

We discuss Growing Neural Gas (GNG; e.g. [6]) as a pre-attentional structure for a fully integrated approach. GNG has not been applied to artificial attention before but exhibits promising features. On the one hand, they can be seen as related to the already mentioned region-based approaches [1,2,22] as pre-attentional structures are employed. On the other hand, they implement a basic perceptual learning in the sense that the current representation is updated with new data instead of recalculating it entirely as in region-based artificial attention. We briefly describe the main concepts of GNG and discuss its potential application to saliency, top-down and affordances calculations.

GNG constitutes an unsupervised learning technique. Nodes (neurons) may be connected by edges and possess attributes representing properties of the state space (e.g., x- and y-positions). Examples are presented to the algorithm and it determines the closest node according to a distance measure on the respective properties. For this node and, with a reduced strength its topological neighbors, the properties are updated. Edges carry an age value, which is increased in every update. A new edge is inserted between the closest node and the runner-up. If such an edge already exists, its age is reset. Edges which are too old are deleted. Furthermore, there is a domain-dependent error value for each node, which must have the characteristic that it is reduced when a new node is inserted in proximity. At fixed intervals, the node with the highest error value is determined, as well as its neighbor with the highest error. A new node is inserted in between and connected with these two nodes, replacing their original connection. The error value is redistributed between all three nodes reducing the probability that the

next insertion is performed nearby, guiding the growth of the network. An additional utility term quantifies a node's usefulness and may result in the deletion of the node to avoid infinite growth of the network.

The algorithm is initialized with two connected nodes with random properties. The dynamically adapting set of nodes with changing neighborhoods can form multiple independent graphs.

Pre-Attentional Structures and Saliency Based on GNG: GNG can potentially be adapted to generate pre-attentional structures. Pixels of the image are chosen uniformly at random and used as examples to train a GNG as described above. Distances to the pixels (in a x-, y- and color-space) can be calculated by using, e.g., a weighted euclidean distance. Figure 5a depicts the result for a simple synthetic test image.

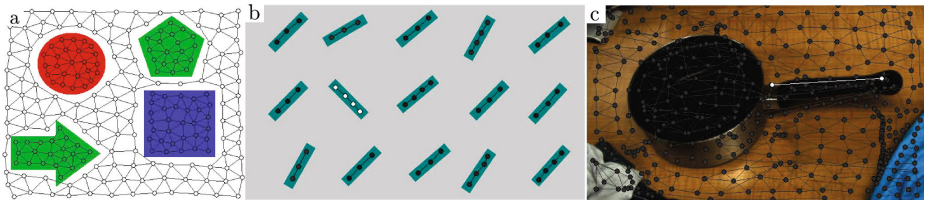


Fig. 5. a: Exemplary results using GNG-based pre-attentional structures. Node colors reflect the represented object. **b:** The result of a saliency computation. Gray levels of the nodes represent the orientation saliency of a graph (background graph removed in calculation). **c:** The GNG may include useful candidates (two-node networks; white) for estimating grasp affordances. Networks with more than two nodes are colored in gray.

For the resulting graphs, feature magnitudes and saliency can be calculated as described for region-based saliency in [1].

The feature *orientation* is used as an example here. As in the aforementioned paper, 2D central moments are calculated for each node $n(x, y)$ in graph G_i:

$$m^i_{1,1} = \sum (x-\overline{x})(y-\overline{y}),\ m^i_{2,0} = \sum (x-\overline{x})^2 \text{ and } m^i_{0,2} = \sum (y-\overline{y})^2\ \forall n(x,y) \in G_i \qquad (1)$$

where $(\overline{x}, \overline{y})$ denotes the center of G_i. The orientation ϕ^i is then computed as

$$\phi^i = \frac{1}{2} tan^{-1}\left(\frac{2m^i_{1,1}}{m^i_{2,0} - m^i_{0,2}}\right), \qquad (2)$$

resulting in an orientation value ϕ^i between 0° and 180°. The orientation saliency s_{ϕ^i} of every graph G_i is then obtained as

$$s_{\phi^i} = \sum_{G_i, i \neq j} \frac{|\phi^i - \phi^j|}{90°}. \qquad (3)$$

The result of this process is an orientation saliency map (saliency value associated with every graph) and shown for an orientation pop-out stimulus in figure 5b.

Applying Top-Down Information in GNG-Based Attention: As argued above, mechanisms from region-based attention can be transferred to GNG structures. Therefore, templates could be used as described in [2,20] for region-based attention.

Furthermore, as the transfer from pixels to the substantially smaller number of neurons provides a simplified problem space, rough heuristics, such as for determining the gist of a scene, may also benefit from a GNG-based representation.

Including Object Affordances in GNG-Based Attention: Pre-attentional structures obtained from GNG may prove sufficiently stable to apply appearance-based affordance estimation in local and global contexts as proposed by [16]. Furthermore, GNG obtained with the described procedure may provide easily identifiable candidates for graspable elements. Figure 5c shows GNG structures obtained for a picture of a cooking pot extracted from Song et al. [16]'s figure 4b (the pink dot was removed). Highlighting only two-node networks in agreement with the ground-truth for such handles (see Song et al. [16]'s figure 2a), successfully detects the pot's handle. Such rough heuristics could be directly useful for generating local graspability estimates which can then be fused with global estimates [16], or provide candidates for computationally more expensive follow-up processing.

4 Conclusion

The results of our first experiment further support the idea that object affordances are important for the spatial deployment of visual attention [22]. In the second experiment we did not find additional enhancements by combining saliency and affordance. This is in line with another change blindness study [18], where saliency did not further enhance the detection of changes in objects which are shown in unusual contexts. Early attention appears to be strongly influenced by the environment represented in the scene. The second experiment, however, did also fail to replicate the advantage of the affordance-based predictions over the saliency-based predictions. This may arise from limitations of the prototypical model (see section 2.2). Alternative possibilities are discussed in [23].

Hence, an important next step in this line of research is a deeper integration of affordance in attention systems. The fact that affordance-based advantages are present in 2D images presented to humans, which depict foregrounds and background (experiment 1 and experiments reported in [21]), proves that binocular cues are not necessary for the effect in biological vision. Thus, a 2D dimensional retinotopical structure would provide a useful domain for such a fully integrated approach. We discussed Growing Neural Gas as a framework for this in section 3. These may allow to integrate appearance-based affordance estimation as suggested by Song et al. [16] with bottom-up and top-down attention in future work to allow more sensitive experiments and practical evaluation in a robot.

References

1. Aziz, M.Z., Mertsching, B.: Fast and Robust Generation of Feature Maps for Region-Based Visual Attention. IEEE Transactions on Image Processing **17**(5), 633–644 (2008)
2. Aziz, M.Z., Mertsching, B.: Visual search in static and dynamic scenes using fine-grain top-down visual attention. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 3–12. Springer, Heidelberg (2008)
3. Backer, M., Tünnermann, J., Mertsching, B.: Parallel k-means image segmentation using sort, scan and connected components on a GPU. In: Keller, R., Kramer, D., Weiss, J.-P. (eds.) Facing the Multicore-Challenge III. LNCS, vol. 7686, pp. 108–120. Springer, Heidelberg (2013)
4. Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B.: Using Object Affordances to Improve Object Recognition. IEEE Transactions on Autonomous Mental Development **3**(3), 207–215 (2011)
5. Detry, R., Kraft, D., Kroemer, O., Bodenhagen, L., Peters, J., Krüger, N., Piater, J.: Learning Grasp Affordance Densities. Paladyn **2**(1), 1–17 (2011)
6. Fritzke, B.: A self-organizing network that can follow non-stationary distributions. In: Gerstner, W., Germond, A., Hasler, M., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 613–618. Springer, Heidelberg (1997)
7. Garrido-Vásquez, P., Schubö, A.: Modulation of Visual Attention by Object Affordance. Frontiers in Psychology **5**, 59 (2014)
8. Gibson, J.J.: The theory of affordances. In: Perceiving, Acting, and Knowing
9. Gijsberts, A., Tommasi, T., Metta, G., Caputo, B.: Object recognition using visuo-affordance maps. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1572–1578 (2010)
10. Handy, T.C., Grafton, S.T., Shroff, N.M., Ketay, S., Gazzaniga, M.S.: Graspable Objects Grab Attention When the Potential for Action is Recognized. Nature Neuroscience **1**, 1–7 (2003)
11. Itti, L., Koch, C.: Feature Combination Strategies for Saliency-Based Visual Attention Systems. Journal of Electronic Imaging **10**(1), 161–169 (2001)
12. Navalpakkam, V., Itti, L.: A goal oriented attention guidance model. In: Bülthoff, H.H., Lee, S.-W., Poggio, T.A., Wallraven, C. (eds.) BMCV 2002. LNCS, vol. 2525, pp. 453–461. Springer, Heidelberg (2002)
13. Oliva, A., Torralba, A.: Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research **155**, 23–36 (2006)
14. Popović, M., Kootstra, G., Jørgensen, J.A., Kragic, D., Krüger, N.: Grasping unknown objects using an early cognitive vision system for general scene understanding. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 987–994 (2011)
15. Rensink, R.A., O'Regan, J.K., Clark, J.J.: To See or Not to See: The Need For Attention to Perceive Changes in Scenes. Psychological Science **8**(5), 368–373 (1997)
16. Song, H.O., Fritz, M., Gu, C., Darrell, T.: Visual grasp affordances from appearance-based cues. In: IEEE ICCV Workshops, pp. 998–1005 (2011)
17. Stark, L., Bowyer, K.: Generic recognition through qualitative reasoning about 3-D shape and object function. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 251–256 (1991)
18. Stirk, J.A., Underwood, G.: Low-Level Visual Saliency Does not Predict Change Detection in Natural Scenes. Journal of Vision **7**, 1–10 (2007)

19. Tseng, P., Tünnermann, J., Roker-Knight, N., Winter, D., Scharlau, I., Bridgeman, B.: Enhancing Implicit Change Detection Through Action. Perception **39**(10), 1311–1321 (2010)
20. Tünnermann, J., Born, C., Mertsching, B.: Top-Down visual attention with complex templates. In: International Conference on Computer Vision Theory and Applications, pp. 370–377 (2013)
21. Tünnermann, J., Krüger, N., Mertsching, B., Mustafa, W.: Affordance Estimation Enhances Artificial Visual Attention: Evidence from a Change Blindness Study (in review)
22. Tünnermann, J., Mertsching, B.: Region-Based Artificial Visual Attention in Space and Time. Cognitive Computation **6**(1), 125–143 (2014)
23. Tünnermann, J., Mertsching, B.: Saliency and affordance in artificial visual attention. In: RSS 2014, First Workshop on Affordances: Affordances in Vision for Cognitive Robotics (2014)
24. Varadarajan, K.M., Vincze, M.: Affordance based part recognition for grasping and manipulation. In: ICRA Workshop on Autonomous Grasping (2011)
25. Yao, B., Ma, J., Fei-Fei, L.: Discovering object functionality. In: IEEE International Conference on Computer Vision, pp. 2512–2519 (2013)
26. Zhao, Y., Zhu, S.C.: Scene parsing by integrating function, geometry and appearance models. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3119–3126 (2013)

Modelling Primate Control of Grasping for Robotics Applications

Ashley Kleinhans[1(✉)], Serge Thill[2], Benjamin Rosman[1], Renaud Detry[3], and Bryan Tripp[4]

[1] CSIR, Pretoria, South Africa
akleinhans@csir.co.za
[2] University of Skövde, Skövde, Sweden
[3] University of Liège, Liège, Belgium
[4] University of Waterloo, Waterloo, Canada

Abstract. The neural circuits that control grasping and perform related visual processing have been studied extensively in macaque monkeys. We are developing a computational model of this system, in order to better understand its function, and to explore applications to robotics. We recently modelled the neural representation of three-dimensional object shapes, and are currently extending the model to produce hand postures so that it can be tested on a robot. To train the extended model, we are developing a large database of object shapes and corresponding feasible grasps. Finally, further extensions are needed to account for the influence of higher-level goals on hand posture. This is essential because often the same object must be grasped in different ways for different purposes. The present paper focuses on a method of incorporating such higher-level goals. A proof-of-concept exhibits several important behaviours, such as choosing from multiple approaches to the same goal. Finally, we discuss a neural representation of objects that supports fast searching for analogous objects.

Keywords: Grasping · Affordances · Macaque · Robotics · AIP · F5

1 Introduction

The neurophysiology that underlies primate grasping has been studied most extensively in macaque monkeys. In macaques, grasping is controlled by an extensive brain network that includes many parts of the visual, parietal, and frontal cortices. A network of dorsal visual and parietal areas detects affordances and may partially parameterize multiple potential movements [1]. Ventral visual and prefrontal areas help to select movements that are consistent with object identities and goals [2]. Our general aim is to translate this rich neurophysiological knowledge into a bio-plausible robotic grasp controller. Specifically, we want to develop a system that uses a robotic hand to grasp a wide range of objects, while reproducing many features of grasp-related neural activity recorded from monkeys.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 438–447, 2015.
DOI: 10.1007/978-3-319-16181-5_33

In pursuit of our goal, we recently developed a neural model [3] that reproduced a variety of electrophysiology data from the caudal and anterior intraparietal areas (CIP and AIP, respectively). These areas encode three-dimensional shape features, and are essential for accurate hand shaping. This model reproduced responses of visual-dominant object-responsive AIP neurons from the macaque literature using a model of CIP activity as input. We parameterized AIP responses using both superquadric parameters and the parameters of an Isomap reduction of the depth map. We found that both the match with AIP data and the performance of the CIP-AIP mapping were better with Isomap parameters. However, it is not yet clear whether such parameters provide a good basis for grasp planning. For example, in contrast to Isomap, superquadrics support a pose-invariant mapping to some gripper parameters.

To address this question, we have recently started to extend the model to frontal area F5 (which encodes hand postures [4]) so that its applicability to robotic grasp control can be tested. We plan to build a database of grasp examples in order to train and test this extended model. The models trained using such a database will be tested with a real-world robot platform and real objects. We will compare the performance of the neural model to a conventional kernel regression machine, and to state-of-the-art robotics heuristics for grasp planning. We hope to show that a neural model trained on large numbers of examples can provide a practical grasp controller, and that its internal signals are consistent with the literature on neural activity in monkey AIP and F5.

Finally, the main focus of the present paper is on how to further extend the above models to account for how higher-level goals and intentions from prefrontal areas can influence the decision of which affordances to attend to (and therefore which hand shape to select). The following sections briefly present our approach and a proof-of-concept model. A notable feature of this proof-of-concept is that is expressed entirely in vector operations.

2 Methods

Often, different grips are appropriate for manipulating an object for different purposes. For example, if one's goal is to put a hammer in a toolbox, there are many ways in which the hammer can be grasped. However, if the hammer is to be used to hit nails there is essentially one way. To model such influences we are forced to consider a much larger network that includes the prefrontal cortex.

The prefrontal cortex is less well understood than the visual cortex, so for these areas the data-driven approach that we previously adopted to model CIP, AIP, and F5 may be less practical. We are instead pursuing a top-down approach based on two key methods. The first is the Neural Engineering Framework [5], which provides a way to map systematically between high-level function and neural activity. The second is Holographic Reduced Representations [6], which are used in cognitive modelling. Recently, these two methods were used together to develop a spiking neural model of the brain with complex cognitive abilities [7]. The methods are described briefly below. For robotics applications, there

are various ways to run large models of this type in real time, e.g. surrogate population models on FPGAs [8].

Neural Engineering Framework. An NEF model is specified in terms of vector variables that are taken to be encoded by the activity of neuron populations, maps between these vectors, and physiological neuron properties (e.g. time constants). The encoding of a vector by a neural activity is typically modelled as

$$r_i = G\left[\mathbf{e}_i^T \mathbf{x} + b_i\right], \tag{1}$$

where r_i is the spike rate of the i^{th} neuron, $\mathbf{x}$ is the encoded vector, $\mathbf{e}$ is the direction in the encoded space in which the neuron spikes fastest (the "preferred direction"), b_i is a static bias, and G is a physiological nonlinearity. The encoded vector $\mathbf{x}$ can be approximately recovered, or "decoded" from the spike rates as

$$\hat{\mathbf{x}} = \sum_i \mathbf{d}_i r_i, \tag{2}$$

where $\mathbf{d_i}$ is called the neuron's "decoding vector", and is chosen to minimize $\mathbf{x} - \hat{\mathbf{x}}$. Furthermore, functions $\mathbf{f}(\mathbf{x})$ of the vector can also be decoded by choosing different decoding weights that minimize $\mathbf{f}(\mathbf{x}) - \hat{\mathbf{f}}(\mathbf{x})$. This is the basis of NEF models of neural-network computation. Specifically, if one population encodes $\mathbf{x}$ and a second population encodes $\mathbf{y} = \hat{\mathbf{f}}(\mathbf{x})$, the synaptic weights that produce this mapping can be determined by substituting $\hat{\mathbf{f}}(\mathbf{x})$ into (1). The result is that the synaptic weight between the i^{th} presynaptic and j^{th} postsyaptic neuron is $w_{ij} = \mathbf{e}_j^T \mathbf{d}_i$. Thus, a model can be developed systematically, beginning with a high-level description of encoded variables and how they are transformed.

Holographic Reduced Representations. HRRs represent concepts as vectors. They support operations that are useful for cognitive models including binding (associating concepts, e.g. associating "dog" with the role of "actor" in the sentence "dog bites man"); unbinding (e.g. extracting the fact that the "actor" is "dog"), and bundling (combining multiple bound and/or unbound concepts into a single vector). HRRs use circular convolution for binding and unbinding, and vector addition for bundling. HRR operations are lossy, e.g. "actor" bound to "dog" has the same vector dimension as "actor" or "dog". Eliasmith [9] showed that HRRs can be encoded and manipulated using NEF neural models, and that HRRs of a few hundred dimensions can store tens of thousands of concepts.

2.1 Proof-of-Concept Cognitive Model

As a first step in exploring the application of the NEF and HRRs to grasping, we developed a simplified model that uses basic drives and knowledge of the environment to choose a goal, and to influence hand posture in a manner consistent with that goal. To simplify the prototype we used abstract HRR vectors and sigmoidal units, given that the the NEF provides a systematic method to

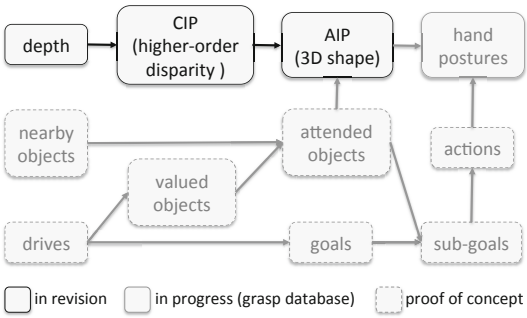


Fig. 1. Proof-of-concept model and its relationship to our other work. Dashed boxes indicate HRR populations and a winner-take-all "actions" population. Also shown are past work (black boxes) and other current work (solid gray box; see Introduction).

develop a spiking neural model from a vector model (this does not work with all vector models, but experience with the NEF suggests that the present model is a good candidate).

Grasping decisions were modelled in the space of the first two principal components of gripper parameters. A grid of sigmoidal units corresponded to different postures in this space. Decisions were made using a diffusion-to-bound mechanism [10], wherein each unit integrates its inputs until one unit's activity crosses a threshold, at which point the winning unit (corresponding to a single posture) inhibits all others. (In future work, this model could be elaborated so that decisions could are made through a distributed consensus across multiple areas [11].) Each input to this network corresponded to the influence of a different brain area on the posture decision, and consisted of a drive pattern across the posture grid. Input from a ten-dimensional object-shape representation was modelled as decoded functions $[f_{ij}(\mathbf{s})]$, where $\mathbf{s}$ is the shape parameters and i and j are grid indices. Desired actions were represented in a 200-dimensional HRR. Different actions were nearly orthogonal in this space, so we used a simple linear map, $\left[\sum_k \alpha_{ijk} \mathbf{a_k}^T \mathbf{a}\right]$, where $\mathbf{a}$ are action vectors and k is an index over possible actions.

We modelled a scenario in which an agent wants a drink of water given two potential sources: a bottle and a faucet. The agent must decide which source to use and the appropriate hand posture for grasping it. While the scope of this example is somewhat broader than grasp control, we wanted to verify that the basic approach was suitable for such examples. The input to the model included a basic "thirst" drive and a list of the objects in the environment (in a more complete system we take it that these would be detected visually and stored in working memory). We used HRR binding to associate water with both the bottle and the faucet. Furthermore, we used several similar vectors to represent different kinds of water, including cold spring water, warm spring water, and cold tap water. We used linear maps between HRRs to cause a "thirst" concept in the

"drives" HRR to probe the "environment" HRR for cold spring water, resulting in selection of the "bottle" concept. Further linear maps between HRRs led to an "action" HRR encoding "grasp" while the "attended object" HRR encoded "bottle". A final linear map from the binding of these two concepts influenced the posture network to choose a posture appropriate for grasping the bottle in order to pour from it.

We also further explored HRR encoding of objects as structures of bound and bundled concepts. Depending on their structure, the similarity between pairs of such HRRs may resemble the degree to which humans consider the corresponding items to be analogous or similar. Plate [6] showed this for both short sentences and simple spatial arrangements of shapes. This is relevant to grasping, in that humans often grasp objects that are functionally similar to known objects, but not identical to them. Humans can also think about substitutes if the ideal object for a certain purpose is not available. In a robotics application, analogies to a given object could be searched for in a large HRR memory simply by multiplying the object's vector with all the vectors in memory, and sorting any products that are above a threshold.

We encoded objects by bundling HRRs for their parts, shapes, structures (i.e. relationships between parts), affordances, and related constraints on grasping. As an example, we encoded a generic coffee mug as

$$\Big\langle parts \circledast \langle inside + cup\_side + opening + bottom + rim + handle \rangle + shape \circledast \langle cylinder\_like + curved\_handle \rangle + structure \circledast \langle inside\_opening + rim\_side + rim\_opening + bottom\_inside + handle\_side \rangle + affordances \circledast \langle drink\_from + pick\_and\_place + fill + pour\_from + hang \rangle + constraints \circledast drink\_from \circledast (do\_not\_cover \circledast opening + prefer\_grasp \circledast handle) \Big\rangle, \tag{3}$$

where most of the variables (e.g. *parts*, *inside*) are random base vectors, $\circledast$ is binding (circular convolution), $+$ is bundling (vector addition), and $<>$ indicates normalization of the vector inside the brackets. The terms that are bound to *structure* correspond to physical relationships between parts, and themselves contain further structures of random base vectors. For example,

$$inside\_opening = \big\langle attached \circledast (above \circledast opening + below \circledast inside) \big\rangle. \tag{4}$$

This expresses the knowledge that the inside of a mug (where the liquid sits) is connected with its opening (through which the liquid passes in and out). There are many reasonable ways to encode information about a given object in an HRR. However, a few variations on the above structure produced similar results, suggesting that these results are not very sensitive to such differences.

Finally, we also examined the accuracy with which grasp constraints could be extracted from such HRRs through unbinding. Specifically, we verified that similarity with a correct constraint vector was well separated from similarity with other vectors.

3 Results

3.1 Grasp Selection Network

Figure 2 shows a snapshot of activity in the hand-posture network, prior to a decision. The insets show two postures of the robot hand that correspond to two potential grips. The one on the left is better suited for lifting the bottle in order to pour from it, and is eventually selected. A different hand posture might be selected if the goal were different (e.g. to put the bottle in a refrigerator) or if the object itself was different.

Simulations of this proof-of-concept model demonstrated promising qualitative properties. First, the model incorporated multiple influences into the selection of a single hand posture. We simulated two specific influences: compatibility with object shape (from AIP); and compatibility with a specified action (from frontal areas). These influences could be arbitrarily broad, narrow, multimodal, etc. Second, the model maps from basic drives to a specific action plan given the objects in the environment. This mapping is oversimplified, but it verifies that such a mapping can be implemented using the NEF and HRRs. Third, the model could choose between multiple routes to the same goal. When we hard-coded the belief that the water bottle was cold, and searched for something similar to cold spring water, attention focused on the bottle. Alternatively, when we hard-coded the belief that the water bottle was warm, attention focused on the faucet instead. We expect that the model could be expanded to include updates based on sensory information.

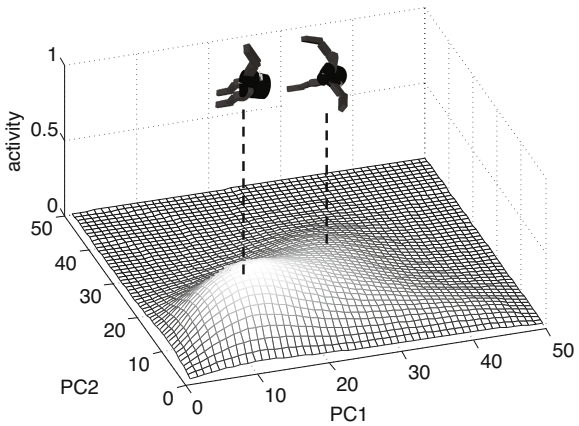


Fig. 2. Activation on a grid over the first two principal components of hand posture, during a decision between postures

3.2 Object Representation

Table 1 shows the similarities (inner products) between composite HRR encodings of four objects, including the *mug* example given in the Methods. The mug and cup are the most similar objects. The mug only differs from the cup in a few respects, e.g. it has a handle, one can hang it by the handle, and it is normally grasped by the handle for drinking. The spoon is not very similar to either the mug, cup, or pot. However in this encoding, it is most similar to the pot.

Table 1. Similarities between various objects encoded as HRRs

	cup	mug	pot	spoon
cup	1.00	0.78	0.55	0.11
mug	0.78	1.00	0.55	0.14
pot	0.55	0.55	1.00	0.21
spoon	0.11	0.14	0.21	1.00

This kind of encoding makes it possible to query rich information directly from the HRR using a series of unbinding and cleanup operations. For example, we queried one of the grasp constraints for drinking from a cup as,

$$cup \circledcirc constraint \circledcirc drink\_from \circledcirc do\_not\_cover, \tag{5}$$

where $\circledcirc$ indicates unbinding. The result is passed through a cleanup memory that replaces it with the most similar known vector, to obtain the result *opening*. (This constraint corresponds to the fact that the opening of a cup should not be covered by the hand when grasping to drink.) The intermediate results were not passed through cleanup memory, so noise (due to non-zero similarity with other parts of the *cup* HRR) was added at each deconvolution step, and the result had a relatively low similarity with the vector *opening* in memory. However, the resulting vector was still distinctly more similar to *opening* than to other vectors in memory, provided the dimension of the HRR was large enough. Figure 3 shows a histogram of similarities of this serial deconvolution with the *opening* vector and all the other vectors in memory with HRR dimension 4096. Target and non-target vectors are well separated.

4 Discussion

Two motivations for this research are: curiosity about the primate visuo-motor systems; and practical interest in robot controllers based on the same principles. While similar in spirit to the models studied in robotics [12–18], our work aims to implement affordances, a popular means of formalizing a robotic agent's interaction with the world [19], via a computational model that is compatible with the mechanisms that govern grasping in the primate brain (see [20] and [21] for

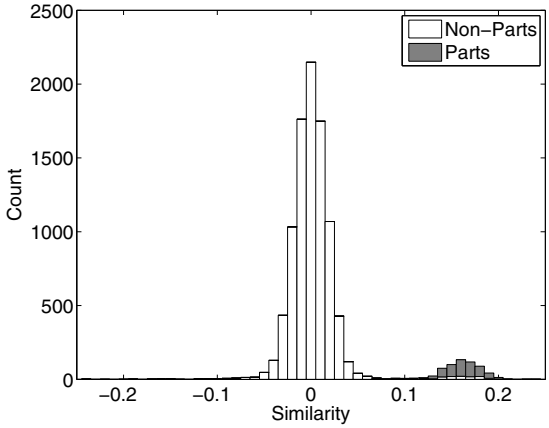


Fig. 3. Similarity of multiple-deconvolution estimate of the *do_not_cover* constraint for drinking from a cup, over 100 runs with random base vectors. Note that the target similarity counts have been multiplied by five (i.e. scaled up vertically), so that they can be seen more easily in the plot. The result of the unbinding has a higher similarity with the correct answer (i.e. *opening*) than with the other vectors, and is therefore reliably cleaned up.

models with similar goals). In other words, the key novelty is the use of a neurologically plausible model that will nonetheless be implemented on a real robot. Previous robotic implementations tend to at best be cast in connectionist terms inspired by neuroscience (for a discussion, see [19,22]). Models of the relevant brain areas similarly tend to be cast in connectionist terms [20,23,24] and analysed for behaviours that resemble that of actual neural circuits. By contrast, the approaches discussed in the present paper can draw more directly from neurophysiological data. Although our work is still at an early stage, this gives us hope that we can both achieve more biologically realistic control and contribute to the understanding of biological control mechanisms in a more in-depth manner than connectionist models can.

As an example, let us highlight that we have cast the model first and foremost in terms of a cognitive architecture for which the NEF provides a systematic way of deriving a neural model. As such, this imposes no a priori assumptions on the type and function of neurons in AIP (or F5 for that matter), instead giving us the freedom to investigate the functional contributions of the organisation of these areas [25] directly in terms of a cognitive architecture.

HRRs are a key component of the Spaun model, which can perform a wide variety of sophisticated tasks such as completing patterns from examples. We take the success of this approach in Spaun to suggest that HRRs provide a practical way to integrate a wide range of cognitive influences (such as verbal instructions) into models of neural visuo-motor systems. Our proof-of-concept model supports this view.

Acknowledgments. This work was supported by the Swedish Foundation for Strategic Research, the Swedish Research Council, the Belgian National Fund for Scientific Research, a DAAD-NRF Scholarship, and the Natural Sciences and Engineering Research Council of Canada.

References

1. Fagg, A.H., Arbib, M.A.: Modeling parietalpremotor interactions in primate control of grasping. Neural Networks **11**(7–8), 1277–1303 (1998)
2. Borra, E., Gerbella, M., Rozzi, S., Luppino, G.: Anatomical evidence for the involvement of the macaque ventrolateral prefrontal area 12r in controlling goal-directed actions **31**(34), 12351–12363 (2011)
3. Rezai, O., Kleinhans, A., Matallanas, E., Selby, B., Tripp, B.: Hierarchical object representations in the visual cortex and computer vision. Frontiers in Computational Neuroscience (in revision)
4. Cerri, G., Shimazu, H., Maier, M.A., Lemon, R.N.: Facilitation from ventral premotor cortex of primary motor cortex outputs to macaque hand muscles **90**(2), 832–842 (2003)
5. Eliasmith, C., Anderson, C.: Neural engineering. MIT Press (2003)
6. Plate, T.A.: Holographic Reduced Representation: Distributed representation for cognitive structures. Center for the Study of Language and Inf. (2003)
7. Eliasmith, C., Stewart, T.C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., Rasmussen, D.: A large-scale model of the functioning brain. Science **338**(6111), 1202–1205 (2012). PMID: 23197532
8. Berzish, M., Tripp, B.: A digital hardware design for real-time simulation of large neural-system models in physical settings. In: CNS (2014)
9. Eliasmith, C.: How to build a brain: A neural architecture for biological cognition. Oxford University Press (2013)
10. Gold, J.I., Shadlen, M.N.: The neural basis of decision making **30**, 535–574 (2007)
11. Cisek, P.: Making decisions through a distributed consensus. Current Opinion in Neurobiology **22**(6), 927–936 (2012)
12. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Learning object affordances: From sensory-motor coordination to imitation. IEEE Transactions on Robotics **24**(1), 15–26 (2008)
13. Stoytchev, A.: Learning the affordances of tools using a behavior-grounded approach. In: Rome, E., Hertzberg, J., Dorffner, G. (eds.) Towards Affordance-Based Robot Control. LNCS (LNAI), vol. 4760, pp. 140–158. Springer, Heidelberg (2008)
14. Sahin, E., Cakmak, M., Dogar, M.R., Ugur, E., Ucoluk, G.: To afford or not to afford: a new formalization of affordances toward affordance-based robot control. In: Adaptive Behavior (2007)
15. Sun, J., Garibaldi, J.: A novel memetic algorithm for constrained optimization. In: IEEE Congress on Evolutionary Computation, pp. 1–8 (2010)
16. Krüger, N., Piater, J., Geib, C., Petrick, R., Steedman, M., Wrgtter, F., Ude, A., Asfour, T., Kraft, D., Omren, D., Agostini, A., Dillmann, R.: Objectaction complexes: grounded abstractions of sensorymotor processes. In: Robotics and Autonomous Systems (2011)
17. Detry, R., Baseski, E., Krüger, N., Popovic, M., Touati, Y., Kroemer, O., Peters, J., Piater, J.: Learning object-specific grasp affordance densities. In: IEEE International Conference on Development and Learning, pp. 1–7 (2009)

18. Kjellström, H., Romero, J., Kragic, D.: Visual object-action recognition: Inferring object affordances from human demonstration. Computer Vision and Image Understanding **115**(1), 81–90 (2011)
19. Thill, S., Caligiore, D., Borghi, A.M., Ziemke, T., Baldassarre, G.: Theories and computational models of affordance and mirror systems: an integrative review. Neuroscience & Bio Behavioral Reviews **37**(3), 491–521 (2013)
20. Caligiore, D., Borghi, A.M., Parisi, D., Baldassarre, G.: Tropicals: a computational embodied neuroscience model of compatibility effects. Psychological Review **117**(4), 1188 (2010)
21. Oztop, E., Imamizu, H., Cheng, G., Kawato, M.: A computational model of anterior intraparietal (aip) neurons. Neurocomputing **69**(10–12), 1354–1361 (2006)
22. Oztop, E., Kawato, M., Arbib, M.A.: Mirror neurons and imitation: A computationally guided review. Neural Networks **19**, 254–271 (2006)
23. Oztop, E., Imamizu, H., Cheng, G., Kawato, M.: Neurocomputing **69**(10–12), 1354–1361 (June 2006)
24. Thill, S., Svensson, H., Ziemke, T.: Modeling the development of goal-specificity in mirror neurons. Cognitive Computation **3**(4), 525–538 (2011)
25. Murata, A., Gallese, V., Luppino, G., Kaseda, M., Sakata, H.: Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip **83**(5), 2580–2601 (2000). PMID: 10805659

OBEliSK: Novel Knowledgebase of Object Features and Exchange Strategies

David Cabañeros Blanco(✉), Ana Belén Rodríguez Arias, Víctor Fernández-Carbajales Cañete, and Joaquín Canseco Suárez

TREELOGIC Telemática y Lógica Racional para la Empresa Europea, Parque Tecnológico de Asturias, Parcela 30, E33428 Llanera, Asturias, Spain
{david.cabaneros,ana.rodriguez,victor.fernandez, joaquin.canseco}@treelogic.com

Abstract. This paper presents the design and development of a system intended for storing, querying and managing all required data related to a fluent human-robot object handover process. Our system acts as a bridge between visual perception and control systems in a robotic setup intended to collaborate with human partners, while the perception module provides information about the exchange environment. In order to achieve these goals, a semantic-ontological approach has been selected favouring system's interoperability and extensibility, complemented with a set of utilities developed ad-hoc for easing the knowledge inference, query and management. As a result, the proposed knowledgebase provides a completeness level not previously reached in related state of the art approaches.

Keywords: Ontologies · Knowledge representation · Handling affordances · Semantic modelling · Assistive robotics

1 Introduction

The work described in this paper comprises the design and development of a knowledgebase about the domain elements involved in the action of exchanging common objects between humans and robotic agents. Our approach requires an in-depth study of the state of the art, inputs from the perception system and a clear definition of the outputs required by the robotic control architecture. The main purpose of this knowledgebase is to model and transfer the acquired knowledge from human-human object exchange experiments to a robotic system, in order to achieve a fluent interaction between human and robotic agents.

The remainder of this paper is organized as follows. The next section (Section 2) introduces the theoretical concepts involved in our work. In Section 3, a current state of the art analysis is performed. Section 4 describes the design and development process of OBEliSK (Object Exchange applied Semantic Knowledgebase). Finally, we present our conclusions in Section 5.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 448–454, 2015.
DOI: 10.1007/978-3-319-16181-5_34

2 Theoretical Frame

Specific-domain knowledge is usually represented by means of ontological systems described by a computational model, composed by a set of entities corresponding to real world items, such as agents, objects or events connected by domain-specific rules [1]. This kind of representation requires an approach far from the "classical" relational database [2], so in this case we have considered using graph databases [3], [4]. The main advantage of such representation is the flexibility that provides when linking related entities, attributes and properties. For representing knowledge through this approach, while extending its usefulness, the Semantic Web [5], provides a common framework for sharing and reusing data and defines a standardized set of technologies, arranged in a hierarchical architecture. Additionally, the RDF (Resource Description Framework) [6] was designed as a method for splitting knowledge into small pieces of data, complemented with a set of rules defining the semantics of these individual fragments of isolated information, relying on RDF/XML syntax for expressing (i.e. serialize) a linked data graph as an XML document. Notation3 [7], Turtle [8] and N-Triples [9] formats were defined in order to ease the reading of RDF documents for humans. The Web Ontology Language (OWL) [10] was also selected in order to represent the envisioned model. A key benefit of the semantic-ontological approach is its reusability empowering, leading to knowledge representations that might be re-used in the development of different systems addressing similar purposes, while bringing interoperatibility between heterogeneous systems according to a consensuated knowledge representation.

As an approach for improving the expressiveness of traditional propositional logic, Description Logic (DL) languages were introduced as knowledge representation methods providing a logical formalism for ontology design, useful for concept representation and reasoning on the of domain-centred terminological knowledge. An axiom, as fundamental modeling element of a DL, is defined by a logical statement composed by a set of concepts, individuals and their relationships. A terminological axiom is defined as

$$C \doteq D \mid C \sqsubseteq D$$

where C and D are concepts. A finite set of terminological axioms is known as T-Box T and is defined using the following descriptions. Note that $I \models C$ stands for *"I models C"*, where I is an interpretation function and C is a concept

$$I \models C \sqsubseteq D \iff C^I \subseteq D^I \qquad I \models T \iff I \models \Phi \ \forall \ \Phi \in T$$

An assertional axiom, representing concepts positively stated, is composed by a set of statements representing basic knowledge about individuals classified within the T-Box hierarchy. An A-Box A is stated according to these definitions:

$$I \models a : C \iff a^I \in C^I \qquad I \models (a, b) : R \iff \left(a^I, b^I\right) \in R^I$$

$$I \models A \iff I \models \Phi \ \forall \ \Phi \in A$$

where a and b are individuals and R is a particular role. Given these formal definitions, a knowledgebase K is an ordered pair of T-Box and A-Box, defined as follows:

$$K = (T, A) \qquad I \models K \iff I = T \wedge I = A$$

3 State of the Art

Before starting with the design of a new ontology for the described problem, we carried out a study and evaluation of several related approaches that could help in the design of OBEliSK, including the following ones:

- **DEXMART**. This project's [11] key objectives were, among others, *i)* the development of original approaches for interpretation, learning and modelling of human object manipulation actions, and *ii)* the design of novel techniques for task planning for conferring the robotic system with self-adapting capabilities.
- **GRASP**. This project [12] had the objective of designing a cognitive system capable of grasping and handling objects in open environments where unexpected events may occur.
- **HANDLE**. Its [13] aim was to understand and replicate human object grasping and skilled hand movements using an anthropomorphic artificial hand by means of object affordances characterization for learning and replicating human handling tasks.
- **RoboEarth**. Designed as a cloud computing service, this robotic-oriented database [14] is focused on making robots capable of learning new abilities from other robots by easing their communication.

As summarized in Table 1, our approach tries to improve some of the shortages found in the previous state of the art study. The main advantage of our knowledgebase design is that provides the required set of mechanisms that makes it suitable for working together with both artificial vision and cognitive control modules, providing the robot with the required skills for achieving a fluent object exchange process.

4 Knowledgebase Design

Within the scope of the CogLaboration project, there is a need for modelling the entities to be handled by the robotic system. Instead of using a traditional relational database system, the decision of modelling the object taxonomy using semantic web based technologies provides the ability of modifying and expanding the knowledgebase in an easy and comprehensive way.

Table 1. Comparative analysis of different capabilities of our approach versus other representative projects introduced in Section 3

Functionality	OBEliSK	DEXMART	GRASP	HANDLE	RoboEarth
Integrated perception	✓	✗	✓	✗	✓
Human-like handling	✓	✓	✗	✓	✗
Obstacle avoidance	✓	✗	✓	✗	✓
Multi-mode grasping	✓	✓	✓	✓	✓
Object handover	✓	✗	✗	✗	✗
Standarized data store	✓	✗	✓	✗	✓
Motion constraints mgt.	✓	✓	✗	✗	✗
Learning capabilities	✓	✓	✗	✗	✗
Data management tool	✓	✗	✗	✗	✗

4.1 Object Perception

The perception system captures a 3-D model of the object using a Kinect sensor. Object models are processed and a set of partial views is extracted from the whole model. These views are then employed for computing feature descriptors to be used in the classification process. Taking into account that is impractical to store these views (161 per object; 30 MB each) in a serialized form in the knowledgebase, views' file paths are stored instead. This is because transmission and deserialization tasks are highly time-consuming and is unacceptable to be used in *i)* the perception subsystem, intended for real-time operation, and *ii)* the robotic cognitive control subsystem, conceived for executing a fluent interaction.

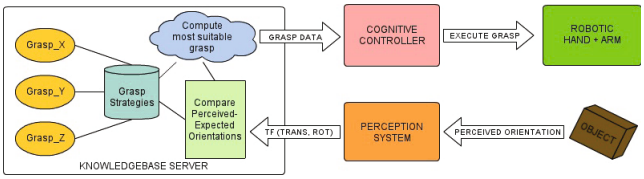


Fig. 1. Grasp strategy selection task flow, from the object's perception to the robotic physical motion execution

4.2 Object Handling

Exchange Properties. Besides the object visual properties representation, it is also crucial to store and manage the set of features describing the way in which each object can be handled during the handover phase of the exchange process. Each object is associated to a set of grasp postures and delivery strategies, defining different ways the robot can handle it. Moreover, in order to ensure a proper manipulation process for certain objects, we have introduced the concepts of

Motion Constraint, standing for the restrictions to be applied to some objects during their handling, and *Symmetry Axis*, representing the object's axial symmetry, if there are any.

Object Affordances. We have also explored the concept of object affordances [15] within the task of modelling the ontology [16]. The taken approach is related to the concept of affordances and based on the idea of categorizing objects based on how they are used. According to Gibson's Theory of Affordances [17], affordances can be seen as the sum of the properties of a situation, including agents, environment and objects, especially those that describe how they can be used to do something [18].

Grasping Phase. The set of grasps to be considered relies on the automatic grasping capabilities provided by the IH2 Azzurra [19] robotic hand. These grasps are based on the taxonomy developed by Cutkosky [20] and modelled under the class *GraspType*. As far as the work developed by Cutkosky is the design of the grasp taxonomy, the modelling process using OWL is made straight from that one to our model, thanks to the hierarchical shape and the classification-oriented vocabulary respectively. Each grasp instance, called *named individual* inside this context, represents an object-specific grasping configuration.

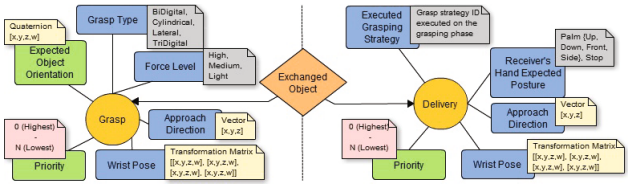


Fig. 2. *Grasp* and *Delivery* model conceptualizations

Delivery Phase. We also considered the idea of improving the knowledgebase value for the project by extending the initial conception of a grasping database to a fully-featured handling knowledgebase, covering in this way the second half of the object exchange process. The control system has to be provided with relevant data about the object handover, being capable to deliver the previously grabbed object to the recipient in a fluent and natural way.

4.3 Knowledgebase Data Management

Data processing and storage in this kind of databases is not trivial. Having a large amount of information and a defined ontology, it is mandatory to fully respect the relational integrity restrictions between entities and their properties. Semantic-ontological data management is usually done through semantic-oriented tools, such as Protégé [21]. With the aim of ease this task, a utility has been developed

Table 2. Summary of grasp and delivery actions model concepts

Concept	Grasping ph.	Delivery ph.
Object involved in the action	✓	✓
Grasp type, from taxonomy primitives	✓	✗
Expected object orientation/receiver's hand posture	✓	✓
Grasp strategy previously executed	✗	✓
Grasp force level to be reached by the fingers	✓	✗
Object's grab/release approach direction	✓	✓
Robotic hand wrist pose	✓	✓
Strategy selection preference (priority)	✓	✓

focused on offering the simplest way to manage the knowledgebase contents. It consists of a web-application acting as interface between the user and the triple store where the ontology data is saved.

5 Conclusions

This paper introduced the design of a robotic handling knowledgebase by means of semantic-ontological technologies, providing an interesting and innovative approach. The developed system meets the expectations and overcomes them, as we extended the initially proposed grasping model to a complete exchange one due to the inclusion of object delivery concepts, improving the object handling fluency of the robotic system, including its adaptability to each particular situation in non-deterministic scenarios.

We are concerned with the interoperability needs of this task among the rest of project's developments, so we have dedicated a considerable amount of our efforts to provide simple, understandable and comprehensive interfaces for both inputs and outputs of this system.

Acknowledgments. This work was supported by the CogLaboration project within the 7th Framework Programme of the European Union, Cognitive Systems and Robotics - Contract Number FP7-287888. See also http://www.coglaboration.eu/.

The authors would also like to thank Dr. Rubén Casado, Dr. Marco Controzzi and Mr. Mario Recio for their valuable comments.

References

1. Presutti, V., Gangemi, A.: Content ontology design patterns as practical building blocks for web ontologies. In: Li, Q., Spaccapietra, S., Yu, E., Olivé, A. (eds.) ER 2008. LNCS, vol. 5231, pp. 128–141. Springer, Heidelberg (2008)
2. Codd, E.F.: A relational model of data for large shared data banks. Communications of the ACM **13**(6), 377–387 (1970)
3. Angles, R., Gutierrez, C.: Survey of graph database models. ACM Computing Surveys (CSUR) **40**(1), 1 (2008)

4. Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., Wilkins, D.: A comparison of a graph database and a relational database: a data provenance perspective. In: Proceedings of the 48th Annual Southeast Regional Conference, p. 42. ACM (2010)
5. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific American **284**(5), 28–37 (2001)
6. Brickley, D., Guha, R.V.: Resource Description Framework (RDF) Schema Specification 1.0: W3C Candidate Recommendation, 27 March (2000)
7. Berners-Lee, T.: Notation 3-an readable language for data on the web (2006)
8. Beckett, D., Berners-Lee, T., Prudhommeaux, E.: Turtle-terse rdf triple language. W3C Team Submission 14 (2008)
9. Beckett, D., Barstow, A.: N-triples. W3C RDF Core WG Internal Working Draft (2001)
10. McGuinness, D.L., Van Harmelen, F., et al.: OWL web ontology language overview. W3C Recommendation **1**0(10) (2004)
11. Siciliano, B.: DEXMART. DEXterous and autonomous dual-arm/hand robotic manipulation with sMART sensory-motor skills: A bridge from natural to artificial cognition (2008–2012). http://www.dexmart.eu
12. Kragic, D.: GRASP. Emergence of Cognitive Grasping through Introspection, Emulation and Surprise (2008–2012). http://www.csc.kth.se/grasp
13. Perdereau, V.: HANDLE: Developmental pathway towards autonomy and dexterity in robot in-hand manipulation (2009–2013). http://www.handle-project.eu
14. Tenorth, M., Bartels, G., Beetz, M.: Knowledge-based specification of robot motions
15. Varadarajan, K.M., Vincze, M.: Knowledge representation and inference for grasp affordances. In: Crowley, J.L., Draper, B.A., Thonnat, M. (eds.) ICVS 2011. LNCS, vol. 6962, pp. 173–182. Springer, Heidelberg (2011)
16. Varadarajan, K., Vincze, M.: Ontological knowledge management framework for grasping and manipulation. In: IROS Workshop on Knowledge Representation for Autonomous Robots (2011)
17. Gibson, J.J.: The theory of affordances. Hilldale, USA (1977)
18. Varadarajan, K.M., Vincze, M.: AfNet: the affordance network. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 512–523. Springer, Heidelberg (2013)
19. Prensilia s.r.l. - Robotic Hands (Self-contained). http://www.prensilia.com/index.php?q=en/node/40 (accessed: July 1, 2014)
20. Cutkosky, M.R.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. Robotics and Automation, IEEE Transactions on **5**(3), 269–279 (1989)
21. Knublauch, H., Fergerson, R.W., Noy, N.F., Musen, M.A.: The Protégé OWL plugin: An open development environment for semantic web applications. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 229–243. Springer, Heidelberg (2004)

How Industrial Robots Benefit from Affordances

Kai Zhou(✉), Martijn Rooker, Sharath Chandra Akkaladevi, Gerald Fritz, and Andreas Pichler

PROFACTOR GmbH, Im Stadtgut A2, 4407 Steyr-Gleink, Austria
{kai.zhou,martijn.rooker,sharath.akkaladevi,
gerald.fritz,andreas.pichler}@profactor.at

Abstract. In this paper we discuss the potential of Gibson's affordance concept in industrial robotics. Recent advances in robotics introduce more and more robots to collaborate with human co-workers in industrial environments, more ambitious development of using mobile manipulators in industrial environments has also received widespread attentions. We investigate how the conventional robotic affordance concept fits the pragmatic industrial robotic applications with the focuses on flexibility, re-purposing and safety.

1 Why Affordances

Majority of today's industrial robots operating in factories are attached to a fixed basement, operate on the various parts passing through a production line. Although they can be reprogrammed with a teach pendant, in many applications (particularly those in the automotive industry) they are programmed once and then fixed behind metal fences, where they repeat that exact same task for years. In recent years, however, collaborative robots have received more attention in manufacturing industry as they can safely work together with human workers in efficient new ways, e.g. to perform the task that requires a robot to do the physical labor while a person does quality-control inspections. High **complexity and uncertainty** of system caused by dealing with a large number of objects, requirement of **fast re-purposing and deployment** for new or swapped tasks and **safety** awareness are three major challenges that are consequent on the utilization of collaborative robots in industry.

The concept of affordances has been coined by J.J. Gibson [1] in his seminal work on the ecological approach to visual perception. Although there are several attempts to formalize the theoretical concept (see [2] for an overview), the idea of a relationship combining *perception*, *action* and *outcome* is innate to most approaches and first formalized in [3]. Mapping the concept of affordance into the domain of industrial robotics could

- reduce the uncertainty and complexity caused by a large number of objects and objects in arbitrary positions/poses in human involved collaboration;

This work was supported by project "Adaptive Produktion 2014", which is funded by European Regional Development Fund (ERDF) and the county of Upper Austria.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 455–458, 2015.
DOI: 10.1007/978-3-319-16181-5_35

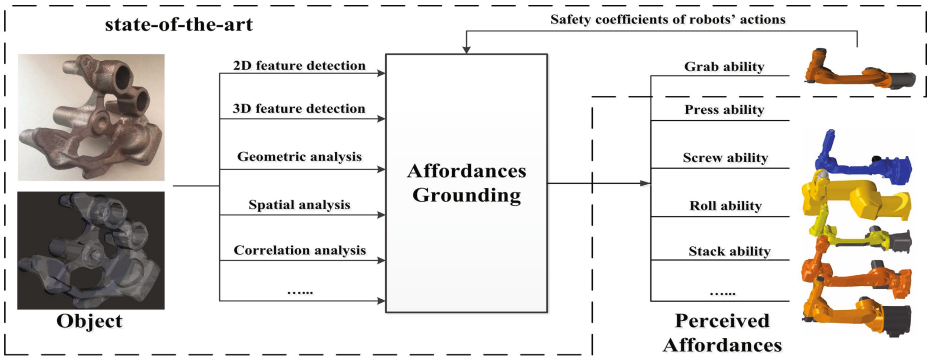


Fig. 1. Schema of using affordances for industrial robotic applications: differing with the traditional affordance-based robotic grasping applications which use one configuration of robot/end effector to handle various objects, we select the most suitable robot/end effector according to the object and its affordable actions.

- increase the flexibility and fast re-purposing of tasks, since affordances naturally rely on actors' abilities and the grounded affordances of object provide attached information about which tool/robot should be used to do various actions;
- provide an alternative safety concept since affordances are always related to actions, which can be assigned to different safety evaluations according to the control parameters of these actions.

Therefore, we propose a new systematic schema, which mediates information of perceived object (e.g. 2D/3D features, geometrical characters etc.) and safety awareness data of actions that could be executed by different robots/end-effectors, to produce perceived affordances that can be safely and effectively used by industrial robots (Fig. 1).

2 How to Use Affordance for Industrial Robotics

While the state-of-the-art affordance-driven robotic grasping is focusing on the solution to find the best grasp points of various objects, we in contrast use affordance to help our industry partners to decide which robot/end effectors are the most suitable for manipulating the specific object in their application (Fig. 2). Also the affordance based evaluation of robot/end effector combinations can provide customers both quantitative and qualitative measurements, thereby facilitating the most suitable solution for trade off hardware cost and system productivity.

For several industrial applications, particularly for the applications involving large part manufacturing such as aerospace industry or shipbuilding industry, large parts are worked on in a stationary production cell. In such a production

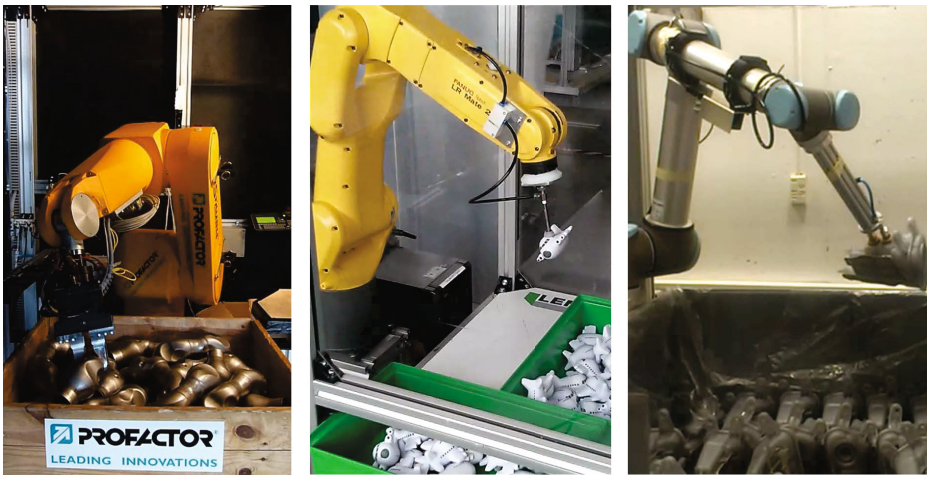


Fig. 2. Various robots/end effectors designed by PROFACTOR GmbH, are picking different objects, systems are designed according to the affordances (based on shape, material, weight etc.) of the objects it faces on.

environment, specialized, stationary robotic systems are not economical and a mobile manipulator is desirable [4]. These mobile manipulators with tool changer can benefit from affordance-based end effector selection when cost of changing tool and safety coefficient of actions are taken into account, i.e. use affordance to evaluation the successful rate of executing one specific action with various end effectors thereby deciding if the tool changing behavior is required at the present time.

Modern vision-based algorithms for feature detection or character analysis normally have quality estimation outputs as part of their results[5]. These quality estimation values can be used in a unified probabilistic framework to discover a best holistic solution [6][7]. We plan to expand this probabilistic framework by combining quality of object analysis/detections and safety estimation of using various robots/end-effectors/tools to execute different action tasks. The maximization of the joint probability will find the safest and most reliable affordance of object which can be manipulated with one specific robotic hardware configuration. Following the work of modeling affordances using Bayesian Network [8], we further include the successful rate of using different tools/robots/end-effectors for various execution tasks, to make the system able to decide whether it requires to change the tool or not, as industrial robots usually are equipped with many tools in order to perform various tasks. Future optimization of tool change frequencies and workflow could also be developed based on this probabilistic framework.

References

1. Gibson, J.J.: The ecological approach to visual perception. Lawrence Erlbaum Associates, Resources for ecological psychology (1986)
2. Şahin, E., Çakmak, M., Doğar, M.R., Uğur, E., Üçoluk, G.: To afford or not to afford: A new formalization of affordances toward affordance-based robot control. Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems **15**(4), 447–472 (2007)
3. Rome, E., Paletta, L., Şahin, E., Dorffner, G., Hertzberg, J., Breithaupt, R., Fritz, G., Irran, J., Kintzler, F., Lörken, C., May, S., Uğur, E.: The MACS project: an approach to affordance-inspired robot control. In: Rome, E., Hertzberg, J., Dorffner, G. (eds.) Towards Affordance-Based Robot Control. LNCS (LNAI), vol. 4760, pp. 173–210. Springer, Heidelberg (2008)
4. Zhou, K., Ebenhofer, G., Eitzinger, C., Zimmermann, U., Oriol, J.N., Castaño, L.P., Hernández, M.A.F., Walter, C., Saenz, J.: Mobile manipulator is coming to aerospace manufacturing industry. In: The 12th IEEE International Symposium on RObotic and SEnsors Environments (ROSE 2014) (2014)
5. Zhou, K., Richtsfeld, A., Zillich, M., Vincze, M.: Coherent spatial abstraction and stereo line detection for robotic visual attention. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011) (2011)
6. Zhou, K., Richtsfeld, A., Zillich, M., Vincze, M., Vrečko, A., Skočaj, D.: Visual information abstraction for interactive robot learning. In: The 15th International Conference on Advanced Robotics (ICAR 2011), Tallinn, Estonia, June 2011
7. Zhou, K., Varadarajan, K.M., Richtsfeld, A., Zillich, M., Vincze, M.: From holistic scene understanding to semantic visual perception: a vision system for mobile robot. In: ICRA 2011 Workshop: Semantic Perception, Mapping and Exploration (SPME), Shanghai, May 2011
8. Montesano, L., Lopes, M., Bernardino, A., Santos-Victor, J.: Modeling affordances using bayesian networks. IROS 2007, pp. 4102–4107, October 2007

The Aspect Transition Graph: An Affordance-Based Model

Li Yang Ku(✉), Shiraj Sen, Erik G. Learned-Miller, and Roderic A. Grupen

School of Computer Science, University of Massachusetts Amherst, Amherst, MA 01003, USA
{lku,shiraj,elm,grupen}@cs.umass.edu

Abstract. In this work we introduce the Aspect Transition Graph (ATG), an affordance-based model that is grounded in the robot's own actions and perceptions. An ATG summarizes how observations of an object or the environment changes in the course of interaction. Through the Robonaut 2 simulator, we demonstrate that by exploiting these learned models the robot can recognize objects and manipulate them to reach certain goal state.

Keywords: Robotic perception · Object recognition · Belief-space planning

1 Introduction

The term affordance first introduced by Gibson [2] has many interpretations, we prefer the definition of affordance as "the opportunities for action provided by a particular object or environment." Affordance can be used to explain how the "value" or "meaning" of things in the environment is perceived. Our models are based on this interactionist view of perception and action that focus on learning relationships between objects and actions specific to the robot. Some recent work [6] [8] [13] in computer vision and robotics extended this concept of affordance and applied it to object classification and object manipulation. Affordances can be associated with parts of an object as, for example in the work done by Varadarajan [16] [15], where predefined base affordances are associated with surface types. In our work, we build models that inform inference in an extension of Gibson's original ideas about direct perception [3] [5].

Affordances describe the interaction between an agent and an object (or environment). For example [2], a chair that is "sittable" for a grown-up might not be "sittable" for a child. In this work we introduce the Aspect Transition Graph (ATG), an affordance-based model that is grounded in the robot's own actions and perceptions. Instead of defining object affordances from a human perspective, they are learned by direct interaction on the part of the robot. Using the Robonaut 2 simulator [1], we demonstrate that by exploiting these learned models the robot can recognize objects and manipulate them to reach goal states.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 459–465, 2015.
DOI: 10.1007/978-3-319-16181-5_36

2 Aspect Transition Graph

Aspect Graphs were first introduced to represent shape [9] [4] in the field of computer vision. An Aspect Graph contains distinctive views of an object captured from a viewing sphere centered on the object. The Aspect Transition Graph introduced in this paper is an extension of this concept. In addition to distinctive views, the object model summarizes how actions change viewpoints or the state of the object and thus, the observation. Besides visual sensors, extensions to tactile, auditory and other sensors also become possible with this representation. The term Aspect Transition Graph was first used in [12] but redefined in this work.

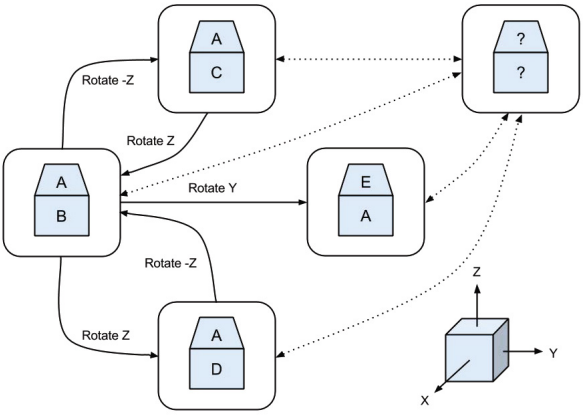


Fig. 1. An example of an incomplete aspect transition graph (ATG) of a cube. Each aspect consists of an observation of two faces of the cube. The lower right figure shows the coordinate frame of the actions and the aspect in the upper right is the "collection node" representing all unknown aspects of the object that may be present. Each solid edge represents a transition between aspects associated with a particular action. Each dotted edge is a transition that may not yet have been observed.

An object in our framework is represented using a directed graph $G = (\mathcal{X}, \mathcal{U})$, composed of a set of aspect nodes $\mathcal{X}$ connected by a set of action edges $\mathcal{U}$ that capture the probabilistic transition between the aspect nodes. Each aspect $x \in \mathcal{X}$ represents the features of an object that are measurable given a set of sensors and their relative geometry to the object. The ATG summarizes empirical observations of aspect transitions in the course of interaction.

The ATG of an object is complete if it contains all possible aspect nodes and node transitions. However, in practice, when ATGs are learned through exploration they are almost always incomplete. In addition, an object might be represented by multiple (incomplete) ATGs. A complete model is more informative but harder to learn autonomously. In this paper, we will focus on handling incomplete object models. Each of our ATG models have a single collection node

representing all unobserved aspects. Figure 1 shows an example of an incomplete ATG on a cube object with a character on each face.

3 Modeling and Recognition

The robot memory $\mathcal{M}$ is defined as a set of ATGs that the robot created through past interaction. Each ATG in the robot memory represents a single object presented to the robot in the past. An ATG is added to the $\mathcal{M}$ only if the presented object is judged to be novel. Although the robot might not have seen all the objects or all the aspects of each object, to simplify this problem we make this very limiting assumption that the robot knows that $|\mathcal{O}|$ objects exist in the environment and each object has $|G|$ aspects. If the robot assumes that there are more objects in the environment or more aspects of an object then there actually are, it will bias the judgment toward novelty.

Let $\mathcal{S}_{T-1}$ denote the set of objects that have been presented to the robot in the first $T-1$ trials. Given a sequence of observations $z_{1:t}$ and actions $a_{1:t}$ during trial T, the probability that the object presented during trial T, O_T, is novel can be calculated;

$$\begin{aligned} & p(O_T \notin \mathcal{S}_{T-1} | z_{1:t}, a_{1:t}, \mathcal{M}) \\ &= \sum_{o_i \notin \mathcal{S}_{T-1}} p(O_T = o_i | z_{1:t}, a_{1:t}, \mathcal{M}) \\ &= \sum_{o_i \notin \mathcal{S}_{T-1}} \sum_{x_t \in \mathcal{X}_i} p(x_t | z_{1:t}, a_{1:t}). \end{aligned} \tag{1}$$

Where o_i is an element of set $\mathcal{O}$ designating all of the objects in the environment. Element x_t of set $\mathcal{X}_i$ describes all the aspects comprising object o_i. The conditional probability $p(x_t|z_{1:t}, a_{1:t})$ of observing an aspect can be inferred using a Bayes filter. Object O_T is classified as novel if $p(O_T \notin \mathcal{S}_{T-1}|z_{1:t}, a_{1:t}, \mathcal{M}) > 0.5$.

If a particular object is judged to be a previously observed object, we associate it with the ATG that is most likely to generate the same set of observations. The posterior probability of object o_i is calculated by summing the conditional probability of observing aspect x_t over all aspects comprising object o_i,

$$p(O_T = o_i | z_{1:t}, a_{1:t}, \mathcal{M}) = \sum_{x_t \in \mathcal{X}_i} p(x_t | z_{1:t}, a_{1:t}). \tag{2}$$

The posterior probability of an aspect $p(x_t|z_{1:t}, a_{1:t})$ is calculated after each measurement and control update using the Bayes Filter Algorithm [14].

4 Action Selection Strategy

The challenge of integrating task-level planners with noisy and incomplete models requires that we confront the partial observability of the state while building

plans. Since the true state of the system cannot be observed, it must be inferred from the history of observations and actions. Our planner belongs to a set of approaches (for example [7][10]) that select actions to reduce the uncertainty of the state estimate maximally with respect to the task.

Object recognition can be viewed as one such task in which the uncertainty over object identities (as quantified by the object entropy) is reduced with each observation. Our task planner selects the action a_t that minimizes the expected entropy of the distribution over elements of set O_T representing the object identity [11];

$$\begin{aligned} &\underset{a_t}{\operatorname{argmin}}\ E(H(O_T|z_{t+1}, a_t, z_{1:t}, a_{1:t-1})) \\ &= \underset{a_t}{\operatorname{argmin}} \sum_{z_{t+1}} H(O_T|z_{t+1}, a_t, z_{1:t}, a_{1:t-1}) \times \\ &\quad p(z_{t+1}|a_t, z_{1:t}, a_{1:t-1}). \end{aligned} \tag{3}$$

Once the object entropy is lower than a threshold, the robot has high certainty regarding the ATG in robot memory that represents the same object. All aspect nodes in this ATG that are reachable from the current aspect node represent the set of aspects that the robot can observe by executing a sequence of actions. If a goal aspect is in one of these aspects, the actions on the shortest path from the current aspect node to the goal aspect node represents an optimal sequence of actions for achieving the goal state.

5 Experiments

We evaluated the capabilities of the proposed affordance models and planner using the Robonaut 2 simulator shown in Figure 2. The simulation contains 100 unique objects called ARcubes that consist of a 28cm cube with unique combinations of ARtags on the six faces; 12 different ARtag patterns are used in this experiment. In an ATG for an ARcube, an aspect consists of ARtag features observed on 2 faces. Each ATG has 24 unique aspects and each aspect has 132 different pattern combinations. For the sake of simplicity, we assume that an object does not have two faces with the same ARtag. The robot can perform 3 different manipulation actions on the object: 1) flip the top face of the cube to the front, 2) rotate the left face of the cube to the front, and 3) rotate the right face of the cube to the front. We emphasize that our model is not restricted to cube like structures and that every inference is based on the combination of observation and action.

Table 1 shows the result of using the planner to recognize the object presented. Each test involves 100 trials and starts with an empty robot memory $\mathcal{M}$. In each trial, the task is to decide which ATG in memory the experiment corresponds to or to declare it to be novel. For each trial, an object is chosen at random and presented to the robot. The robot observes the object and executes an action. This process is repeated 10 times. At the end of each trial the robot

Fig. 2. The simulated Robonaut 2 interacting with an ARcube

determines the likelihood that the presented object is novel and the most likely existing object in memory is identified.

The last row in Table 1 presents the results averaged over all the tests. The success rate is the percentage of objects correctly classified, that is, correctly identified in memory or declared as a novel object. The system correctly recognizes the object 90.7% of the time, and correctly determines if the presented object is novel or not 81.6% of the time.

Table 1. The success rate of an information theoretic planner in recognizing the object (10 actions per trial)

Test	Correct Identification	Correct Recognition	Success Rate
1	80/100	20/21	79%
2	79/100	25/27	77%
3	87/100	21/25	83%
4	78/100	26/28	76%
5	84/100	24/27	81%
average	81.6%	90.7%	79.2%

We also tested the efficiency of the planner against a random policy. The number of actions executed per trial were varied from 4 to 20. Figure 3 shows how the success rate of a test varies with the number of actions executed per trial. As is evident from the plots, the information theoretic planner outperforms a random exploration policy for all cases except when the number of actions per trial is low. Both algorithms perform equally poor when not enough information is provided.

To demonstrate how ATGs can be used to reach certain goal state. We set up an environment where 3 ARcubes are located in front of the simulated Robonaut 2 as shown in Figure 2. The goal is to rotate the cubes till certain faces are observable. The robot starts with a robot memory learned through interacting with 20 different ARcubes including the 3 ARcubes located in the test environment. To achieve the goal state, Robonaut 2 manipulates the object to condense

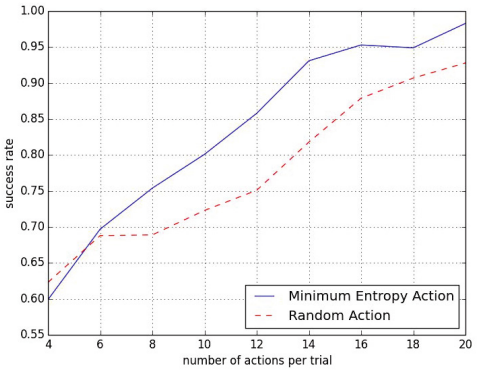


Fig. 3. The plot shows the average success rate of 10 tests as the number of actions per trial are increased. Selecting actions that minimize entropy leads to a higher success rate then selecting actions at random.

belief over objects. Once the object entropy is lower than a threshold, Robonaut 2 tries to execute the sequence of actions that is on the shortest path from the current aspect node to the goal aspect node in the corresponding ATG if such goal aspect node exists. In this experiment, the simulated Robonaut 2 successfully reaches the goal state by manipulating the cubes so that the observed aspects match the given goal aspects.

6 Discussion

This paper introduces an affordance-based model and demonstrates that it can be learned and used to support inference in a simulated environment with discrete actions and observations. To apply this model to real world applications, several challenges need to be addressed. First, in this work we assume that all actions lead to aspect transitions for all objects. A more realistic assumption will relate actions to new aspects probabilistically. Second, unlike ARcubes, real objects do not have a set of unique aspects; metrics such as the deviation of a new observation to past observations can be used to determine if a new aspect is observed. For future work, we plan to address these difficulties and test the ATG model in a more realistic environment. We are also exploring how to represent interactions between multiple objects in the scene and extensions of the idea that can incorporate multi-modal sensory features like tactile data.

Acknowledgments. This material is based upon work supported under Grant NASA-GCT-NNX12AR16A and a NASA Space Technology Research Fellowship.

References

1. Dinh, P., Hart, S.: NASA Robonaut 2 Simulator (2013). http://wiki.ros.org/nasa_r2_simulator (accessed on July 7, 2014)
2. Gibson, J.: The Theory of Affordance. In: Perceiving, Acting, and Knowing: Toward an Ecological Psychology. Lawrence Erlbaum Associates, Michigan (1977)
3. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin, Boston (1979)
4. Gigus, Z., Malik, J.: Computing the aspect graph for line drawings of polyhedral objects. IEEE Transactions on Pattern Analysis and Machine Intelligence **12**(2), 113–122 (1990)
5. Goldstein, E.B.: The ecology of jj gibson's perception, pp. 191–195. Leonardo (1981)
6. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1529–1536. IEEE (2011)
7. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artificial Intelligence **101**(1), 99–134 (1998)
8. Katz, D., Venkatraman, A., Kazemi, M., Bagnell, J.A., Stentz, A.: Perceiving, learning, and exploiting object affordances for autonomous pile manipulation (2013)
9. Koenderink, J.J., van Doorn, A.J.: The internal representation of solid shape with respect to vision. Biological Cybernetics **32**(4), 211–216 (1979)
10. Platt, R., Tedrake, R., Kaelbling, L., Lozano-Perez, T.: Belief space planning assuming maximum likelihood observations. In: Proceedings of Robotics: Science and Systems, Zaragoza, Spain (2010)
11. Sen, S., Grupen, R.: Manipulation planning using model-based belief dynamics. In: Proceedings of the 13th IEEE-RAS International Conference on Humanoid Robots, Atlanta, Georgia (October 2013)
12. Sen, S.: Bridging the gap between autonomous skill learning and task-specific planning. Ph.D. thesis, University of Massachusetts Amherst (2013)
13. Stoytchev, A.: Toward learning the binding affordances of objects: a behavior-grounded approach. In: Proceedings of AAAI Symposium on Developmental Robotics, pp. 17–22 (2005)
14. Thrun, S., Burgard, W., Fox, D.: Probabilistic robotics. MIT Press (2005)
15. Varadarajan, K.M., Vincze, M.: Object part segmentation and classification in range images for grasping. In: 2011 15th International Conference on Advanced Robotics (ICAR), pp. 21–27. IEEE (2011)
16. Varadarajan, K.M., Vincze, M.: Afrob: The affordance network ontology for robots. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1343–1350. IEEE (2012)

W12 - Graphical Models in Computer Vision

MAP-Inference on Large Scale Higher-Order Discrete Graphical Models by Fusion Moves

Jörg Hendrik Kappes(✉), Thorsten Beier, and Christoph Schnörr

Heidelberg Collaboratory for Image Processing, Heidelberg University, Heidelberg, Germany
kappes@math.uni-heidelberg.de

Abstract. Many computer vision problems can be cast into optimization problems over discrete graphical models also known as Markov or conditional random fields. Standard methods are able to solve those problems quite efficiently. However, problems with huge label spaces and or higher-order structure remain challenging or intractable even for approximate methods.

We reconsider the work of Lempitsky et al. 2010 on fusion moves and apply it to general discrete graphical models. We propose two alternatives for calculating fusion moves that outperform the standard in several applications. Our generic software framework allows us to easily use different proposal generators which spans a large class of inference algorithms and thus makes exhaustive evaluation feasible.

Because these fusion algorithms can be applied to models with huge label spaces and higher-order terms, they might stimulate and support research of such models which may have not been possible so far due to the lack of adequate inference methods.

1 Introduction

Many computer vision problems can be cast into optimization problems over discrete graphical models also known as Markov or conditional random fields. While standard methods are able to solve those problems quite efficiently, problems with huge label spaces and or higher-order structure are still challenging and even approximate methods do not scale well.

Consequently, research has focused on models with moderate order and small label spaces [7,19,33], models with huge but decomposable label spaces [12], or higher-order models that can be reformulated into second order models with additional auxiliary variables [6,21,22].

A more generic approach to deal with large label spaces has been suggested by Lempitsky et al. [26]. Starting with an initial labeling, they generate an alternative proposal and search for a better labeling within the subspace of labeling

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-16181-5_37) contains supplementary material, which is available to authorized users.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 469–484, 2015.
DOI: 10.1007/978-3-319-16181-5_37

spanned by the current and the proposed labeling. This step is called *move*, since the current labeling is moved within the subspace without increasing the energy. Except for some special cases, e.g. [25], finding the optimal move for a given proposal is NP-hard. The common way to calculate a move exploits that the problem is binary and QPBO is used to calculate a labeling with a persistency certificate [4,31]. For all persistent variables we can change the current label to the persistent one and do not increase the energy. This procedure has been generalized to higher-oder problems by reducing the higher-order binary subproblems two second-order ones and additional auxiliary variables [10,15,17].

A complementary part of fusion algorithms that need to be specified is the generation of proposal. Proposal generators can be generic or problem specific. As discussed in [26] a good proposal should have a high *quality* and the proposals should be *diverse* among each other to allow various moves.

Except for fusion with simple α-proposals in [19], fusion moves have not been considered in recent benchmarks [7,18,19]. This might be caused by the lack of a publicly available implementations and the option to choose *any* generator. Likewise, in many applications fusion moves with less generic *problem specific* proposal generators have been used.

Contribution: (1) The first *publicly available generic implementation of fusion moves*. It supports user defined proposal generators and is embedded into the OpenGM-Library [1]. **(2)** *Two novel methods for calculation fusion moves* that outperform QPBO in several settings. **(3)** We show how *improved any-time performance* of state-of-the-art methods can be obtained by embedding them into the fusion framework. **(4)** A *detailed evaluation* of proposal generators and fusion algorithms on recent published and new benchmark datasets.

Outline: We start in Sec. 2 with the mathematical formulation of the problem and fusion moves and present in Sec. 3 novel and state-of-the-art methods to calculate them. In Sec. 4 we present some generic proposal generators. In the experimental section 5 we evaluate the performance of fusion-methods and proposal generators on recent benchmark datasets and conclude in Sec. 6.

2 Problem Formulation

We assume that our discrete energy minimization problem is specified on a factor graph $G = (V, F, E)$, a bipartite graph with finite sets of variable nodes V and factors F, and a set of edges $E \subset V \times F$ that defines the relation between those [23,28]. The variable x_a assigned to the variable node $a \in V$ lives in a discrete label-space X_a and notation X_A, $A \subset V$, stands for a Cartesian product $\otimes_{a \in A} X_a$. Each factor $f \in F$ has an associated function $\varphi_f : X_{ne(f)} \rightarrow \mathbb{R}$, where $ne(f) := \{v \in V : (v, f) \in E\}$ defines the variable nodes connected to the factor f. The functions φ_f will also be called *potentials*. We define the order of a factor by its degree $|ne(f)|$, e.g. pairwise factors have order 2, and the order of a model by the maximal degree among all factors. The energy function of the

Algorithm 1. Fusion Based Algorithms

```
1: procedure FUSION-BASED-INFERENCE(GEN,FUSE,J,X)
2:     x^0 ← initial state form X
3:     n ← 0                                          ▷ Number of moves
4:     m ← 0                                          ▷ Number of moves without progress
5:     while m < m_max and n < n_max do
6:         n ← n + 1
7:         x' ← GEN(x^{n-1}, J, X)                    ▷ Generate proposal
8:         if J(x^{n-1}) ≤ J(x') then
9:             x^n ← FUSE(x^{n-1}, x', J)
10:        else
11:            x^n ← FUSE(x', x^{n-1}, J)
12:        end if
13:        if J(x^n) ≤ J(x^{n-1}) then
14:            m ← 0                                  ▷ Reset counter
15:        else
16:            m ← m + 1                              ▷ Increment counter
17:        end if
18:    end while
19:    return x^n
20: end procedure
```

discrete labeling problem is then given as

$$J(x) = \sum_{f \in F} \varphi_f(x_{ne(f)}), \tag{1}$$

where the assignment of the variable x is also known as the labeling. We consider the problem to find a labeling with minimal energy, i.e.

$$\hat{x} \in \arg\min_{x \in X} J(x). \tag{2}$$

This labeling is a maximum-a-posteriori (MAP) solution of a Gibbs distribution $p(x) = \exp\{-J(x)\}/Z$ defined by the energy $J(x)$. Here, Z normalizes the distribution.

To avoid the large labeling space X, fusion moves optimize only over the subspace $X' \subset X$, which is defined as the set of labelings spanned by the current x^{cur} and proposed x^{pro} labeling,

$$X'(x^{\text{cur}}, x^{\text{pro}}) = \{x \in X \,|\, \forall i : x_i \in \{x_i^{\text{cur}}, x_i^{\text{pro}}\}\}\,. \tag{3}$$

The set of all feasible moves, i.e. that decrease the energy, is given by

$$X^{\text{MOVE}}(x^{\text{cur}}, x^{\text{pro}}) = \{x \in X' \,|\, J(x) \leq J(x^{\text{cur}})\}\,. \tag{4}$$

Since finding the optimal move (optimal labeling in X^{MOVE}) is NP-hard we can not expect to find the optimal move in polynomial time. This is why we define and consider fusion-operators $\text{FUSE}(x, x', J)$ which return an element of $X^{\text{MOVE}}(x, x')$.

Given a proposal generator *GEN*, a fusion-operator *FUSE*, an objective function J, and a state-space X we can define the class of *Fusion-Algorithms*, as shown in Alg. 1. They all monotonically decrease the energy. As stopping condition we will use the maximal number of moves $n_{\max}$ as well as the maximal

Algorithm 2. Fusion Moves

Require: $J(x) \leq J(x')$
Ensure: $J(\hat{x}) \leq J(x)$

procedure $\text{FUSE}_{\text{QPBO}}(x, x', J)$
  $\bar{X} \leftarrow \{\bar{x} \in X | \forall i : \bar{x}_i \in \{x_i, x'_i\}\}$ ▷ Build Boolean subspace of X
  $\hat{x} \leftarrow QPBO(J(\cdot), \bar{X})$ ▷ Solve relaxation for persistent states
  $\hat{x}_i \leftarrow x_i \quad \forall \hat{x}_i = \frac{1}{2}$ ▷ Replace non-persistent states
  return $\hat{x}$
end procedure

procedure $\text{FUSE}_{\text{LF2}}(x, x', J)$
  $\bar{X} \leftarrow \{\bar{x} \in X | \forall i : \bar{x}_i \in \{x_i, x'_i\}\}$ ▷ Build Boolean subspace of X
  $LazyFlipper.setStartingPoint \leftarrow x$ ▷ Set starting point
  $LazyFlipper.searchDepth \leftarrow 2$ ▷ Set search depth
  $\hat{x} \leftarrow LazyFlipper(J(\cdot), \bar{X})$ ▷ Lazy Flipper improves the current state
  return $\hat{x}$
end procedure

procedure $\text{FUSE}_{\text{ILP}}(x, x', J)$
  $\bar{X} \leftarrow \{\bar{x} \in X | \forall i : \bar{x}_i \in \{x_i, x'_i\}\}$ ▷ Build Boolean subspace of X
  $RILP.setStartingPoint \leftarrow x$ ▷ Add the current best into the solution pool
  $\hat{x} \leftarrow RILP(J(\cdot), \bar{X})$ ▷ ILP improves the current state
  return $\hat{x}$
end procedure

procedure $\text{FUSE}_{\text{BASE}}(x, x', J)$
  return $\arg\min_{\bar{x} \in \{x, x'\}} J(\bar{x})$
end procedure

length of a sequence of non-improving moves $m_{\max}$. Algorithms in this family are distinguished by the fusion operation and the proposal generator that they employ, which we will discuss in the next two sections.

3 Fusion Move Operators

As discussed in the previous section an elementary part of fusion-algorithms is the fusion-operator *FUSE*. In this section we discuss different operators and present two novel fusion-operators. The corresponding pseudo code is shown in Alg. 2. The returned labeling is guaranteed to have an energy lower or equal to the energy of the current labeling and the proposed labeling.

QPBO Fusion: The standard fusion-operator $FUSE_{QPBO}$ was proposed by Lempitsky et al. [26] and generalized to the higher-order case by Ishikawa [15] and Fix et al. [10], which reduce in a preprocessing step the higher-order subproblem into a second-order one. For the second-order problem the local polytope relaxation is solved by QPBO [31] and persistency is used to improve the current best labeling. While this can be done in polynomial time, there is in general no guaranty that we obtain persistency for any variable. However, empirically this fusion-operator works well and is therefore widely considered as state-of-the-art.

Lazy Flipping Fusion: An alternative ansatz is to improve the current labeling by local flipping. In the case when only one variable is flipped at the same time this boils down to ICM [3]. Lempitsky et al. [26] show that ICM-Fusion does not work well. However, Andres et al. have suggested a generalization of

ICM to multi-variable flipping, called Lazy Flipper [2]. Lazy Flipper can handle higher-order terms directly, hence order reduction is not required. In the present work we use lazy flipping with search depth two defining the fusion-operator $FUSE_{LF2}$ and initialize it with the current best labeling. The initial labeling is sequentially improved by flips of less or equal than two variables until no further improvement is possible. Obviously, the final labeling will not be worse than the initial one. While Lazy Flipping does not require the existence of persistent variables, it stops if improvements can only be obtained by flipping too many variables simultaneously.

Optimal Fusion: Recently, Kappes et al. [20] have shown that many discrete optimization problems in computer vision can be solved exactly by first reducing the problem size by partial optimality and than solving the smaller remaining problem by advanced methods like integer linear programming (ILP). In the case that the remaining problem splits in several connected components, those can be handled independently which gives additional speed up. The fusion-operator $FUSE_{ILP}$ is defined by using QPBO [31] with the reduction of Fix [10] for higher-order models to obtain partial optimality and solving the connected components of the remaining problem by the Cplex ILP-solver [14]. By adding the current best solution in the solution pool of the ILP solver it is guaranteed that the final solution will not be worse. Furthermore, this provides a good starting point and an upper bound. Since the remaining ILPs can still be quite hard, we interrupted the solver after 100 seconds and return the best labeling from the solution pool. Consequently, in our experiments a move is optimal if it is calculated within 100 seconds.

Base Fusion: To determine the impact of fusion-operations, we also define a naive operator $FUSE_{BASE}$, which returns the better of the two labelings

$$\bar{x} = \arg\min_{\bar{x}\in\{x,x'\}} J(\bar{x}). \tag{5}$$

This fusion-operator does only profit from the proposal quality and not from their diversity.

4 Generating Proposals

The second major component of a fusion-algorithm is the generation of proposals. On the one hand, proposals should be of high quality with respect to the energy function $J(\cdot)$. On the other hand, they should be also diverse among each other and cheap to calculate. Proposal generators can be clustered into four groups: (i) inference-based generators, (ii) randomized generators, (iii) deterministic generators, and (iv) application specific generators.

Pseudo-code for (i)–(iii) is given in Alg. 3. We do not consider application specific generators in the present work because they are none generic and require more data than just the objective function.

Inference-Based Generators: For the cartographic label placement problem Lempitsky et al. [26] used the labelings that Loopy Belief Propagation (LBP)

Algorithm 3. Proposal Generators

```
procedure RandomGen(x, J, X)
Require: ∀i ∈ V : P_i(x_i)                          ▷ Shared for all moves
    for i ∈ V do
        x̂_i ∼_{P_i(x_i)} X_i
    end for
    return x̄
end procedure

procedure InfGen(x, J, X)
Require: INF ← INF(J, X)                            ▷ Shared for all moves
    INF.runOneStep
    X̄ ← INF.getLabeling
    return x̄
end procedure

procedure DeterministicGen(x, J, X)
Require: n ← 0                                      ▷ Shared for all moves
    X̄ ← gen(x, n, X)
    n ← n + 1
    return x̄
end procedure
```

generates after each iteration as proposals. They obtained a result superior to state-of-the-art for this problem instance.

This result was not further generalized or tested for other problems in later work. However, it is very appealing since methods based on linear programming relaxations like TRWS [24], MPLP [11] or approximative message passing methods like LBP [8], BPS [24] provide after each iteration good proposals close to the optimal one. The diversity is generated by the heuristic rounding procedure. Fusion moves can profit from this diversity and overcome failures caused by greedy rounding if this failures are not present in all iterations.

We use the visitor concept of OpenGM [1] and inject the fusion operation after each algorithmic unit. This allows using any OpenGM-inference method as proposal generator with a few lines of code. In the present work we show results for TRWS, MPLP, BPS and LBP with different damping. MPLP and LBP can also deal with higher-order problems.

Randomized Generators: A general way to generate diverse proposals is to sample those from a distribution P. The disadvantage of such generators is that the proposals usually have bad quality. One can try to alleviate this by prior knowledge. We consider the following sampling distributions, which all defined independently for each variable. For problems with arbitrary structure we consider *uniform random distributions* (P_U)

$$P_i(x_i) = \frac{1}{|X_i|}, \tag{6}$$

and *local marginal approximations* (P_L) which estimate for a given temperature T first order marginals from unary terms $\bar{f}_i$ by

$$P_i(x_i) \propto \exp\{-T \cdot \bar{f}_i(x_i)\}. \tag{7}$$

For $T \to 0$ the distribution becomes uniform and for $T \to \infty$ all its mass concentrated in the local mode. When local data terms a weak or misleading the distribution is not helpful.

We also follow the idea used in [10,15], which blur the current labeling on the image grid and sample proposals around the "blurred labeling". Of course this is only useful if labels have the same meaning for all variables. Empirically we observe no advantage by repeating the blurring in each iteration if the standard variation of the Gaussian blur is large. We suggest to blur the unary terms instead of the labeling, this is also more robust to missing unary terms and uncertain information. Furthermore, blurring has to be done only once. For each variable we obtain a Gaussian blurred unary term label-wise

$$\bar{f}_i^B(x_i) = GaussianBlur_\sigma(\bar{f}(x_i))_i, \tag{8}$$

$$\bar{x}_i^B = \arg\max_{x_i} \bar{f}_i^B(x_i). \tag{9}$$

As in [10,15] we sampling uniformly (P_{UB}) from

$$P_i(x_i) \propto \begin{cases} 1 & \text{if even round or } x_i \in [\bar{x}_i^B - 1.5\sigma, \bar{x}_i^B + 1.5\sigma] \\ 0 & \text{else} \end{cases} \tag{10}$$

Alternatively we can use the blurred unaries for a *local blurred marginal approximations* (P_{LB}) as in the non-blurred case

$$P_i(x_i) \propto \exp\{-T \cdot \bar{f}_i^B(x_i)\}. \tag{11}$$

Deterministic Generators: Deterministic generators provide very simple proposals with low workload. The proposals depend on the current labeling x and iteration n. For deterministic generators we determine the number of moves with no improvements $m_{\max}$ for which immediate termination will have no effect on the final solution. An example is the generalization of α-*Expansion* [5] where $m_{\max} = \max_{i \in V} |X_i|$. The proposal $\hat{x}$ in iteration n takes the label $\alpha(n) = n \mod m_{\max}$ if possible, i.e.

$$\hat{x}_i = \begin{cases} \alpha(n) & \text{if } \alpha(n) \in X_i \\ x_i & \text{else} \end{cases} \tag{12}$$

Another example are $\alpha\beta$-*Swaps* [5] which can be generalized to arbitrary discrete problems. Here in each step n variables that have the labels $\alpha(n)$ and $\beta(n)$ are changed to $\beta(n)$ and $\alpha(n)$ if possible, respectively. Here $m_{\max} = 0.5 \cdot \max_{i \in V} |X_i| \cdot (\max_{i \in V} |X_i| - 1)$.

$$\hat{x}_i = \begin{cases} \alpha(n) & \text{if } x_i = \beta(n) \text{ and } \alpha(n) \in X_i \\ \beta(n) & \text{if } x_i = \alpha(n) \text{ and } \beta(n) \in X_i \\ x_i & \text{else} \end{cases} \tag{13}$$

5 Evaluation

We compare the combination of fusion operations and proposal generators for different graphical models benchmarks [7,18,19] and the FoE-dataset [30]. All this instances are or will be made publicly available in the OpenGM-format.

Table 1. Overview of the used models and the number of variables (# variables), number of labels (# labels), model order (order), number of instances (# instances) and temperature used for determine local marginals

modelname	# variables	# labels	order	# instances	used temperature
Field of Experts	38801	256	4	100	0.1
MRF Inpainting	65536	256	2	2	0.001
Protein Folding	1972	503	2	21	0.1
Protein Prediction	14441	2	3	8	1
DTF Inpainting	17856	2	2	100	0.1
Matching	21	21	2	4	0.1
Cell Tracking	41134	2	9	1	0.1

We run all combinations for 1000 iterations ($n_{\max} = 1000$) and maximal 900 seconds on a Core i7-2600K with 3.40 GHz single-threaded. We stop after 50 moves without improvement ($m_{\max} = 50$). Stopping condition of deterministic methods are the deterministic default. Due lack of space we add the complete results as supplementary material and show only selected combinations here. The used temperature parameter for the sampling distributions and an overview of the models is given in Tab. 1.

We report the energy value, averaged over all model instances, of the best labeling after 10, 60 and 600 seconds as well as for the final labeling. Additionally we report the mean runtime and the number of iterations or moves. The best value among all fusion-algorithms in each time slot is marked green, and the fusion-operation which give the best mean energy for a given proposal-generator blue. Additionally, we add results of state-of-the-art-methods to the tables, if those results were available. If the best of those beats all fusion algorithm it is marked red.

Table 2. For *field of experts* instances $\text{FUSION}_{\text{LF2}}$ overall performs best

algorithm	value				time	it
	(10 sec)	(60 sec)	(600 sec)	(end)	(end)	(end)
α-Exp-$\text{FUSION}_{\text{ILP}}$	115331.95	112908.90	108011.80	105001.75	941.28 sec	27.30
α-Exp-$\text{FUSION}_{\text{LF2}}$	109604.69	76950.74	35553.15	34958.88	709.57 sec	999.88
α-Exp-$\text{FUSION}_{\text{QPBO}}$	113027.96	107330.34	56267.95	54571.25	900.91 sec	541.81
P_{UB}-$\text{FUSION}_{\text{ILP}}$	107585.42	105930.25	37603.67	35351.69	903.09 sec	220.24
P_{UB}-$\text{FUSION}_{\text{LF2}}$	71918.21	38631.97	32925.36	32848.61	695.85 sec	1000.00
P_{UB}-$\text{FUSION}_{\text{QPBO}}$	97796.97	47536.08	33481.48	33090.60	872.46 sec	899.96
P_L-$\text{FUSION}_{\text{ILP}}$	87010.93	41320.95	32779.26	32637.81	899.81 sec	806.71
P_L-$\text{FUSION}_{\text{LF2}}$	54960.13	35583.31	32619.64	32586.99	701.58 sec	1000.00
P_L-$\text{FUSION}_{\text{QPBO}}$	57337.32	35918.37	32646.95	32613.20	688.93 sec	1000.00
P_U-$\text{FUSION}_{\text{ILP}}$	81230.66	41289.75	32936.93	32779.52	806.42 sec	999.44
P_U-$\text{FUSION}_{\text{LF2}}$	64828.40	38662.98	32882.46	32782.16	736.44 sec	996.45
P_U-$\text{FUSION}_{\text{QPBO}}$	63305.54	38500.68	32871.21	32797.15	699.14 sec	1000.00

Field of Experts: Field of experts were introduced by Roth and Black [30], which use higher-order terms to expressive image priors that capture the statistics of natural scenes. Field of export models have become a standard benchmark

Table 3. For the *inpainting* problems fusion-algorithms improve the performance of TRWS. Random generators do not work well here.

algorithm	value				time	it
	(10 sec)	(60 sec)	(600 sec)	(end)	(end)	(end)
TRWS	26481554.50	26465539.50	26464769.50	26464759.00	632.40 sec	944.50
TRWS-LF2	∞	∞	∞	26463829.00	3009.52 sec	–
P_U-FUSION$_{\text{BASE}}$	420556187.50	420556187.50	420556187.50	420556187.50	2.40 sec	50.00
P_U-FUSION$_{\text{ILP}}$	60296247.50	38570409.50	34890334.50	34890334.50	196.09 sec	1000.00
P_U-FUSION$_{\text{LF2}}$	100770607.50	45696051.50	35241978.50	34985385.50	501.38 sec	1000.00
P_U-FUSION$_{\text{QPBO}}$	50696441.50	36367339.50	34904322.00	34904322.00	119.94 sec	1000.00
TRWS-FUSION$_{\text{BASE}}$	26481554.50	26465534.50	26465416.50	26465416.50	103.48 sec	163.00
TRWS-FUSION$_{\text{ILP}}$	26476904.00	26464727.50	26464158.00	26464158.00	217.59 sec	318.50
TRWS-FUSION$_{\text{LF2}}$	26482403.50	26465290.00	26464904.50	26464904.50	206.98 sec	276.00
TRWS-FUSION$_{\text{QPBO}}$	26476820.00	26464728.50	26464158.00	26464158.00	214.05 sec	318.50

for fusion moves [10,15]. We follow the experimental setup used in [10,15] and take the 100 test images from the BSD300 [27], downscale them by a factor of two and add Gaussian noise with standard deviation $\sigma = 20$. The energy function includes unary terms penalize the L_1-distance of the 256 labels/colors to the noisy pixel color and fourth order experts learned and kindly provided by Roth and Black [30].

Classical QPBO-based fusion is clearly inferior to LazyFlipper-based, c.f. Tab. 2 and Fig. 1(a). For the α-expansion generator QPBO-fusion does a bad job as reported in [15]. When we switch to LazyFlipper-based fusion it is still not best but comparable to other combinations. Using optimal moves does not improve the results significantly. The moves are only marginal better but slower. Overall best results are obtained when sampling from the distributions base on non-blurred unary terms.

Inpainting: We consider the two inpainting problems from [33] which have 256 labels. For these instances TRWS followed by local search is currently the

Table 4. For the *protein folding* instances BPS-FUSION leads to better results and is more than ten times faster than BPS

algorithm	value				time	it
	(10 sec)	(60 sec)	(600 sec)	(end)	(end)	(end)
BPS	−5958.72	−5958.72	−5958.72	−5958.72	25.34 sec	1000.00
LBP	−5817.90	−5841.98	−5872.91	−5872.91	183.53 sec	1000.00
TRWS	−5735.86	−5799.52	−5846.86	−5846.86	118.17 sec	675.48
CombiLP	∞	∞	−5822.45	−5911.12	568.86 sec	–
BPS-FUSION$_{\text{BASE}}$	−5958.37	−5958.37	−5958.37	−5958.37	1.63 sec	57.24
BPS-FUSION$_{\text{ILP}}$	−5959.82	−5959.82	−5959.82	−5959.82	1.69 sec	57.05
BPS-FUSION$_{\text{LF2}}$	−5959.48	−5959.48	−5959.48	−5959.48	1.70 sec	57.05
BPS-FUSION$_{\text{QPBO}}$	−5959.82	−5959.82	−5959.82	−5959.82	1.61 sec	57.05
LBP-0.5-FUSION$_{\text{BASE}}$	−5926.10	−5944.87	−5944.87	−5944.87	16.95 sec	86.24
LBP-0.5-FUSION$_{\text{ILP}}$	−5928.60	−5946.35	−5946.35	−5946.35	16.19 sec	80.67
LBP-0.5-FUSION$_{\text{LF2}}$	−5926.10	−5944.87	−5944.87	−5944.87	16.99 sec	86.24
LBP-0.5-FUSION$_{\text{QPBO}}$	−5928.60	−5945.28	−5945.28	−5945.28	16.11 sec	81.86

Table 5. For the *protein-prediction* problems the FUSION$_{\text{ILP}}$ leads to better results even with random proposals

algorithm	value				time	it
	(10 sec)	(60 sec)	(600 sec)	(end)	(end)	(end)
LBP-0.5	53407.52	52974.98	52974.98	52974.98	60.97 sec	766.88
LBP-LF2	∞	∞	52942.95	52942.95	69.86 sec	–
LBP-0.5-FUSION$_{\text{BASE}}$	52971.53	52971.53	52971.53	52971.53	6.22 sec	110.50
LBP-0.5-FUSION$_{\text{ILP}}$	52827.89	52821.38	52821.38	52821.38	9.64 sec	110.12
LBP-0.5-FUSION$_{\text{LF2}}$	52971.53	52971.53	52971.53	52971.53	6.22 sec	110.50
LBP-0.5-FUSION$_{\text{QPBO}}$	52826.64	52826.64	52826.64	52826.64	6.20 sec	124.12
P_U-FUSION$_{\text{BASE}}$	97071.97	97071.97	97071.97	97071.97	0.76 sec	50.00
P_U-FUSION$_{\text{ILP}}$	95886.12	95787.15	55531.88	55509.32	380.11 sec	689.25
P_U-FUSION$_{\text{LF2}}$	58622.95	58622.81	58622.81	58622.81	13.62 sec	87.25
P_U-FUSION$_{\text{QPBO}}$	75582.10	66164.13	65933.02	65933.02	58.27 sec	955.00

leading method [19]. These methods make use of the convex regularizer and apply distance transform [9] for good any time performance. Fusion algorithms did not work well within 1000 iterations except TRWS is used as generator. This agrees with the results reported in [33] where α-expansion also needed much more iterations and has a simple explanation. The unaries and the regularizer are based on squared differences. This make them very picky and selective. This limits the set of improving moves for random proposals.

Protein Folding: The protein folding instances [34] have a moderate number of variables, but are fully connected and have for some variables huge label spaces. Recently it has been shown [18], that sequential Belief Propagation (BPS) gives very good results near optimality. Using BPS as generator fusion obtain better

Table 6. For the *DTF Chinese characters* fusion based methods has not beaten LSA-TR. However, we get quite close and improve standard methods. **Results was taken from the original papers and not reproduced.*

algorithm	value				time	it
	(10 sec)	(60 sec)	(600 sec)	(end)	(end)	(end)
TRWS	−49512.31	−49514.04	−49514.06	−49514.06	112.37 sec	856.13
BPS-TAB	−49536.02	−49537.63	−49538.16	−49538.16	78.65 sec	1000.00
LSA-TR*	−49547.61	−49547.61	−49547.61	−49547.61	0.21 sec	−−
MCBC-pct*	−−	−−	−−	−49550.10	2053.89 sec	−−
α-Exp-FUSION$_{\text{BASE}}$	−49434.39	−49434.39	−49434.39	−49434.39	0.01 sec	2.00
α-Exp-FUSION$_{\text{ILP}}$	−49434.39	−49434.39	−49527.97	−49528.00	273.90 sec	4.40
α-Exp-FUSION$_{\text{LF2}}$	−49495.76	−49496.83	−49496.83	−49496.83	13.39 sec	3.50
α-Exp-FUSION$_{\text{QPBO}}$	−49499.09	−49501.69	−49501.69	−49501.69	7.63 sec	11.53
BPS-FUSION$_{\text{BASE}}$	−49535.10	−49535.10	−49535.10	−49535.10	5.17 sec	81.73
BPS-FUSION$_{\text{ILP}}$	−49504.33	−49504.36	−49542.08	−49543.30	447.73 sec	40.79
BPS-FUSION$_{\text{LF2}}$	−49535.69	−49535.69	−49535.69	−49535.69	6.27 sec	75.30
BPS-FUSION$_{\text{QPBO}}$	−49535.82	−49535.82	−49535.82	−49535.82	4.90 sec	74.58
TRWS-FUSION$_{\text{BASE}}$	−49512.19	−49512.21	−49512.21	−49512.21	8.59 sec	71.63
TRWS-FUSION$_{\text{ILP}}$	−49476.91	−49482.33	−49535.98	−49537.55	543.30 sec	41.83
TRWS-FUSION$_{\text{LF2}}$	−49528.15	−49529.41	−49529.41	−49529.41	16.06 sec	63.29
TRWS-FUSION$_{\text{QPBO}}$	−49531.64	−49532.29	−49532.29	−49532.29	17.03 sec	69.55

Table 7. For *matching* problems results could only be marginally improved, since the feasible move space is small in most iterations

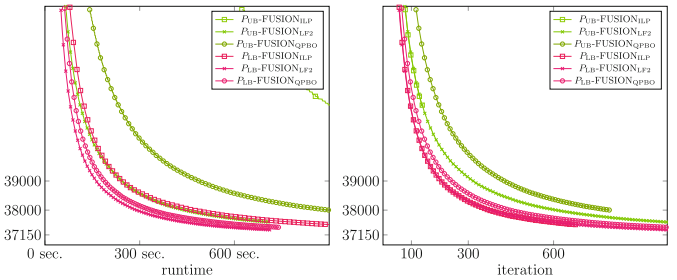
algorithm	value				time	it
	(10 sec)	(60 sec)	(600 sec)	(end)	(end)	(end)
TRWS	43.38	43.38	43.38	43.38	0.35 sec	253.00
MPLP-C	21.22	21.22	21.22	21.22	4.63 sec	145.25
LBP-0.5-FUSION$_{\text{BASE}}$	26.87	26.87	26.87	26.87	0.16 sec	77.25
LBP-0.5-FUSION$_{\text{ILP}}$	24.56	24.56	24.56	24.56	0.19 sec	78.25
LBP-0.5-FUSION$_{\text{LF2}}$	26.87	26.87	26.87	26.87	0.16 sec	77.25
LBP-0.5-FUSION$_{\text{QPBO}}$	27.80	27.80	27.80	27.80	0.13 sec	66.00
P_U-FUSION$_{\text{ILP}}$	43.36	43.36	43.36	43.36	1.04 sec	243.00
P_U-FUSION$_{\text{LF2}}$	55.22	55.22	55.22	55.22	0.30 sec	216.00
P_U-FUSION$_{\text{QPBO}}$	50.78	50.78	50.78	50.78	0.03 sec	232.25
TRWS-FUSION$_{\text{BASE}}$	43.38	43.38	43.38	43.38	0.08 sec	59.25
TRWS-FUSION$_{\text{ILP}}$	40.97	40.97	40.97	40.97	0.37 sec	67.75
TRWS-FUSION$_{\text{LF2}}$	42.00	42.00	42.00	42.00	0.13 sec	67.25
TRWS-FUSION$_{\text{QPBO}}$	40.97	40.97	40.97	40.97	0.09 sec	67.75

and faster results than BPS alone and advanced combinatorial methods like CombiLP [32]. For other generators the results are worse but still comparable with other methods and always improve the baseline significantly, c.f. Tab.4 and Fig. 1(c).

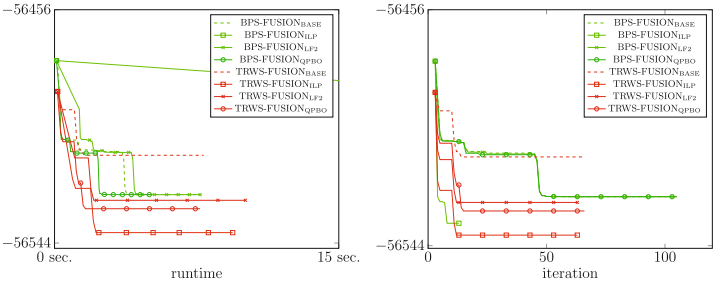
Protein Prediction: The protein prediction instances [16] include sparse third-order binary models. We beat the best performing method from the benchmark [18] which is LBP with damping 0.5 followed by Lazy Flipping of search depth 2, by using damped LBP as generator and QPBO or ILP for fusion, c.f. Tab.5 and Fig. 1(d).

Table 8. For the *cell-tracking* instance we obtain faster good results only marginally worse than the optimum

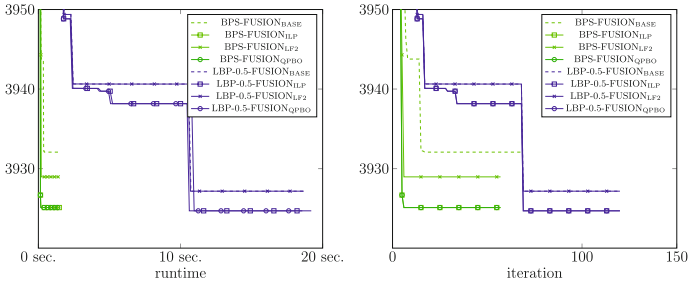
algorithm	value				time	it
	(10 sec)	(60 sec)	(600 sec)	(end)	(end)	(end)
LBP	107515639.76	107515319.56	107515319.56	107515319.56	80.70 sec	1000.00
ILP	45364196.24	7514421.21	7514421.21	7514421.21	13.78 sec	0.00
LBP-0.5-FUSION$_{\text{BASE}}$	7822517.15	7822517.15	7822517.15	7822517.15	10.00 sec	89.00
LBP-0.5-FUSION$_{\text{ILP}}$	7518000.15	7514751.98	7514751.98	7514751.98	26.69 sec	234.00
LBP-0.5-FUSION$_{\text{LF2}}$	7822517.15	7822517.15	7822517.15	7822517.15	9.83 sec	89.00
LBP-0.5-FUSION$_{\text{QPBO}}$	10324281.39	10314354.13	10314354.13	10314354.13	24.96 sec	227.00
LBP-FUSION$_{\text{BASE}}$	7518099.53	7518099.53	7518099.53	7518099.53	11.83 sec	111.00
LBP-FUSION$_{\text{ILP}}$	7515318.79	7515029.55	7515029.55	7515029.55	17.49 sec	145.00
LBP-FUSION$_{\text{LF2}}$	7518099.53	7518099.53	7518099.53	7518099.53	12.11 sec	111.00
LBP-FUSION$_{\text{QPBO}}$	7516031.12	7515029.55	7515029.55	7515029.55	15.96 sec	145.00
P_U-FUSION$_{\text{BASE}}$	58794439.99	58794439.99	58794439.99	58794439.99	2.17 sec	50.00
P_U-FUSION$_{\text{ILP}}$	14033539.27	7791724.31	7531572.24	7531572.24	643.67 sec	304.00
P_U-FUSION$_{\text{LF2}}$	9281131.45	9278699.79	9278699.79	9278699.79	18.91 sec	109.00
P_U-FUSION$_{\text{QPBO}}$	11217379.70	9008429.54	8437145.94	8437145.94	156.95 sec	1000.00



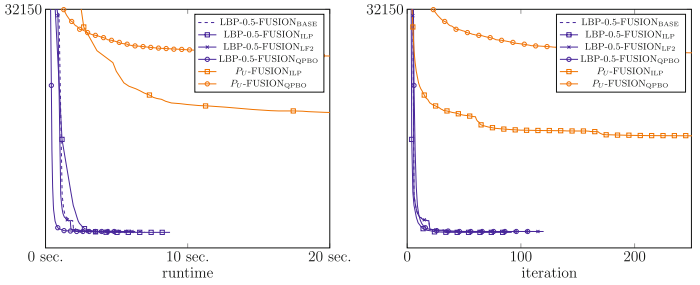
(a) FoE - instance: 101085



(b) DTF Chinese Characters - instance: 0001



(c) Protein Folding - instance: pdb1b25



(d) Protein Prediction - instance: 1

Fig. 1. Energy improvement for selected instances and methods over time **(left)** and over iterations **(right)**

DTF Chinese Characters: A challenging second-order binary problem is using decision tree fields (DTF) for inpainting [19,29]. While advanced combinatorial solvers (MCBC) [20] give best performance [19], they are slow. The best fast solver in [19] was sequential belief propagation (BPS). Recently, Gorelick et al. presented a fast and accurate alternative based on local submodular approximations with trust region terms (LSA-TR) [13]. While we do not beat LSA-TR we improve other methods significantly. This indicates that fusion algorithms are also useful for hard problems – especially if ILP-Fusion is used – and improve final solutions and any-time performance, c.f. Tab. 6 and Fig. 1(b). Note that contrary to MCBC and LSA-TR, Fusion algorithms are not limited to binary models.

Matching: We also consider the matching instances from [19] which are small but very hard. In [19] it has been shown that α-expansion proposals are not an adequate proposal choice. This is no longer true for other proposals including random ones. However fusion moves often run into a labeling which is hard to escape. If such a labeling is feasible, i.e. represents a one-to-one match, a proposal has to support a cyclic swap of the labels in order to fulfill the one-to-one matching constraint and improve the energy in order to escape. Consequently, it is less likely to find global optimal solutions.

Cell-Tracking: The tracking model considered in [19] include binary variables and terms of order up to 9. While ILP-solvers solves this instance to optimality very efficiently one should not expect that this will hold for larger models. In such scenarios relaxations would be an alternative but those suffer from the soft-constraints and labelings generated by rounding might violate those. In such situations Fusion can help a lot and provide early close-to-optimal solutions.

6 Conclusions

Fusion algorithms are very powerful and their performance on discrete graphical models has been apparently underestimated in the past. We showed that the performance of any inference method can be improved by embedding it as a proposal generator into a fusion algorithm. This leads to better solutions as well as to better any-time performance by compensating rounding artefacts, c.f. Fig. 1. The additional computational costs are usually negligible.

Concerning proposal generators, inference based generators are overall superior, since the proposals are of high quality. However, for large scale or higher-order models they are sometimes no longer applicable, e.g. for field of experts, or much slower, e.g. for protein folding, than random or deterministic ones. Here randomized generators work often reasonable. Application specific or more advanced generators might be able to further close this gap with small additional computational costs.

The quality of fusion algorithms can be also improved by fusion operators different from QPBO-Fusion. We presented two powerful alternatives: Integer linear programming solvers can be used to calculate the optimal moves in each step.

This can lead to much better results when the persistency of QPBO is small, e.g. DTF or protein prediction. Lazy Flipping based fusion does also not suffer from small persistency but requires that the global move can be obtained by a sequence of local moves. When this is the case, as for the field of expert instances, Lazy Flipping fusion gives the best trade-off between runtime and energy improvement. Another interesting observation in this context is that optimal moves are not always desirable. Contrary to non-optimal moves optimal moves, can tend to run into "dead ends" for which only a small number of proposals generate moves which allow to escape. Such a proposal might not be generated within $m_{\max}$ iterations and the algorithm stops too early. Furthermore, fusion is a greedy procedure and an optimal fusion move might not be optimal in the long run. For example for some protein folding instances QPBO fusion is sometimes marginal better than ILP fusion for the same number of iterations. However, except for these outliers and on average optimal moves performs better than QPBO-based moves – at least in the long run.

Finally we would like to remark that contrary to the standard QPBO-based fusion-operator the presented alternatives can deal with more than one proposal. Consequently the subproblems would be multi-label problems and X' larger, which allows more powerful moves.

References

1. Andres, B., Beier, T., Kappes, J.H.: OpenGM2 (2012). http://hci.iwr.uni-heidelberg.de/opengm2/
2. Andres, B., Kappes, J.H., Beier, T., Köthe, U., Hamprecht, F.A.: The lazy flipper: Efficient depth-limited exhaustive search in discrete graphical models. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 154–166. Springer, Heidelberg (2012)
3. Besag, J.: On the Statistical Analysis of Dirty Pictures. Journal of the Royal Statistical Society. Series B (Methodological) **48**(3), 259–302 (1986)
4. Boros, E., Hammer, P.L.: Pseudo-boolean optimization. Discrete Appl. Math. **123**(1–3), 155–225 (2002)
5. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11), 1222–1239 (2001)
6. Delong, A., Osokin, A., Isack, H., Boykov, Y.: Fast approximate energy minimization with label costs. International Journal of Computer Vision **96**, 1–27 (2012)
7. Elidan, G., Globerson, A.: The probabilistic inference challenge (PIC 2011). http://www.cs.huji.ac.il/project/PASCAL/
8. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. Int. J. Comput. Vision **70**(1), 41–54 (2006). http://dx.doi.org/10.1007/s11263-006-7899-4
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. International Journal of Computer Vision **70**(1), 41–54 (2006)
10. Fix, A., Gruber, A., Boros, E., Zabih, R.: A graph cut algorithm for higher-order Markov random fields. In: ICCV (2011). http://dx.doi.org/10.1109/ICCV.2011.6126347

11. Globerson, A., Jaakkola, T.: Fixing max-product: Convergent message passing algorithms for map lp-relaxations. In: NIPS (2007)
12. Goldluecke, B., Strekalovskiy, E., Cremers, D.: Tight convex relaxations for vector-valued labeling. SIAM Journal on Imaging Sciences **6**(3), 1626–1664 (2013)
13. Gorelick, L., Boykov, Y., Veksler, O., Ayed, I.B., Delong, A.: Submodularization for binary pairwise energies. In: CVPR. IEEE (2014) (in press)
14. IBM: ILOG CPLEX Optimizer (2013). http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/
15. Ishikawa, H.: Transformation of general binary mrf minimization to the first-order case. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(6), 1234–1249 (2011)
16. Jaimovich, A., Elidan, G., Margalit, H., Friedman, N.: Towards an integrated protein-protein interaction network: A relational markov network approach. Journal of Computational Biology **13**(2), 145–164 (2006)
17. Kahl, F., Strandmark, P.: Generalized roof duality. Discrete Applied Mathematics **160**(16–17), 2419–2434 (2012)
18. Kappes, J.H., Andres, B., Hamprecht, F.A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B.X., Kröger, T., Lellmann, J., Komodakis, N., Savchynskyy, B., Rother, C.: A comparative study of modern inference techniques for structured discrete energy minimization problems. CoRR abs/1404.0533 (2014)
19. Kappes, J.H., Andres, B., Hamprecht, F.A., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B.X., Lellmann, J., Komodakis, N., Rother, C.: A comparative study of modern inference techniques for discrete energy minimization problems. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
20. Kappes, J.H., Speth, M., Reinelt, G., Schnörr, C.: Towards efficient and exact MAP-inference for large scale discrete computer vision problems via combinatorial optimization. In: CVPR (2013)
21. Kim, S., Nowozin, S., Kohli, P., Yoo, C.D.: Higher-order correlation clustering for image segmentation. In: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS) (2011)
22. Kohli, P., Ladicky, L., Torr, P.H.: Robust higher order potentials for enforcing label consistency. International Journal of Computer Vision **82**(3), 302–324 (2009). http://dx.doi.org/10.1007/s11263-008-0202-0
23. Koller, D., Friedman, N.: Probabilistic Graphical Models: Principles and Techniques. MIT Press (2009)
24. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(10), 1568–1583 (2006)
25. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part III. LNCS, vol. 2352, pp. 65–81. springer, heidelberg (2002)
26. Lempitsky, V., Rother, C., Roth, S., Blake, A.: Fusion moves for markov random field optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(8), 1392–1405 (2010)
27. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV (2001)
28. Nowozin, S., Lampert, C.H.: Structured learning and prediction in computer vision. Foundations and Trends in Computer Graphics and Vision **6**(3–4), 185–365 (2011)

29. Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B., Kohli, P.: Decision tree fields. In: ICCV, pp. 1668–1675. IEEE (2011)
30. Roth, S., Black, M.J.: Fields of experts. International Journal of Computer Vision **82**(2), 205–229 (2009)
31. Rother, C., Kolmogorov, V., Lempitsky, V.S., Szummer, M.: Optimizing binary MRFs via extended roof duality. In: CVPR (2007)
32. Savchynskyy, B., Kappes, J.H., Swoboda, P., Schnörr, C.: Global MAP-optimality by shrinking the combinatorial search area with convex relaxation. In: NIPS (2013)
33. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. IEEE PAMI **30**(6), 1068–1080 (2008). http://dx.doi.org/10.1109/TPAMI.2007.70844
34. Yanover, C., Schueler-Furman, O., Weiss, Y.: Minimizing and learning energy functions for side-chain prediction. Journal of Computational Biology **15**(7), 899–911 (2008)

Feedback Loop Between High Level Semantics and Low Level Vision

Varun K. Nagaraja(✉), Vlad I. Morariu, and Larry S. Davis

University of Maryland, College Park, MD, USA
{varun,morariu,lsd}@umiacs.umd.edu

Abstract. High level semantic analysis typically involves constructing a Markov network over detections from low level detectors to encode context and model relationships between them. In complex higher order networks (e.g. Markov Logic Networks), each detection can be part of many factors and the network size grows rapidly as a function of the number of detections. Hence to keep the network size small, a threshold is applied on the confidence measures of the detections to discard the less likely detections. A practical challenge is to decide what thresholds to use to discard noisy detections. A high threshold will lead to a high false dismissal rate. A low threshold can result in many detections including mostly noisy ones which leads to a large network size and increased computational requirements. We propose a feedback based incremental technique to keep the network size small. We initialize the network with detections above a high confidence threshold and then based on the high level semantics in the initial network, we incrementally select the relevant detections from the remaining ones that are below the threshold. We show three different ways of selecting detections which are based on three scoring functions that bound the increase in the optimal value of the objective function of network, with varying degrees of accuracy and computational cost. We perform experiments with an event recognition task in one-on-one basketball videos that uses Markov Logic Networks.

1 Introduction

Computer vision systems are generally designed as feed-forward systems where low level detectors are cascaded with high level semantic analysis. Low level detectors for objects, tracks or short activities usually produce a confidence measure along with the detections. The confidence measures can sometimes be noisy and hence a multitude of false detections are fed in to subsequent analysis stages. To avoid these false detections, it is common practice to discard some detections that are below a particular confidence threshold. Unfortunately, it is difficult to reliably select a threshold a priori given a particular task. The threshold is generally selected to achieve a "reasonable" trade-off between detector precision and

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-16181-5_38) contains supplementary material, which is available to authorized users.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 485–499, 2015.
DOI: 10.1007/978-3-319-16181-5_38

recall, since it is generally not possible to find all true detections (high recall) without also hallucinating false alarms (low precision).

High level analysis integrates multiple low level detections together using semantics to discard false detections rather than simply thresholding detector scores. For example, in an event recognition system for basketball, the low level detections like shot missed and rebound events are related by high level rules of the game which say that a shot missed event is followed by a rebound event. The analysis of high level interactions between detections can improve the confidence in the detections.

High level analysis typically involves constructing a Markov network over the detections, where contextual relationships corresponding to high level knowledge about the image or video are encoded as factors over combinations of detections [1,2,9,10,15]. A detection usually corresponds to one or more nodes in the network and relationships between detections correspond to a factor. In Markov networks of high order, each detection can be part of exponentially many instantiations of a factor and the network size grows rapidly as a function of the number of detections. The problem is further exacerbated by the inference process, whose computational cost is related exponentially to the network complexity. When many detections are hypothesized at low precision, the size of the Markov network becomes unnecessarily high since the inference process sets most of the detections to false.

We tackle the problem of keeping the network size small by incrementally adding only those detections that are most likely to be inferred as true while the rest of them are kept false. We achieve this by adding a feedback loop between the high level and low level stages, where the high level semantics guides the selection of relevant low level detections. There are several advantages to this feedback loop. First, it can locally adjust the thresholds for low level detectors based on the neighboring context. Second, it keeps the network size small and the inference procedure tractable. And third, we can potentially save computation by selectively running the low level procedures like feature extraction and classification only when needed.

The goal of our feedback based incremental technique is to perform inference and obtain the optimal solution of the objective function corresponding to the *full network* (the network obtained when we include all the detections) by unclamping only the relevant detections. We start with detections above a high confidence threshold and clamp the remaining detections to false based on the closed world assumption, the assumption that what is not known to be true is false. We then incrementally select from the remaining detections below the threshold to add to the network. Our proposed feedback loop involves a principled mechanism by which we identify the detections that are most likely to improve the objective function. Motivated by cluster pursuit algorithms [13] for inference, we derive three scoring functions that bound the increase in the objective function with varying degrees of accuracy and computational cost. The first score function yields the exact increase in the objective function, but it requires that the detector has been run everywhere and that inference can be performed

exactly; the second bounds the change in the objective function, relaxing the inference requirements; the third provides an even looser bound, but it is least computationally intensive and does not require the low level detector to have processed the candidate detections (which is why we call it the *Blind Score*).

We perform experiments on an event recognition task using one-on-one basketball videos. Morariu and Davis [10] used Markov Logic Networks (MLNs) on this dataset to detect events like Shot Made, Shot Missed, Rebound etc. The inputs are a set of event intervals hypothesized from low level detectors like the tracks of objects. Using the feedback loop technique we show that we can successfully select the most relevant event intervals that were earlier discarded due to thresholding. The experiments show that our score functions can reach the optimal value in fewer iterations with smaller network sizes when compared with using just the low level confidence measures.

2 Related Work

High level context plays an important role in many vision systems like scene segmentation [8], object detection in 2D [2,14] and 3D [9] and event recognition [1,10,15]. Usually these systems hypothesize a set of candidate detections using low level detectors and then feed them into the high level model which assigns a label to the candidate detections based on the context. Since low level detectors are not perfect, a multitude of false positives propagate from the low level to the high level. So a high level system is faced with the choice of either dealing with a large model size or having a threshold for the inputs so that model size is contained, but only by discarding true detections that happen to have low confidences.

While many inference techniques work in an incremental fashion to tackle the complexity issues, they do not necessarily behave as a feedback loop and hence do not present with the advantages mentioned earlier. We mention few works here that iteratively add detections while performing inference. In a scene segmentation task, Kumar and Koller [8] hypothesize a set of regions in an image through multiple bottom-up over-segmentations and exploit the high level energy function to iteratively select input regions that are relevant for the task. Zhu et al. [16] use the greedy forward search technique of Desai et al. [3] for inference in their event recognition system. The inference algorithm of Desai et al. first sets the output label for the inputs to the background class. Each input is then scored based on the change in the objective function if it were allowed to be labelled as a non-background class. The top scoring inputs are then iteratively added until convergence. Our feedback loop technique is based on the same idea of greedily reaching the MAP value as quickly as possible but we provide a principled mechanism to performing inference in higher order networks. Also we do not use it just as an incremental technique, but extract more insight from the high level semantics to save computation for the low level module. An interesting characteristic of our feedback technique is that we can potentially run low level processes only when required during the inference.

Apart from the advantages of keeping the inference tractable, a feedback loop can also be useful in other ways. Sun et al. [14] apply a feedback loop for object detection with geometrical context. They jointly infer about the location of an object, the 3D layout of the scene and the geometrical relationships between the object and the 3D layout. The speciality of their feedback loop is that the object detector module adaptively improves its accuracy in the confidence measures of detections based on the feedback from the scene layout.

The idea of incrementally building a network can be approached in principled ways, including Cutting Plane Inference (CPI) and Cluster Pursuit Algorithms. Many inference problems can be cast as an Integer Linear Program (ILP) which is well suited for CPI. CPI employs an iterative process where the ILP is kept small by adding only the most violated constraints. However, CPI cannot be used for our feedback loop technique where we need to selectively set some detections to false. Sontag et al. [13] propose a cluster pursuit algorithm, an alternative formulation that incrementally adds cliques of variables (called *clusters*) and optimizes the dual function, an objective function obtained through Lagrangian relaxation that is an upper bound on the original (or *primal*) objective function. Their score function for clusters is an approximation to the decrease in the dual value of the objective function after adding a cluster, which is derived from the message passing updates of Globerson and Jaakkola [4]. We use this idea of cluster pursuit algorithm and derive a feedback technique for higher order Markov networks. Our scoring functions use the dual value to calculate approximations for the increase in the primal MAP value after adding a particular cluster.

3 Incremental Inference with Feedback Loop

We consider Markov networks defined over binary nodes $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ with factors $\theta_c(\boldsymbol{x}_c)$ defined over cliques of nodes $\boldsymbol{x}_c$ such that $c_1, \ldots, c_k \subset \{1, \ldots, n\}$. The Maximum A Posteriori (MAP) problem is defined as finding an assignment $\boldsymbol{x}^*$ that maximizes the function

$$\Phi(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{c \in \mathcal{C}} \theta_c(\boldsymbol{x}_c) \tag{1}$$

The nodes x_i are instantiated over candidate detections that are hypothesized by low level detectors. For example, they can be object detections obtained from running single-object detectors. The detector confidence scores output along with the detections are used as unary factors for the nodes. The factors θ_c that involve more than one detection represent the relationships between the detections. For example, they can be spatial relationships like the placement of an object on top of other objects. We obtain a MAP solution by performing inference, that will ultimately label the hypothesized detections as true positives or false positives.

In Markov networks of high order, every newly added detection can become combinatorially linked to other detections through the higher order factors. When many detections are hypothesized at low precision, the size of the Markov

network becomes exponentially large and the inference process becomes computationally expensive even though many of the detections are going to be inferred as false.

The goal of our incremental approach for inference is to maximize the function in (1) while keeping the network size small. We achieve this by unclamping only those detections that are most likely to be labeled as true by the inference. The rest of the detections are clamped to false, and while they always participate in the objective function over the iterations, they are excluded from the network during inference. We first perform inference with an initial network constructed from high confidence detections while the rest are clamped to false. We then calculate scores for the remaining detections based on the initial network. The scores measure the change in the MAP value after adding a detection to the current network. These scores are equivalent to locally adding an offset to the low level detector confidences, based on the feedback, so that the detections appear above the threshold. Another way to interpret this is that the thresholds get locally modified to select the detections that are below the threshold. We then unclamp a selected number of top detections and the process is repeated. When the incremental procedure is stopped, the MAP solution to the current network provides the true/false labels to the active detections and the remaining set of detections are labeled as false.

3.1 Clusters under Closed World Assumption

We show that incrementally unclamping detections is equivalent to adding clusters of factors. First we partition the Markov network into three clusters as shown in Figure (1). Let f be the set of active detections that are currently in to the network and $\boldsymbol{x_f}$ be the nodes that are instantiated over only the detections from f. The factor θ_f is defined over just the nodes $\boldsymbol{x_f}$. Let g be the set of one or more detections that is to be unclamped in a given iteration and $\boldsymbol{x_g}$ be the nodes instantiated over at least one detection from g and any other detections from f. The factor θ_g is defined over nodes $\boldsymbol{x_g}$ and other nodes from $\boldsymbol{x_f}$ that it shares with θ_f. Let h be the remaining set of detections and $\boldsymbol{x_h}$ be the nodes that are grounded over at least one detection from h and any other detections from $f \cup g$. The factor θ_h is defined over $\boldsymbol{x_h}$ and the other shared nodes with θ_f and θ_g. The overall objective function expressed as a sum of these clusters is

$$\begin{aligned}\Phi(\boldsymbol{x}) = \theta_f(x_{f1}, x_{f2}, x_{f3}, x_{f4}) + \theta_g(x_{g1}, x_{g2}, x_{f2}, x_{f3}) \\ + \theta_h(x_{h1}, x_{g2}, x_{f3}, x_{f4})\end{aligned} \tag{2}$$

Under the closed world assumption, any detection that is not included in the Markov network due to thresholding is assumed to be false. To satisfy this condition during the incremental process, we need to repartition the objective function (2). During every iteration of the process, we have a Markov network that includes a set f of active detections. The remaining detections from g and h are not yet added and hence the nodes instantiated over these detections must be clamped to false. The associated factors are projected on to the current network

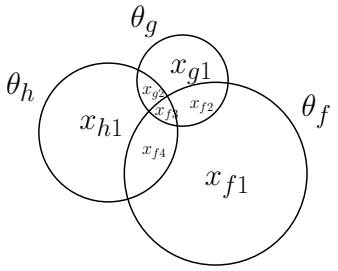


Fig. 1. The shared nodes between clusters in a partitioning of a Markov network. The set f contains active detections that are currently in the network and $\boldsymbol{x_f}$ are the nodes that are instantiated over only the detections from f. The set of factors $\theta_f(\boldsymbol{x_f})$ is defined over the nodes $\boldsymbol{x_f}$. Similarly, g is the set of detections to be unclamped at an iteration and h is the set of detections that are still clamped to false.

after setting the nodes of the excluded detections to false. The resulting objective function is

$$\begin{aligned}\Phi_{\text{cur}}(\boldsymbol{x}_{\text{cur}}) = \theta_f(x_{f1}, x_{f2}, x_{f3}, x_{f4}) + \theta_g(x_{g1} = 0, x_{g2} = 0, x_{f2}, x_{f3}) \\ + \theta_h(x_{h1} = 0, x_{g2} = 0, x_{f3}, x_{f4})\end{aligned} \tag{3}$$

To calculate a score for the set of detections in g, we need the objective function to include these detections in the active set while all other remaining detections from h are still clamped to false. This gives rise to the objective function

$$\begin{aligned}\Phi'(\boldsymbol{x}) = \theta_f(x_{f1}, x_{f2}, x_{f3}, x_{f4}) + \theta_g(x_{g1}, x_{g2}, x_{f2}, x_{f3}) \\ + \theta_h(x_{h1} = 0, x_{g2}, x_{f3}, x_{f4})\end{aligned} \tag{4}$$

Hence, the cluster of factors that need to be added to the current network during an iteration is given by

$$\Phi_{\text{new}}(\boldsymbol{x}_{\text{new}}) = \Phi' - \Phi_{\text{cur}}(x_{\text{cur}}) \tag{5}$$

$$\begin{aligned}= \theta_g(x_{g1}, x_{g2}, x_{f2}, x_{f3}) - \theta_g(x_{g1} = 0, x_{g2} = 0, x_{f2}, x_{f3}) \\ - \theta_h(x_{h1} = 0, x_{g2} = 0, x_{f3}, x_{f4}) + \theta_h(x_{h1} = 0, x_{g2}, x_{f3}, x_{f4})\end{aligned} \tag{6}$$

We now propose three score functions that measure the change in the MAP value after adding the cluster $\Phi_{\text{new}}(x_{\text{new}})$ to $\Phi_{\text{cur}}(x_{\text{cur}})$, with varying degrees of accuracy and computational cost.

3.2 Detection Scoring Function

We define the score for a detection based on the change in the MAP value after adding the detection to the current network. If we are adding the detection in g, the score is given by

$$score(g)_{exact} = \Delta\Phi = \max\left[\Phi_{\text{cur}}(x_{\text{cur}}) + \Phi_{\text{new}}(x_{\text{new}})\right] - \max\left[\Phi_{\text{cur}}(x_{\text{cur}})\right] \tag{7}$$

We also propose an upper bound to the exact score - $score(g)_{upper}$, that is derived based on the ideas of cluster pursuit algorithm of Sontag et al. [13]. We first obtain a dual of the MAP problem through Lagrangian relaxation. The MAP problem is now equivalent to minimizing the dual objective function since the dual value is an upper bound on the primal MAP value. We then use the message passing algorithm of Globerson et al. [4] to obtain the message update equations for the dual variables. Similar to Sontag et al. [13], we obtain an approximation to the new dual value after adding a cluster to the current network, by performing one iteration of message passing. Since the dual value is an upper bound on the primal MAP value, the new decreased dual value gives an upper bound for the exact score.

Proposition 1 (Upper Bound Score). *An upper bound on the change in the MAP value (7) after adding a cluster is given by*

$$\Delta\Phi \leq score(g)_{upper} \tag{8}$$

$$= \frac{1}{|s|} \sum_{i \in s} \max_{x_i} \left(\max_{x_{cur \setminus i}} \Phi_{cur}(x_{cur}) + \max_{x_{new \setminus i}} \Phi_{new}(x_{new}) \right) - \max \Phi_{cur}(x_{cur}) \tag{9}$$

where s is the set of nodes in the intersection of the sets x_{cur} and x_{new}.

The proof can be found in the supplementary material. The first term in the upper bound score is equivalent to averaging the MAP values obtained by enforcing same assignment for one shared node at a time. The upper bound score can be calculated efficiently using an inference algorithm that calculates max-marginals with only a little computation overhead (eg. dynamic graph cuts [6]) and hence can avoid performing repeated inference to calculate the exact score.

We derive another approximation to the score function called the *Blind Score* since it is dependent only on the max-marginals of the current network and does not involve the max-marginals of the new cluster to be added. It is obtained as a lower bound to the upper bound score (not the exact score).

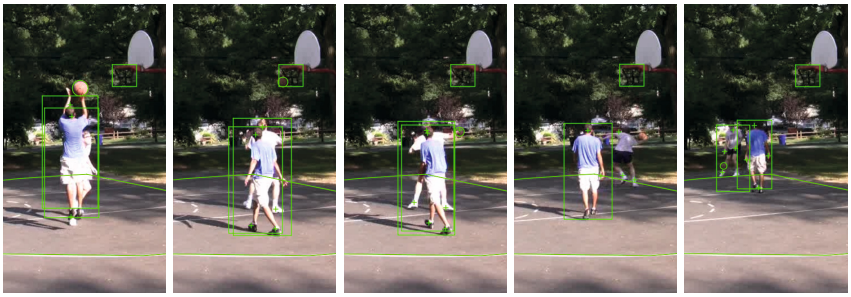
Proposition 2 (Blind Score). *A lower bound to the upper bound score (9) is given by*

$$score(g)_{upper} \geq score(g)_{blind} \tag{10}$$

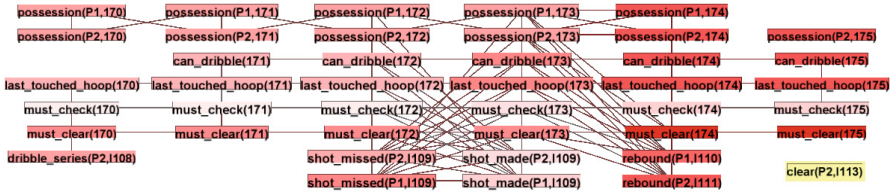
$$= \frac{-1}{|s|} \sum_{i \in s} \left| \max_{x_i = 0, x_{cur \setminus i}} \Phi_{cur}(x_{cur}) - \max_{x_i = 1, x_{cur \setminus i}} \Phi_{cur}(x_{cur}) \right| \tag{11}$$

where s is the set of nodes in the intersection of the sets x_{cur} and x_{new}.

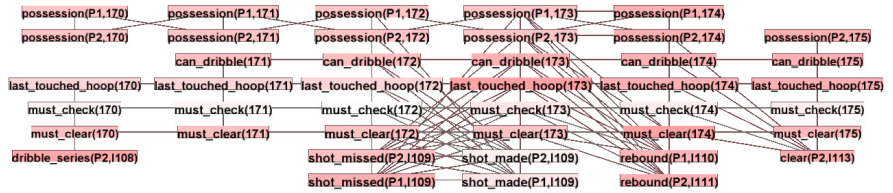
The proof can be found in the supplementary material. This score measures the average of the difference in max-marginals of the shared nodes. It indicates the susceptibility of the shared nodes in the current network to change their values when a new cluster is added. The score is low if the absolute difference in the max-marginals of the shared variables is high. This indicates that the current network has low uncertainty (or strong belief) in the assigned values



(a) A sequence of events which shows a shot being missed by Player1 and the rebound received by Player2. When Player2 is clearing the ball, the track goes missing for a while and hence the confidence measure for that clear event is low.



(b) Applying an initial threshold for Clear events does not include the highlighted Clear event. However the corresponding Shot Missed event by Player1 is included in the network. The absolute difference in the max-marginals represents *certainty* of a node assignment and hence the negative of that difference represents *uncertainty*. Here, darker colors indicate high uncertainty. When the Clear event is missing, the network is highly uncertain right after the Shot Missed event.



(c) The node assignments become more certain after adding the missing Clear event.

Fig. 2. Visualization of the Feedback Loop

to the shared variables. Similarly the score is high if the absolute difference in the max-marginals is low. This indicates that the network has high uncertainty in the assignments to the shared variables and that is where we need more evidence/observations.

Since the blind score is independent of max-marginals of the new cluster, it does not need the confidence score of a detection which is usually used as a unary potential in the new cluster. This can save computation for the low level detectors by avoiding expensive procedures like feature extraction and classification throughout an image/video and instead run them only when it is needed by the inference. However, the blind score needs to know the shared variables

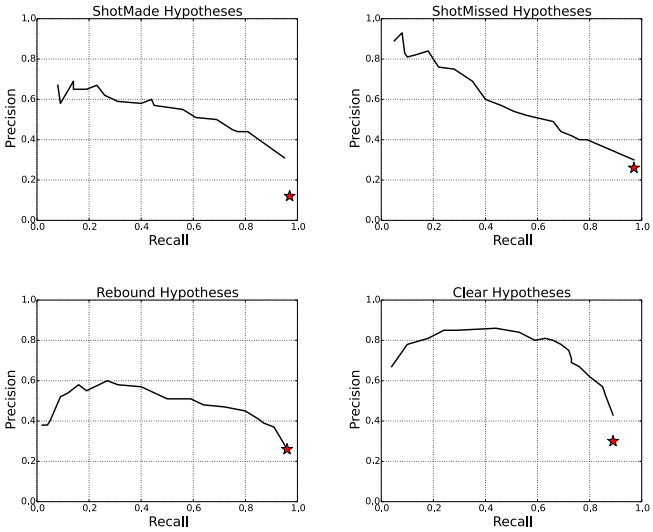


Fig. 3. PR curves for the newly hypothesized events with continuous confidence measures. The red star shows the operating point of Morariu et al. [10] in their feed-forward approach.

(s) between the new cluster and the current network. This corresponds to determining the locations where the detector would be run and these are usually easy to obtain for sliding-window approaches. For example, to perform 3D object detection, Lin et al. [9] first generate candidate cuboids without object class information which fixes the structure of their network and hence tells us the shared variables for any cluster. They then extract features for generating unary potentials and use it in a contextual model to assign class labels to the hypothesized cuboids. If we use the blind score during the inference, we can potentially save computation by not extracting features for cuboids that are likely to be labeled as false. Figure (2) illustrates our feedback loop technique using an example from the basketball dataset of Morariu et al. [10].

4 Experiments

4.1 One-on-One Basketball Dataset

The one-on-one basketball dataset used by Morariu et al. [10] contains tracks of players and ball along with court annotations for seven basketball videos. The are eight events of interest: Check, Out Of Bounds, Shot Made, Shot Missed, Rebound, Clear, Dribble and Steal. They use a Markov Logic Network (MLN) [12] to represent high level rules of the game which interrelates the various events. The inputs to the MLN are candidate events hypothesized by low level detectors which use the tracks of players and the ball.

	Morariu et al. [10]			Ours		
	P	R	F1	P	R	F1
Check	0.84	0.89	0.87	**0.86**	**0.90**	**0.90**
Clear	**0.86**	0.61	0.71	0.81	**0.82**	**0.82**
Dribble	**0.81**	0.75	0.78	0.79	**0.82**	**0.80**
OutOfBounds	**0.88**	**0.66**	**0.75**	0.80	0.62	0.70
Rebound	0.62	0.72	0.67	**0.82**	**0.84**	**0.83**
ShotMade	0.64	0.86	0.73	**0.87**	**0.87**	**0.87**
ShotMissed	0.67	0.79	0.72	**0.81**	**0.85**	**0.83**
Steal	0.08	**0.50**	**0.13**	**0.25**	0.25	0.12
Overall	0.72	0.75	0.74	**0.81**	**0.83**	**0.82**

Table 1. Comparison of MLN Recognition Performance using all the hypothesized intervals without thresholding. We can see that the continuous confidence measures for input events play a significant role in improving the performance.

4.2 Hypothesizing Candidate Events

In the MLN used by Morariu et al. [10], each event was hypothesized with just two discrete confidence values. However, continuous confidence measures are required for the events to better tie them to reality. We hypothesize a new set of candidates with continuous confidence measures for the Shot Made, Shot Missed, Rebound and Clear events and copied the other events (Check, Dribble, Out Of Bounds, Steal) from their dataset. The confidences are obtained based on observations like ball near a player, ball seen inside the hoop, player being inside the two point area, etc. The PR curves of the event hypotheses is shown in Figure (3). Since our modified observation model introduces higher uncertainty in event interval endpoints, we also make few minor modifications to the original MLN to make it robust to the overlapping endpoints of different event intervals.

We first test the importance of continuous confidences in the feed-forward setting by feeding in all the hypothesized intervals to the MLN without thresholding. The confidence measures are used as unary potentials for event predicates in the MLN. Inference is then performed to obtain a MAP assignment for the ground MLN, which labels the candidate events as true or false based on the high level context of the game. The results are shown in Table (1). We see that the confidence measures play a significant role in improving the event recognition performance.

We have implemented our system as an extension of Alchemy [7], a software package for probabilistic logic inference. The MAP problem for MLNs is framed as an Integer Linear Program (ILP) [11] and we integrated our system with the Gurobi ILP solver [5] for performing inference.

4.3 Incrementally Adding Events with Feedback Loop

We demonstrate the feedback loop technique by incrementally adding one type of event, the Clear event. The confidence values for the Clear event are scaled between 0.5 and 1. We initialize the network with all the event intervals except

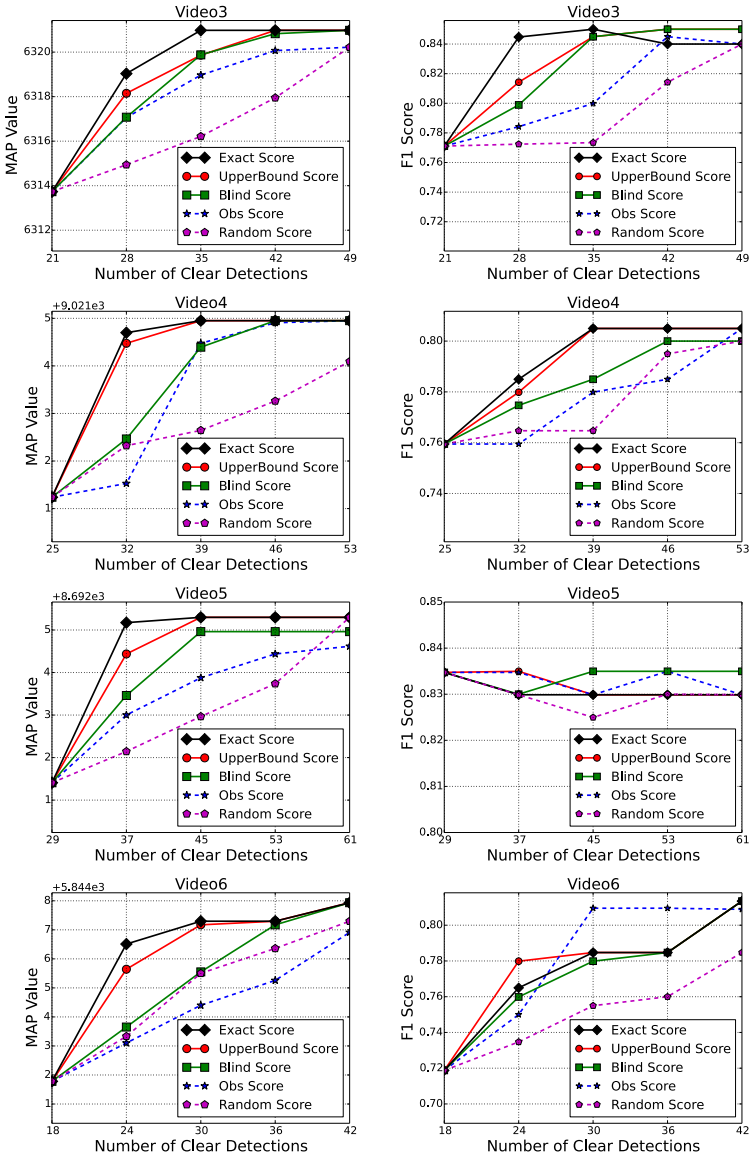


Fig. 4. Feedback based scores achieve better solutions with fewer detections; We apply an initial threshold on the Clear events and incrementally add the remaining events using the feedback based scores. We measure the exact MAP value of the Markov network along with the $f1$ score corresponding to the ground truth. The plots start at the same initial value for all the five scoring methods since the initial network contains the same set of events. Our feedback based scores achieve better solutions with fewer detections than the baselines - observation score and random score.

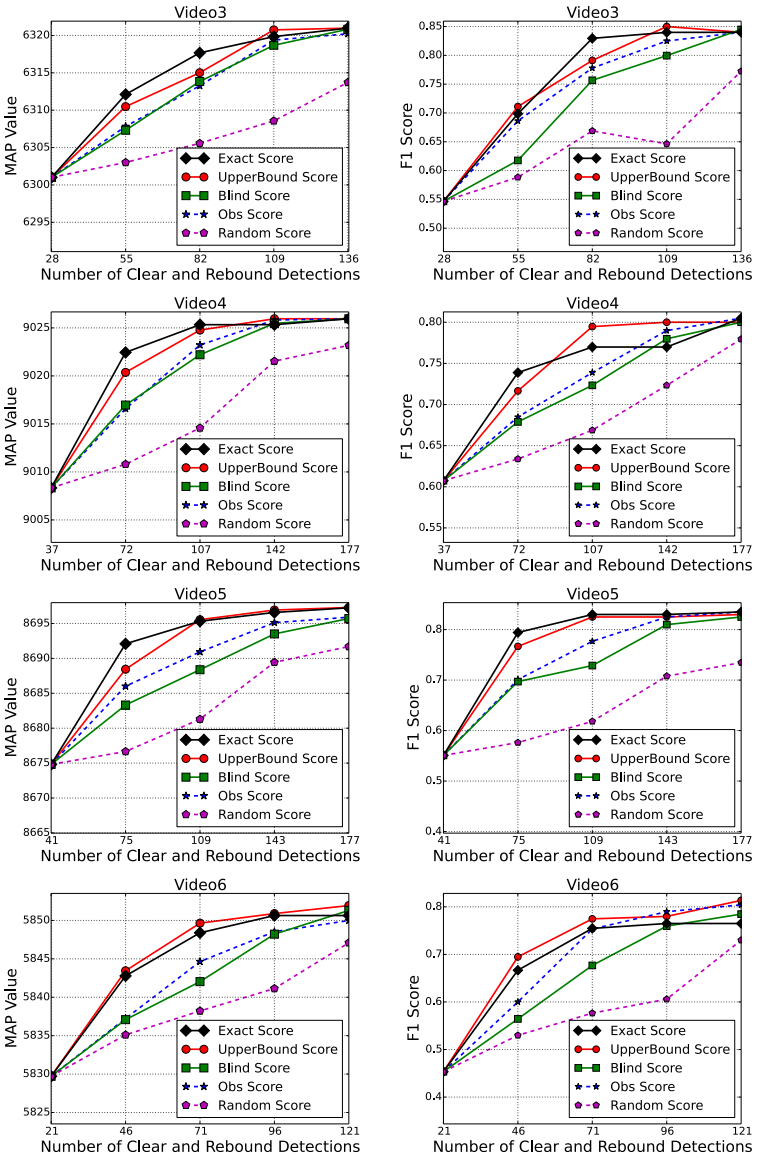


Fig. 5. We apply threshold on both the Rebound and Clear events for initial network and then incrementally add both events at every iteration. We still see that the exact score and the upper bound score reach better solutions with fewer detections than the observation score. However, the blind score falls slightly below the observation score since it depends only on the current network and the context in the current network is weak due to fewer events.

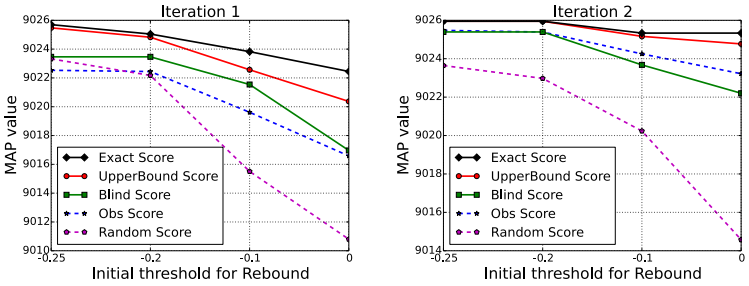
for Clear which is thresholded at 0.75. We then run four iterations of the feedback loop and in each iteration, we add a certain number of top ranking Clear events from the remaining set. There are five different kinds of scores that we experiment with: $score(g)_{exact}$, $score(g)_{upper}$, $score(g)_{blind}$, observation score and random score. The observation and random scores are baseline approaches to incrementally adding constants without using a feedback loop. The observation score is the confidence measure that comes from the low level detectors. By adding constants based on their observation score, we are effectively reducing the threshold uniformly throughout the video. The random score is basically selecting a certain number of Clear events randomly and adding them without looking at either the confidence measures or the context in the main network.

The results are shown in Figure (4). Among the seven videos from the dataset, four of the them are large enough to add intervals in an iterative manner. We show the plots of MAP value and also the $f1$ scores against the number of Clear detections in the current network. The plots start at the same initial value for all the five scoring methods since the initial network contains the same set of detections. The goal of our feedback technique is to reach the final MAP value in few iterations by adding only the relevant detections while keeping the rest of them false. The MAP values increase faster with all of our three feedback based score functions when compared to the observation score. The exact score is the quickest followed by the upper bound score and then the blind score. The plots of $f1$ scores also show that we can reach the best possible value with fewer detections using feedback based score functions implying that they select the most relevant events from the missing ones. We observe that the blind score performs well when compared with the observation score. This indicates that the context in the main network has a huge impact on what needs to be added to improve the MAP value.

We also experiment with jointly thresholding the Rebound event along with the Clear event. The Rebound events are scaled between -0.25 to 0.1 and we choose a threshold of 0 for the initial network. The Clear events are scaled between 0.5 to 1 and we choose a threshold of 0.75. We then proceed to iteratively add the remaining Rebound and Clear events. The results in Figure (5) show that the exact score and upper bound score can reach the best possible MAP value and $f1$ score by adding fewer detections. However the plot for blind score falls below that of the observation score. By increasing the threshold on the Rebound event, the strength of context in the main network is weakened and hence the blind score which is dependent on just the current network starts to perform poorly.

4.4 Effect of Initial Threshold

To observe the effect of initial threshold, we experimented with four different initial thresholds for the Rebound event. Like before, the Rebound events are scaled between -0.25 to 0.1 and the Clear events are scaled between 0.5 to 1. We choose a threshold of 0.75 for Clear events and vary the initial threshold for Rebound events starting from the lowest, which is -0.25 (includes all the



(a) First iteration of adding Rebound and Clear events

(b) Second iteration of adding Rebound and Clear events

Fig. 6. Effect of initial threshold for the Rebound event in video 4; The confidence scores for the Clear events are scaled between 0.5 to 1 and the Rebound events between -0.25 to 0.1. We fix the initial threshold for Clear event at 0.75 and vary the threshold for Rebound from -0.25 to 0. We observe that a higher threshold for Rebound event in the initial network decreases the MAP value that is achieved in the first iteration of adding Rebound and Clear events to the initial network. The blind score continues to perform poorly in later iterations at higher initial threshold due to weak context in the initial network. However, the exact score and the upper bound score are still stable with respect to the initial threshold.

Rebound events) and increase up to the value 0 which is high enough to weaken the context. As the initial threshold is increased for the Rebound events, the initial network becomes sparse weakening the context in the initial network. Figure (6a) shows that a higher threshold decreases the MAP value achieved in the first iteration of adding events to initial network. The blind score is affected the most since it is dependent only on the current network. It continues to perform poorly in later iterations (Figure (6b)) at higher initial threshold for the Rebound event. Hence, it is important to select a reasonably high threshold that allows enough number of events in the initial network without increasing the network size.

5 Conclusion

We propose a computational framework for a feedback loop between high level semantics and low level detectors in a computer vision system, where we use the information in the high level model to select relevant detections from a set of candidate hypotheses. We start with high confidence detections and then iteratively add only those detections to the model that are most likely to be labeled as true by the high level model. This helps us keep the model size small especially in the presence of many noisy detections. We develop the framework for higher order Markov networks and propose three feedback based scoring functions to rank the detections. We show through our experiments on an event

recognition system that the feedback loop can construct smaller networks with fewer detections and still achieve the best possible performance.

Acknowledgement. This research was supported by contract N00014-13-C-0164 from the Office of Naval Research through a subcontract from United Technologies Research Center.

References

1. Brendel, W., Fern, A., Todorovic, S.: Probabilistic event logic for interval-based event recognition. In: CVPR (2011)
2. Choi, M., Torralba, A., Willsky, A.: A tree-based context model for object recognition. PAMI **34**, 240–252 (2012)
3. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV (2009)
4. Globerson, A., Jaakkola, T.: Fixing max-product: convergent message passing algorithms for MAP LP-relaxations. In: NIPS (2007)
5. Gurobi-Optimization-Inc.: Gurobi Optimizer Reference Manual (2013). http://www.gurobi.com
6. Kohli, P., Torr, P.: Measuring uncertainty in graph cut solutions - efficiently computing min-marginal energies using dynamic graph cuts. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 30–43. Springer, Heidelberg (2006)
7. Kok, S., Sumner, M., Richardson, M., Singla, P.: The Alchemy System for Statistical Relational (2009). http://alchemy.cs.washington.edu/
8. Kumar, M.P., Koller, D.: Efficiently selecting regions for scene understanding. In: CVPR (2010)
9. Lin, D., Fidler, S., Urtasun, R.: Holistic scene understanding for 3D object detection with RGBD cameras. In: ICCV (2013)
10. Morariu, V., Davis, L.: Multi-agent event recognition in structured scenarios. In: CVPR (2011)
11. Noessner, J., Niepert, M., Stuckenschmidt, H.: RockIt: exploiting parallelism and symmetry for map inference in statistical relational models. In: AAAI (2013)
12. Richardson, M., Domingos, P.: Markov logic networks. Machine Learning, January 2006
13. Sontag, D., Meltzer, T., Globerson, A.: Tightening LP relaxations for map using message passing. In: UAI (2008)
14. Sun, M., Bao, S.Y., Savarese, S.: Object detection using geometrical context feedback. IJCV, August 2012
15. Tran, S.D., Davis, L.S.: Event modeling and recognition using markov logic networks. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 610–623. Springer, Heidelberg (2008)
16. Zhu, Y., Nayak, N., Chowdhury, A.R.: Context-aware activity recognition and anomaly detection in video. In: CVPR (2013)

How to Supervise Topic Models

Cheng Zhang(✉) and Hedvig Kjellström

Computer Vision and Active Perception Lab, Centre for Autonomous Systems,
KTH Royal Institute of Technology, Stockholm, Sweden
{chengz,hedvig}@kth.se

Abstract. Supervised topic models are important machine learning tools which have been widely used in computer vision as well as in other domains. However, there is a gap in the understanding of the supervision impact on the model. In this paper, we present a thorough analysis on the behaviour of supervised topic models using Supervised Latent Dirichlet Allocation (SLDA) and propose two factorized supervised topic models, which factorize the topics into signal and noise. Experimental results on both synthetic data and real-world data for computer vision tasks show that supervision need to be boosted to be effective and factorized topic models are able to enhance the performance.

Keywords: Topic modeling · SLDA · LDA · Factorized supervised topic models

1 Introduction

Topic modelling, as one of the most important machine learning tools, has been successfully applied to in computer vision [5,8,11,15,22,24], as well as other domains. It is a type of generative latent structure model that represent the underlying structure of data as topics. Hence, it has advantages on handling missing data and reasoning the data structure, which are desired properties in many computer vision tasks.

In many applications, not least in computer vision, the learning task is often to estimate a label from a piece of data. Hence, supervised topic models have drawn a lot of attention. Although several supervised topic models have been proposed [2,8,12,27], very little work has been done to study the impact of supervision on the latent representation itself. In this paper, we will perform such a study, analysing the behaviour of one type of supervised topic model, Supervised Latent Dirichlet Allocation (SLDA) [2,22], and propose a number of enhancements that could potentially improve the performance.

Supervised topic models are especially important for computer vision since classification is one of the most common tasks in this domain. Popular supervised

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-16181-5_39) contains supplementary material, which is available to authorized users.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 500–515, 2015.
DOI: 10.1007/978-3-319-16181-5_39

models include: the above mentioned SLDA [2,22], which models the joint likelihood of the class label and the observed data in a principled Bayesian framework; Labeled LDA (LLDA)[1] [8], which optimizes the hyperparameter for each class, but no direct dependence between the class label and the observed data is modeled; Discriminative LDA (DiscLDA) [12], which models the conditional likelihood of the data on the class label through matrix transformation; Max-Entropy Discrimination LDA (MedLDA) [27,28], which utilizes max-margin principle to learn the topic space using regularised Bayesian inference. Among these models, SLDA is the most popular one, since it is the most principled and straightforward Bayesian framework. Hence, we will focus on SLDA based topic models in this paper.

Very few studies have been done on understanding the behaviour of topic models, although various new topic models have been proposed every year. Semantic consistency of learned topics is studied [6] for classic unsupervised topic models which are pLSI [10], LDA [3] and Correlated Topic Model (CTM) [1]; and the behaviour of LDA with respect to the size of observed data (length of documents and number of documents) has been studied recently [19]. However, there is still a big gap in the understanding of the behaviour of supervised topic models. In computer vision, similar classification results have been achieved using standard LDA with a separate SVM for classification, as with SLDA [15], which raise the question of how effectively the topic space in an SLDA model is adapted to the class labels. How much impact the supervision has in the model has been discussed [18], but never been studied. In this paper, we will address this question and present analysis on the behaviour using SLDA and an adaptation, Power SLDA (P-SLDA), where the effect of the class label is boosted.

A latent representation that is able to capture the difference among data from different classes is the key to achieve good classification performance. The goal of using supervised topic models is to learn a better latent representation of the data that is suitable for the given task. Intuitively, only part of the information from the data is relevant for the classification task. For example, given the task to classify mugs from books, the shape of the object is relevant, which is the signal, and the pattern printed on the objects is not relevant, which is the noise. Classic topic models model the entire data together. Hence, the performance suffers when the data have low signal-noise ratio using classic topic models. Several works for different applications have considered this problem and allow the model to have different strategies to handle noise [11,16,21,25]. Three of these, [11,16,21], are designed for specific (non-classification) tasks, while the fourth, [25], is heuristic in the sense that it introduces an entropic regularizer. In this paper, we propose two variations of SLDA which are probabilistically principled framework. To explore a better way to supervise topic models, these proposed models will be studied together with SLDA and P-SLDA on how they can influence the learning of topics.

[1] LLDA indicates the model from [8] which is deigned for natural scene classification. The other popular model termed LLDA is from [17] and is designed for using multiple tags rather than class label.

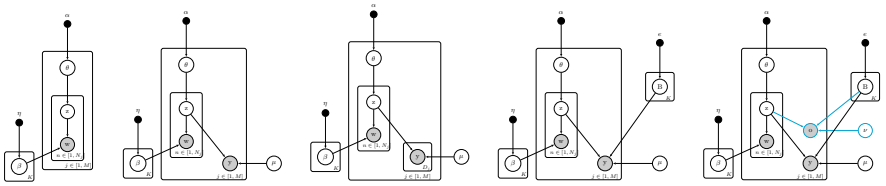
We summarize our contributions as follows:

1. *A thorough analysis of the supervision effect of SLDA compared to LDA are presented.*
   Experimental results shows that the impact of supervision is limited on the learning of the latent space compared to the LDA due to the imbalance of the model.
2. *Power SLDA (P-SLDA) which maps the class label to higher dimension to boost supervision is contracted for further analysis on supervision behaviours.*
   Clear impact can be observed with boosted supervision, however, the benefit of supervision with P-SLDA is data dependent.
3. *Two novel factorized topic models are designed to learn better latent representation for classification tasks and provide better interpretation of topics.*
   Experimental results show that these factorized models are able to factorize topics into signal and noise and are more robust compared to SLDA and P-SLDA.

The paper is organised as follows: all the models that are involved in the paper are described in Section 2; experimental evaluation and analysis are presented in Section 3; finally, we conclude the paper in Section 4.

2 Methods

In this section, we will firstly present all the models that are used in this paper. Then, we will briefly present the inference algorithms and classification methods. The derivation of inference algorithms and implementation details will be presented in the supplementary document.



(a) LDA (b) SLDA (c) P-SLDA (d) NUF-SLDA (e) NCF-SLDA

Fig. 1. Graphical representation of topic models studied in this paper. M indicates the number of documents; N_j indicates the number of words in the document j; K indicates the number of topics; α and β are hyper-parameters; w indicates the observed words; y indicates the class label;$\theta \sim Dir(\alpha)$ is the topic distribution for each document; $z \sim Mult(\theta)$ is the topic assignment for each word; $\beta \sim Dir(\eta)$ indicates the topics which are distributed over words.

2.1 Models

Topic models encode latent structure as topics, which assume that each piece of information is composed of latent topics. LDA [3], shown in Figure 1(a), is the cornerstone of topic models which was originally applied in information retrieval.

LDA assumes a generative process where each document is modeled as a distribution over topics, and each topic is modeled as a distribution over words. LDA is the basic framework which has been evolved and applied for different tasks and all these models are called topic models. They can be applied on different types of data. For example, in computer vision, a document can be an image and a word can be a visual word from the bag-of-words representation. In this paper, we concentrate on supervised topic models that can be applied for classification tasks which is extremely important for computer vision applications. In this section, we will present all the models in the evolutionary order. Supervision is designed to be a part of the model for all the models but the first model, which learn the topics and classification parameters in separate stages.

LDAC. LDA [3] is an unsupervised model. To perform classification tasks, the simplest way is to use the topics that are learned from LDA and apply an additional classifier on the topics. We call it LDA for Classification (LDAC) in this paper. Hence, LDAC is the same on modeling the topics as LDA. LDA/LDAC can be present as:

$$p(w, z, \theta, \beta|\alpha, \eta, \mu) = p(\beta|\eta) \prod_{j=1}^{M} \Big(p(\theta_j|\alpha) \prod_{n=1}^{N} \left(p(w_{jn}|z_{jn}, \beta) p(z_{jn}|\theta_j) \right) \Big). \tag{1}$$

The graphical representation is shown in Figure 1 (a). A standard softmax regression is used for the classification tasks for a fair comparison in this paper. Note that the training of LDA and the softmax regression are done in separate steps. The label information is not involved in the learning of the topics.

SLDA. For classification tasks, a topic representation that leads to better separation between classes is preferred. Hence, a supervised model is preferred for classification tasks. SLDA [2,22], shown in Figure 1 (b), is the most straightforward and the most commonly used supervised topic modelling framework. Compared to LDA, the supervision is modelled as a response [2] to the topic assignments of each document. Hence, the topics are used to generate both the words and the label. SLDA models the words and the label jointly, which means that the inference will optimize the joint likelihood of the words and the labels. SLDA can be represented as:

$$p(w, z, \theta, \beta, y|\alpha, \eta, \mu) = p(\beta|\eta) \prod_{j=1}^{M} \Big(p(\theta_j|\alpha) p(y_j|z_{1:N}, \mu) \prod_{n=1}^{N} \left(p(w_{jn}|z_{jn}, \beta) p(z_{jn}|\theta_j) \right) \Big), \tag{2}$$

where y_j is a one-dimensional label and $p(y_j = l|z_j, \mu) = \frac{\exp(\mu_l^T(\bar{z}_j))}{\sum_{c=1}^{C} \exp(\mu_c^T(\bar{z}_j))}$, where $\bar{z}_j = \left(\frac{1}{N} \sum_{n=1}^{N} \mathfrak{z}_{jn} \right)$, in which $\mathfrak{z}_{jn}$ is the vector representation of the topic assignment indicator z_{jn}.

[2] The response can be any type with generalized linear model [2]. In this paper, we mainly focus on the case when the response is the class label [22].

P-SLDA. Topic assignments are used to explain both the words and the label using SLDA. However, for a document, the words are N_j-dimensional and the class label y_j is one-dimensional. It is not balanced between these two views due to different dimensionality. Hence, the supervision may not be sufficient in the model. Power-SLDA (P-SLDA), shown in Figure 1 (c), is a model that we construct to study the effectiveness of supervision with SLDA. It is a variation of SLDA. Compared to SLDA, in which the response is drawn only once for each document, P-SLDA draw the response D_j times. Hence, P-SLDA allows the label to be mapped to D_j dimensional. By varying D_j, we can study how much the supervision influence the learning of the topics. P-SLDA can be presented as:

$$p(w, z, \theta, \beta, y|\alpha, \eta, \mu) = p(\beta|\eta) \prod_{j=1}^{M} \Big(p(\theta_j|\alpha) p(y_j|z_{1:N}, \mu)^{D_j} \prod_{n=1}^{N} \left(p(w_{jn}|z_{jn}, \beta) p(z_{jn}|\theta_j) \right) \Big). \quad (3)$$

Comparing to SLDA, the difference lies in the power index D_j on $p(y_j|z_{1:N}, \mu)$. Since documents may have different length, we define $D_j = \frac{N_j}{s}$, where s is the scaling parameter. When $s = N_j$, PSLDA becomes SLDA. When $s = 1$, the label is mapped to the same dimension as the words $D_j = N_j$.

NUF-SLDA and P-NUF-SLDA. SLDA and P-SLDA model all the data together as many other topic modelling framework based on SLDA. However, the data are noisy, and the noise in the data may be inconsistent with the label which will cause poor performance when the data has low signal-noise ratio. The concept, which use factorized representation for information that can be shared among different views and information that cannot be shared between different views, has been applied to different frameworks with a long history [4,7,20]. We will adopt the same concept for supervised topic models. In NUF-SLDA, we assume that only part of the topics should be shared between the observed words and the label, which are used for generating the words and generating the label; and the other part of the model is only used to model the rest of words which are not relevant for the classification task. We call the shared topics as signal topics and the ones not shared as noise topics. As the graphical representation of the model shown in Figure 1 (d), we introduce a signal-noise indicator B in the SLDA model which indicates whether the topic is used to model signal or noise, where $B \sim Bern(e)$. Comparing to SLDA, the main difference is that the class label y only respond to the topics which are indicated as signal (with $B_k = 1$). NUFSLDA can be represented as:

$$p(w, z, \theta, \beta, y|\alpha, \eta, \mu) = p(\beta|\eta) p(B|e) \prod_{j=1}^{M} \Big(p(\theta_j|\alpha) p(y_j|z_{1:N}, \mu, B) \prod_{n=1}^{N} \left(p(w_{jn}|z_{jn}, \beta) p(z_{jn}|\theta_j) \right) \Big). \quad (4)$$

Differently from SLDA, the softmax regression in the NUF-SLDA is defined by[3]

$$p(y_j = l|z_j, B, \mu) = \frac{\exp(\mu_l^T (\bar{z}_j \otimes B))}{\sum_{c=1}^{C} \exp(\mu_c^T (\bar{z}_j \otimes B))}. \quad (5)$$

[3] "$\otimes$" is used to indicate the element product.

Similarly to P-SLDA, P-NUF-SLDA with boosted supervision can be constructed as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)p(B|e)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu,B)^{D_j}\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{6}$$

NCF-SLDA and P-NCF-SLDA. The previous model, NUF-SLDA, factorizes signal and noise. In this section, we adjust the model to constrain the noise to be *structured*, which share the same assumption as in[25]. Compared to NUF-SLDA, the key difference is that the noise part responds to a noise class. With this constraint, all the noise has the same label, hence, it is the structured noise. The graphic representation of NCF-SLDA is shown in Figure 1 (e). The noise response variable o is introduced, which is marked cyan in Figure 1 (e). NCF-SLDA can be represented as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)p(B|e)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu,B)p(o|z_{1:N},B,\nu)\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{7}$$

The additional noise response term is modeled as $p(o_j|z,B,\nu)=\frac{\exp(\sum_{k=1}^{K}\nu_k\bar{z}_{jk}(1-B_k))}{\exp(\sum_{k=1}^{K}\nu_k\bar{z}_{jk}(1-B_k))+1}$.

Similarly, P-NCF-SLDA can be presented as:

$$p(w,z,\theta,\beta,y|\alpha,\eta,\mu) = p(\beta|\eta)p(B|e)\prod_{j=1}^{M}\Big(p(\theta_j|\alpha)p(y_j|z_{1:N},\mu,B)^{D_j}p(o|z_{1:N},B,\nu)\prod_{n=1}^{N}\big(p(w_{jn}|z_{jn},\beta)p(z_{jn}|\theta_j)\big)\Big). \tag{8}$$

2.2 Inference

Variational inference and sampling based methods are the two main classes of methods that are generally used in the inference of topic models. Variational inference is known for its fast convergence and it is easy to adapt batch variational inference to an online setting [9,23,26]. In this work, we will use the standard batch mean field variational inference for all the models in this paper. Fully factorized variational distribution is used as [2,3,22]. Derivation details of variational inference for NUF-SLDA and NCF-SLDA are presented in the supplementary document.

2.3 Classification

For classification, we would like to estimate $p(y_j|z_j,\mu)$ or $p(y_j|z_j,B,\mu)$ for the test document j. The estimated label is the one with highest probability. In this case, the variational approximation for the true posterior is used. Hence, the prediction rule for LDAC, SLDA and PSLDA is:

$$\hat{y}_j = argmax_{l\in\{1,\ldots,C\}}\mathop{\mathbb{E}}_{q}[\mu_l^T\bar{z}] = argmax_{l\in\{1,\ldots,C\}}\mu_{y_j}^T\Big((\frac{1}{N}\sum_{n=1}^{N}\phi_{jn})\Big), \tag{9}$$

and the prediction rule for both NUF-SLDA and NCF-SLDA is:

$$\hat{y}_j = argmax_{l\in\{1,\dots,C\}} \mathop{\mathbb{E}}_{q}[\mu_l^T(\bar{z}\otimes B)] = argmax_{l\in\{1,\dots,C\}} \mu_{y_j}^T \Big((\frac{1}{N}\sum_{n=1}^{N}\phi_{jn})\otimes f\Big). \quad (10)$$

Note that μ is learned during the inference of the model and is able to affect the learning of topic assignment for all models but LDAC. In LDAC, μ is learned as a separate parameter where all the topics are learned using LDA.

3 Experiments

In this section, we will present our experimental results and discussion on these results. The experiments are carried out on three datasets: synthetic dataset[4], KTH video dataset and natural scene image dataset. The synthetic dataset is constructed to analysis the behaviour of the model in a controlled manner. The other two datasets are real world datasets, among which KTH dataset present a low signal- noise ratio case and the natural scene dataset present a high signal-noise ratio case. For each dataset, experiments are performed to evaluate two aspects: the effectiveness of the supervised topic models; and the performance of factorized topic models . The experimental results are ordered by datasets and the discussion will be presented in the end of this section.

3.1 Classification on Synthetic Dataset

We construct a synthetic dataset to study the behaviour of the models in a controlled manner. Figure 2 shows how synthetic data is generated. In the experiment, 200 documents for each class are generated for training and 40 documents

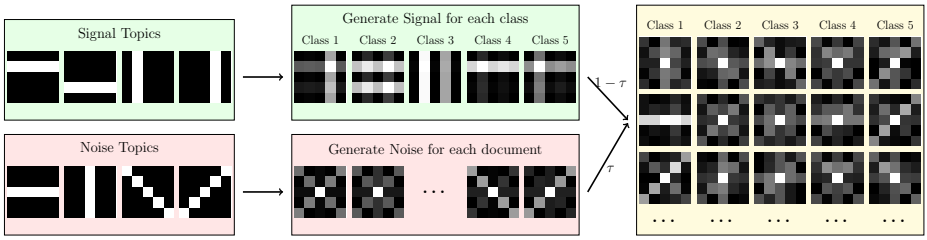
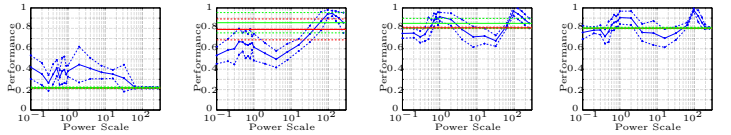
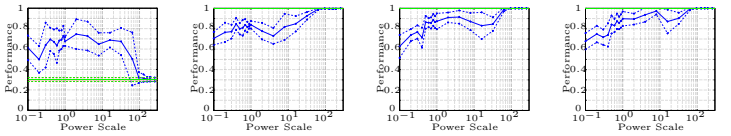


Fig. 2. The generation of the synthetic data. Eight topics are set first where four of them are defined as signal topics and the other four are defined as noise topics. Signal for five classes which are convex combination of the signal topics are generated. Then we add noise, which are random convex combination of noise topics, to generate the dataset. For each document/image, the noise is generated independently. The noise level is control by the parameter τ. The final document/image is generated by $(1-\tau)\times Signal + \tau \times Noise$. The noise level $\tau = 0.8$ is used for the example documents above.

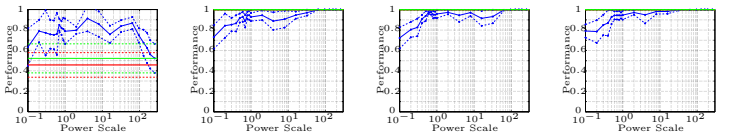
[4] The synthetic dataset and our implementation for all the novel models will be published.



(a) $K = 4$, $\tau = 0.9$ (b) $K = 8$,$\tau = 0.9$ (c) $K = 16$,$\tau = 0.9$ (d) $K = 20$,$\tau = 0.9$



(e) $K = 4$, $\tau = 0.8$ (f) $K = 8$,$\tau = 0.8$ (g) $K = 16$,$\tau = 0.8$ (h) $K = 20$,$\tau = 0.8$

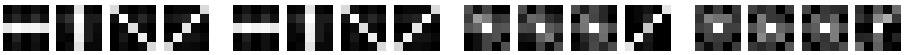


(i) $K = 4$, $\tau = 0.7$ (j) $K = 8$,$\tau = 0.7$ (k) $K = 16$,$\tau = 0.7$ (l) $K = 20$,$\tau = 0.7$

Fig. 3. Performance Evaluation for LDAC, SLDA and P-SLDA with difference power scale s under different number of topics K and different noise level τ. All experiments are run 9 times over different random seeding. The mean is represented using solid line and the standard deviation is represented using dashed line. The blue curve in these plots shows the performance of P-SLDA with different label dimension D. The x axis is the supervision power scale s which is plotted in the log scale ranging from 0.1 to 256, which indicates that D range from 2560 to 1 from left to right. While $s = N = 256$, P-SLDA becomes SLDA. The performance of SLDA is marked with green dot and the performance of LDAC is plotted in red.

for each class are generated for testing, which yields 1000 training and 200 testing documents in total.

Supervision Effectiveness. Firstly, we compare the classification performance on synthetic datasets using LDAC, SLDA and P-SLDA with different power scales s. Figure 3 shows the performance of these models with different number of topics and different noise levels. Hyperparameters $\alpha = 0.5$ and $\eta = 0.1$ are used in these experiments. All experiments are run 9 times with different random seeds for initialization. The mean and standard deviation are reported. LDAC and SLDA have similar performance over all different settings, although better performance is expected from SLDA over LDAC since the learning process is supervised. By boosting the supervision using P-SLDA, clear change of performance can be observed. This shows that the supervision is not effective using SLDA on learning of the latent space. Figure 3(a) (e) (i) show the performance with $K = 4$ with different noise levels. We can see that the improvement on classification results is significant through all different noise levels when the number of topics is small. Figure 3 (a) (b), (c), (d), show that the the performance can be clearly improved



(a) LDA (b) SLDA (c) P-SLDA D=8 (d) P-SLDA D: 256

Fig. 4. The learned topics ($\hat{\beta}$) from different models with number of topics $K = 4$ and $\tau = 0.8$. (a) The topics learned using LDA (b) The topics learned with SLDA (c) The topics learned with PSLDA with power scale $s = 32$ which is $D = 8$ (d) The topics learned with P-SLDA with power scale $s = 1$ which is $D = N = 256$

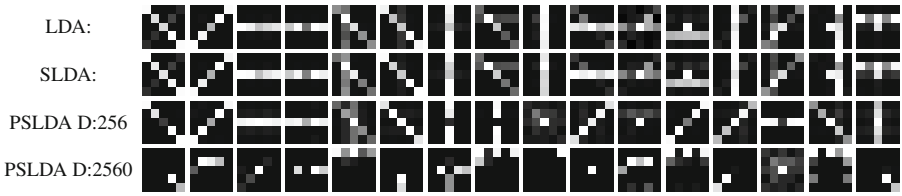
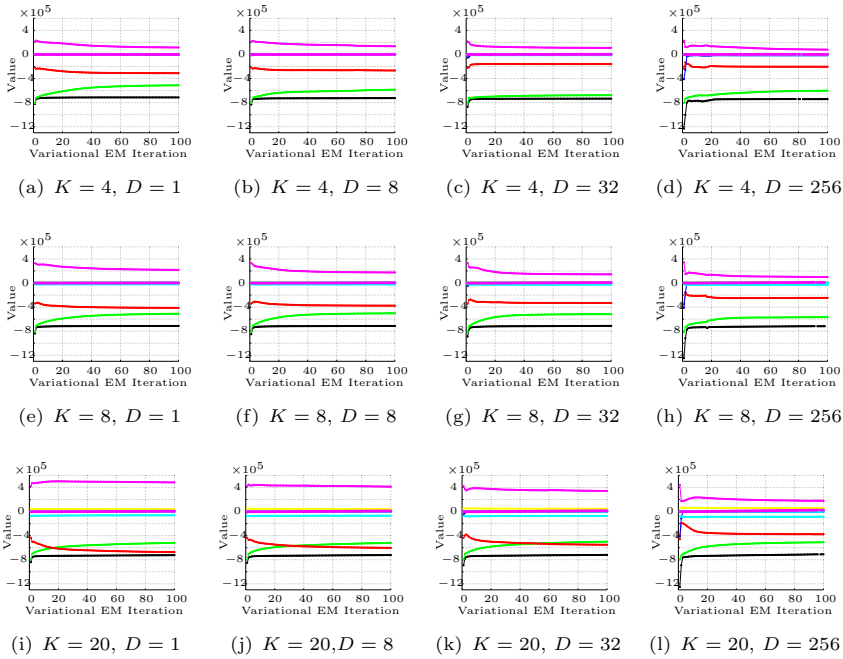


Fig. 5. The learned topics ($\hat{\beta}$) from different models with number of topics $K = 16$ and $\tau = 0.8$

with all different number of topics when the noise level is high ($\tau = 0.9$). However, the lower-right plots in Figure 3 show different levels of drop in performance where the number of topics are large and the data is less noisy.

To further understand the phenomenon in Figure 3, we visualize the topics that are learned with different models. We present two typical cases using the noise level $\tau = 0.8$ to analyze the reason for the performance change. Figure 4 shows the topics learned with different models when $K = 4$, which corresponds to Figure 3 (e). As expected, LDA/LDAC only learn the topics to represent noise, since noise is dominant in the data. Topics learned using SLDA are almost the same as the topics that is learned with LDA, which shows the way to model the class label in SLDA is not effective to supervise the model to learn a better representation for classification. This also explains that SLDA has similar performance as LDAC, since the learned latent structures are similar. By mapping the class label to D dimension using P-SLDA, we can observe that the learned topics start to differ from LDA. As shown in Figure 4(c) and (d), the larger the D is, the more impact the supervision has on the model. Since the topics are used to explain both signal and noise in P-SLDA and there are limited topics, the learned topics become mixed with signal and noise even with boosted supervision. However, P-SLDA is still able to catch the signal compared to LDA, hence, the performance is improved in this case.

Figure 5 shows the topics learned with different models when $K = 16$, which correspond to Figure 3(g). Topics learned with SLDA and LDA appear to be the same as in the previous case. However, both signal and noise are captured when the topics space is large through all the models, which explains the good performance by both SLDA and LDAC in Figure 3 (g). By boosting the supervision, the learned topics start to change. However, the changes are minor compared to the previous case when $K = 4$. When we boost the supervision in an extreme



(a) $K = 4,\ D = 1$ (b) $K = 4,\ D = 8$ (c) $K = 4,\ D = 32$ (d) $K = 4,\ D = 256$

(e) $K = 8,\ D = 1$ (f) $K = 8,\ D = 8$ (g) $K = 8,\ D = 32$ (h) $K = 8,\ D = 256$

(i) $K = 20,\ D = 1$ (j) $K = 20, D = 8$ (k) $K = 20,\ D = 32$ (l) $K = 20,\ D = 256$

Fig. 6. Likelihood Analysis for supervision effectiveness. Legend: $-*-\ \mathcal{L}_{P-SLDA}$; $-*-\ \mathbb{E}_q[\log p(w|z,\beta)]$; $-*-\mathbb{E}_q[\log p(z|\theta)]$;$-*-\ \mathbb{E}_q[\log p(\theta|\alpha)]$; $-*-\ \mathbb{E}_q[\log p(y|z,\mu)]$; $-o-\ \mathbb{E}_q[\log p(\beta|\eta)]$; $-*-\ -\mathbb{E}_q[\log q(\theta)]$;$-*-\ -\mathbb{E}_q[\log q(z)]$; $-o-\ \mathbb{E}_q[\log q(\beta)]$

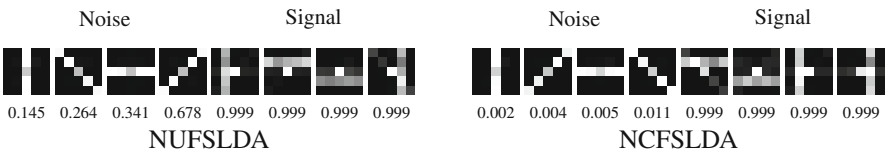


Fig. 7. Learned topics by NUF-SLDA and NCF-SLDA with $\tau = 0.8$. $\mathbb{E}_q[B]$ is marked below each topic which indicates the confidence for the topic to be signal

case, $D = 10N$, the learned topics start to break into small fragments, which is due to over-fitting and causes the performance drop.

To further study the supervision effectiveness, we observe the likelihood behaviour during the learning process. Recall the Evidence Lower Bound (ELBO) of SLDA is

$$\begin{aligned}\mathcal{L}_{SLDA} =& \mathop{\mathbb{E}}_{q}[\log p(w|z,\beta)] + \mathop{\mathbb{E}}_{q}[\log p(z|\theta)] + \mathop{\mathbb{E}}_{q}[\log p(\theta|\alpha)] + \mathop{\mathbb{E}}_{q}[\log p(\beta|\eta)] \\ &+ \mathop{\mathbb{E}}_{q}[\log p(y|z,\mu)] - \mathop{\mathbb{E}}_{q}[\log q(\beta)] - \mathop{\mathbb{E}}_{q}[\log q(z)] - \mathop{\mathbb{E}}_{q}[\log q(\theta)].\end{aligned} \tag{11}$$

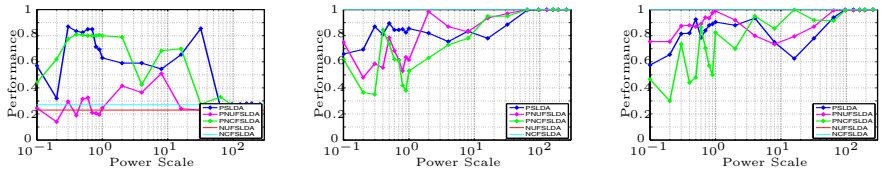
The same form holds for P-SLDA. The difference lies in the dimension of y. In SLDA, y is 1-dimensional, whereas in P-SLDA, y is D-dimensional. The value of each item over each iteration is plotted in Figure 6 for different cases. It is clear that $\mathbb{E}_q[\log p(z|\theta)]$ becomes higher and $\mathbb{E}_q[\log q(z)]$ becomes lower while

the supervision is boosted. $\mathbb{E}_q[\log p(w|z,\beta)]$ drops slightly as well with boosted supervision. The topic assignments z are used to explain both the words and the label in a document. By boosting the supervision, the label will get better explained by the topic assignments, however, the cost is that the words get less well explained by the topic assignments, which is confirmed by the drop of $\mathbb{E}_q[\log p(w|z,\beta)]$. The drop of the entropy term $\mathbb{E}_q[\log q(z)]$ shows that the topic assignment distribution becomes more sparse with boosted supervision. This is caused by the fact that different topics are used to explain different classes and topics become less shared among different classes. This shows the tradeoff between the use of latent space to explain the words and the label, when data are noisy.

3.2 Factorized Models

In this part, we evaluate factorized topic models NUF-SLDA, P-NUF-SLDA, NCF- SLDA and P-NCF-SLDA. The same experimental setting is used as in the previous section. Due to space limitation, we only show results with $\tau = 0.8$ in this part. Firstly, whether the factorized model is able to learn the correct factorization is evaluated. Figure 7 shows the learned topics with $K = 8$ using NUF-SLDA and NCF-LDA. Both models are able to correctly factorize the topics.

The performances of these factorized models compared with P-SLDA are shown in Figure 8. $e = 0.2$ is used through all these experiments. We can see that P-NUF-SLDA is more robust compared to P-SLDA and P-NCF-SLDA, when there is sufficient amount of topics. P-SLDA could achieve better performance when the number of topics is small.Because factorized model separate the topic space to signal part and noise part. When the number of topics are not sufficient, factorizing the model makes the number of topics describing the signal even smaller.



(a) $K = 4$ (b) $K = 8$ (c) $K = 16$

Fig. 8. Performance comparison of factorized topic models for synthtic dataset with $\tau = 0.8$

3.3 Video Action Classification

We use threes action classes: boxing, hand clapping and hand waving from KTH action dataset [11,13] for the action classification experiment. Intuitively, only the human movement is the signal and all the variances in the background and video shooting settings are noise. Hence, this is a real world dataset where signal-noise ratio is low. There are around 100 video clips for each action of which 80%

are randomly selected for training and the remaining 20% are for testing. Bag-of-STIP [13] is use to represente visual words in each video clip which is treated as a document. $\alpha = 0.1$, $\eta = 0.1$ are used through all the experiments using this dataset.

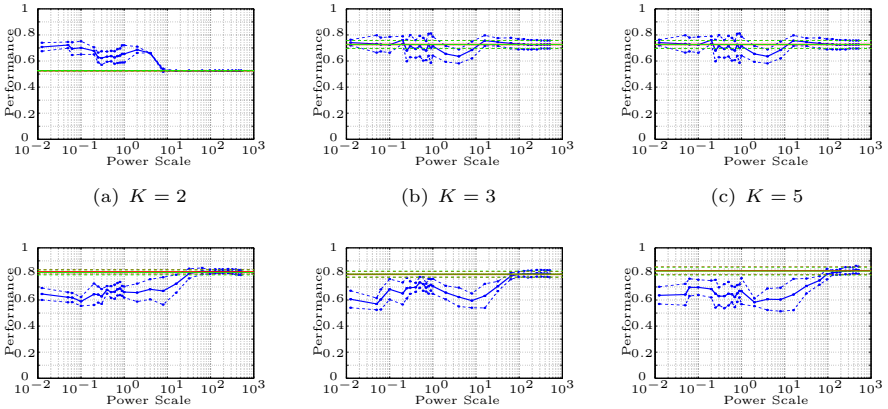
Supervision Effectiveness. Figure 9 shows the performance of LDAC, SLDA and P-SLDA with different power scale s, which are consistent with the one using synthetic dataset. When the number of topics is small, boosting the supervision can improve the performance significantly as in Figure 9 (a), (b), (c). When the number of topics is more than sufficient, boosting the supervision may disturb the classification performance. The result is not only interesting for understanding the supervised effectiveness, it is also significant from a application perspective. The number of topics is essential for computational complexity in the inference. By boosting the supervision, using a small number of topics will be able to achieve similar level of performance as using a large number of topics, but the computation time will be significantly reduced.

Factorized Models. Performances of factorized models are evaluated and compared in Figure 10. This dataset has low signal-noise ratio, hence $e = 0.3$ is used through all the experiments in this part. We can see that P-NUF-SLDA is more robust to the change of supervision level and it has potential to overperform P-SLDA as in Figure 10 (c) (d). However, since it uses less topics to models the signal. The performance maybe worse when the number of topics is not enough to factorize. NCF-SLDA is in general not as robust with boosted supervision, we believe that it is caused by that the structured noise assumption is too strong and the boosting effect is doubled in P-NCF-SLDA with the additional noise label.

3.4 Natural Scene Classification

Four classes of natural scene images are used in this experiment as [8,26]. Intuitively, all the information from natural scene is useful to judge the scene category. Hence, this is a real world dataset where the signal-noise ratio is high. There are more than 300 images per class of which 80% of the data are randomly selected for training and the remaining 20% for testing. Bag-of-SIFT [14] is used to represent visual words in each image which is treated as a document. $\alpha = 0.1$, $\eta = 0.1$ are used through all the experiments in this section.

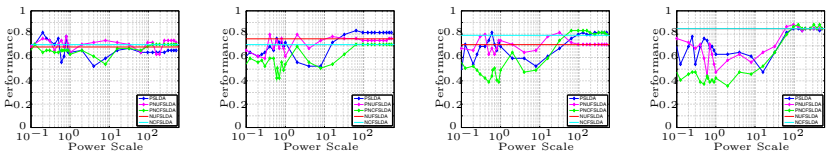
Supervision Effectiveness. Figure 11 shows the performance of LDAC, SLDA and P-SLDA with different power scale s for natural scene classification. LDAC and SLDA have the same performance with different number of topics as previous experiments. Differently from the previous experiments, the classification performance does not change as much by boosting the supervision. The performance gets worse when the supervision is boosted too much, which is caused by overfitting. This shows that when the signal-noise ratio is high, the optimum for both unsupervised model and supervised model are similar, since the label is



(a) $K = 2$ (b) $K = 3$ (c) $K = 5$

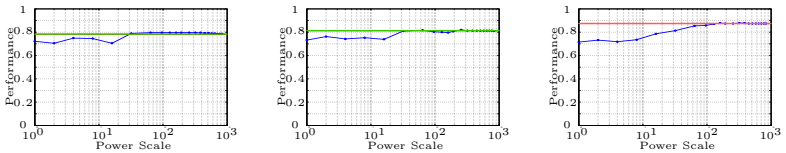
(d) $K = 10$ (e) $K = 20$ (f) $K = 40$

Fig. 9. Action classification performance. The x axis is the power scale s and the y axis shows the classification performance. All the experiments are repeated 8 times with different random seeds for initialization. The mean, solid line, and the standard deviation, dashed line, are shown in the plot. P-SLDA with different power scale is plotted with the blue curve. LDAC is marked with the green line and SLDA is marked with the red line.



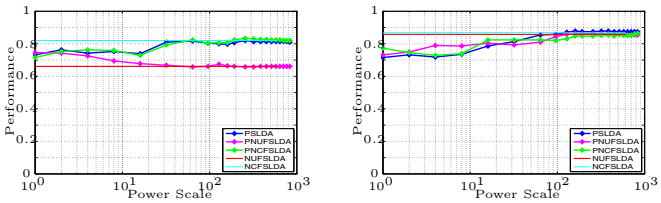
(a) $K = 5$ (b) $K = 10$ (c) $K = 20$ (d) $K = 30$

Fig. 10. Performance comparison of factorized topic models



(a) $K = 5$ (b) $K = 10$ (c) $K = 30$

Fig. 11. Scene classification performance. P-SLDA with different power scale is plotted with the blue curve. LDAC is marked with the green line and SLDA is marked with the red line.



(a) $K = 10$ (b) $K = 30$

Fig. 12. Performance comparison of factorized topic models for natural scene classification. The x axis is the power scale s (ranging from 10^0 to 10^3) and the y axis shows the performance (ranging from 0 to 1).

consistent with the words. A little improvement can still be observed by boosting the supervision as in Figure 11 (a), since the data is not ideal.

Factorized Models. Figure 12 shows the performance of factorized models on the natural scene classification task. Consistent with the previous experiments, the performance of P-NUF-SLDA is more robust with respect to the boosting of the supervision and shows better performance when the label is mapped to high dimension compared to P-SLDA. P-NUF-SLDA does not perform as good as P-SLDA when the number of topics is small. Because with factorisation, only around half of the topics are used to model the signal which is not enough when the total number of the topics is small. P-NCF-SLDA perform on par with P-SLDA. All the models are more robust with this dataset since the data has high signal-noise ratio.

3.5 Discussion

To sum up the experiments with three different datasets, we will present a discussion in three points. Firstly, all the experiments above show that SLDA and LDAC have similar performance through all different settings. Further analysis with the synthetic data shows that the topics learned by SLDA and LDA are similar. This confirms that supervision on LDA using SLDA is not effective on learning of topics. Secondly, P-SLDA is able to boost the supervision, which makes the supervision affect the learning of topics. Experiments on different settings show that whether boosting the supervision can be beneficial is highly dependent on the data and the parameter setting. When the data is noisy (low signal-noise ratio), as in the first two experiments, boosting supervision is able to increase the performance, especially when the number of topics is small. When the data is informative (high signal-noise ratio), boosting the supervision is not able to clearly affect the classification performance since the label and words information are consistent. Over boosting the supervision can harm the performance since it makes the model biased towards the label and causes overfitting. Thirdly, factorized models are able to recognize the signal topics and noise topics correctly, which improves the interpratation of the learned topics. They also have more robust performance with the boosting of the supervision.

4 Conclusions

In this paper, we have presented a thorough study on the behaviour of supervision on topic models, which fills the gap in the understanding of supervised topic models; and we have proposed two types of alternative factorized supervised topic models which improve the interpretation of topics and enhance the model performance. Variational inference has been used and fully derived for the proposed models. All the models have been evaluated with both synthetic data and real world data. We conclude in the study that: supervision is not effective using SLDA on the learning of the topics; balancing the model using P-SLDA can boost the supervision, which provide further improvements in case of noisy data; factorized models can increase the performance robustness.

We will continue our research in two directions. Firstly, we will analyze and compare a wider range of supervised topic models, such as DiscLDA [12] and MedLDA [27,28], to have a deeper insight on the behaviours of all different supervised topic models. Secondly, we will continue working on factorized topic models, since most models can not deal with noise sufficiently well. We will both apply the factorization on different supervised topic modeling framework and use more effective factorization prior.

References

1. Blei, D.M., Lafferty, J.: Correlated topic models. In: NIPS (2006)
2. Blei, D.M., McAuliffe, J.D.: Supervised topic models, arxiv:1003.0783 (2010)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
4. Browne, M.W.: The maximum-likelihood solution in inter-battery factor analysis. British Journal of Mathematical and Statistical Psychology **32**(1), 75–86 (2011)
5. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: ICCV (2007)
6. Chang, J., Boyd-Graber, J., Wang, C., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: NIPS (2009)
7. Damianou, A., Ek, C.H., Titsias, M.K., Lawrence, N.D.: Manifold relevance determination. In: ICML, pp. 145–152 (2012)
8. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: CVPR (2005)
9. Hoffman, M.D., Blei, D.M., Bach, F.: Online learning for latent Dirichlet allocation. In: NIPS (2010)
10. Hofmann, T.: Probabilistic latent semantic analysis. In: ACM SIGIR (1999)
11. Hospedales, T.M., Gong, S.G., Xiang, T.: Learning tags from unsegemented videos of multiple human actions (2011)
12. Lacoste-Julien, S., Sha, F., Jordan, M.I.: DiscLDA: discriminative learning for dimensionality reduction and classification. In: NIPS (2008)
13. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: MacLean, W.J. (ed.) SCVMA 2004. LNCS, vol. 3667, pp. 91–103. Springer, Heidelberg (2006)
14. Lowe, D.G.: Object recognition from local scale-invariant features. In: ICCV (1999)
15. Niu, Z., Hua, G., Gao, X., Tian, Q.: Semi-supervised relational topic model for weakly annotated image recognition in social media. In: CVPR (2014)

16. Rabinovich, M., Blei, D.M.: The inverse regression topic model. In: ICML (2014)
17. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Conference on Empirical Methods in Natural Language Processing (2009)
18. Rasiwasia, N., Vasconcelos, N.: Latent Dirichlet allocation models for image classification. PAMI **35**(11), 2665–2679 (2013)
19. Tang, J., Meng, Z., Nguyen, X., Mei, Q., Zhang, M.: Understanding the limiting factors of topic modeling via posteriror contraction analysis. In: ICML (2014)
20. Tucker, L.R.: An Inter-Battery Method of Factory Analysis. Psychometrika **23**, June 1958
21. Wang, C., Blei, D.: Collaborative topic modeling for recommending scientific articles. In: ACM SIGKDD (2011)
22. Wang, C., Blei, D.M., Fei-Fei, L.: Simultaneous image classification and annotation. In: CVPR (2009)
23. Wang, C., Paisley, J., Blei, D.: Online variational inference for the hierarchical Dirichlet process. In: AISTATS (2011)
24. Weinshall, D., Levi, G., Hanukaev, D.: Latent Dirichlet allocation topic model with soft assignment of descriptors to words. In: ICML (2013)
25. Zhang, C., Ek, C.H., Damianou, A., Kjellström, H.: Factorized topic models. In: International Conference on Learning Representations (2013)
26. Zhang, C., Ek, C.H., Gratal, X., Pokorny, F.T., Kjellström, H.: Supervised hierarchical Dirichlet processes with variational inference. In: ICCV Workshop on Inference for Probabilistic Graphical Models (2013)
27. Zhu, J., Ahmed, A., Xing, E.P.: Medlda: maximum margin supervised topic models for regression and classification. In: ICML (2009)
28. Zhu, J., Chen, N., Perkins, H., Zhang, B.: Gibbs max-margin supervised topic models with fast sampling algorithms. In: ICML (2013)

W14 - Light Fields for Computer Vision

Barcode Imaging Using a Light Field Camera

Xinqing Guo[1(✉)], Haiting Lin[1], Zhan Yu[1], and Scott McCloskey[2]

[1] University of Delaware, Newark, DE, USA
xinqing@udel.edu
[2] Honeywell ACS Labs, Minneapolis, MN, USA

Abstract. We present a method to capture sharp barcode images, using a microlens-based light field camera. Relative to standard barcode readers, which typically use fixed-focus cameras in order to reduce mechanical complexity and shutter lag, employing a light field camera significantly increases the scanner's depth of field. However, the increased computational complexity that comes with software-based focusing is a major limitation on these approaches. Whereas traditional light field rendering involves time-consuming steps intended to produce a focus stack in which all objects appear sharply-focused, a scanner only needs to produce an image of the barcode region that falls within the decoder's inherent robustness to defocus. With this in mind, we speed up image processing by segmenting the barcode region before refocus is applied. We then estimate the barcode's depth directly from the raw sensor image, using a lookup table characterizing a relationship between depth and the code's spatial frequency. Real image experiments with the Lytro camera illustrate that our system can produce a decodable image with a fraction of the computational complexity.

Keywords: Light field camera · Barcode imaging · Spatial frequency

1 Introduction

A barcode is an optical machine-readable representation of data relating to the object to which it is attached. Nowadays the ubiquitous barcodes found on product packaging significantly improve the speed and accuracy of computer data entry. A traditional 1D barcode scanner uses a line of photocells to detect the reflected light from the barcode. These linear imagers need to be well aligned with the barcode to produce accurate results and therefore the scanning process is not fully automatic. More recent 2D imagers address the automation issue by capturing the image of the entire barcode and then automatically orienting the image for decoding.

2D scanners are fundamentally low-cost cameras, and capture is limited by well-known tradeoffs between noise and blur: if the camera uses a small aperture to acquire the barcode image, the result will be corrupted by noise; if it uses a wide aperture, the result will be less noisy but the depth of field is reduced. Active illumination is used in 2D scanners using small apertures, but strict

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 519–532, 2015.
DOI: 10.1007/978-3-319-16181-5_40

price and power budgets typically limit this to low-power LEDs. When using wide apertures, conventionally a user would need to manually move the barcode towards or away from the scanner to ensure it is within the depth of field of the scanner. Alternatively, the scanner can conduct a focal sweep and select the proper focal slice to decode. However, implementing focal sweep requires adding moving parts to the scanner, which reduces robustness to mechanical shock. The overwhelming majority of purpose-built scanners are fixed focus for these reasons.

In this paper, we present a novel barcode scanning system by using the recent commercial light field camera. A light field camera such as Lytro and Raytrix uses a microlens array to capture multiple views of the scene in a single shot. The rich set of rays captured by the light field camera enables the user to conduct post-capture refocusing, *i.e.*, focal stack can be synthesized after the capture. This reduces the mechanical complexity of moving parts in exchange for increased computational complexity in the form of a refocusing algorithm.

The focal stack defines the extended depth of field of a light field camera. A straightforward way to utilize a light field camera for barcode scanning would be to simply apply barcode detection and decoding to images in the focal stack. However, synthesizing the complete focal stack requires applying computationally expensive light field rendering schemes. In order to reduce the time from capture to decoding, we present a much simpler scheme based on the frequency characteristics of barcodes. We speed up the process by first segmenting out the barcode region, which we detect from a sub-sampled version of the raw sensor image. Then, we directly estimate the depth of the barcode by analyzing the variance of pixel intensities in the lenslet images formed behind each microlens. Finally, we conduct refocusing only at the estimated depth.

Compared to 2D imagers, our system only involves two extra steps: depth estimation and barcode image rendering. With little computational cost, we gain a system with its range of depth of field nearly triples that of a conventional camera. Comprehensive experiments demonstrate our new light-field based barcode scanner system is fast, accurate and robust to barcode orientation, size variation, and lighting.

2 Related Work

Barcode Imaging. Recently, there has been an emerging interest on barcode reading using 2D imagers. Barcode reading consist of two distinct stages: localization and decoding. Tremendous efforts have been made to enhance the performance of both stages. Muniz *et.al.* [9] apply hough transform to the image to locate the barcode and find its optimal orientation for further decoding. Zhang *et.al.* [16] jointly analyze the texture and shape information to search for the barcode. Chai and Hock [1] improve the barcode localization by using morphological operator to identify parallel line patterns at block level. Gallo and Manduchi [5] employ a deformable template matching method and enforce global spatial coherence to correctly read barcodes in difficult situations. Xu and McCloskey

[14] describe a system for localizing and deblurring motion-blurred image using a flutter shutter camera. In contrast to their methods, our system features a better light efficiency and aims at reducing the defocus blur of the barcode image.

Light Field Photography and Depth Estimation. Integral or light field photography captures a rich set of rays to describe the visual appearance of the scene. A distinct advantage of light field photography is the ability to render an image after exposure with a desired focal plane. Modern light field rendering is introduced by Levoy and Hanrahan [8] and Gortler *et.al.* [7]. Early approaches [12] utilize camera arrays to capture a light field with high spatial resolution. However, the system tends to be bulky and impractical for daily use. Ng [10] designs a hand-held light field camera where a microlens array is placed on top of the sensor to optically sort the rays by direction onto the pixels underneath. In addition to its refocusing capability, light field is also applicable to depth estimation. Several methods [2,13] exploit the epipolar-plane image (EPI) to extract the disparity map. Others use correspondences [4] or combined with depth from defocus technique [11] to achieve similar result. In contrast to their methods for general scenes which are geometrically complex, our work focuses on barcode imaging and only extract the depth of barcode region based on its unique frequency characteristics, thus largely reducing the computational cost. Similarly, our rendering approach also prefer speed to quality. We utilize basic ray tracing for rendering a correct image, without using other image enhancement techniques such as [4] since they won't benefit barcode decoding. In this paper, we use Lytro camera to validate our algorithm, but note that our methods apply to most microlens-based light field camera.

3 Frequency Characteristics

Conventional barcodes are composed of high contrast black and white bars or patches, which facilitate the localization process. Several approaches have been proposed and optimized to take advantage of the texture information for localization. However, the imaging mechanism of light field camera will distort and deteriorate these features, making existing approaches less effective, even unusable. The structure of light field camera is similar to a conventional camera, except that it adds a microlens array in front of the sensor to further diverge the rays based on their directions. Thus, the resultant raw light field image consists of hundreds of thousands of lenslet images, as shown in Fig. 1. Directly locating the barcode on the raw light field image would be extremely challenging: each lenslet image contains at most 10×10 pixels; and the high contrast in the boundary region of each lenslet image will fail gradient based detection algorithms.

3.1 Barcode Localization

In order to address these issues, we aim to first localize the barcode on a sub-aperture image instead of the raw image. A sub-aperture image is a normal

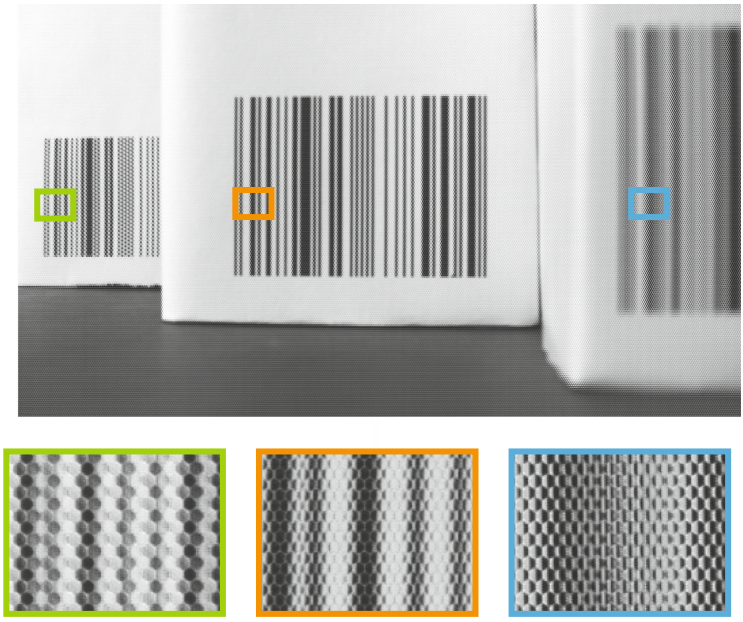


Fig. 1. Lenslet image pattern changes with the depth of the barcode

2D image composed by stitching together the same pixels underneath each microlens. It can be thought of as an image taken by a virtual camera with its center of projection in front of the main lens. In our case, we pre-calibrate the center of each lenslet image and pick the center pixels to generate a central sub-aperture image. Interpolation is required since the lenslet arrangement is hexagonal.

Although the sub-aperture image is of low resolution (about 328×378 for Lytro) which inhibits direct decoding, it is detailed enough for barcode localization. We extend the method proposed in [5] by incorporating the barcode orientation into the feature computation, and analyse the shape of the region with high average feature responses for robust localization. For each angle $\theta \in \{-90, -85, ..., 90\}$, feature response $I_e^\theta(p) = |I_{x_\theta}(p)| - |I_{y_\theta}(p)|$ is evaluated at each pixel p, where $I_{x_\theta}(p)$ and $I_{y_\theta}(p)$ are the image gradient along orthogonal directions $x_\theta(\cos\theta, \sin\theta)$ and $y_\theta(-\sin\theta, \cos\theta)$ respectively. A box filter is applied to I_e^θ to get locally averaged feature response $\bar{I}_e^\theta$. The potential barcode region is identified by a connected region of constantly high average response $\bar{I}_e^{\theta^*}$ with θ^* maximizing the mean of $\bar{I}_e^\theta(p)$'s within the region. The shape of this region is also required to be tightly bounded by an oriented rectangle. Within this rectangle, we compute the size of the candidate barcode as the distance between the first and the last black bars. In order to eliminate the effects of illumination variations, the input sub-aperture image is preprocessed using local histogram equalization. Fig. 2 shows an example of our barcode localization algorithm.

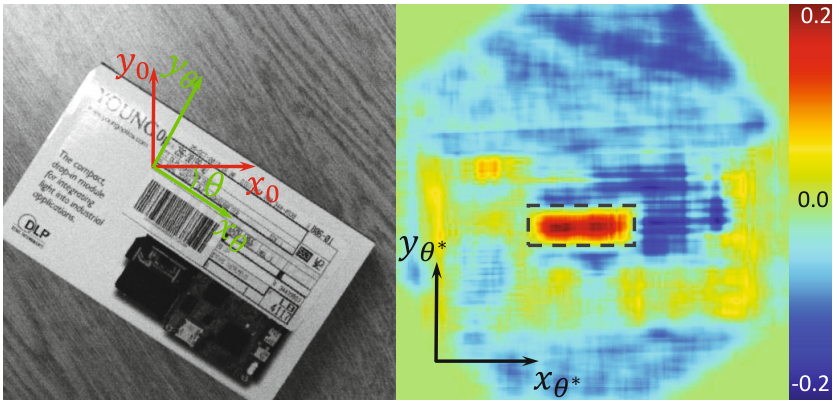


Fig. 2. A barcode localization example. An optimal rotation angle θ^* is determined maximizing the mean feature response of the potential barcode region.

Note that our localization method is designed for 1D barcode. We refer the reader to [14] and other related work for 2D barcode localization. After we locate the barcode in the sub-aperture image, we can continue to crop the corresponding barcode region in the raw light field image and only process this region to speed up our following ray tracing algorithm.

3.2 Spatial Frequency *vs.* Depth

We first study the correlation between the spatial frequency of the raw barcode region and its depth. Here we assume that the barcode is approximately frontal parallel to the camera so we only consider one depth value. As shown in Fig. 1, barcodes positioned at different depth exhibits different lenslet image patterns. In the first inset, each lenslet image shows uniform color, indicating the image plane of the main lens coincides with the plane of the microlens array. As the barcode moves nearer to the camera, increasing intensity variations are evident in lenslet images. Therefore, our intuition is to use this statistical characteristics of barcode for depth estimation.

To better illustrate our algorithm, we simplify the barcode as evenly distributed black and white bars. The spatial frequency of the barcode is defined as the number of line pairs per unit length. Fig. 3 shows two cases of formation of lenslet images. In the first case, the image plane of main lens falls in front of the microlens array, where each lenslet image is a real image. On the contrary, when the image plane is behind the microlens array, a virtual image will be observed. Given the spatial frequency of the barcode X_1, we apply thin lens equation to compute the spatial frequency at the image plane of the main lens X_2 as:

$$X_2 = \frac{a}{b} \cdot X_1 = \frac{a - F}{F} \cdot X_1 \tag{1}$$

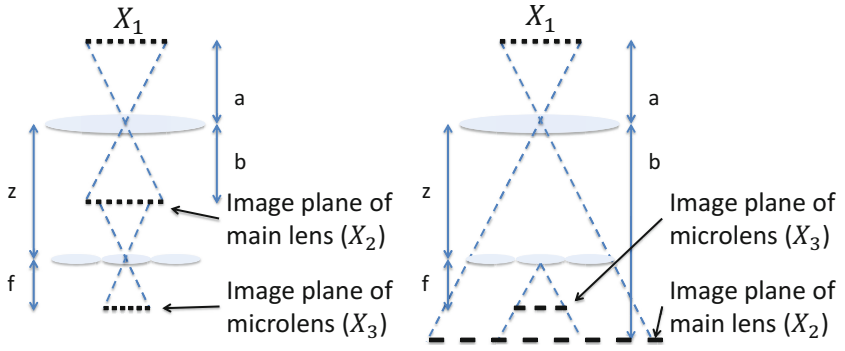


Fig. 3. Spatial frequencies of the barcode image at different image planes

where a is the object distance and F is the focal length of the main lens. We repeat this process to obtain the spatial frequency of the barcode image at the sensor plane X_3 as:

$$X_3 = \frac{z-b}{f} \cdot X_2 = \frac{a(z-F)-zF}{Ff} \cdot X_1 \tag{2}$$

when the main lens image plane is in front of the microlens and

$$X_3 = \frac{z-b}{f} \cdot X_2 = \frac{a(F-z)+zF}{Ff} \cdot X_1 \tag{3}$$

when the image plane is behind the microlens. Here z represents the distance between the main lens and the microlens, b is the image distance and f is the focal length of the microlens. In both cases a linear relationship between the barcode's spatial frequency at the sensor and its depth can be observed.

3.3 Variance *vs.* Depth

Although we can mathematically compute the sensor plane's spatial frequency X_3, it is very challenging to robustly measure this frequency since each lenslet image is only of size 10×10 pixels–*i.e.* a very small portion of the barcode, with its boundary region corrupted by vignetting. In our experiments, we observe at most two color transitions inside each lenslet image. Therefore, we instead use variance to represent the spatial frequency of each lenslet image. Specifically, we define a window around each lenslet center and measure the variance of pixel intensities within the window. Our intuition is that the higher the spatial frequency, the larger the chance to observe intensity transitions inside the window. We then compute the overall variance as the spatial frequency measurement by averaging the variances from the lenslet images inside the barcode region.

To formulate the correlation between variance and depth, we make following assumptions based on observation that at most two intensity transitions appear

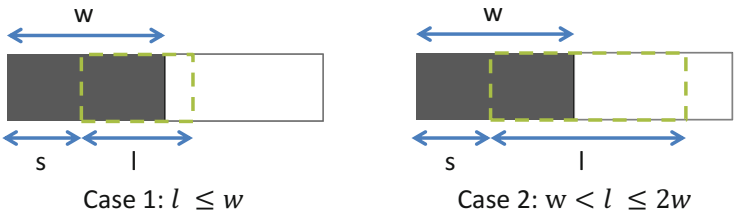


Fig. 4. Lenslet images function as a sliding window across the barcode region

within each lenslet image. Next, we regard the light field camera as a relay imaging system, which consist of mainlens and microlenses as pinhole cameras. We first analyze the image captured by the microlens, then extend our analysis to the whole system.

First we want to define variance σ^2. Suppose our target is evenly distributed black/white bars. Our pinhole camera has N pixels and the captured image contains m white pixels and n black pixels. And we further denote the intensity of the white pixel as 1 and that of black pixel as 0. Then we can get

$$\mu = \frac{m}{m+n} \tag{4}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \frac{mn}{(m+n)^2} \tag{5}$$

where μ is the mean value of the image and x_i is the pixel value.

Next we only consider the lenslet image. As each lenslet image only observes a very small portion of the barcode, its variance changes with its relative positions with the bar. As shown in Fig. 4, we denote the bar width of the image as w, the sensor size at the barcode image plane as l and the distance between the starting point of the lenslet image and a intensity transition as s. Then we continue our analysis in two cases:1) If $l \leq w$, then

$$\sigma^2 = \begin{cases} 0, & s \leq w - l \\ \frac{-s^2+(2w-l)s+lw-w^2}{l^2}, & w - l < s \leq w \end{cases} \tag{6}$$

We only compute the variance σ^2 as a function of s ranging from 0 to w because it is a periodic function. Since the lenslets are hexagonally arranged, their images can be considered as a sliding windows across the entire barcode image. From the distribution of σ^2, we can get the average variance $\bar{\sigma}^2$ as:

$$\bar{\sigma}^2 = \frac{\int_0^w \sigma^2 ds}{w} = \frac{1}{w}\left(\int_0^{w-l} \sigma^2 ds + \int_{w-l}^{w} \sigma^2 ds\right) = \frac{l}{6w} \tag{7}$$

It is evident that average variance $\bar{\sigma}^2$ is linearly relates to l. We can further map l through the mainlens to the real barcode as L. By using similar triangles, we

have $L = \frac{al}{b} = \frac{A}{Ff}[(z-F)a - zF] or \frac{A}{Ff}[(F-z)a + zF]$ and $l = \frac{A(z-\frac{aF}{a-F})}{f}$, where A is the size of the sensor and a, b, F, f, z are defined in last section. Therefore, each lenslet image covers an area of l on the barcode image through mainlens, and an area of L on the real barcode. Because l increases monotonically with the increase of a, we can obtain a one-on-one mapping between the depth a and average variance $\bar{\sigma}^2$.

2) if $w < l \leq 2w$, we have

$$\sigma^2 = \begin{cases} \frac{-s^2+(2w-l)s+lw-w^2}{l^2}, & 0 < s \leq 2w-l \\ \frac{lw-w^2}{l^2}, & 2w-l < s \leq w \end{cases} \tag{8}$$

Similarly, we compute its average variance $\bar{\sigma}^2$ as:

$$\bar{\sigma}^2 = \frac{\int_0^w \sigma^2 ds}{w} = \frac{1}{w}(\int_0^{2w-l} \sigma^2 ds + \int_{2w-l}^{w} \sigma^2 ds) = \frac{w^2}{3}l^{-2} - \frac{1}{6w}l - wl^{-1} + 1 \tag{9}$$

To prove $\bar{\sigma}^2$ monotonically increases with l, we compute its first and second order derivative as $(\bar{\sigma}^2)' = -\frac{2w^2}{3}l^{-3} - \frac{1}{6w} + wl^{-2}$ and $(\bar{\sigma}^2)'' = 2w^2 - 3wl$. Since $w < l \leq 2w$, $(\bar{\sigma}^2)'' < 0$. We further examine the value of $(\bar{\sigma}^2)'$ at $l = w$ and $l = 2w$, they are both larger than 0. Therefore, we can prove that $(\bar{\sigma}^2)' > 0$, so $\bar{\sigma}^2$ monotonically increases with l. Similar to the first case, we can also obtain a one-to-one mapping between the depth and average variance.

4 Efficient Refocusing

Our analysis above reveals that we can quickly use the variance to determine the depth of the barcode. This allows us to conduct refocusing with high efficiency.

4.1 Barcode Depth Estimation

To validate our use of variance as a depth cue, we measure the average variance of several randomly selected UPC-A barcodes over a range of distances from the camera. Fig. 5(a) shows the average results using different window sizes for variance computation. Clearly we can see valley shaped curves with two approximately linear regions. The bottom of the curve indicates the main lens image plane falls on the microlens, so the lenslet image gets uniform intensity which results in a minimum overall variance. Here one variance value may correspond to two different depths. To resolve this two-fold ambiguity, we only use the left linear region in our experiments, as barcodes of practical sizes at depths in the right linear side are resolution limited even when properly focused. If necessary, the right linear side can be used similarly to estimate another depth in the case that the depth from the left side leads to a undecodable result. Note that due to defocus blur and resolution limitation [6] in the lenslet image, the curve fluctuates in both ends, making these regions unusable. For robustness reasons, we estimate three depth values independently based on different window sizes 3×3, 5×5 and 7×7, and compute the mean of the corresponding depths as the final estimation.

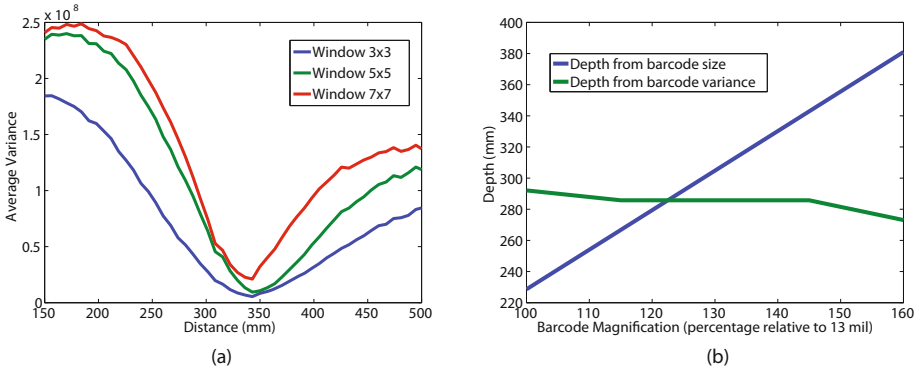


Fig. 5. (a) The average variances of the barcode image using different window sizes *vs.* its depth. (b) The depth of the barcode region is determined jointly by the variance and the size of the detected barcode region.

The variance *vs.* depth curve in Fig. 5(a) is for standard 13 mil barcodes. Scaling the size of the overall barcode will change the underlying spatial frequency X_1, and change the relationship between depth and variance. This is inevitable since product manufacturers tend to adjust the size of the barcode to suit the package. Our solution is first to build a look-up table indexed by variances per barcode size. Then we jointly determine the final depth based on both the variance and the size of the detected barcode region in the central sub-aperture image. From projective geometry, we obtain the relationship between the barcode image size s and the depth d as $s \propto S/d$, where S is the original size of the barcode. Fig. 5(b) illustrates our depth determination strategy. Given a detected barcode size, the larger the barcode's original size, the further its distance. Given a measured variance, another size *vs.* depth curve is formed by collecting depths from the look-up tables for corresponding barcode sizes. The ground truth original barcode size and the depth are therefore indicated by the intersection of these two lines/curves.

4.2 Refocusing

The final step in our light field barcode imaging system renders a focused image of the barcode region, using the depth estimated from the variance and size of this region. We set out to perform ray tracing to generate the in focus barcode image. Ray tracing mimics the physical process of image formation. The intensity of a point on the target image plane (virtual plane) is computed by integrating all the rays of different directions passing through it.

As shown in Fig. 6(a), adopting two parallel plane parameterization (2PP) [8], a ray can be indexed by $(\mathbf{s}, \mathbf{u})$, where $\mathbf{s}$ and $\mathbf{u}$ are the 2D intersections with the target image plane $\Pi_{\mathbf{s}}$ and the microlens plane $\Pi_{\mathbf{u}}$ respectively. The formation process of the target image I' can be summarized as:

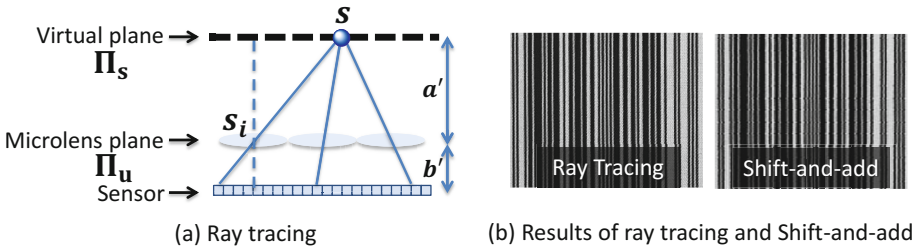


Fig. 6. (a)High quality barcode rendering by ray tracing. (b)Results from two implementations of refocusing algorithm.

$$I'(\mathbf{s}) = \int r(\mathbf{s}, \mathbf{u}) d\mathbf{u}, \tag{10}$$

where $r(\mathbf{s}, \mathbf{u})$ is the irradiance of the corresponding ray and is recorded by the sensor. As shown in the Fig. 6(a), the directions of the rays are discretized through the lenslets. Let $\mathbf{s}_i$ denote the location of the optical center of lenslet m_i, a' the distance from $\Pi_\mathbf{s}$ to $\Pi_\mathbf{u}$ and b' the distance from $\Pi_\mathbf{u}$ to the sensor plane, Eq. 10 can be rewritten discretely as:

$$I'(\mathbf{s}) = \sum_i I((\mathbf{s}_i - \mathbf{s})\frac{b'}{a'} + \mathbf{s}_i), \tag{11}$$

where I is the raw image on the sensor.

In our experiments, we first adopt the method proposed by [3] and use preloaded white images from Lytro camera to locate the lenslet centers $\mathbf{s}_i$ according to the camera's focal length setting. The target image plane is then determined based on the estimated depth and is discretized into pixels. Next we conduct ray tracing for each pixel $\mathbf{s}$ to gather the recorded irradiance of the rays and apply bilinear interpolation to achieve a better approximation of the pixel value. Note that there is a tradeoff between the resolution of the barcode image and its computational cost. The ray tracing technique provides the flexibility to vary the resolution by simply changing the sampling rate on the virtual plane. In our experiments, we render a barcode image of approximately 200×200 pixels to balance these two factors. Compared to the shift-and-add refocusing algorithm in [10], which requires rectified light field images (lenslet images arranged on grids), our method produces sharper rendering results as shown in Fig. 6(b). The blur artifacts in the shift-and-add result is due to the interpolation operation conducted when generating the rectified light field image from Lytro data. Generating images with even higher quality is still possible [13,15], but impractical due to its high computational cost.

5 Experiments

We use Lytro camera as our prototype light field camera. The raw images are preprocessed according to the metadata from Lytro's proprietary file format [3]

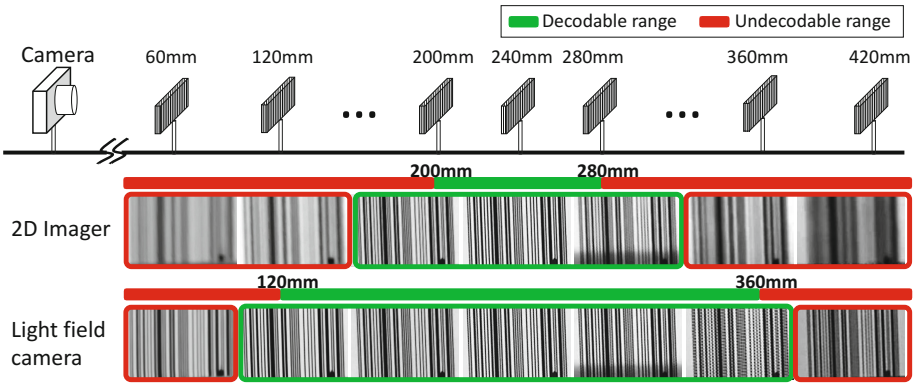


Fig. 7. Barcode images captured at variant depths using different devices. Light field camera largely extends the decodable range while keeping the noise level low.

and the vegnetting effects are removed using the pre-stored calibration images in Lytro camera. Demosaicing is then applied to get the final raw light field image. While capturing, we keep both the focal length and focal plane unchanged to simulate a light field camera without active parts.

Depth of Field. Our first experiment is to determine the amount of extended depth of field the light field camera has over a conventional camera. We collect a set of images of the barcode positioned at 60 mm to 420 mm from the camera with a incremental step of 6.9 mm. Using Lytro's desktop application, we generate two groups of images using the same focal length and aperture size: 1) one with focal plane coincides with the moving barcode and 2) the other one with a fixed focal plane simulating the conventional scanner. We test the decodability of the barcode images with a proprietary decoder. Results show that images from conventional camera is only decodable within a range of 80 mm due to the defocus blur. On the contrary, the images from light field camera features extended depth of field, with a decodable region of 240 mm, which nearly triples the range of conventional camera. Fig. 7 shows the comparison of the decodable range of 2D scanner and the light field camera, as well as the sharpness of their resultant images.

Depth Estimation and Image Rendering. Our subsequent experiments are to validate our barcode localization and depth estimation algorithm. We set our recognition target to be the standard 13 mil UPC-A barcode with 1.0x, 1.15x, 1.3x, 1.45x and 1.6x magnifications. Our variance *vs.* depth look up tables and size *vs.* depth curves are calibrated based on training data of random UPC-A codes. Barcodes with codes different from the training data are used for test. Fig. 8 shows the comparison between the estimated depths and the ground truth depths for barcodes of different sizes. The estimation errors are less than 50 mm which is within the decodable range. Fig. 9 shows our rendering results for barcodes on real products. Note that our algorithm is robust to different

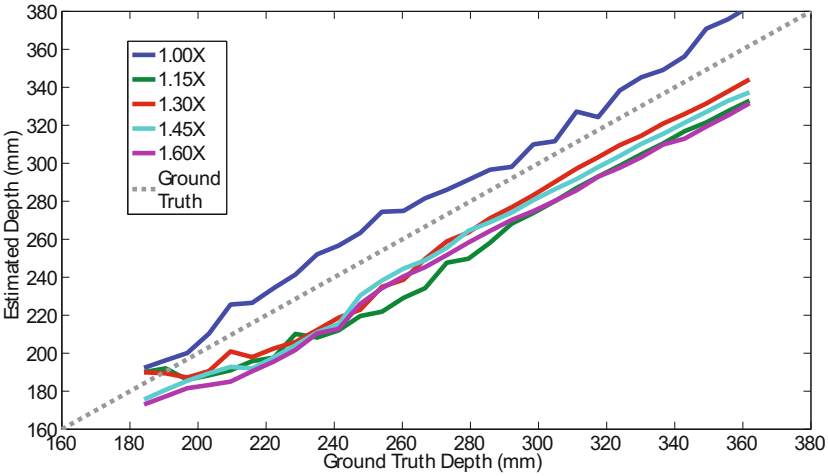


Fig. 8. Comparison between measured depths and the ground truth depths for barcodes of different sizes

Fig. 9. Rendering results of real barcodes using our scanning system. The full image on the left of each barcode example is the in focus image at the ground truth depth. Our rendering results are shown with orange boundary, while the ground truth are shown with green boundary for comparison.

sizes, orientations and nonuniform lighting conditions. Although we assume the barcode is approximately frontal parallel to the camera, our algorithm is tolerant

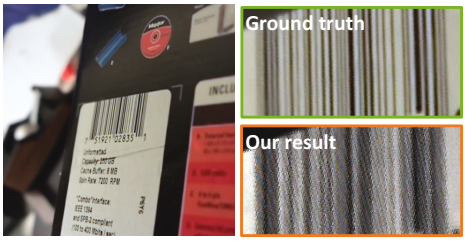


Fig. 10. An example where our algorithm fails

of slight projective distortion as shown in the last example in Fig. 9. However severe distortions result in failure cases as shown in Fig. 10. The main reason for this failure case is that our barcode localization algorithm detects a rectangle rather than a tight parallelogram only encloses the barcode. The non-barcode region inside our rectangle pollutes the variance estimation for depth estimation.

Running Time. We compare the processing speed/time of our system and a 2D scanner. A 2D scanner directly locates and decodes the barcode after exposure, while our system requires two extra steps: depth estimation and rendering of the barcode region. In our C++ implementation, the extra steps take around 0.2s for each light field image. Note that the result is not fully optimized. With application-specific integrated circuit (ASIC), as is implemented in most scanners, the overall processing time can be further reduced.

6 Conclusions and Future Work

In this paper, we present a novel, extended depth of field barcode scanning system based on a commercial light field camera. While a purpose-built light field scanner would likely use a smaller aperture than the Lytro camera, our emphasis has been on algorithmic improvements that would apply to such hardware. Our efficient, high quality barcode image rendering technique first segments the barcode and then estimates its depth in order to render only the necessary focal slice. The depth estimation is based on the spatial frequency and the size of barcode region, and is implemented by employing calibrated look up tables. Real barcode imaging experiments demonstrate the effectiveness of our scanning system. Depending on the size of the barcode in the image, and on the depth complexity of the scene, these improvements can dramatically reduce the amount of time needed to produce a decodable image. We will extend our system to 2D barcode scanning for our future work.

References

1. Chai, D., Hock, F.: Locating and decoding ean-13 barcodes from images captured by digital cameras. In: 2005 Fifth International Conference on Information, Communications and Signal Processing, pp. 1595–1599 (2005)
2. Dansereau, D., Bruton, L.: Gradient-based depth estimation from 4d light fields. In: Proceedings of the 2004 International Symposium on Circuits and Systems ISCAS 2004, vol. 3, pp. 549–552 (2004)
3. Dansereau, D., Pizarro, O., Williams, S.: Decoding, calibration and rectification for lenselet-based plenoptic cameras. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1027–1034 (2013)
4. Fiss, J., Curless, B., Szeliski, R.: Refocusing plenoptic images using depth-adaptive splatting. In: International Conference on Computational Photography (ICCP 2014). IEEE Computer Society (2014)
5. Gallo, O., Manduchi, R.: Reading 1d barcodes with mobile phones using deformable templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(9), 1834–1843 (2011)
6. Georgiev, T., Yu, Z., Lumsdaine, A., Goma, S.: Lytro camera technology: theory, algorithms, performance analysis. In: Proc. SPIE 8667 (2013)
7. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: SIGGRAPH 1996, pp. 43–54 (1996)
8. Levoy, M., Hanrahan, P.: Light field rendering. In: SIGGRAPH 1996, pp. 31–42 (1996)
9. Muniz, R., Junco, L., Otero, A.: A robust software barcode reader using the hough transform. In: Proceedings of the 1999 International Conference on Information Intelligence and Systems, pp. 313–319 (1999)
10. Ng, R., Levoy, M., Bredif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Stanford University Computer Science Tech. Report **2**, 1–11 (2005)
11. Tao, M.W., Hadap, S., Malik, J., Ramamoorthi, R.: Depth from combining defocus and correspondence using light-field cameras. In: Proceedings of the 2013 IEEE International Conference on Computer Vision, pp. 673–680 (2013)
12. Vaish, V., Wilburn, B., Joshi, N., Levoy, M.: Using plane + parallax for calibrating dense camera arrays. In: 2004 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2–9 (2004)
13. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(3), 606–619 (2014)
14. Xu, W., McCloskey, S.: 2d barcode localization and motion deblurring using a flutter shutter camera. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 159–165 (2011)
15. Yu, Z., Guo, X., Ling, H., Lumsdaine, A., Yu, J.: Line assisted light field triangulation and stereo matching. In: Proceedings of the 2013 IEEE International Conference on Computer Vision, pp. 2792–2799. ICCV 2013 (2013)
16. Zhang, C., Wang, J., Han, S., Yi, M., Zhang, Z.: Automatic real-time barcode localization in complex scenes. In: 2006 IEEE International Conference on Image Processing, pp. 497–500 (2006)

Depth Estimation for Glossy Surfaces with Light-Field Cameras

Michael W. Tao[1(✉)], Ting-Chun Wang[1], Jitendra Malik[1], and Ravi Ramamoorthi[2]

[1] University of California, Berkeley, USA
mtao@berkeley.edu
[2] University of California, San Diego, USA

Abstract. Light-field cameras have now become available in both consumer and industrial applications, and recent papers have demonstrated practical algorithms for depth recovery from a passive single-shot capture. However, current light-field depth estimation methods are designed for Lambertian objects and fail or degrade for glossy or specular surfaces. Because light-field cameras have an array of micro-lenses, the captured data allows modification of both focus and perspective viewpoints. In this paper, we develop an iterative approach to use the benefits of light-field data to estimate and remove the specular component, improving the depth estimation. The approach enables light-field data depth estimation to support both specular and diffuse scenes. We present a physically-based method that estimates one or multiple light source colors. We show our method outperforms current state-of-the-art diffuse and specular separation and depth estimation algorithms in multiple real world scenarios.

1 Introduction

Light-fields [1,2] can be used to refocus images [3]. Cameras that can capture such data are readily available in both consumer (e.g. Lytro) and industrial (e.g. RayTrix) markets. Because of its micro-lens array, a light-field camera enables effective passive and general depth estimation [4,5]. This makes light-field cameras point-and-capture devices to recover shape. However, current depth estimation algorithms support only Lambertian surfaces, making them ineffective for glossy surfaces, which have both specular and diffuse reflections. In this paper, we present the first light-field camera depth estimation algorithm for *both diffuse and specular* surfaces using the consumer Lytro camera (Fig. 1).

We build on the dichromatic model introduced by Shafer [6]. Diffuse and specular reflections behave differently in different viewpoints (Fig. 2). As shown in Eqn. 2, the surface color contributes to the diffuse reflectance component, while only light source color contributes to the specular component. Both diffuse and specular color remain fixed for all views; only specular intensity changes.

We present a novel algorithm that uses light-field data to exploit the dichromatic model to estimate depth of scenes involving glossy objects, with *both* diffuse and specular reflections and *one or multiple* light sources. We use the full

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 533–547, 2015.
DOI: 10.1007/978-3-319-16181-5_41

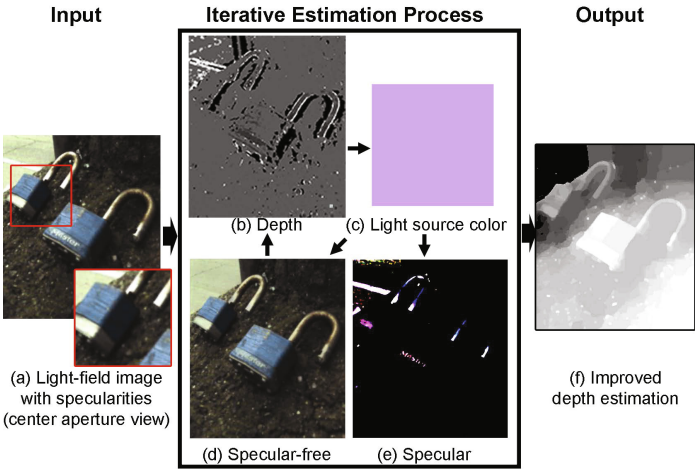


Fig. 1. Iterative Depth Estimation for Glossy Surfaces. Our input is a light-field image with both specular and diffuse reflections. Here we have an outdoor scene with glossy metallic locks in the foreground and road reflectors in the background (a). In our method, we iteratively exploit the light-field data to estimate depth (b); estimate the light source color (c); and generate the specular-free image (d) and generate the remaining components (e). Note: throughout this paper, we increased the contrast for the specular component for readability. We show that this approach improves depth estimation from (b) to our final depth estimation output (f). Darker represents farther and lighter represents closer in depth maps.

extent of the light-field data by shearing the 4D epipolar image to refocus and extract multiple viewpoints. In Fig. 3, we show that the rearrangement allows the diffuse and specular dichromatic analysis. Because no additional correspondence is needed, the analysis robustly estimates the light source color, extracting the specular-free image and estimating depth.

The algorithm uses three core routines iteratively: *first*, we exploit the 4D epipolar image (EPI) extracted from the light-field data to generate the specular-free image and estimate depth [4]; *second*, to estimate the light source color, we exploit the refocusing ability of the light-field data to extract multiple viewpoints for color variance analysis as shown in Fig. 2; and finally, *third*, to extract the specular-free image, we exploit the complete light-field angular information to improve robustness, giving consistent high quality results in synthetic, controlled, and natural real-world scenes.

We show that our algorithm works robustly across many different light-field images captured using the Lytro light-field camera, with both diffuse and specular reflections. We compare our specular and diffuse separation against Mallick et al. [7] and Yoon et al. [8], and our depth estimation against Tao et al. [4]. Our main contributions are

1. Light-field depth estimation with glossy surfaces. This will be the first published light-field depth estimation algorithm that supports both diffuse and

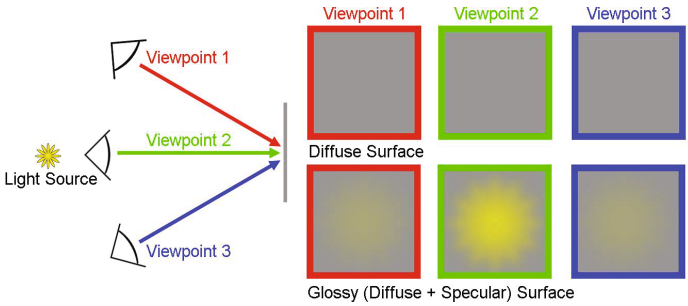


Fig. 2. Diffuse vs. Glossy Surfaces. This simple three view example shows that a diffuse surface will have minimal color changes. In a glossy surface, we can see color intensity changes that are correlated to the light source position and color. We use this property in this paper to estimate the light source color.

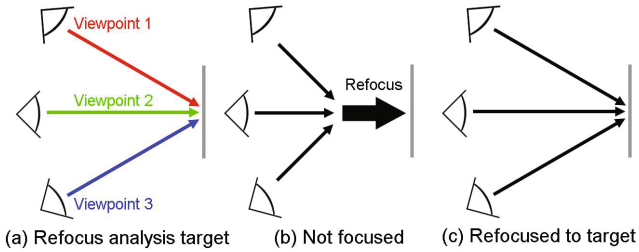


Fig. 3. The Light-Field Advantage. To perform the analysis as shown in Fig. 2, we first refocus to the plane of interest and then extract multiple views. Both processes are made possible by rearranging the light-field data. Because no additional correspondence is needed, the analysis robustly estimates the light source color, improving diffuse-specular separation and depth estimation.

glossy surfaces. Upon publication of this work, image datasets and code will be released.

2. 4D EPI light source color estimation. We perform the multiple viewpoint light source analysis by using and rearranging the light-field's full 4D EPI to refocus and extract multiple-viewpoints. Because of the light-field data's small baseline, shearing the light-field EPI gives us the refocusing ability. The framework distinguishes itself from the traditional approach of specular and diffuse estimation for conventional images by providing better results and supporting multiple light source colors.

3. Specular-free image. We use the light source color estimation to create a specular-free image by using the full 4D EPI for robustness (Algorithm 1).

4. Iterative depth estimation. We develop an iterative framework that uses the specular-free image to improve depth estimation.

2 Previous Work

Estimating depth and separating diffuse-specular components have been studied extensively. In our work, by using the full light-field data, we show that the two can work hand-in-hand to improve each others' results.

Defocus and correspondence depth estimation. Depth estimation has been studied extensively through multiple methods. Depth from defocus requires multiple exposures [9,10]; stereo correspondence finds matching patches from one viewpoint to another viewpoint(s) [11–13]. The methods are designed for Lambertian objects and fail or degrade for glossy or specular surfaces, and also do not take advantage of the full 4D light-field data.

Light-field depth estimation. More recent works have exploited the light-field data by using the epipolar images [4,5,14]. Because all these methods assume Lambertian surfaces, glossy or specular surfaces pose a large problem. In our work, we use the full 4D light-field data to perform specular and diffuse separation and depth estimation. The iterative approach directly addresses the problems at specular regions. In our comparisons, we show that specularities cause instabilities in the confidence maps computed in Tao et al. [4]. Specular regions retain incorrect depth values with high confidence, often causing the regularization step by Markov Random Fields (MRF) to fail or produce incorrect depth in most places, even when specularities affect only a part of the image (Figs. 7 and 8).

Multi-view stereo with specularity. Exploiting the dichromatic surface properties in Fig. 2 has also been studied through multi-view stereo. Lin et al. [15] propose a histogram based color analysis of surfaces. However, to achieve a similar surface analysis as Fig. 2, accurate correspondence and segmentation of specular reflections are needed. Noise and large specular reflections cause inaccurate depth estimations. Jin et al. [16] propose a method using a radiance tensor field approach to avoid such correspondence problems, but, as discussed in the paper, real world scenes do not follow their tensor rank model. In our implementation, we avoid the need of accurate correspondence of real scenes by exploiting the refocusing and multi-viewpoint abilities in the light-field data as shown in Fig. 3.

Diffuse-specular separation and color constancy. Separating diffuse and specular components by transforming from the RGB color space to the SUV color space such that the specular color is orthogonal to the light source color has been effective; however, these methods require an accurate estimation of or known light source color [7,17,18]. Without multiple viewpoints, most diffuse and specular separation methods assume the light source color is known [7,8,19–23]. As noted by Artusi et al. [24], these methods are limited by the light source color, prone to noise, and work well only in controlled or synthetic settings. To alleviate the light source constraint, we use similar specularity analyses as proposed by Sato and Ikeuchi and Nishino et al. [25,26]. However, prior to our work, the methods require multiple captures and robustness is dependent on the number of captures. With fewer images, the results become prone to noise. We avoid both of these problems by using the complete 4D EPI of the light-field data to enable a single capture that is robust against noise (Fig. 5). Estimating light source

color (color constancy) exhibits the same limitations and does not exploit the full light-field data [27,28]; however, these analyses are complementary to Eqn. 5. Since we are using the full light-field data, we can also independently estimate the light source color that each pixel is reflecting, enabling us to estimate more than just one light source color (see supplementary).

3 Theory and Algorithm

In this section, we explain the relationship between the dichromatic reflectance model and light-field data. The relationship enables us to estimate the light source color(s). We will then describe our algorithm that uses the light source color to improve depth estimation.

3.1 Background

Dichromatic reflection model. The basis of the algorithm revolves around diffuse and specular properties where diffuse is independent of view angle changes while specular is dependent. We use the dichromatic model for the bidirectional reflectance distribution function (BRDF) [6]. The dichromatic BRDF surface model, f, has the following expression,

$$f(\lambda, \Theta) = g_d(\lambda) f_d + g_s f_s(\Theta) \tag{1}$$

where λ is the wavelength and Θ represents the camera viewing angle and incoming light direction. g_d is the spectral reflectance and f_d and f_s are the diffuse and specular surface reflection multipliers respectively. Because we are dealing with dielectric materials, g_s is wavelength independent. The image captured by the camera can then be rewritten as

$$\begin{aligned} I_k &= (D_k f_d + S_k f_s(\Theta))\mathbf{n} \cdot \mathbf{l} \\ I_k &= (D_k f_d + L_k \bar{f}_s(\Theta))\mathbf{n} \cdot \mathbf{l} \end{aligned} \tag{2}$$

$\mathbf{n}$ and $\mathbf{l}$ are the surface normal and light source direction with k as the color channel. D is the diffuse color multiplied by the light source color, while S is proportional to the light source color. The top equation rewrites the dichromatic surface model (Eqn. 1) in terms of the surface normal and light direction. $S_k f_s(\Theta)$ can be rewritten as $L_k \bar{f}_s(\Theta)$, where $\bar{f}_s = g_s \cdot f_s(\Theta)$ and L is the light source color. Note that the diffuse component only depends on surface normal and light source direction. However, the specular component depends on Θ, or the camera viewpoint, making the color intensity view angle dependent. We will exploit the two properties through the following steps of our algorithm. We drop the $\mathbf{n} \cdot \mathbf{l}$ term for simplicity because the term acts as a modulator, and does not affect the color, on which our separation algorithm is based.

Light-field data and the dichromatic model. The light-field image encodes both spatial and angular information of the scene. The light-field image is represented

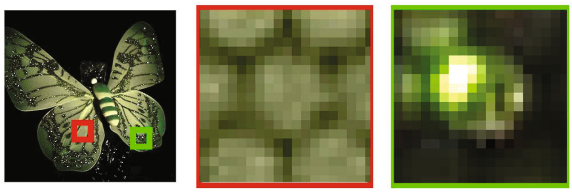


Fig. 4. Micro-lens With Diffuse and Specular Surfaces. In a scene with both specular and diffuse surfaces (a), the light-field image consists of different micro-lens behavior for specular and diffuse. For diffuse surfaces, the micro-lenses have consistent pixels (b). For specular surfaces, the lenses consist of different pixels that are influenced by the specular reflection term (c). This is consistent with our proposed analysis in Fig. 3, which we use to estimate the light source color.

by x, y, u, v, where x, y and u, v represent the spatial and angular domains respectively. With a light-field image, rearranging the pixels enables refocusing while extracting pixels from the micro-lens array gives multiple-views [3], described in Eqn. 3.

Rearranging the pixels to refocus allows us to perform the analysis in Fig. 3, 4. When the light-field is rearranged to focus to a certain point, the viewing directions all converge to that point. In such cases, diffuse surfaces will be registered the same from all viewpoints because the diffuse component is independent of Θ in Eqn. 2. In specular cases, since Θ is changed with the viewpoint, we estimate L by analyzing the color differences. The goal of our algorithm is to estimate L by exploiting this property of the light-field data. This differentiates our work from previous works of estimating L because we avoid the use of accurate correspondence and have pixel-based light source support, enabling estimation of multiple light source colors.

3.2 Algorithm

Our algorithm consists of three steps (Algorithm 1). The input is the light-field image captured by the Lytro camera, I. The first step (line 7) estimates depth, Depth, from the light-field image. The second step (line 8) estimates the light source color, L, for each pixel by using the refocusing and multi-perspective viewpoint with the depth estimation from the first step. The third step (line 9) separates the specular-free image, I_D, from the original light-field image input. Because depth estimation is reliable with Lambertian diffuse surfaces, the specular-free estimation improves depth estimation. We iteratively use the result from the separation to re-perform the computations of steps 1 to 3 (lines 7-9). The estimations of I_D, Depth, and L show improvements over the iterations (see supplementary). We then regularize the depth estimation with a MRF technique (line 14) presented by Janoch et al. [29].

Algorithm 1. Specular-Diffuse Separation for Depth

```
1: procedure SEPARATION(I)
2:     initialize  I_D, L_p
3:       I_D  = I                                  ▷ Diffuse as input LF image
4:       L_p  = 1/√3 [1,1,1]                       ▷ [R,G,B]; Light source is white
5:       L_Δ  = ∞
6:     while L_thres < L_Δ do
7:         Depth          = DepthEstimation(I_D)
8:         L, M_I, M_D    = LightSourceEstimation(I, Depth)
9:         I_D            = SpecularFree(I, L, M_I, M_D)

10:        L_Δ            = |L − L_p|
11:        L_p            = L
12:    end while
13:    return  I_D, Depth, L
14: end procedure
```

In the beginning of the algorithm, we initialize the diffuse buffer, I_D, as the original light-field input, I; the estimated light source color, L_p as $\frac{1}{\sqrt{3}}[1,1,1]$ in $[R,G,B]$ vector form (we normalize the L color vector as explained in Eqn. 6); and L_Δ as ∞. The iterations stop when the current L estimation has a root mean squared difference from the previous iteration, L_p, that is less than a threshold, L_Δ.

Depth estimation for refocusing (Line 7). Before we can estimate L at each given pixel, refocusing to each of the pixels in the scene is required to perform the analysis as shown in Fig. 2. We use the recent algorithm by Tao et al. [4], which is one of the first published depth estimation methods for the Lytro camera, and combines defocus and correspondence. However, other approaches such as Kim et al. [14] could also be used as we are using I_D, the specular-free estimation, as input. After the depth is computed, at each pixel, we have an approximation of where to refocus the image. Depth(x, y) registers the depth of each image pixel.

Exploiting refocus and multiple views for light source color estimation (Line 8). To estimate L, we will use the depth map that was generated to refocus and create multiple views. $L(x, y)$ is the estimate of the light source color at each pixel.

For each depth, we have the light-field input image, $I(x, y, u, v)$, where x, y and u, v represent the spatial and angular domains respectively. As explained by Ng et al. [3], we remap the light-field input image given the desired depth as follows,

$$I_\alpha(x, y, u, v) = I\left(x + u(1 - \frac{1}{\alpha}), y + v(1 - \frac{1}{\alpha}), u, v\right) \tag{3}$$

$\alpha = 0.2 + 0.007 \times$ Depth, where Depth ranges from 1 to 256. The α and range are scene dependent; however, we found these parameters to work for most of our examples.

For each depth value, we compute the color intensity changes within u, v of each x, y to perform the analysis shown in Fig. 2. Within u, v, we cluster them into specular-free (diffuse only) pixels, and specular pixels. By looking at the difference in centroids between the clusters (the specular intensity may vary at different views, but the color remains consistent), we classify two sets of pixels: n pixels with both diffuse and specular, $Df_d + L\bar{f}_s$, and m pixels with just specular-free, Df_d. The number of angular pixels u, v in each x, y equals to $n + m$. We perform a k-means clustering across the u, v pixels of each x, y to estimate the two. For simplicity, the two centroids of the two clusters will be denoted as ($\langle . \rangle$ denotes the expected value),

$$\begin{aligned} M_I(x, y) &= \langle Df_d + L\bar{f}_s \rangle (x, y, n) \\ M_D(x, y) &= \langle Df_d \rangle (x, y, m) \end{aligned} \tag{4}$$

In our implementation, the k-means uses 10 iterations. To compute $L\bar{f}_s$, we subtract the two centroids

$$L\bar{f}_s(x, y) = M_I(x, y) - M_D(x, y) \tag{5}$$

The M_D characterizes the specular-free term, and, if specular variations occur, $M_I - M_D$ characterizes the specular term. The specular term is proportional to the light source color intensity. Because $L\bar{f}_s$ represents the light source color with a multiplier, we will normalize each channel, k, of $L\bar{f}_s(x, y)$ to find $L_k(x, y)$,

$$L_k(x, y) = \frac{L_k \bar{f}_s(x, y)}{||L\bar{f}_s(x, y)||} \tag{6}$$

For pixels without the specular term, $|M_I(x, y) - M_D(x, y)| \approx 0$ because $L_k \bar{f}_s$ is close to zero, while pixels with the specular term or occlusions will not be zero. To differentiate between specular and occlusion areas in the light source color estimation, we want higher confidence in regions where the brightness of M_I and the distance between the two centroids are high. We characterize the confidence value for each $L(x, y)$ as follows,

$$C_L(x, y) = e^{-\beta_0 / |M_I(x,y)| - \beta_1 / |M_I(x,y) - M_D(x,y)| + \beta_2 / R} \tag{7}$$

where R is the average intra-cluster distance, β_0 is a constant parameter that changes the exponential fall off for the brightness term, β_1 is the fall off parameter for the centroid distance term, and β_2 for the robustness of the clustering. In our implementation, we used 0.5 for both β_0 and β_1 and 1 for β_2.

We can now separate the light source color at each pixel. However, for greater consistency, we perform a weighted average. For a scene with one light source, we average the light source estimation buffer, L, with the confidence map:

$$\text{Light Source Color} = \langle C_L(x, y) L(x, y) \rangle \tag{8}$$

where the expected value is normalized by the sum of $C_L(x, y)$.

For more than one light source, we perform a k-means cluster to the number of light sources. For each cluster, we perform the same weighted average to compute the light source colors. In our supplementary, we show two examples of two different light sources. The left shows an example with two highly glossy cans lit by two different light sources. The right shows a scene with two semi-glossy crayons lit by the same two light sources. In both cases, our algorithm estimates both light source colors accurately.

Discussion We find the correct light source color, but as with most similar bilinear problems involving a product of illumination and reflectance, we do not recover the actual intensity of the light source. If the specularity component is saturated throughout all (u, v), M_D does not represent specular-free color, causing the metric to fail. When f_s is small or the specular term is not present, the metric is unreliable. In both of these cases, the confidence level, C_L, is low. These pixels are shown as zero (see supplementary). However, the pixels with high confidence suffice to estimate one or more light source colors, and create the specular-free image and depth map.

Generating the specular-free image (Line 9). Using the L buffer from the previous step, we can compute a specularity-free image by using the full light-field data. For each pixel, x, y, u, v, of the light-field image, we subtract the specular term $L\bar{f}_s$, which is represented by $M_I - M_D$. For robustness, we search through a small neighborhood around x, y, u, v and compute an average of the specular term. Since not all pixels in the image contain $L\bar{f}_s$, we weight the subtracted specular value by favoring higher confidence of the light source estimation, C_L, and smaller difference between the pixel color, $I(x, y, u, v)$, and the neighbor's M_I (which represents $Df_d + L\bar{f}_s$). We use the following equation to compute the specular-free image:

$$\begin{aligned} Df_d(x, y, u, v) &= I(x, y, u, v) - \langle W \times (M_I(x', y') - M_D(x', y')) \rangle \\ W &= e^{-\gamma/(C_L(x', y') \times |I(x,y,u,v) - M_I(x', y')|)} \end{aligned} \tag{9}$$

where x', y' are within the search window around x, y, u, v. We normalize the value by the sum of the weights. In our implementation, we use a 15×15 search window and $\gamma = 0.5$.

4 Results

We verified our results with synthetic images, where we have ground truth for the light source, and diffuse and specular components. For all real images in the paper, we used the Lytro camera. We tested the algorithms across images with multiple camera parameters, such as exposure, ISO, and focal length, and in controlled and natural scenes.

Quantitative validation. In Figs. 5 and 6, we generated a scene using PBRT [30] with a matte material red wood textured background and a similarly textured sphere with Kd as the texture, Ks of color value, $[1, 1, 1]$, and roughness of 0.01.

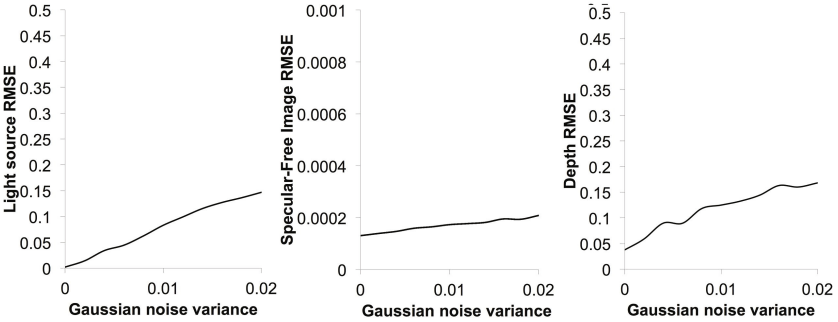


Fig. 5. Quantitative Synthetic Results. We use synthetic light-field inputs to verify our light source estimation, specular-free image, and depth estimation. We added Gaussian noise with zero mean and variance as the variable parameter. We compute the RMSE of our results against the ground truth light source color, diffuse image, and depth map. In the left, the light source estimation error is linear with the Gaussian noise variance, while yielding low error. In the middle, because we use the complete 4D-EPI to remove specularity, our specular-free result RMSE is very low. In the right, the RMSE for depth estimation also performs favorably to increased noise. At variance of 0.02, the input image exhibits high noise throughout the image, but our method performs well, even qualitatively (Fig. 6).

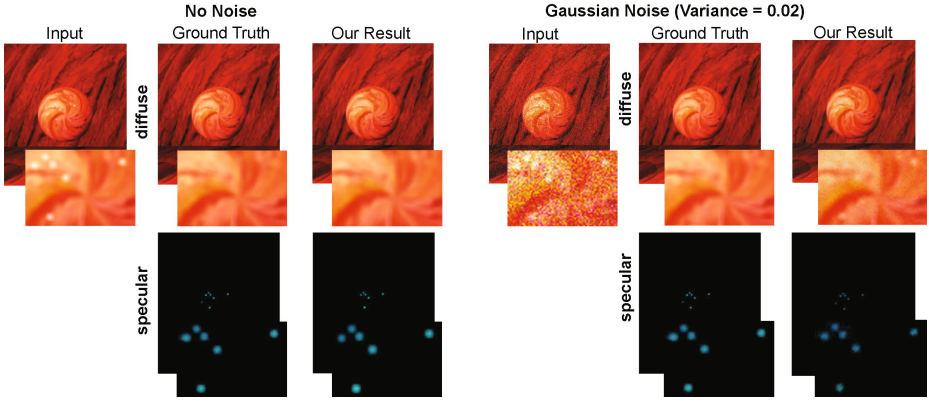


Fig. 6. Qualitative Synthetic Results. Using the zero noise and a high Gaussian noise with a variance of 0.02 as inputs, we can see that our specular-free image is very close to the ground truth, showing our algorithm's robustness against noise and successfully removing the six specular reflections on the sphere.

We have six point light sources scattered throughout the scene behind the camera with the same normalized color of $[0.03, 0.63, 0.78]$. We added Gaussian noise to the input image with mean of 0 and variance between 0 and 0.02. Our light source estimation, diffuse, and depth estimation errors increase linearly with noise variance (Fig. 5). Qualitatively, our algorithm is still robust with high noisy inputs (Fig. 6).

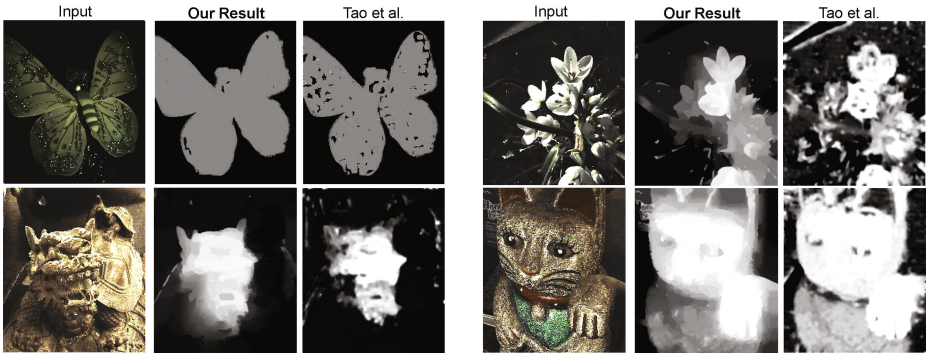


Fig. 7. Depth Map Comparisons. We compare our results against Tao et al. [4]. On the top left, the butterfly is placed perpendicular to the camera direction. Our depth estimation shows more consistent depth registration. Tao et al. shows spikes and instabilities in glossy regions. On the top right, we have a glossy plant, where our result still produces consistent results and Tao et al. show inconsistent depth registrations. On the bottom two, we have two different complex sculptures with different specular properties. The glossy surface creates instabilities in Tao et al.'s algorithm, which fails to estimate both depth and confidence correctly. Even in this complex glossy scene, our algorithm produces reasonable results that far outperform Tao et al.

To measure the accuracy of our L color estimation, we took two examples of controlled scenes. Both contain two light sources: one with highly glossy cans and the other with semi-glossy crayons. The light source estimations are consistent with the ground truth colors (see supplementary). Pixels that are indicated as black have low confidence values, C_L. We used a complex scene with multiple colors and materials with one known light source color (see supplementary). The light source estimation converges to the ground truth light source color. We tested the result by using a far-off initial light source estimate, $[0, 0, 1]$. After 15 iterations, the light source estimation is $[0.62, 0.59, 0.52]$, which converges to the ground truth value of $[0.60, 0.61, 0.52]$.

Depth map comparisons. To qualitatively assess depth improvement, we compare our work against Tao et al. [4]. We also compare against Wanner et al. [5], and Sun et al. [31] in our supplementary. We tested our algorithm through multiple scenarios involving specular highlights and reflections. In Fig. 7, the top left shows an example of a glossy butterfly. Our result is not thrown off by the specular surfaces. Tao et al. shows inconsistent depth registrations at specular surfaces because these regions have incorrect depths with high confidence values, which is also apparent in Fig. 8. The top right is an example of a glossy plant. Our algorithm generates a much more reasonable depth map while Tao et al. fails due to instability in confidence and depth estimation in glossy regions. In both sculpture examples, we have sculptures with different specular properties and complex shapes under low light. Our method is able to recover the surfaces. In Fig. 8, our method correctly estimates the shape of the dinosaur in a complex

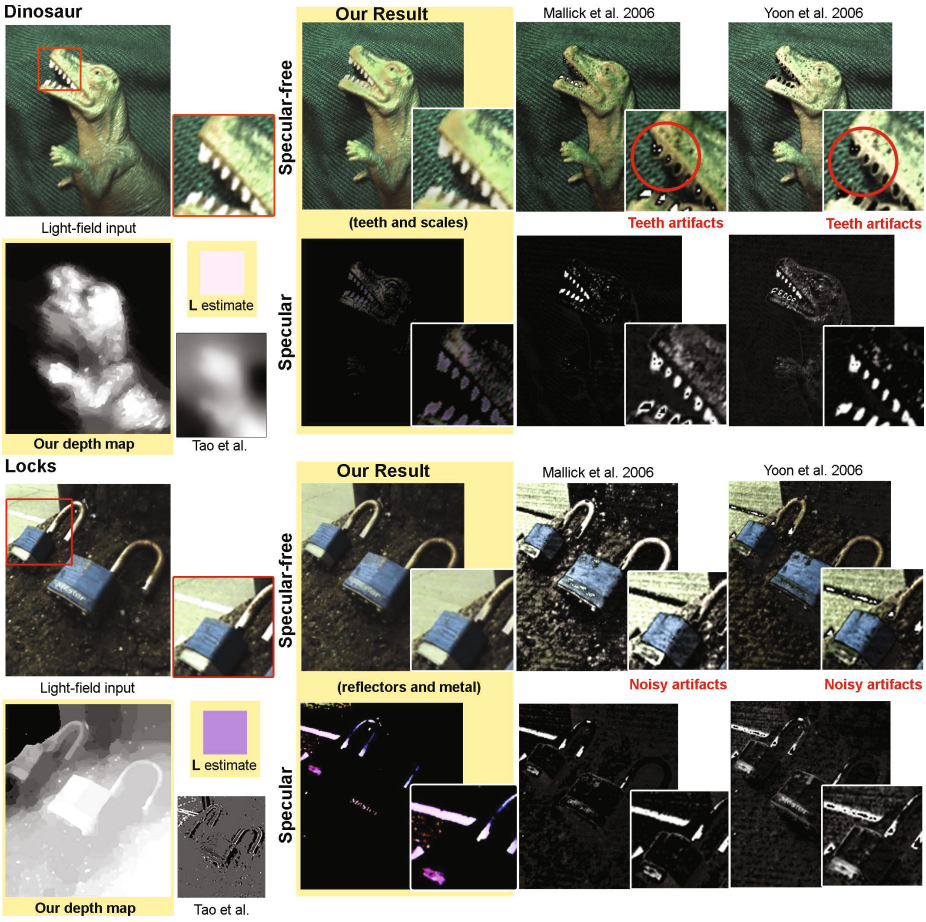


Fig. 8. Specular-Free Comparison. We compare our separation results against Mallick et al. [7] and Yoon et al. [8] and depth results against Tao et al. [4]. Our outputs are highlighted in yellow. Our method uses the depth maps to estimate L, which provides a significant benefit in generating our specular-free image. In the dinosaur example, our method's diffuse result shows reduced reflections on the very glossy teeth and semi-glossy cloth and scales of the dinosaur while the other methods result in artifacts. Because of the glossiness of the whole scene, Tao et al. fail dramatically due to the MRF instability in glossy surfaces, where confidence is high and depth is inaccurate. In the bottom, we have a natural outdoor scene with locks and street reflectors in the background. Both the metallic areas of the lock and the street reflector are correctly removed, but the other methods show hole artifacts. Both Mallick et al. and Yoon et al. exhibit noisy artifacts in the results, and incorrectly estimate the light source color as close to white. Note: our result does not completely remove saturated highlights, which is discussed in limitations and discussion. The results are best seen electronically and in our supplementary materials.

scene where both the dinosaur and cloth have glossy surfaces; the locks example also benefited from our specular separation. In both cases, we outperform Tao et al.'s depth maps.

Specular-free image comparisons. We verified specular reflection separation improvements over iterations (see supplementary). The specular color, after multiple iterations, is close to the light source color. We also compare our work against Mallick et al. [7] and Yoon et al. [8]. In Fig. 8, we tested the algorithms on two difficult cases. In the dinosaur example, we chose a glossy cloth for the background, and a glossy dinosaur with highly glossy teeth. Our result removes the reflections correctly while the other methods produce heavy artifacts and fail to remove most of the cloth's glossiness. In the locks example, our method correctly removes the glossiness from the metallic locks and road reflectors in the background. The other methods result in heavy artifacts. This is clearly shown in the specular components of the other methods. Both Mallick et al. and Yoon et al. bias the specular estimation close to white; while in real world scenarios, light sources are not always white.

Limitations and Discussion. Because of the small-baseline nature of light-field data, the light source cannot be too close to the reflective surface. In these situations, the light source cannot be easily detected as it will not move too much with respect to the viewpoint change. Saturated highlights also cannot be completely removed. As explained in Eqn. 5, in these cases, $M_D(x, y)$ does not represent the specular-free color, making the estimation hard. However, our confidence measure prevents this from affecting results and is further alleviated through our window search as described in our specular-free image generation. As with most specular-diffuse separation methods, our method does not perform well with mirrors and other highly specular surfaces. By using the dichromatic model described in Eqn. 1, our algorithm supports dielectric materials only, and will not work as well for metallic or highly specular surfaces, where highlights also take on the material color.

5 Conclusion and Future Work

In this paper, we present an iterative approach that uses light-field data to estimate and remove the specular component, improving the depth estimation. The method is the first to exploit light-field data depth estimation to support both specular and diffuse scenes. Our light-field analysis uses a physically-based method that estimates one or multiple light source colors. Upon publication, image datasets and source code will be released. The source code will allow ordinary users to acquire depth maps using a $400 consumer Lytro camera, in a point-and-shoot passive single-shot capture, including of specular and glossy materials. For future work, we will expand our analysis to more general reflection models to separate specular components for dielectric materials and incorporate shading information to improve robustness of the depth map regularization.

Acknowledgements. We acknowledge support from ONR grants N00014-09-1-0741 and N00014-14-1-0332, support from Adobe, Nokia, Samsung and Sony, and NSF and Berkeley graduate fellowships.

References

1. Gortler, S., Grzeszczuk, R., Szeliski, R., Cohen, M.: The lumigraph. In: ACM SIGGRAPH (1996)
2. Levoy, M., Hanrahan, P.: Light field rendering. In: ACM SIGGRAPH (1996)
3. Ng, R., Levoy, M., Bredif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photographhy with a hand-held plenoptic camera. CSTR 2005–02 (2005)
4. Tao, M., Hadap, S., Malik, J., Ramamoorthi, R.: Depth from combining defocus and correspondence using light-field cameras. In: ICCV (2013)
5. Wanner, S., Goldluecke, B.: Globally consistent depth labeling of 4D light fields. In: CVPR (2012)
6. Shafer, S.: Using color to separate reflection components. Color research and applications (1985)
7. Mallick, S., Zickler, T., Kriegman, D., Belhumeur, P.: Beyond lambert: reconstructing specular surfaces using color. In: CVPR (2005)
8. Yoon, K., Choi, Y., Kweon, I.: Fast separation of reflection components using a specularity-invariant image representation. In: IEEE Image Processing (2006)
9. Wantanabe, M., Nayar, S.: Rational filters for passive depth from defocus. IJCV (1998)
10. Xiong, Y., Shafer, S.: Depth from focusing and defocusing. In: CVPR (1993)
11. Horn, B., Schunck, B.: Determining optical flow. Artificial Intelligence (1981)
12. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Imaging Understanding Workshop (1981)
13. Min, D., Lu, J., Do, M.: Joint histogram based cost aggregation for stereo matching. PAMI (2013)
14. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.: Scene reconstruction from high spatio-angular resolution light fields. In: SIGGRAPH (2013)
15. Lin, S., Li, Y., Kang, S.B., Tong, X., Shum, H.-Y.: Diffuse-Specular Separation and Depth Recovery from Image Sequences. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part III. LNCS, vol. 2352, pp. 210–224. Springer, Heidelberg (2002)
16. Jin, H., Soatto, S., Yezzi, A.J.: Multi-view stereo beyond lambert. In: CVPR (2003)
17. Mallick, S.P., Zickler, T.E., Belhumeur, P.N., Kriegman, D.J.: Specularity Removal in Images and Videos: A PDE Approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 550–563. Springer, Heidelberg (2006)
18. Park, J.: Efficient color representation for image segmentation under nonwhite illumination. SPIE (2003)
19. Bajscy, R., Lee, S., Leonardis, A.: Detection of diffuse and specular interface reflections and inter-reflections by color image segmentation. IJCV (1996)
20. Tan, R., Ikeuchi, K.: Separating reflection components of textured surfaces using a single image. PAMI (2005)
21. Tan, R., Quan, L., Lin, S.: Separation of highlight reflections fromEtextured surfaces. CVPR (2006)

22. Yang, Q., Wang, S., Ahuja, N.: Real-Time Specular Highlight Removal Using Bilateral Filtering. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 87–100. Springer, Heidelberg (2010)
23. Kim, H., Jin, H., Hadap, S., Kweon, I.: Specular reflection separation using dark channel prior. In: CVPR (2013)
24. Artusi, A., Banterle, F., Chetverikov, D.: A survey of specularity removal methods. Computer Graphics Forum (2011)
25. Sato, Y., Ikeuchi, K.: Temporal-color space analysis of reflection. JOSAA (1994)
26. Nishino, K., Zhang, Z., Ikeuchi, K.: Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. ICCV (2001)
27. Finlayson, G., Schaefer, G.: Solving for color constancy using a constrained dichromatic reflection model. IJCV (2002)
28. Tan, R., Nishino, K., Ikeuchi, K.: Color constancy through inverse-intensity chromaticity space. JOSA (2004)
29. Janoch, A., Karayev, S., Jia, Y., Barron, J., Fritz, M., Saenko, K., Darrell, T.: A category-level 3D object dataset: putting the kinect to work. In: ICCV (2011)
30. Pharr, M., Humphreys, G.: Physically-based rendering: from theory to implementation. Elsevier Science and Technology Books (2004)
31. Sun, D., Roth, S., Black, M.: Secrets of optical flow estimation and their principles. In: CVPR (2010)

Accurate Disparity Estimation for Plenoptic Images

Neus Sabater(✉), Mozhdeh Seifi, Valter Drazic,
Gustavo Sandri, and Patrick Pérez

Technicolor, 975 Av. des Champs Blancs, 35576 Cesson-Sévigné, France
Neus.Sabater@technicolor.com

Abstract. In this paper we propose a post-processing pipeline to recover accurately the views (light-field) from the raw data of a plenoptic camera such as Lytro and to estimate disparity maps in a novel way from such a light-field. First, the microlens centers are estimated and then the raw image is demultiplexed without demosaicking it beforehand. Then, we present a new block-matching algorithm to estimate disparities for the mosaicked plenoptic views. Our algorithm exploits at best the configuration given by the plenoptic camera: (i) the views are horizontally and vertically rectified and have the same baseline, and therefore (ii) at each point, the vertical and horizontal disparities are the same. Our strategy of demultiplexing without demosaicking avoids image artifacts due to view cross-talk and helps estimating more accurate disparity maps. Finally, we compare our results with state-of-the-art methods.

Keywords: Plenoptic camera · Raw-data conversion · Disparity estimation

1 Introduction

Plenoptic cameras are gaining a lot of popularity in the field of computational photography because of the additional information they capture compared to traditional cameras. Indeed, they are able to measure the amount of light traveling along each ray bundle that intersects the sensor, thanks to a microlens array placed between the main lens and the sensor. As a result, such cameras have novel post-capture processing capabilities (e.g., depth estimation and refocusing). There are several optical designs for plenoptic cameras depending on the distance between the microlens array and the sensor. If this distance is equal to the microlenses focal length it is called a type 1.0 plenoptic camera [17]; and type 2.0 (or focused) plenoptic camera [16] otherwise. In the first case the number of pixels per rendered view[1] is equal to the number of microlenses (only one pixel per microlens is rendered on each view). In the second case, the rendered views have a higher spatial resolution, but that comes at the cost of decreasing the angular resolution. Depending on the application, one camera or another would be preferred. In this paper we focus on type 1.0 plenoptic cameras.

[1] The terms *view* and *sub-aperture image* are equally used in the literature.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 548–560, 2015.
DOI: 10.1007/978-3-319-16181-5_42

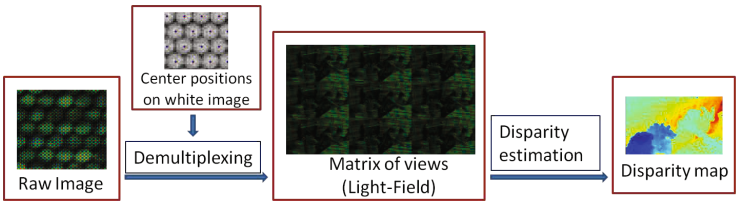


Fig. 1. Pipeline of our method. For visualization purposes only a part of the subimages and the views are shown. The LF is obtained by demultiplexing mosaicked data using the center subimage positions. Then the accurate disparity map for a reference view is estimated from the LF.

The concept of *integral photography*, which is the underlying technology in plenoptic cameras was introduced in [15] and then brought up to computer vision in [3], but it has recently become practical with the hand-held cameras that Lytro[2] and Raytrix[3] have put on the market for the mass market and professionals respectively. Since then, the scientific community has taken an interest in the LF (Light-Field) technology. Recent studies in the field address the bottleneck of the plenoptic cameras, namely the resolution problem ([10], [5], [18] and [24]). Besides super-resolution, depth estimation has also been investigated as a natural application of plenoptic images ([5], [24] and [22]). Indeed, the intrinsic information of the LF has the advantage to allow disparity computation without the image calibration and rectification steps required in classic binocular stereo or multi-view algorithms, making it an enormous advantage for 3D applications. However, the last cited works consider the sampled LF (the set of demultiplexed views) as input for their disparity estimation methods, meaning that they do not study the process that converts the raw data acquired by the plenoptic camera into the set of demultiplexed views. In this paper we show that such processing, called *demultiplexing*, is of paramount importance for depth estimation.

The contributions of this paper are twofold. First, we model the demultiplexing process of images acquired with a Lytro camera and then we present a novel algorithm for disparity estimation specially designed for the singular qualities of plenoptic data. In particular, we show that estimating disparities from mosaicked views is preferred to using views obtained through conventional linear demosaicking on the raw data. Therefore, for the sake of accurate disparity estimation, demosaicking is not performed in our method (see our pipeline in Fig. 1). To the best of our knowledge this approach has never been proposed before.

2 Related Work

The closest works to our demultiplexing method have been published recently. In [7] a demultiplexing algorithm followed by a rectification step where lens

[2] http://www.lytro.com
[3] http://www.raytrix.de

distortions are corrected using a 15-parameter camera model is proposed. In [6], the authors also proposed a demultiplexing algorithm for the Lytro camera and studied several interpolation methods to superresolve the reconstructed images. On the contrary, [9] recovers the refocused Lytro images via splatting without demultiplexing the views.

Considering disparity estimation for plenoptic images, several works have proposed a variational method ([24], [4], [5] , [13] and [23]). In particular, [24] uses the epipolar plane image (EPI), [4] and [5] propose an antialiasing filtering to avoid cross-talk image artifacts and [13] combines the idea of Active Wavefront Sampling (AWS) with the LF technique. In fact, variational methods better deal with the noise in the images but they are computationally expensive. Given the large number of views on the LF, such approaches are not suitable for many of applications. In addition to variational approaches, other methods have been proposed for disparity estimation. [14] estimates disparity maps from high spatio-angular LF with a fine-to-coarse algorithm where disparities around object boundaries are first estimated using an EPI-based method and then propagated. [22] proposes an interesting approach that combines defocus and correspondence to estimate the scene depth. Finally, [25] presents a Line-Assisted Graph-Cut method in which line segments with known disparities are used as hard constraints in the graph-cut algorithm.

In each section we shall discuss the differences between our method and the most related works on demultiplexing and disparity estimation methods on Lytro data. While demosaicking is not the goal of this paper, note that [10] already pointed out artifacts due to raw plenoptic data demosaicking and that a practical solution was proposed by [26] for type 2.0 plenoptic data.

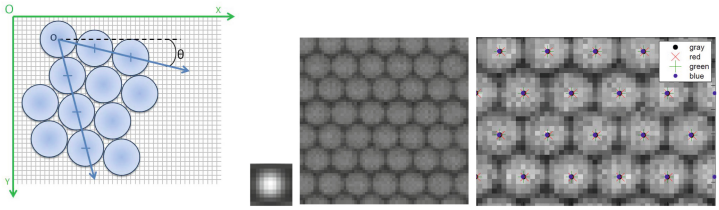
3 Demultiplexing RAW Data

Demultiplexing (also called "decoding" [7] or "calibration and decoding" [6]) is data conversion from the 2D raw image to the 4D LF, usually represented by the two-plane parametrization [12]. In particular, demultiplexing consists in reorganizing the pixels of the raw image[4] in such a way that all pixels capturing the light rays with a certain angle of incidence are stored in the same image creating the so-called *views*. Each view is a projection of the scene under a different angle. The set of views create a block matrix where the central view stores the pixels capturing the light rays perpendicular to the sensor. In fact, in plenoptic type 1.0, the angular information of the light rays is given by the relative pixel positions in the *subimages*[5] with respect to the subimage centers. After demultiplexing, the number of restored views (entries of the block matrix) corresponds to the number of pixels covered by one microlens and each restored view has as many pixels as the number of microlenses.

Estimating Subimage Centers: In a plenoptic camera such as Lytro the microlens centers are not necessarily well aligned with the pixels of the sensor.

[4] We use the tool in [1] to access the raw data from Lytro.

[5] The image that is formed under a microlens and on the sensor is called a subimage.



(a) (b) (c) (d)

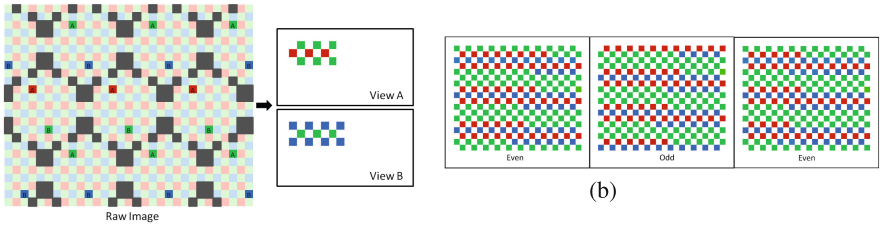
Fig. 2. (a) Microlenses projected on the sensor plane in a hexagonal arrangement. The green and blue axes represent the two CSs. There is a rotational offset θ and a translational offset $O - o$. (b) Mask used to locally estimate subimage center positions. (c) Lytro raw image of a white scene. (d) Estimated center positions. They coincide when estimated from one color channel only or from all the pixels in the raw image (gray).

There is a rotation offset between the sensor and the microlens plane. Also, the microlens diameter does not cover an integer number of pixels. Finally, the microlenses are arranged on a hexagonal grid to efficiently sample the space. Thus, in order to robustly estimate the microlens centers, we estimate the transformation between two coordinate systems (CS), the Cartesian CS given by the sensor pixels and K, the microlens center CS. K is defined as follows: the origin is the center of the topmost and leftmost microlens and the basis vectors are the two vectors from the origin to the adjacent microlens centers (see Fig.2-(a)). Formally, if $\mathbf{x}$ and $\mathbf{k}$ are respectively the coordinates on the sensor and microlens CSs, then, we estimate the system transformation matrix $\mathbf{T}$ and the offset vector between the origins $\mathbf{c}$ such that $\mathbf{x} = \mathbf{T}\mathbf{k} + \mathbf{c}$, and

$$\mathbf{T} = \begin{pmatrix} 1 & 1/2 \\ 0 & \sqrt{3}/2 \end{pmatrix} \begin{pmatrix} d_h & 0 \\ 0 & d_v \end{pmatrix} \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}, \tag{1}$$

where the first matrix accounts for the orthogonal to hexagonal grid conversion due to the microlens arrangement, the second matrix deals with the vertical and horizontal shears and the third matrix is the rotation matrix. Thus, estimating the microlens model parameters $\{\mathbf{c}, d_h, d_v, \theta\}$ gives the microlenses center positions.

In practice, the subimage centers are computed from a *white image* depicted in Fig. 2-(c), that is an image taken through a white Lambertian diffuser. Actually, the subimage centers $\mathbf{x}_i$ of the i-th microlens in the raw image are computed as the local maximum positions of the convolution between the *white image* and the mask shown in Fig. 2-(b). Then, given $\mathbf{x}_i$ and the integer positions $\mathbf{k}_i$ in the K CS, the model parameters (and consequently $\mathbf{T}$ and $\mathbf{c}$) are estimated as the solution of a least square error problem from the equations $\mathbf{x}_i = \mathbf{T}\mathbf{k}_i + \mathbf{c}$. Thus, in this paper, the final center positions used in the demultiplexing step are the pixel positions given by $\mathbf{c}_i := \text{round}(\mathbf{T}\mathbf{k}_i + \mathbf{c})$. However, more advanced approaches



(a) (b)

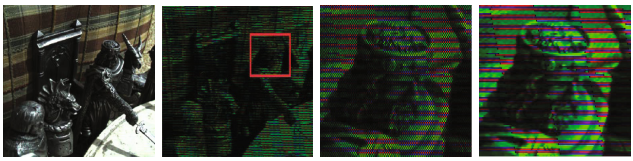
Fig. 3. (a) Demultiplexing. Pixels with the same relative position w.r.t. the subimage centers are stored in the same view. Only two views are illustrated for visualization. Color corresponds to sensor color on original Bayer pattern, and is carried over to assembled raw views. (b) Color patterns of three consecutive mosaicked views (even, odd and even positions of a line of the matrix of views) for a Lytro camera ($\sim 10 \times 10$ pix. per microlens). Color patterns from the views at even positions are very similar while the color pattern at the odd position is significantly different although there are horizontal color stripes too. White (empty) pixels are left to avoid aliasing.

can take into account the sub-pixel accuracy of the estimated centers and re-grid the data on integer spatial coordinates of the Cartesian CS. Fig. 2-(d) shows the subimage center estimation obtained with the method described above. Since the raw white image has a Bayer pattern, we have verified that the center positions estimated by considering only red, green or blue channel, or alternatively considering all color channels, are essentially the same. Indeed, demosaicking the raw *white image* does not create image cross-talk since the three color channels are the same for all pixels in the center of the subimages.

Reordering Pixels: In the following, we assume that the raw image has been divided pixel-wise by the white image. This division considerably corrects the vignetting [6] which is enough for our purposes. We refer to [7] for a precise vignetting modeling in plenoptic images. Now, in order to recover the different views, pixels are organized as illustrated in Fig. 3-(a). In order to preserve the pixel arrangement in the raw image (hexagonal pixel grid), empty spaces are left between pixels on the views as shown in Fig. 3-(b). Respecting the sampling grid avoids creating aliasing on the views. Notice that, since the raw image has not been demosaicked, the views inherit new color patterns. Because of the shift and rotation of the microlenses w.r.t. the sensor, the microlens centers (as well as other relative positions) do not always correspond to the same color. As a consequence, each view has its own color pattern (mainly horizontal monochrome lines in Lytro).

After demultiplexing, the views could be demosaicked without risking to fuse pixel information from different angular light rays. However, classic demosaicking algorithms are not well adapted to these new color patterns, specially on

[6] Light rays hitting the sensor at an oblique angle produce a weaker signal than other light rays.



(a) (b) (c) (d)

Fig. 4. (a) Lytro image (for visualization purposes). (b) One mosaicked view. (c) Zoomed red rectangle in view (b). (d) Same zoom with horizontal interpolation of empty (black) pixels, when possible. This simple interpolation does not create artifacts since all the pixels in a view contain same angular information.

high frequencies. For the sake of disparity estimation, we simply fill the empty pixels in a color chanel (white pixels in Fig. 3) when the neighboring pixels have the color information for this chanel (see Fig. 4). For example, if an empty pixel of the raw data has a green pixel on the right and on the left, then the empty pixel is filled with a green value by interpolation (1D Piecewise Cubic Hermite interpolation). Other empty pixels are left as such.

Differences with State-of-the-Art: The main difference with the demultiplexing method in [7] is the fact that in their method the raw data of a scene is demosaicked before being demultiplexed. This approach mixes information from different views and, as we will show in the next section, it has dramatic consequences on the disparity estimation. Besides, the method in [7] estimates the microlenses centers similarly to us but it does not force the center positions to be integer as we do in our optimization step. Instead, the raw image is interpolated to satisfy this constraint. Even if theoretically this solution should provide a more accurate LF, interpolating the raw data implies again mixing information from different views which creates image cross-talk artifacts. The method for estimating the center positions in [6] differs considerably from ours since the centers are found via local maxima estimation in the frequency domain. First, the raw image is demosaicked and converted to gray and the final center positions are the result of fitting the local estimation on a Delaunay triangular grid. Moreover, the second step to render the views is coupled with super-resolution providing views of size 1080×1080 (instead of 328×328, which is the number of microlenses).

The goal of this paper is to estimate accurately the disparity on plenoptic images, but we have observed that the processing needed before doing that is of foremost importance. So, even if the works in [7] and [6] are an important step forward for LF processing, we propose an alternative processing of the views which is better suited to subsequent disparity estimation.

4 Disparity Estimation

In this section, we present a new block-matching disparity estimation algorithm adapted to plenoptic images. We assume that a matrix of views is available

(obtained as explained in the previous section) such that the views are horizontally and vertically rectified, i.e., satisfying the epipolar constraint. Therefore, given a pixel in a reference view, its corresponding pixels from the same row of the matrix are only shifted horizontally. Similar reasoning is valid for the vertical pixel shifts among views from the same column of the matrix. Furthermore, consecutive views have always the same baseline a (horizontally and vertically). As a consequence, for each point, its horizontal and vertical disparities with respect to nearest views are equal provided the point is not occluded. In other words, given a point in the reference view, the corresponding point in its consecutive right view is displaced horizontally by the same distance than the corresponding point in its consecutive bottom view is displaced vertically. By construction, the plenoptic camera provides a matrix of views with small baselines, which means that the possible occlusions are small. In fact, each point of the scene is seen from different points of views (even if it is occluded for some of them). Thus, the horizontal and vertical disparity equality is true for almost all the points of the scene. To the best of our knowledge, this particular property of plenoptic data has not been exploited before.

Since the available views have color patterns as in Fig. 3, we propose a block matching method in which only pixels in the block having the same color information are compared. We propose to use a similarity measure between blocks based on the ZSSD (Zero-Mean Sum of Squared Differences). Formally, let I^p be a reference view of the matrix of views and I^q be a view belonging to the same matrix row as I^p. Let $a_{p,q}$ be the respective baseline (a multiple of a). Then, the cost function between I^p and I^q at the center (x_0, y_0) of a block B_0 in I^p is defined as a function of the disparity d:

$$CF_0^{p,q}(d) = \frac{1}{\sum\limits_{(x,y)\in B_0} W(x,x',y)} \sum_{(x,y)\in B_0} W(x,x',y)\Big(I^p(x,y) - \overline{I_0^p} - I^q(x',y) + \overline{I_0^q}\Big)^2, \tag{2}$$

where $x' := x + a_{p,q}\, d$, $\overline{I_0^p}$ and $\overline{I_0^q}$ are the average values of I^p and I^q over the block centered at (x_0, y_0) and $(x_0 + a_{p,q}\, d, y_0)$ respectively and W is the window function

$$W(x,x',y) = G_0(x,y) \cdot S(x,x',y)\,,$$

where G_0 is a Gaussian function centered at (x_0, y_0) and supported in B_0 and S is the characteristic function controlling that only pixels in the block with same color information are compared in the cost function: $S(x,x',y) = 1$ if $I^p(x,y)$ and $I^q(x',y)$ have the same color information, and 0 otherwise. Note that the cost function is similarly defined when I^p and I^q are views from the same matrix column. In practice, we consider blocks of size 13×13.

Now, our algorithm takes advantage of the multitude of views given by the LF and estimates the disparity through all the rows and columns of the matrix. Let Θ be the set of index-view pairs such that the disparity can be computed horizontally or vertically w.r.t. the reference view I^p. In other words, Θ is the set of index-view pairs of the form (I^p, I^q), where I^q is from the same row or the same column as I^p. In fact, consecutive views are not considered in Θ

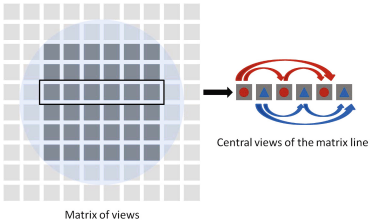


Fig. 5. On the left: LF (matrix of views). Views in the center get more radiance than views of the border of the matrix (pixels coming from the border of the microlenses). The 6×6 central views among the 10×10 are used. On the right: 6 central views from the same row of the matrix. Odd and even views have different color patterns between them (but very similar patterns between odd views and even views). This is represented with a red circle and a blue triangle. The index-view pairs in Θ corresponding to this matrix row are represented with the red and blue arrows.

since consecutive color patterns are essentially different because of the sampling period of sensor's Bayer pattern. Besides, views on the borders of the matrix are strongly degraded by the vignetting effect of the main lens. So, it is reasonable to only consider the 8×8 or 6×6 matrix of views placed in the center for the Lytro camera. Fig. 5 depicts the pairs of considered images for disparity estimation in a matrix row. Finally, given a reference view I^p, the disparity at (x_0, y_0) is given by

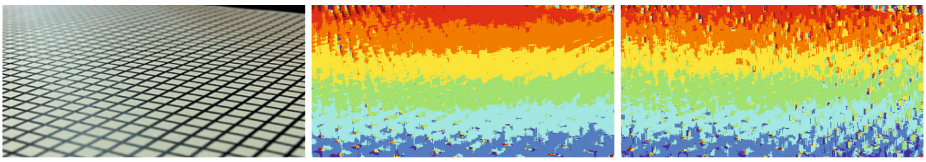
$$d(x_0, y_0) = \mathrm{Med}_{(p,q)\in\Theta}\left\{ \arg\min_d CF^{p,q}_{B_0}(d) \right\}, \tag{3}$$

where Med stands for the 1D median filter. This median filter is used to remove outliers that may appear on a disparity map computed for a single pair of views, specially in low-textured areas. It should be noted that through this median filtering, all the horizontally and vertically estimated disparities are considered to select a robust estimation of disparity which is possible thanks to the horizontal and vertical disparity equality mentioned beforehand.

Removing Outliers: Block-matching methods tend to provide noisy disparity maps when there is a matching ambiguity, e.g., for repeated structures in the images or on poorly textured areas. Inspired by the well-known cross-checking in binocular stereovision [20] (i.e., comparing left-to-right and right-to-left disparity maps), our method can also remove unreliable estimations comparing all possible estimations. Since a large amount of views are available from a LF, it is straightforward to rule out inconsistent disparities. More precisely, points (x_0, y_0) are considered unreliable if

$$\mathrm{Std}_{(p,q)\in\Theta}\left\{ \arg\min_d CF^{p,q}_{x_0,y_0}(d) \right\} > \varepsilon, \tag{4}$$

where Std stands for standard deviation and ε is the accuracy in pixels. In practice, we consider an accuracy of an eight of a pixel, $\varepsilon = \frac{1}{8}$.



(a) (b) (c)

Fig. 6. (a) Lytro Image of the scene. (b) Disparity estimation without raw image demosaicking. (c) Disparity estimation with raw image demosaicking. The cost function is the same but the characteristic function is equal to one for all the points since the views are in full RGB. For the sake of accurate analysis no sub-pixel refinement has been performed. Errors due to image cross-talk artifacts are tremendous on disparity maps.

Sub-Pixel Disparity Estimation: By construction, the baseline between the views is small, specially between views with close positions in the matrix. So the disparity estimation for plenoptic images must achieve sub-pixel accuracy. Such precision can be achieved in two different ways: either by upsampling the views or by interpolating the cost function. Usually the first method achieves better accuracy but at a higher computational burden, unless GPU implementations are used [8]. For this reason, the second method (cost function interpolation) is usually used. However, it has been proved [19] that block-matching algorithms with a quadratic cost function as in Eq. (2) achieve the best trade-off between complexity and accuracy only by first upsampling the images by a factor of 2 and then interpolating the cost function. We follow this rule in our disparity estimation algorithm.

Differences with State-of-the-Art: The closest disparity estimation method for plenoptic images compared to ours is the method presented in [5] but there are several differences between both methods. First, our method properly demultiplexes the views before estimating the disparity, whereas the method in [5] considers full RGB views and proposes an antialiasing filter to cope with the weak prefilter in plenoptic type 2.0. Then, the energy defined in [5] (compare Eq. 3 of this paper with Eq. 3 in [5]) considers all the possible pairs of views even if in practice, for complexity reasons, only a subset of view pairs can be considered. In [5], no criteria is given to define such subset of view pairs while a reasonable subset is given with respect to the color pattern in our views. Finally, the proposed energy in [5] considers a regularization term in addition to the data term and the energy is minimized iteratively using conjugate gradients. In another state-of-the-art method, [22] combines spatial correspondence with defocus. More precisely, the algorithm uses the 4D EPI and estimates correspondence cues by computing angular variance, and defocus cues by computing spatial variance after angular integration. Both cues are combined in an MRF global optimization process. Nevertheless, their disparity estimation method does not take care of the demultiplexing step accurately. Their algorithm not only demosaicks the raw

image, but it stores it using JPEG compression. So, the resulting LF is affected by image cross-talk artifacts and compression artifacts. In next section, we shall compare our results with this method. Unfortunately, a qualitative comparison with [5] is not possible since the authors work with different data: mosaicked views from a focused or type 2.0 plenoptic camera.

5 Experimental Results

In this section we show the results obtained with our algorithm. First of all, we have compared the disparity maps obtained with and without demosaicking the raw image. Intuitively one can think that demosaicking the raw image will get better results since more information is available on the views. However this intuition is rejected in practice (see for instance Fig. 6). Therefore, we claim that accurate disparity estimation should consider only the raw data on the views. Unfortunately, experimental evaluation with available benchmarks with ground-truth [24] as in [13] is not possible because all LF in the benchmark are already demosaicked.

Fig. 7 compares our disparity maps from Lytro using [2] and the disparity map from [22] using the code provided by the authors and the corresponding microlenses center positions for each experiment. The algorithms have been tested with images from [22] and images obtained with our Lytro camera. The poor results from [22] with our data show a strong sensitivity to parameters of their algorithm. Also, their algorithm demosaicks and compresses (JPEG) the raw image before depth is estimated. On the other hand, Lytro disparity maps are more robust but they are strongly quantized which may not be sufficiently accurate for some applications. All in all, our method has been tested on a large number of images from Lytro with different conditions and it provides robust and accurate results compared to state-of-the-art disparity estimation method for plenoptic images.

Obviously, other approaches could be considered for disparity estimation. For instance, our cost function can be regarded as the data term in a global energy minimization approach as in [25]. However, for the sake of computational speed we have preferred a local method. Specially, because a multitude of disparity estimations can be performed at each pixel. Moreover, other approaches using EPI's as in [24] could be used but we have observed that EPI's from Lytro are highly noisy and only disparities on object edges are reliable (EPI from Lytro is only ~ 10 pixels width).

In this paper we propose to not perform demosaicking on the raw image to avoid artifacts but full RGB images are needed for some applications (i.e., refocusing). In that case we suggest to recover the lacking colors by bringing the color information from all the corresponding points in all views using the estimated disparity information as in [21]. Indeed, one point in the reference view seen with one color channel is seen in the other views with another color. Fig. 8 shows disparity-guided demosaicking results. We show that our approach avoids color artifacts compared with the method in [22] that demosaicks raw images. So,

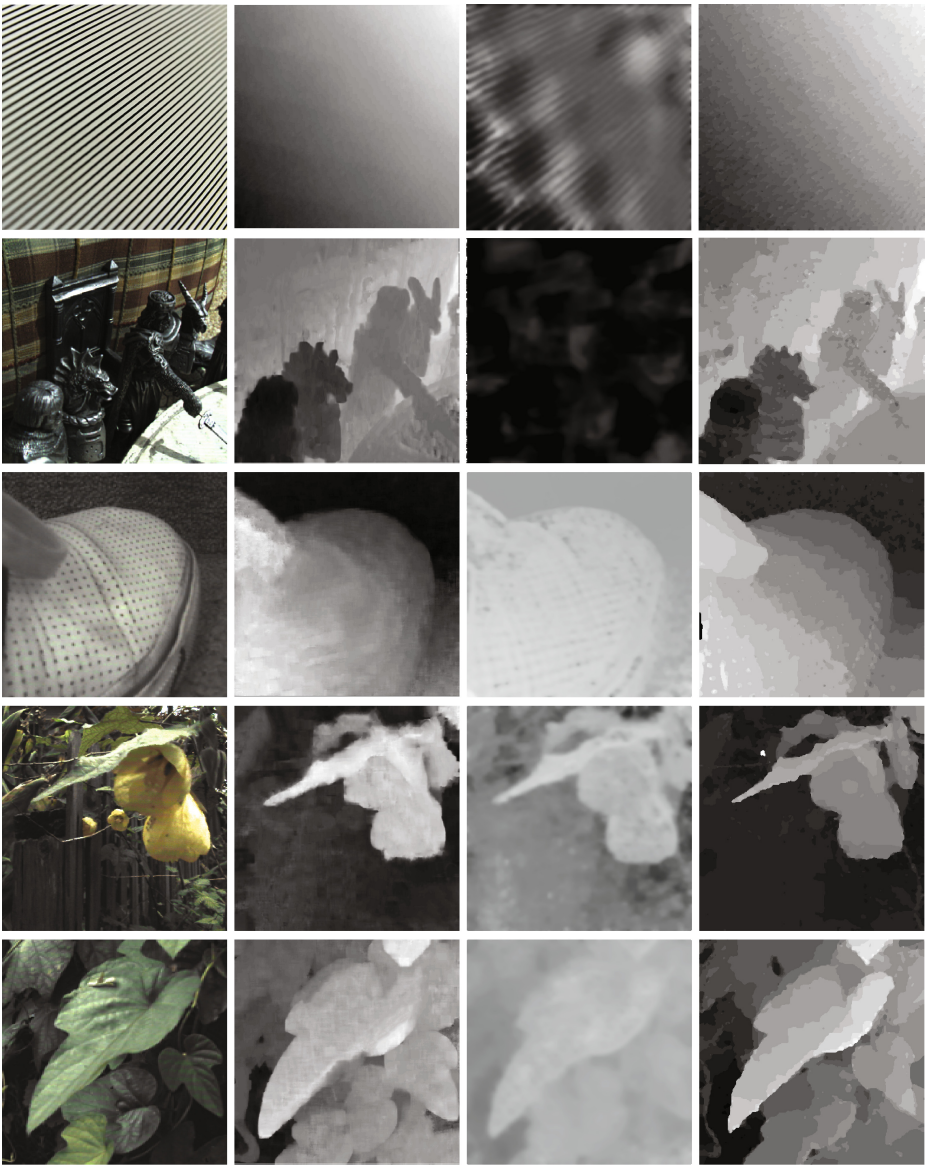


Fig. 7. (a) Original data. The three last images are published in [22]. (b) Our disparity map results. (c) Results from [22]. The authors have found a good set of parameters for their data but we have found poor results using their algorithm with our data. (d) Depth map used by Lytro, obtained with a third party toolbox [2].

Fig. 8. Comparison of RGB views. Left: Our result. Right: Result of demosaicking the raw data as in [22]. Besides of a different dynamic range certainly due to a different color balance, notice the reddish and greenish bands on the right flower (best seen on PDF).

our demultiplexing mosaicked data strategy not only avoids artifacts on disparity maps but also on full RGB view rendering.

It shall be pointed out that we assume the Lytro camera to be a plenoptic type 1.0. Although not much is officially available about its internal structure, our observation of the captured data and the study in [11] support this assumption. However, the assumption on the camera type only changes the pixel reordering in the demultiplexing step, and the proposed method can be easily generalized to the case of plenoptic type 2.0.

Finally, even if our method only considers central views of the matrix of views, we have observed slightly bigger errors on the borders of the image. Pushing further the correction of vignetting and of other chromatic aberrations could be profitable to accurate disparity estimation. This is one of our perspectives for future work.

6 Conclusion

Plenoptic cameras are promising tools to expand the capabilities of conventional cameras, for they capture the 4D LF of a scene. However, specific image processing algorithms should be developed to make the most of this new technology. There has been tremendous effort on disparity estimation for binocular stereovision [20], but very little has been done for the case of plenoptic data. In this paper, we have addressed the disparity estimation problem in plenoptic data and we have seen that it should be studied together with demultiplexing. In fact, the proposed demultiplexing step on mosaicked data is a simple pre-processing that has clear benefits for disparity estimation and full RGB view rendering since they do not suffer from view cross-talk artifacts.

References

1. http://code.behnam.es/python-lfp-reader/
2. http://optics.miloush.net/lytro/

3. Adelson, E., Wang, J.: Single lens stereo with a plenoptic camera. TPAMI **14**(2), 99–106 (1992)
4. Bishop, T.E., Favaro, P.: Full-Resolution Depth Map Estimation from an Aliased Plenoptic Light Field. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 186–200. Springer, Heidelberg (2011)
5. Bishop, T.E., Favaro, P.: The light field camera: Extended depth of field, aliasing, and superresolution. TPAMI **34**(5), 972–986 (2012)
6. Cho, D., Lee, M., Kim, S., Tai, Y.W.: Modeling the calibration pipeline of the lytro camera for high quality light-field image reconstruction. In: ICCV (2013)
7. Dansereau, D.G., Pizarro, O., Williams, S.B.: Decoding, calibration and rectification for lenselet-based plenoptic cameras. In: CVPR (2013)
8. Drazic, V., Sabater, N.: A precise real-time stereo algorithm. In: ACM Conf. on Image and Vision Computing New Zealand, pp. 138–143 (2012)
9. Fiss, J., Curless, B., Szeliski, R.: Refocusing plenoptic images using depth-adaptive splatting. In: ICCP (2014)
10. Georgiev, T., Chunev, G., Lumsdaine, A.: Superresolution with the focused plenoptic camera. In: SPIE Electronic Imaging (2011)
11. Georgiev, T., Yu, Z., Lumsdaine, A., Goma, S.: Lytro camera technology: theory, algorithms, performance analysis. In: SPIE Electronic Imaging, pp. 86671J–86671J (2013)
12. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The lumigraph. In: Conf. on Computer Graphics and Interactive Techniques (1996)
13. Heber, S., Ranftl, R., Pock, T.: Variational Shape from Light Field. In: Heyden, A., Kahl, F., Olsson, C., Oskarsson, M., Tai, X.-C. (eds.) EMMCVPR 2013. LNCS, vol. 8081, pp. 66–79. Springer, Heidelberg (2013)
14. Kim, C., Zimmer, H., Pritch, Y., Sorkine-Hornung, A., Gross, M.: Scene reconstruction from high spatio-angular resolution light fields. ACM Trans. Graph. **32**(4), 73 (2013)
15. Lippmann, G.: Epreuves reversibles donnant la sensation du relief. J. Phys. Theor. Appl. **7**(1), 821–825 (1908)
16. Lumsdaine, A., Georgiev, T.: The focused plenoptic camera. In: ICCP (2009)
17. Ng, R.: Digital light field photography. Ph.D. thesis, Stanford University (2006)
18. Perez, F., Perez, A., Rodriguez, M., Magdaleno, E.: Fourier slice super-resolution in plenoptic cameras. In: ICCP (2012)
19. Sabater, N., Morel, J.M., Almansa, A.: How accurate can block matches be in stereo vision? SIAM Journal on Imaging Sciences **4**(1), 472–500 (2011)
20. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV **47**(1–3), 7–42 (2002)
21. Seifi, M., Sabater, N., Drazic, V., Perez, P.: Disparity-guided demosaicing of light-field images. In: ICIP (2014)
22. Tao, M., Hadap, S., Malik, J., Ramamoorthi, R.: Depth from combining defocus and correspondence using light-field cameras. In: ICCV (2013)
23. Tulyakov, S., Lee, T., H., H.: Quadratic formulation of disparity estimation problem for light-field camera. In: ICIP (2013)
24. Wanner, S., Goldluecke, B.: Variational light field analysis for disparity estimation and super-resolution. TPAMI (2014) (to appear)
25. Yu, Z., Guo, X., Ling, H., Lumsdaine, A., Yu, J.: Line assisted light field triangulation and stereo matching. In: ICCV (2013)
26. Yu, Z., Yu, J., Lumsdaine, A., Georgiev, T.: An analysis of color demosaicing in plenoptic cameras. In: CVPR (2012)

SocialSync: Sub-Frame Synchronization in a Smartphone Camera Network

Richard Latimer(✉), Jason Holloway, Ashok Veeraraghavan, and Ashutosh Sabharwal

Rice University, Houston, TX, USA
rplatimer@gmail.com

Abstract. SocialSync is a sub-frame synchronization protocol for capturing images simultaneously using a smartphone camera network. By synchronizing image captures to within a frame period, multiple smartphone cameras, which are often in use in social settings, can be used for a variety of applications including light field capture, depth estimation, and free viewpoint television. Currently, smartphone camera networks are limited to capturing static scenes due to motion artifacts caused by frame misalignment. Because frame misalignment in smartphones camera networks is caused by variability in the camera system, we characterize frame capture on mobile devices by analyzing the statistics of camera setup latency and frame delivery within an Android app. Next, we develop the SocialSync protocol to achieve sub-frame synchronization between devices by estimating frame capture timestamps to within millisecond accuracy. Finally, we demonstrate the effectiveness of SocialSync on mobile devices by reducing motion-induced artifacts when recovering the light field.

Keywords: Multiple viewpoints · Camera array · Camera network · Synchronization · Smartphone · Mobile device

1 Introduction

Smartphones, and by extension smartphone cameras, have been predicted to approach 1 billion units in annual sales by the end of 2014 [9]. The rapid rise of readily available cameras has drastically increased the number of pictures that are taken each day, while the advent of social media and image sharing websites (e.g., Facebook, Flickr, and Picasa) allows for easier image dissemination than ever before.

While sharing images has become a common activity in social interactions – Facebook sees an average of 350 million images uploaded to its servers daily [7] – capturing images remains an individual activity. Despite collectively viewing, sharing, and commenting on images, photographers remain as islands; each taking pictures independently and ignoring the resources of other nearby smartphone cameras. Our goal is to synchronize image captures using mobile devices during *social image acquisition*, whereby users can collaboratively capture images

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-16181-5_43) contains supplementary material, which is available to authorized users.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 561–575, 2015.
DOI: 10.1007/978-3-319-16181-5_43

(a) Present day: Individual imaging for social sharing

(b) Future: Illustration of SocialSync for social imaging

Fig. 1. (a) While the flood of mobile devices has become ubiquitous during major historical events, as seen during the election of Pope Francis, each user effectively operates independently. Image credit: Michael Sohn Associated Press; (b) Synchronizing the image capture times across mobile phones, a group of people working together will be able to capture rich information of an event, even with dynamic motion present in the scene.

which, when taken together, are of greater value than the collection of individual photographs.

1.1 Why Social Image Acquisition?

It is common to see many smartphones hoisted aloft capturing images at public events. For example, Fig. 1(a) shows St. Peter's square in the Vatican as the election of Pope Francis was announced. Mobile devices are ubiquitous throughout the square, as people take pictures and video. The sheer number of cameras at such events presents an opportunity to recover rich data about the scene, far exceeding what is available with a single camera. Applications include capturing light fields for post-capture processing, free viewpoint video, and computing depth maps for scene reconstruction and modeling.

1.2 Problem Definition

Efforts such as Photo Tourism from Snavely et al. [20] (later commercialized by Microsoft into PhotoSynth[1]) and its extension by Agarwal et al. [1] use images taken from many cameras to reconstruct a 3D model of a target. A reasonable facsimile of public objects and scenes can be rendered by scouring image aggregation and sharing sites, such as Flickr, and by using geometric constraints provided from the disparate viewpoints. Users can zoom into an object, fly around buildings, and remotely tour faraway locales. The limitation is that the scene must be static, since the images have been taken at different times. Such an

[1] www.photosynth.net

approach works well with buildings, natural monuments, and landscapes, but not so well for fast moving scenes, such as sports venues or concerts. Capturing a dynamic scene requires that cameras be synchronized to an accuracy that is a fraction of the duration of a frame.

Synchronizing consumer cameras is a challenging task, even more so for smartphone cameras. Mobile phones do not accept external hardware trigger signals and software triggers do not offer tight enough bounds to capture images simultaneously. In order for picture taking to become a communal experience, as illustrated in Fig. 1(b), a protocol for synchronizing smartphone cameras must overcome the variability caused by the camera system when triggering frame capture and delivery.

1.3 Contributions

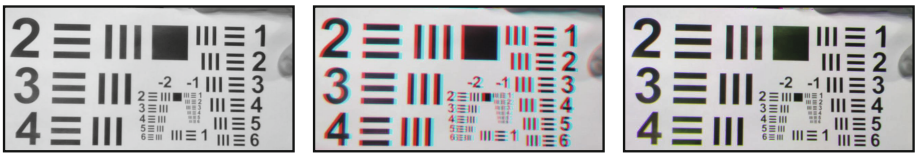
We demonstrate a protocol and highlight the necessity for highly accurate synchronization of frames across mobile devices both for indoor and outdoor dynamic scenes. To address the temporal challenges present when using mobile devices for single snapshot social image acquisition, we use the HTC One (M7) and Nexus 5 to:

1. Characterize the variables associated with relative latency that cause temporal differences between frames captured from different mobile devices. We identify the setup latency of the camera service as the main cause of variability when synchronizing frame capture.
2. Develop SocialSync, a sub-frame synchronization protocol, using additional measurements of frame rate and frame delivery to estimate the timestamp of a frame captured with millisecond accuracy. Compared to a naive synchronization implementation, where frames are aligned to the duration of a frame, our implementation achieves sub-frame alignment by requiring a duration of time to achieve synchronization before the frame capture request.
3. Demonstrate our ability to reduce motion artifacts using SocialSync when recovering the light field from a smartphone camera network. Compared to a naive synchronization implementation, SocialSync considerably reduces visible artifacts in the fused light field.

2 Background

2.1 Related Work

Multiple camera image capture: Many imaging tasks can be performed easily using multiple cameras, whether the cameras are arranged in a calibrated array or arranged randomly. For example, camera arrays can be used to capture the light field of a scene [10,23,25,26], record high speed video [18,19,24], and improve image resolution [19], while distributed cameras have been used to construct virtual cities from online photo repositories [1,20] and synthesize 3



(a) Reference image (b) Unsynchronized phones (c) Synchronized phones

Fig. 2. Motion artifacts manifest when aligning unsynchronized frame sequences. (a) A grayscale image of a planar resolution chart moving to the right taken from Fig. 3. Grayscale images from three (b) unsynchronized and (c) synchronized cameras are warped using homographies to the true depth of the moving resolution chart. The aligned images are shown as an RGB image where misaligned edges present as color artifacts. Notice that without synchronization (b) the bars in the resolution chart are misaligned by 10 pixels while the synchronized images have errors of at most 1 pixel.

dimensional models of buildings [6]. State-of-the-art snapshot light-field acquisition methods which may be used in smartphones require specialized hardware [10,14,23]. Furthermore, mask-based systems [14] reduce light throughput while camera arrays such as the PiCam [23] require hardware synchronization to ensure each element of the array captures images simultaneously. Fig. 2 highlights the need for synchronization in dynamic scenes. A planar resolution chart translates to the right in front of unsynchronized and synchronized cameras (Fig. 2(b) and Fig. 2(c) respectively). Aligning images using homographies shows that the unsynchronized images have motion artifacts of approximately 10 pixels while the synchronized cameras have error less than 1 pixel.

Using multiple cameras to capture a scene enables many benefits over single viewpoint imaging. Applications include:

Light field: Light field cameras, such as Lytro [8] and Raytrix[2], can be used for digital refocusing, but sacrifice spatial resolution. Compared to single camera techniques, various works have demonstrated light field recovery using camera arrays [5,26].

Free Viewpoint Television: Free-viewpoint television uses multiple cameras for viewing a 3D scene by changing viewpoints [21]. In addition, an array of smartphones could be used for a variety of special effects such as bullet time [25].

3D and Depth: Camera arrays are also useful when recovering 3D and depth from a scene [17,22].

2.2 Android Camera Library

The android camera library provides access to camera functions, such as locking exposure, focus, zoom, and capturing images or video on demand. By abstracting the camera utilities for the developer, the camera library hides the details of binding to the Android camera service and operating the sensor hardware. An application activates the camera by calling `startPreview()` to begin streaming a

[2] http://www.raytrix.de/

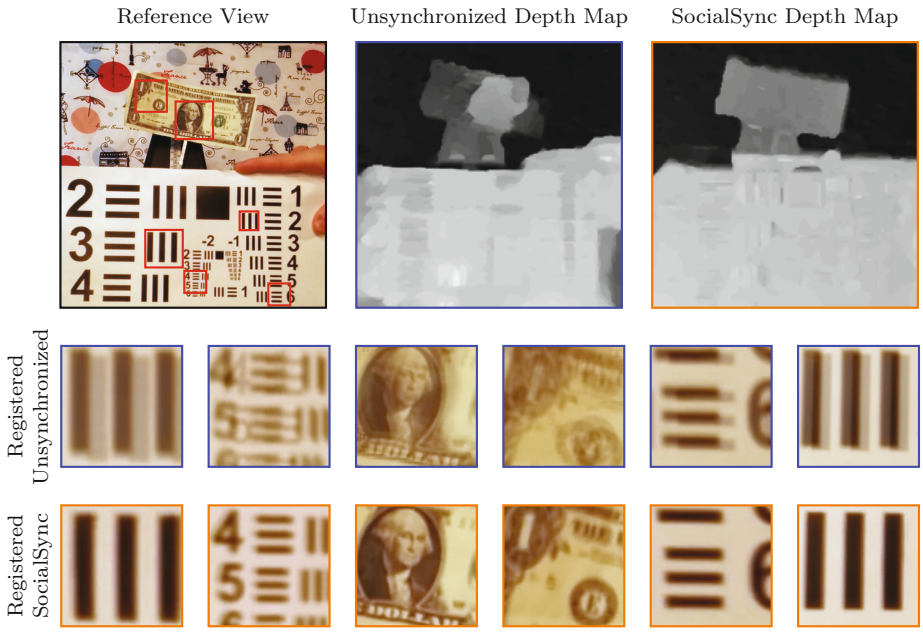


Fig. 3. Computing depth maps to register 4 cameras to a reference view. Depth estimation with unsynchronized images (top center) is challenging as the images are never truly aligned (see Fig. 2(b)). Depth estimation is more accurate when using our SocialSync protocol (top right). Outset show the average of the 4 registered images using timestamps to synchronize (middle row) and SocialSync (bottom row).

sequence of image frames. A developer can specify a callback function to trigger when a preview frame is available, either for processing or for saving to disk. Both the Nexus 5 and HTC One support a variety of preview sizes. In our setup we set both devices to capture 1920×1080 pixel images.

2.3 Time Synchronization Protocols

Due to manufacturing differences, each smartphone's system clock will drift at slightly different rates, creating misalignment between recorded timestamps. An important and well studied problem, clock synchronization achieves a consistent global time across all devices in the network. Our solution uses the Network Time Protocol (NTP) [15,16] to perform clock synchronization among devices. The maximum clock synchronization error is bounded by the round-trip time of the network. Because our WiFi access point is capable of round-trip times (RTTs) of less than 2 ms to our time server, NTP permits clocks synchronization to be within 1 ms.

2.4 Latency

A camera network's response to a request for an image capture is limited by two sources of latency:
Network Latency: Events sent between devices incur an end-to-end network latency. Our measurements demonstrated two devices sharing the same WiFi access point had a mean round trip latency of around 3 ms as well as an outlier RTT of 75 ms.
Camera I/O Latency: There is a non-deterministic latency from the time the software issues a command to take a picture and the time the hardware captures a frame due to the variables in mobile OS resource management. In our measurements, we found that the average camera I/O latency is specific to particular device models. Fig. 3 shows the necessity of compensating for I/O latency when estimating depth from independent smartphone cameras with synchronized clocks. Notice that the depth map for the unsynchronized cameras contains errors for the dynamic scene elements while the SocialSync cameras give an accurate depth map.

3 Camera Characterization

We reduce the problem of synchronizing frame capture to that of the I/O camera latency associated with triggering frame capture and delivery. Our implementation uses network clock synchronization to devices clocks and requires that requests for frame capture reach each mobile device before the capture event.

3.1 Camera Timestamps

To characterize the latency through the system, we define the following:

- **Frame Capture** $T_C(i)$**:** The time image exposure ends for the ith frame.
- **Frame Delivery** $T_D(i)$**:** The time the application receives the ith frame.
- **Camera Setup Latency** $T_C(0)$**:** The setup time to capture the 0^{th} frame.
- **Frame Rate** $(T_C(i) - T_C(i\text{-}1))^{-1}$**:** The rate of capturing consecutive frames.
- **Frame Delay** $T_D(i) - T_C(i)$**:** System delay when delivering the ith frame.

We use the mobile system timestamp on the preview callback to obtain $T_D(i)$, since preview frames in Android do not contain EXIF millisecond meta data timestamps. As the capture timestamp is not accessible through the mobile operating system, we build a characterization setup to measure $T_C(i)$.

3.2 Camera Characterization Setup

We capture the frame latency with an experimental setup in order to recover the frame capture timestamps precisely. For further details regarding our smartphone app implementation and rolling shutter measurements, we direct the interested reader to our supplementary material [12].

Characterization Smartphone App: The camera object runs on a dedicated background thread to prevent resource conflicts with the foreground activity. Auto exposure and white balance are locked, putting the camera system in a mode that enables rapid capture. To streamline memory allocation, the application pre-allocates preview frames into a circular buffer queue. The focus of each camera is fixed at infinity.

Image Timestamp from Visible Clock: To obtain a timestamp of a frame capture $T_C(i)$, we use a camera scene that includes a visible clock. For accuracy, we built an 8×8 array of LEDs, sequentially triggered at precise time intervals by a Raspberry Pi (RPi). The RPi sequentially lights each column of LEDs on the array for 1 ms. When the camera takes an image of the LED clock, the position of the illuminated LEDs on the image serve as a timestamp for the image. Because rows of pixels are read out at different times due to the rolling shutter, $T_C(i)$ indicates the time when reading the 1st row from the image sensor. Further details regarding our measurement setup for calculating rolling shutter speed and $T_C(i)$ are described in [12].

Timing Precision of the Visible Clock: The RPi acts as a global reference clock. It is synchronized via a wired GPS clock to minimize clock drift. `loopstats` in the NTP protocol reports the resulting clock jitter of the RPi as 5μ. The pre-synchronization clock drifts for the smartphones were small enough for characterization purposes, drifting less than 60μ after 1 second of elapsed time. The smartphones wirelessly synchronizes their clocks with the RPi, repeating synchronization attempts until the RTT is less than 2 ms and clock error is less than 1 ms.

3.3 Characterization Measurements

We characterize the camera setup latency, frame rate, and delay when delivering preview frames for a Nexus 5 and HTC One.

Camera Setup Latency $T_C(0)$: On Android, before capturing an image, the camera must first be activated by starting the preview image sequence. The variability in setting up the camera service, sensor, and preview sequence limits the ability to synchronize frames. By measuring the latency from launching the preview sequence to the capture of the first frame $T_C(0)$, we see launching the camera preview sequence at the same time is insufficient to achieve accurate synchronization because of the randomness in the latency. The camera setup time for a Nexus 5 has a sample mean of $\mu = 283.3$ ms and may deviate with a standard deviation of $\sigma = 9.4$ ms. The distribution shown in Fig. 4 (left) is representative of the variability in setting up image capture on a mobile device.[3]

Frame Rate $(T_C(i) - T_C(i\text{-}1))^{-1}$: Although the capture time of a frame is stochastic, the time between frames is deterministic. By knowing the time interval between image capture timestamps, all frame capture timestamps can be determined as long as one timestamp is known. The difference between subsequent capture timestamps is inversely proportional to the frame rate of the image

[3] $T_C(0)$ will vary between devices.

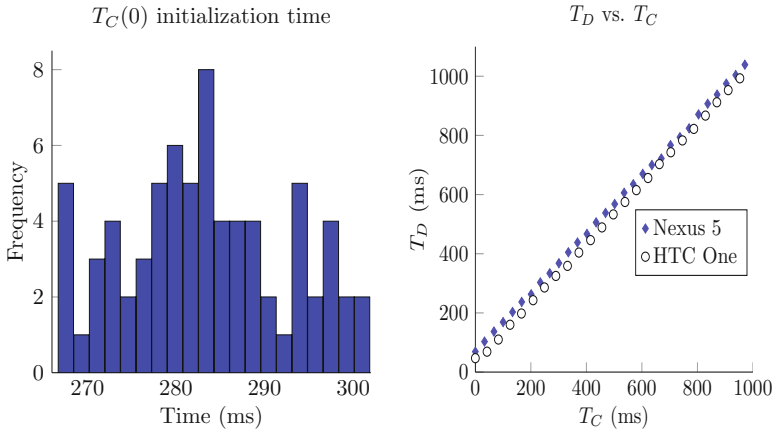


Fig. 4. (Left) Camera Setup: Android phones require an activated preview image sequence prior to capturing a photo. Therefore, frame synchronization between devices is based on the offset between setting up the camera and capture the first frame $T_C(0)$. We show that for Nexus 5 camera, simultaneous launches of the camera have a setup time with a mean of $\mu = 283\,\text{ms}$ and a standard deviation of $\sigma = 9.4\,\text{ms}$; (Right) The delivery time T_D of a frame to an application is highly correlated with its capture time T_C. The relationship between delivery time and capture time provides the basis for estimating $T_C(0)$.

sequence. Because Android devices provide various ranges for setting frame rates, in our setup we locked the frame rate to a valid range supported by the Android devices and then measured the frame rate using our LED clock. Upon locking the auto exposure, the frame rate became constant at $f = 29.8497 \pm 0.0001$ fps for a Nexus 5 and $f = 24.1513 \pm 0.0002$ fps for an HTC One.

Frame Delay $T_D(i) - T_C(i)$: For a fixed frame rate image sequence, $T_C(i)$ is highly correlated with $T_D(i)$, the time for delivering a frame to the application as shown in Fig. 4(b). By measuring latency between capturing a frame and delivering a frame, we will be able to build a model for estimating $T_C(i)$. The frame delay can be represented as a stationary stochastic variable with a normal distribution N_F whose mean $\mu_F = 36.83\,\text{ms}$ and standard deviation $\sigma_F = 4.68\,\text{ms}$ for an HTC One and $\mu_F = 66.67\,\text{ms}$ and $\sigma_F = 4.48\,\text{ms}$ for a Nexus 5.[4] The large difference between the two data sets is because the Nexus 5 passes two frames before delivering the captured frame, while the HTC One delivers the captured frame after only one frame has passed.

4 SocialSync Protocol

SocialSync achieves highly accurate synchronization across a diverse range of Android devices in a network by (1) estimating capture timestamps based on

[4] Assumption of normal distribution is valid because $\sigma_F \ll \mu_F$.

the delivery timestamps of previously delivered frames and (2) using repeated attempts at launching the preview image sequence until a set of frames is obtained for which the computed timestamps align (frames are in sync).[5]

4.1 Capture Timestamp Estimation

In single camera tasks, frames recorded by the camera are sequential and evenly spaced, specified by the frame rate. In multi-camera tasks, knowing the exact capture timestamp is required to align frames from different cameras, as the relative position of a frame from one camera is unknown with respect to the frame from a second camera. If the camera frame rates are known, then the calibration task is simplified by providing a common time origin and measuring the offset to each camera's first frame. Therefore, the precision in estimating the capture timestamp of a frame is based strictly on the estimation of $T_C(0)$, the setup capture timestamp.

For a fixed frame rate f, the time the ith frame is captured is related to the camera setup latency $T_C(0)$ according to

$$T_C(i) = T_C(0) + (1/f) \cdot i. \tag{1}$$

Let T_N be a random variable representing the frame delay following the normal distribution N_F. $T_C(i) = T_D(i) - T_N(i)$, where $T_D(i)$ provides a sample for estimating $T_C(0)$. Therefore, $T_C(0)$ can be expressed as Gaussian random variable with a distribution N_F such that

$$T_C(0) \approx T_D(i) - 1/f \cdot i - T_N(i). \tag{2}$$

Camera setup latency $T_C(0)$ is estimated by taking multiple measurements of $T_D(i)$, determining the distribution of the frame delay, and calculating the average. The timestamp of $T_C(0)$ is the center of the Gaussian frame delay distribution. A standard error calculation of $T_C(0)$ provides a method for estimating the sample mean within a desired confidence interval. Therefore, to obtain a 95% confidence interval of less than δ ms with the number of samples frames n is

$$\frac{\sigma_F}{\sqrt{n}} \cdot 1.96 \leq \delta. \tag{3}$$

Therefore, an estimate of $T_C(0)$ at a 95% confidence interval, and all subsequent capture timestamps, to within 2 ms requires the delivery of at least 22 preview frames and to within 1 ms requires the delivery of at least 85 frames for both an HTC One and Nexus 5.

4.2 Frame Synchronization Upper Bound

Camera I/O latency $\Delta T_C(i)$ is the delay between a request for a frame capture and the execution of the event at $T_C(i)$. Because each frame's capture timestamp

[5] In the protocol, we assume a global reference clock, such as one obtained using NTP.

can be estimated precisely using the results of Sec. 4.1, a mobile app can deliver the most recently captured frame $T_C(i)$ for each request. Because a periodic sequence of images has a fixed frame rate f, a captured frame $T_C(i)$ closest to the time of an arbitrary request will result in $\Delta T_C(i)$ being uniformly distributed between 0 and $\tau = 1/f$ seconds. Therefore, the upper bound synchronization error between frames from multiple devices is the frame sequence with the longest interval τ, i.e. the inverse of the lowest frame rate.

4.3 Obtaining Sub-Frame Synchronization

By estimating capture timestamps, the SocialSync protocol achieves sub-frame image capture through launching the smartphone preview image sequence stream repeatedly until frame sequences are aligned[6]. Under the hood, synchronization is achieved by estimating capture timestamps to successfully predict the image sequence frame setup time, thereby capturing a frame at a desired request time within a specified tolerance.

Suppose a user requires that the camera I/O latency $\Delta T_C(i)$ for frame capture is within the range $(0, t)$, where $t \leq \tau$. The probability the phone will fail (p_f) to capture a frame at a time within the range $(0, t)$ is $p_f = 1 - t/\tau$.

Repeated attempts at starting the image sequence would improve the odds of starting within the desired synchronization range. Using our capture timestamp estimation technique described in Sec. 4.1, we can determine whether an image sequence is in the desired synchronization range. By successfully estimating whether a given image sequence will succeed or fail, sub-frame synchronization is based on following equations:

Single Camera Sync Probability: The probability that a single phone will start the continuous image sequence in the range $(0, t)$ after k attempts is $P_k = 1 - (p_f)^k$.

Multiple Cameras Sync Probability: The probability that n phones will start the continuous image sequence in the range $(0, t)$ after k attempts is $(P_k)^n$.

Expected Number of Sync Cameras: The expected number of phones to start the continuous image sequence in the range $(0, t)$ after k attempts of n phones is nP_k.

5 Evaluation

To demonstrate the advantages of the SocialSync sub-frame synchronization protocol, we capture images of dynamic scenes and demonstrate improvements in recovering the light field by reducing motion artifacts. To reduce errors not associated with synchronization misalignment, we constrain our evaluation to a structured camera array consisting of Nexus 5 devices shown in Fig. 5. The cameras are calibrated using the Caltech calibration toolbox [2] and further refined using bundle adjustment [13].

[6] With a large number of smartphones a subset of synchronized cameras could be used without the need to restart the preview streams.

Maximum difference in capture times for synchronized smartphone cameras

	NaiveSync	SocialSync
4 Cameras	23 ms	5 ms
8 Cameras	35 ms	6 ms

Fig. 5. Camera array setup used for evaluation. (Left) Up to 9 cameras are placed in an rigid array to minimize errors not associated with scene motion. (Right) Camera synchronization timings measured in evaluation for NaiveSync (i.e. timestamp comparison) and SocialSync. SocialSync offers tighter synchronization than NaiveSync.

Reference View Depth Map Reference View Depth Map

(a) NaiveSync (b) SocialSync

Fig. 6. Eight cameras capture a dynamic indoor scene. A drinking bird provides angular motion while a toy truck translates across the scene. (a) Depth estimates of scene using the NaiveSync protocol exhibit artifacts for dynamic scene elements. (b) SocialSync achieves accurate depth map recovery including dynamic regions such as the truck window and drinking bird.

5.1 Recovering the Light Field

We use the SocialSync protocol to synchronize cameras within 6 milliseconds (shown in Fig. 5). We compare our results against a naive frame synchronization implementation (called NaiveSync), which only saves the frame with the closest delivery timestamp. We collect indoor and outdoor datasets using 8- and 4-camera arrays respectively. Depth maps recovered from the disparate views allow for post-capture refocusing. Point correspondences are computed using a plane sweep algorithm and a window-based normalized cross correlation cost function. We use the graph cuts implementation of [3,4,11] to impose a smoothness penalty between neighboring pixels and recover our depth estimates.

Indoor Scene with an 8-Camera Array: In the scene shown in Fig. 6, dynamic scene elements (the angular motion of the drinking bird and translation motion of the truck) require image synchronization to compute accurate depth maps. Using NaiveSync, which saves the frames with the closest delivery

(a) All in focus (b) Far Focus (c) Mid Focus (d) Near Focus

Fig. 7. Post-capture refocusing using the accurate depth map of Fig. 6(b) captured using SocialSync. (a) The captured image is refocused in the (b) far, (c) middle, and (d) near ground of the scene post-capture. Please view digitally to see details.

(a) Reference (b) Depth from NaiveSync (c) Depth from SocialSync (d) Refocus with NaiveSync (e) Refocus with SocialSync

Fig. 8. SocialSync provides advantages in dynamic outdoor scenes. Seven phones are divided into two groups of 4 phones with one overlapping phone. One group uses our SocialSync protocol and the other group uses with NaiveSync. (a) Reference view of two people tossing a stuffed toy. (b) The depth recovered using NaiveSync has motion artifacts not present when (c) computing the depth using SocialSync. (d) Proper post-capture refocusing cannot be achieved with NaiveSync. Notice that the thrower's face and shorts are incorrectly blurred when focusing on the thrower's body. (e) SocialSync allows for accurate blurring for the thrower's entire body.

timestamps, results in synchronization of 35 ms while our SocialSync protocol reduces the error to 6 ms. The two data sets are captured independently. Note that the depth map recovered when using SocialSync, Fig. 6(b), is free of the artifacts present when using NaiveSync, Fig. 6(a). In particular, dynamic scene elements such as the drinking bird and the truck's wheels and window remain blurred when using NaiveSync.

The accurate depth map provided by using SocialSync in Fig. 6(b) allows users greater artistic license when viewing captured images. Fig. 7 shows the indoor scene refocused post-capture on the near, middle, and far planes.

Outdoor Scene with a 4-Camera Array: Figure 8 shows a scene taken outdoors of two people throwing a toy bird. Seven cameras captured the scene with one chosen as a reference camera. Four cameras were synchronized using SocialSync (including the reference) while the remaining three are unsynchronized with respect to each other and the reference. The four SocialSync cameras are synchronized to within 5 ms while the four NaiveSync cameras have a 23 ms spread. Note that the depth map recovered from the SocialSync cameras (Fig. 8(c)) accurately captures the depth of the scene while the depth computed using the NaiveSync cameras (Fig. 8(b)) has many artifacts. Fig. 8(d) highlights the inability to refocus on the thrower properly when using the NaiveSync depth map, while refocusing using SocialSync (Fig. 8(e)) has no such limitation.

6 Conclusions

Our work highlights and addresses the sub-frame synchronization challenge when using smartphones for multi-viewpoint light field recovery. Without sub-frame synchronization between mobile devices, light field acquisition is limited to static scenes due to motion artifacts caused by frame misalignment. As the first step towards multi-viewpoint image capture of dynamic scenes using smartphone camera networks, we characterized the camera setup, frame rate, and frame delay on an HTC One and Nexus 5. Next, we introduced SocialSync, a sub-frame synchronization protocol, based on an estimation of frame capture timestamps. Finally, we evaluated the benefit of using SocialSync by comparing it to the best existing smartphone camera synchronization method and demonstrating improvements in depth map estimation and digital refocusing.

As a limitation, sub-frame synchronization of smartphone cameras is only effective for capturing a single snapshot or a few frames, due to variability in frame rates caused by clock drift and manufacturing quality. Furthermore, due to the stochastic nature of synchronization, increasing the number of devices requires more synchronization attempts. Therefore, as future work to address scalability issues with large social events, we would explore methods for grouping subsets of smartphones, which would be naturally synchronized within the group.

Acknowledgments. We would like to thank the LF4CV reviewers and Robert LiKamWa for useful discussions regarding this work. The authors were partially supported by NSF Grants CNS 1012921, CNS 1161596, IIS 1116718, and CCF 1117939.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)
2. Bouguet, J.Y.: Camera calibration toolbox for matlab (2008)
3. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(9), 1124–1137 (2004)

4. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11), 1222–1239 (2001)
5. Buehler, C., Bosse, M., McMillan, L., Gortler, S., Cohen, M.: Unstructured lumigraph rendering. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 425–432. ACM (2001)
6. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 11–20. ACM (1996)
7. Facebook, Ericsson, Qualcomm: A focus on efficiency. Tech. rep. (September 2013) http://internet.org white paper
8. Georgiev, T., Yu, Z., Lumsdaine, A., Goma, S.: Lytro camera technology: theory, algorithms, performance analysis. In: IS&T/SPIE Electronic Imaging, pp. 86671J–86671J. International Society for Optics and Photonics (2013)
9. Gupta, A., Cozza, R., Lu, C.: Market share analysis: Mobile phones, worldwide, 4q13 and 2013. Tech. rep., Gartner, Inc. (February 2014) (white paper)
10. Heptagon Advanced Micro Optics. http://www.hptg.com/products/imaging (2014), (Online accessed March 31, 2014)
11. Kolmogorov, V., Zabin, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(2), 147–159 (2004)
12. Latimer, R., Holloway, J., Veeraraghavan, A., Sabharwal, A.: Supplementary material for SocialSync: Sub-frame synchronization in a smartphone camera network (2014), Computer Vision-ECCV 2014. LF4CV submission. Supplied as additional material
13. Lourakis, M.A., Argyros, A.: SBA: A software package for generic sparse bundle adjustment. ACM Trans. Math. Software **36**(1), 1–30 (2009)
14. Marwah, K., Wetzstein, G., Bando, Y., Raskar, R.: Compressive light field photography using overcomplete dictionaries and optimized projections. ACM Transactions on Graphics (TOG) **32**(4), 46 (2013)
15. Mills, D.L.: Network time protocol (ntp). Network (1985)
16. Mills, D.L.: Computer Time Synchronization: The Network Time Protocol on Earth and in Space, 2 edn. CRC Press (2010)
17. Naemura, T., Tago, J., Harashima, H.: Real-time video-based modeling and rendering of 3d scenes. IEEE Computer Graphics and Applications **22**(2), 66–73 (2002)
18. Nayar, S., Ben-Ezra, M.: Motion-based motion deblurring. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(6), 689–698 (2004)
19. Shechtman, E., Caspi, Y., Irani, M.: Space-time super-resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(4), 531–545 (2005)
20. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings, pp. 835–846. ACM Press, New York (2006)
21. Tanimoto, M.: Overview of free viewpoint television. Signal Processing: Image Communication **21**(6), 454–461 (2006)
22. Tsai, R.Y.: A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. IEEE Journal of Robotics and Automation **3**(4), 323–344 (1987)

23. Venkataraman, K., Lelescu, D., Duparré, J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R., Nayar, S.: Picam: an ultra-thin high performance monolithic camera array. ACM Transactions on Graphics (TOG) **32**(6), 166 (2013)
24. Wilburn, B., Joshi, N., Vaish, V., Levoy, M., Horowitz, M.: High-speed videography using a dense camera array. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2, pp. II-294. IEEE (2004)
25. Wilburn, B., Joshi, N., Vaish, V., Talvala, E.V., Antunez, E., Barth, A., Adams, A., Horowitz, M., Levoy, M.: High performance imaging using large camera arrays. ACM Transactions on Graphics (TOG) **24**(3), 765–776 (2005)
26. Zhang, C., Chen, T.: A self-reconfigurable camera array. In: ACM SIGGRAPH 2004 Sketches. p. 151. ACM (2004)

Depth and Arbitrary Motion Deblurring Using Integrated PSF

Takeyuki Kobayashi(✉), Fumihiko Sakaue, and Jun Sato

Department of Computer Science and Engineering,
Nagoya Institute of Technology, Nagoya, Japan
takeyuki-kobayashi@kabuata.com

Abstract. In recent years, research for recovering depth blur and motion blur in images has been making a significant progress. In particular, the progress in computational photography enabled us to generate all-in-focus images and control depth of field in images. However, the simultaneous recovery of depth and motion blurs is still a big problem, and recoverable motion blurs are limited.

In this paper, we show that by moving a camera during the exposure, the PSF of the whole image becomes invariant, and motion deblurring and all-in-focus imaging can be achieved simultaneously. In particular, motion blurs caused by arbitrary multiple motions can be recovered. The validity and the advantages of the proposed method are shown by real image experiments and synthetic image experiments.

Keywords: Coded imaging · PSF · Deblur · Motion Blur · All-in-Focus

1 Introduction

Deblurring depth and motion blurs is very important in many applications. In order to deblur depth blur and motion blur, various methods have been studied in recent years. Many methods use specific models of PSF (Point Spread Function) for representing the blur. By using the PSF, blurred images can be represented by convolution of the PSF and the sharp (all-in-focus) images. Thus, deblurring of the image can be achieved by deconvolution of the PSF. However, the PSF is in general not unique for a whole image, since the PSF depends on the depth of objects. In order to suppress the variation of PSFs, some methods based on the light field computation were proposed[1,4] in recent years. Although these methods can deblur observed images with various depth, we need to obtain multiple images which are captured under different blurring conditions.

The image deblurring has also been studied in computational photography in recent years[8,9]. Veeraraghavan et al.[9] proposed the coded aperture for image deblurring. They focused on the zero-cross in the frequency characteristics of PSF in coded aperture, and optimized the coded aperture by decreasing the zero-cross. Nagahara et al.[6] proposed focus sweep imaging for expanding the depth of field. In their method, the image sensor (image plane) in a camera moves during exposure. By this movement, the PSF on an image plane becomes

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 576–587, 2015.
DOI: 10.1007/978-3-319-16181-5_44

approximately invariant under change in depth. Thus, we can deblur observed images easily by using a single PSF all over the image. However, we need to move the image sensor in a camera device quickly, and thus it is not easy to implement by using ordinary camera systems.

The motion blur occurred by relative motions between cameras and objects has also been studied. Raskar et al.[7] proposed coded exposure for deblurring the motion blur accurately. They proposed a method for controlling a shutter during exposure, i.e. coded exposure. By using the coded exposure, the quality of deblurred images can be improved. However, the obtained images become darker, since the exposure time becomes a half of the original exposure time, and the S/N ratio of obtained image becomes worse. Furthermore, we have to obtain the image motion beforehand in order to optimize the coded exposure. Levin et al.[5] showed that the PSF of motion blur becomes invariant under image motions, if the camera moves along with a parabolic orbit. Although the method works well when we know the orientation of the object motions, arbitrary unknown motions cannot be deblurred. Cho et al.[3] proposed an imaging technique which enables us to obtain invariant motion blurs under arbitrary 2D image motions and deblur them. However, we need to obtain two different images moving the camera with parabolic motions in two orthogonal directions. Bando et al.[2] proposed a method for estimating motion blur by using circular motion of image sensor. Although we can estimate PSF of motion blur by using this method, the method cannot be applied when we have complex motions in the scene.

In this paper, we propose a method for deblurring depth and motion blurs simultaneously. In particular we propose a method for deblurring motion blurs caused by complex multiple motions of objects. In this research, we use the focus sweep technique proposed by Nagahara et al.[6], and show that we can recover not only depth blurs but also motion blurs simultaneously by using the focus sweep technique. We also show that it enables us to deblur not only a single motion in images, but also arbitrary multiple motions in images simultaneously. Furthermore, we clarify the condition of deblurring the mixture of depth and motion blurs in images, which is very useful for designing the imaging systems. The method is tested by using real images and synthetic images generated by a lens simulator.

2 Lens Model

In this paper, we first consider a bilateral telecentric lens in order to simplify the explanation of our method, and then generalize it to ordinary perspective lens systems.

We first explain the characteristics of a bilateral telecentric lens, which is shown in Fig. 1. The focal lengths of lens 1 and lens 2 are f_1 and f_2, and these lenses are placed at their respective focal distances from the aperture to form a bilateral telecentric lens system. Let a be the diameter of aperture, p be the distance between the image sensor and Lens 2, u be the distance between the object and Lens 1. Then, all the incident lights are concentrated at point A,

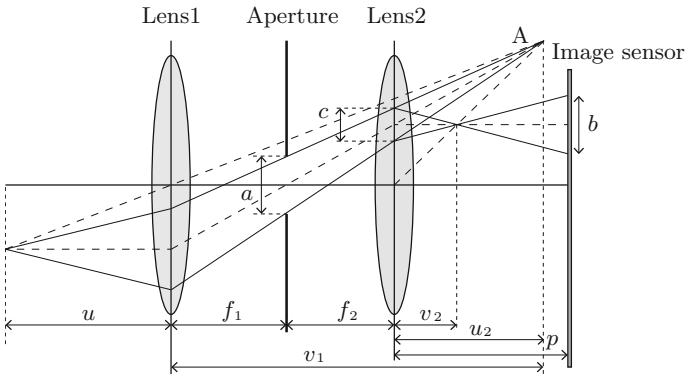


Fig. 1. Bilateral telecentric lens

whose distance is v_1 from Lens 1, and u_2 from Lens 2, and are finally concentrated at a point whose distance is v_2 from Lens 2 as shown in Fig. 1. Then, the following equations hold for these lenses.

$$\frac{1}{f_1} = \frac{1}{u} + \frac{1}{v_1} \tag{1}$$

$$\frac{1}{f_2} = -\frac{1}{u_2} + \frac{1}{v_2} \tag{2}$$

By using these equations and geometric relationships shown in Fig. 1, we have the following equation, which shows the diameter of a blurred circle b introduced by the lens system.

$$b = a\left|\frac{f_2 u}{f_1^2} + \frac{p}{f_2} - \frac{f_2}{f_1} - 1\right| \tag{3}$$

3 IPSF

By using the telecentric lens model shown in the previous section, we next consider the PSF of an image under focus sweep imaging. In this method, image plane moves along with light axis during exposure, and thus, observed PSF can be described by the integration of PSF which changes according to the image plane motion. In this paper, we call the integrated PSF as IPSF following [6].

Let us consider the case where a 3D point $\mathbf{X}$ is projected to $\mathbf{m}$ in the image. If we have image blur, the point in the image is spread, and the observed intensity at $\mathbf{x} = [x, y]^\top$ can be described by the pill box function as follows:

$$P(r, u, p) = \frac{4}{\pi b^2}\Pi\left(\frac{r}{b}\right) \tag{4}$$

where r is the distance between $\mathbf{x}$ and $\mathbf{m}$, and b denotes the radius of the image blur. Note, the radius b is determined by u and p, which are the distance between

the lens and the object, and the distance between the lens and the image sensor, as shown in Eq.(3). The function $\Pi(w)$ is a pill box function, which is described as follows:

$$\Pi(w) = \begin{cases} 1, & |w| < \frac{1}{2} \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

The Eq.(4) indicates that observed PSF depends on the distance u, and thus, the PSF is not unique for whole image because the depth u changes pixel by pixel in general scene.

Now, let us consider the case where the distance u or p changes linearly during image exposure. Let us denotes the distance u at time t as $u(t)$, and p at time t as $p(t)$. Then, the integrated PSF (IPSF) can be described as follows:

$$IP(r) = \int_{T_1}^{T_2} P(r, u(t), p(t))dt \tag{6}$$

where T_1 denotes shutter opened time and T_2 denotes shutter closed time, and thus the exposure time is $T = T_2 - T_1$.

For example, if the camera translates along with light axis with a uniform speed s_u, then the changes of distance u can be described as follows:

$$u(t) = u_0 + s_u t \tag{7}$$

where u_0 indicate a distance from the camera to the object at $t = 0$. In this case, the change in size of blur is constant, and thus Eq.(3) can be rewritten as follows:

$$b(t) = |2s_b t| \tag{8}$$

where, s_b denotes the speed of the change in radius of blur.

4 Invariance of IPSF Under 3D Motions

We next show that IPSF under focus sweep imaging is invariant against speed, direction and depth if some conditions are satisfied. In this section, we analyze the characteristics of IPSF and derive the conditions in which the IPSF becomes invariant.

Let us consider the case where a moving 3D point is projected onto the image plane. The 3D point at t is denoted by $\mathbf{X}(t)$ and the projected point is denoted by $\mathbf{m}(t) = [u, v]^\top$. In this case the PSF of a projected point at t can be described as follows:

$$P(\mathbf{x}, u, p) = \frac{4}{\pi b(t)^2} \Pi \left(\frac{||\mathbf{x} - \mathbf{m}(t)||}{b(t)} \right) \tag{9}$$

In this equation, we described the PSF by using the observed point $\mathbf{x}$ and the center of blur $\mathbf{m}(t)$, since the center of blur $\mathbf{m}(t)$ also moves in time. From the

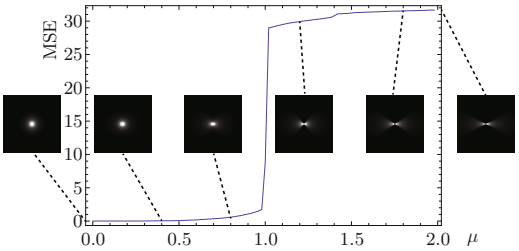


Fig. 2. The relationship between IPSF and the speed ratio μ

integration of this function with respect to t, the IPSF can be computed as follows:

$$IP(\mathbf{x}) = \int_{T_1}^{T_2} \frac{4}{\pi b(t)^2} \Pi \left(\frac{||\mathbf{x} - \mathbf{m}(t)||}{b(t)} \right) dt \tag{10}$$

Suppose the motion of the projected point $\mathbf{m}(t)$ can be described linearly by using its speed s_m and direction $\mathbf{v}$ on the image sensor. Then, the motion of the point $\mathbf{m}(t)$ can be described as follows:

$$\mathbf{m}(t) = s_m t\, \mathbf{v} \tag{11}$$

Now, if the camera moves linearly with a speed s_u along with the light axis, the size of blur changes linearly as shown in Eq.(8). Thus, the IPSF can be computed from Eq.(8), Eq.(11) and Eq.(10) as follows:

$$IP(\mathbf{x}) = \frac{1}{\pi s_b^2} \left\{ \lambda_0 \left(\frac{1}{|t_0|} - \frac{2}{T} \right) + \lambda_1 \left(\frac{1}{|t_1|} - \frac{2}{T} \right) \right\} \tag{12}$$

where t_0 and $t_1 (|t_1| > |t_0|)$ are the solutions of a quadratic equation $|\mathbf{x} - s_m t\, \mathbf{v}|^2 = s_b^2 t^2$, and λ_0 and λ_1 are variables which are described as follows:

$$t_{0,1} = \frac{-s_m \mathbf{x} \cdot \mathbf{v} \pm \sqrt{s_m^2 (\mathbf{x} \cdot \mathbf{v})^2 + |\mathbf{x}|^2 (s_b^2 - s_m^2)}}{s_b^2 - s_m^2} \tag{13}$$

$$\lambda_0 = \begin{cases} 1, & |t_0| < \frac{T}{2} \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

$$\lambda_1 = \begin{cases} 1, & (|t_1| < \frac{T}{2}) \wedge (t_0 t_1 < 0) \\ -1, & (|t_1| < \frac{T}{2}) \wedge (t_0 t_1 > 0) \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

If the exposure time of camera is sufficiently small, arbitrary motions can be approximated by linear motions. Thus, Eq.(12) covers all the motions.

By using the IPSF model described in Eq.(12), we next consider the relationship between the IPSF and a speed ratio μ. The speed ratio μ is defined as follows:

$$\mu = \left| \frac{s_m}{s_b} \right| \tag{16}$$

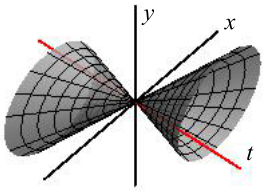


Fig. 3. Changes in PSF ($\mu = 0.8$)

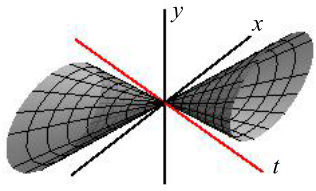


Fig. 4. Changes in PSF ($\mu = 1.4$)

and it represents the relative speed between the speed of projected point s_m and the speed of the change in radius of blur s_b.

Fig. 2 shows the relationship between the speed ratio μ and the change in IPSF, that is the difference between the IPSF of a static object and the IPSF of a moving object. Fig. 2 shows that the IPSF is almost unchanged when μ is smaller than 1, although it changes drastically when μ is larger than 1. Furthermore, the IPSF is almost isotropic when μ is smaller than 1, while it is unisotropic when μ is larger than 1. Thus, we find that the IPSF can be considered as invariant when the following condition holds:

$$\mu \leq 1 \tag{17}$$

Thus, if this condition holds, we can deblur image blurs caused by arbitrary motions just by deblurring with a uniform IPSF all over the image.

Let us consider the reason why invariance breaks when the speed ratio μ becomes larger than 1. The weight of PSF at each time in IPSF is not equivalent in general. For example, the value of $P(\mathbf{x}, u, p)$ becomes extremely large when the target object is at the focus position. On the other hand, the value of $P(\mathbf{x}, u, p)$ becomes very small when the target object is far from the focus position. Thus, the IPSF heavily depends on the PSF at focus position. When object speed s_m is smaller than the speed of blur s_b, the changes in PSF (which is pill box function) can be described as shown in Fig. 3. In this case, the time axis t (which is the center of blur) is in a cone which represents the PSF, and thus, the center of blur is always in the PSF during the motion. Thus, IPSF is approximately invariant, even if the motions are different. However, if s_m becomes larger than s_b, the time axis t is out of the PSF cone as shown in Fig. 3. Thus, the IPSF changes drastically depending on the motions. As a result, Eq.(17) is the critical condition for the invariance of IPSF.

5 Invariant IPSF By Using Ordinary Lens

We next generalize our analysis into ordinary perspective lens from bilateral telecentric lens. When we use ordinary lenses for image projection, the position of projected points depends on not only horizontal and vertical positions of object, but also the depth u of object. Therefore, if the image plane or the whole

camera moves along with the light axis, projected points also moves even if the 3D point is static.

However, the motion of a projected point by using the ordinary lens can be regarded as a radial motion of target object under telecentric lens. Thus, the IPSF becomes invariant, if the radial motion satisfies the condition described in the previous section.

Let us describe the image motion of the projected point caused by the change in depth u by using the direction $\mathbf{u}$ and the speed s_z. Then, the motion of the projected point can be described by the summation of the motion $s_z\mathbf{u}$ caused by the change in depth and the motion $s_m\mathbf{v}$ caused by the object motion as follows:

$$s_a\mathbf{w} = s_m\mathbf{v} + s_z\mathbf{u} \tag{18}$$

where, $\mathbf{w}$ denotes the direction of the combined motion, and s_a denotes the speed of the combined motion.

Now, we define the speed ratio μ as follows:

$$\mu = \left| \frac{s_a}{s_b} \right| \tag{19}$$

Then, the focus point is always in the blur circle, when the following condition holds:

$$\mu \leq 1 \tag{20}$$

Thus, we find that the IPSF is invariant under ordinary perspective lens systems, when Eq.(20) is satisfied, and we can deblur images by using a uniform IPSF, even if we have arbitrary multiple motions in the scene.

6 Experimental Result Using Synthesized Images

In this section, we evaluate the proposed method by using synthetic images. We made a lens simulator which can simulate arbitrary lens systems, and generated synthetic images of objects under the focus sweep. The object is a planar surface, and some characters are printed on this plane. For comparison, images taken by the method proposed by Levin et al.[5] were also synthesized. The synthesized images are shown in Fig. 5 and Fig. 6. These images were synthesized under various motions of object surface, i.e. (a) static, (b) horizontal motion, (c) vertical motion, (d) diagonal motion, (e) rotational motion and (f) zoom. In the proposed method, the camera moved 11mm during exposure. The telecentric lens was used and its focus lengths are $f_1 = 60$ mm and $f_2 = 60$ mm respectively. The size of aperture was $a = 10$ mm. In Levin's method, the camera moved in horizontal parabolic orbit. The Gaussian noises with the std of $\sigma = 0.1$ were added to each image intensity. The maximum value of the speed ratio μ was (a) 0, (b) 0.45, (c) 0.46, (d) 0.65, (e) 0.98 and (f) 0.86, thus all of them satisfied the proposed condition. Wienner filter was used for deconvolution in both methods. The PSNR of deblurred images were computed and they are represented under the each result.

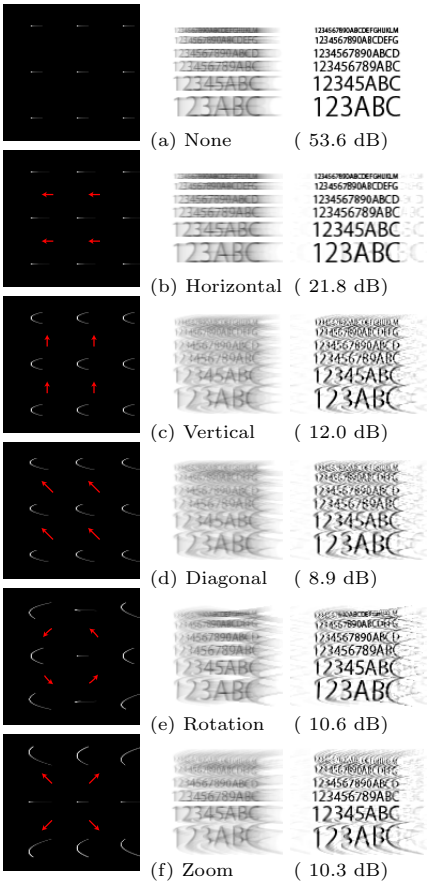


Fig. 5. Images deblurred by Levin's method: IPSF (left), observed images (center) and deblurred results (right)

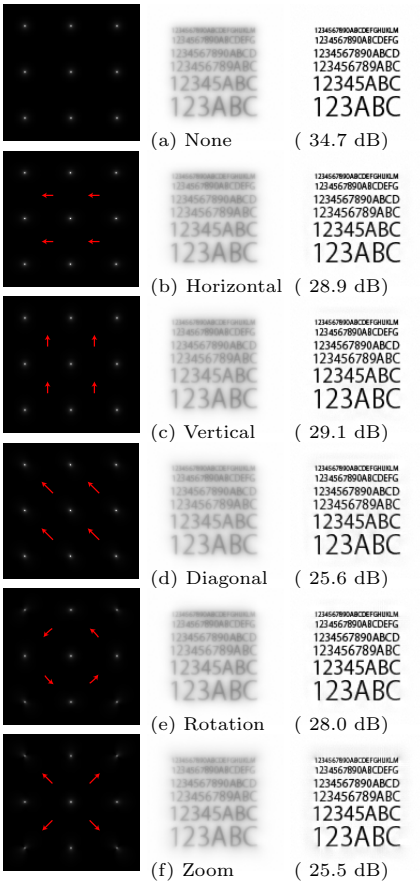


Fig. 6. Image deblurred by the proposed method: IPSF (left), observed images (center) and deblurred result

As shown in Fig. 5, although Levin's method provides us good results in horizontal motion, it does not work well in other motions. This is because the horizontal parabola was used in Levin's method, and it cannot deblur images under other motions. In contrast, the proposed method provides us very good results under all the motions as shown in Fig. 6. In particular, change of scale and rotation, which cannot be deblurred properly by the ordinary deblurring methods can be deblurred properly by our method. From these results, we find that the proposed method can deblur arbitrary unknown motions when the condition described in Eq.(20) holds.

We next evaluate the robustness of the proposed method against changes in object speed. In this experiment, the speed ratio μ was changed from 0 to 2.0, and observed images were deblurred by the proposed method. Fig. 7 shows the

(a) μ=0 (b) μ=0.5 (c) μ=0.9 (d) μ=1.1 (e) μ=1.5 (f) μ=2.0

Fig. 7. Image deblurring results under various speed ratio μ. When μ is smaller than 1.0, images can be deblurred properly.

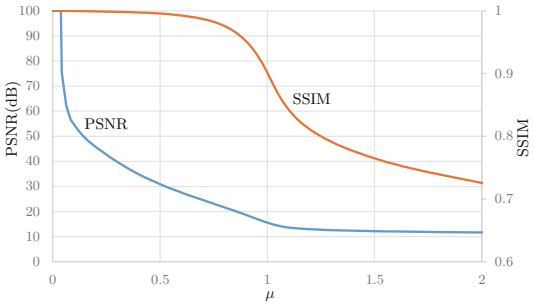


Fig. 8. Relationship between speed ratio and PSNR/SSIM of deblurred images

deblurred images under each speed ratio μ. As shown in Fig. 7, the image blur was recovered properly in (a), (b) and (c), while it was not recovered properly in (d), (e) and (f). This is because the speed ratios of (d), (e) and (f) are over 1.0, and they do not satisfy the condition for image deblurring. These results show that the deblurring condition derived in this paper is valid.

Fig. 8 shows the relationship between the speed ratio μ and the accuracy of deblurring. In this figure, the vertical axis shows the PSNR and SSIM[10] of deblurred images. This figure shows that the quality of image deblurring depends on the speed ratio. In particular, SSIM of deblurred images changes drastically at around $\mu = 1.0$, and thus, we find that the limitation of μ exists at around 1.0.

7 Experimental Results by Using Real Devices

We next show experimental results from real camera systems. We first show results when we used telecentric lens for the camera system. The camera system and a target object were fixed on a translation/rotation stage as shown in Fig. 9. The target object was moved horizontally, vertically and rotationally. The camera system was moved in the proposed method, and the speed of the motion was $s_u = 300$ mm/sec. The exposure time of the camera was 0.5 sec. For comparison, the camera was moved according to Levin's method and took images. The zoom

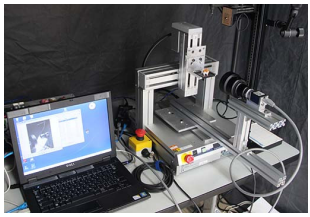
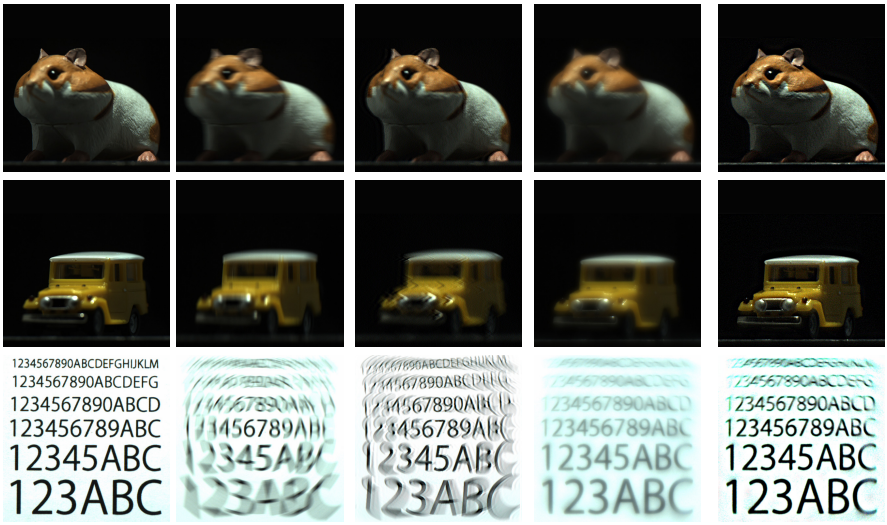


Fig. 9. The camera system for obtaining images



(a) No motion (b) With motion (c) Levin's method (d) IPSF image (e) Deblurred image

Fig. 10. Experimental results by using a real camera system: Target objects were moved horizontally (top row), vertically (middle row) and rotationally (bottom row)

of the telecentric lens was 0.17, W.D = 113 mm, depth of field is 11 mm and F=4.0. The size of CCD of the camera was 1/2".

Fig. 10 shows the observed images and the deblurred images. (a) shows the observed images from fixed cameras and fixed objects, (b) shows the images of a moving objects taken from a fixed camera, (c) shows the deblurred result by Levin's method, (d) shows the IPSF images taken by a moving camera and (e) shows the deblurred result by the proposed method. Note, the depth of the target object is larger than 11 mm, which is the depth of field of the camera, and thus, some pixels have depth blur even if the target object is static. In the results of Levin's method shown in (c), although the horizontal motion blur could be recovered, the depth blur remains in the image. In addition, vertical and rotational motion blur could not be recovered. In contrast, the proposed method could deblur properly in any motion blurs as long as the blur satisfies the

(a) Experimental system (b) IPSF image (c) Deblurred image (d) Normal imaging

Fig. 11. Experimental results by using the ordinary perspective lens. (a) shows translation stage used for obtaining images, (b) is the observed IPSF image, (c) is the deblurred image derived from the proposed method and (d) is the image taken by the ordinary camera system.

condition. Furthermore, the depth blur could also be recovered by the proposed method.

We next show deblurring results when we used an ordinary perspective lens for a camera system. In this case, our method can deblur image when Eq.(20) is satisfied. In this experiment, the camera system was moved by using a translation stage as shown in Fig.11 (a), and the IPSF images were obtained by using the moving camera system. The example of the observed image is shown in Fig. 11 (b) and the deblurred result is shown in (c). For comparison, the normal exposure image is shown in Fig. 11. As shown in this image, the proposed method can deblur images, even if they include many different motion blurs. The result indicates that our proposed method can deblur arbitrary motion blurs as long as the blur satisfies the condition Eq.(20).

8 Conclusion

In this paper, we proposed a method for obtaining invariant IPSF image by using camera motion during exposure. By using the method, the IPSF of the image becomes uniform, and thus, we can deblur whole image by using a single IPSF. Furthermore, we analyze properties of the IPSF and derived the condition for obtaining invariant IPSF image. The method can apply not only a camera with a telecentric lens, but also an ordinary perspective lens. The experimental results show that the proposed method can deblur not only arbitrary motion blur but also depth blur.

References

1. Levin, A., Durand, F.: Linear view synthesis using a dimensionality gap light field prior. IEEE CVPR, pp. 1831–1838 (2010)
2. Bando, Y., Chen, B.Y., Nishita, T.: Motion deblurring from a single image using circular sensor motion. Computer Graphics Forum (Proceedings of Pacific Graphics) **30**(7), 1869–1878 (2011)

3. Cho, T., Levin, A., Durand, F., Freeman, W.T.: Motion blur removal with orthogonal parabolic exposures. IEEE International Conf. on Computational Photography (ICCP) pp. 1–8 (2010)
4. Kodama, K., Mo, H., Kubota, A.: Simple and fast all-in-focus image reconstruction based on three-dimensional/two-dimensional transform and filtering. IEEE ICASSP **1**, 769–772 (2007)
5. Levin, A., Sand, P., Cho, T., Durand, F., Freeman, W.: Motion-invariant photography. ACM Trans. Graphics, SIGGRAPH (2008)
6. Nagahara, H., Kuthirummal, S., Zhou, C.Y., Nayar, S.K.: Flexible Depth of Field Photography. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 60–73. Springer, Heidelberg (2008)
7. Raskar, R., Agrawal, A., Tumblin, J.: Coded exposure photography: Motion deblurring using fluttered shutter. ACM Trans. Graphics **25**, 795–804 (2006)
8. Subbarao, M., Surya, G.: Depth from defocus: A spatial domain approach. International Journal of Computer Vision **13**(3), 271–294 (1994)
9. Veeraraphavan, A., Raskar, R., Agrawal, A., Mohan, A., Tumblin, J.: Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. ACM Trans., Graphics (2007)
10. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)

Acquiring 4D Light Fields of Self-Luminous Light Sources Using Programmable Filter

Motohiro Nakamura[1], Takahiro Okabe[1(✉)], and Hendrik P.A. Lensch[2]

[1] Kyushu Institute of Technology, Fukuoka Prefecture, Japan
okabe@ai.kyutech.ac.jp
[2] Tübingen University, Tübingen, Germany

Abstract. Self-luminous light sources in the real world often have non-negligible sizes and radiate light inhomogeneously. Acquiring the model of such a light source is highly important for accurate image synthesis and understanding. In this paper, we propose a method for measuring 4D light fields of self-luminous extended light sources by using a liquid crystal (LC) panel, *i.e.* a programmable filter and a diffuse-reflection board. The proposed method recovers the 4D light field from the images of the board illuminated by the light radiated from a light source and passing through the LC panel. We make use of the feature that the transmittance of the LC panel can be controlled both spatially and temporally. The proposed method enables us to utilize multiplexed sensing, and therefore is able to acquire 4D light fields more efficiently and densely than the straightforward method. We implemented the prototype setup, and confirmed through a number of experiments that the proposed method is effective for modeling self-luminous extended light sources in the real world.

Keywords: Self-luminous light source · Extended light source · 4D light field · Programmable filter · Multiplexed sensing

1 Introduction

The appearance of an object depends not only on the geometric and photometric properties of the object but also on light sources illuminating the object. Therefore, acquiring the models of self-luminous light sources is highly important in the fields of computer graphics and computer vision, in particular for photorealistic image synthesis and accurate image-based modeling.

Conventionally, in the field of computer vision, ideal light sources such as directional light sources (point light sources at infinity) and isotropic point light sources are mostly assumed for photometric image analysis. Unfortunately, however, this is not the case; an object of interest is illuminated by nearby light sources, and more importantly, self-luminous light sources in the real world often have nonnegligible sizes, *i.e.* they are considered to be extended light sources and radiate light inhomogeneously. This means that the illumination distribution seen from a point on an object surface varies over the surface. Therefore,

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 588–599, 2015.
DOI: 10.1007/978-3-319-16181-5_45

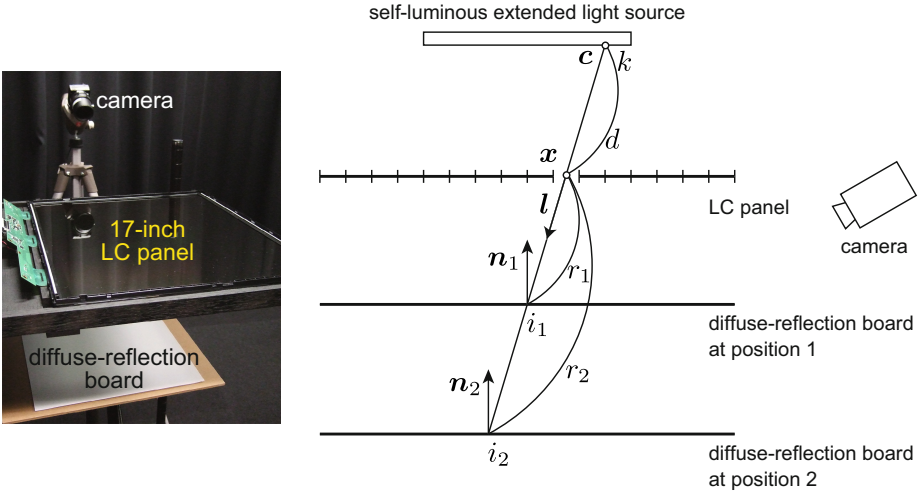


Fig. 1. Our proposed setup consisting of an LC panel and a diffuse-reflection board (left) and the sketch of its cross section (right)

in order to analyze the shading observed on an object surface under real-world extended light sources, we need to acquire the radiant intensity distributions of the light sources.

The difficulty in acquiring the radiant intensity distribution of an extended light source, which is described by a 4D light field [1], is that we need to measure a wide range of the light field. Note that consumer light field cameras are not suitable for such a purpose because their measurement ranges are limited. To cope with this problem, Goesele *et al.* [4] propose a setup consisting of a static optical filter and a diffuse-reflection board, and demonstrate the effectiveness of the setup for modeling extended light sources.

In this paper, we propose a method for acquiring the radiant intensity distribution of a self-luminous extended light source in the real world by using an LC panel, *i.e.* a programmable filter and a diffuse-reflection board as shown in Fig. 1. The key idea of the proposed method is to make use of the feature that the transmittance of the LC panel can be controlled both spatially and temporally. Specifically, the proposed method changes the transmittance patterns of the LC panel dynamically, and recovers the 4D light field from the images of the board illuminated by the light radiated from a light source and passing through the LC panel. In particular, the proposed method utilizes multiplexed sensing [10,11,15], which is a well-known technique for increasing signal-to-noise ratio (SNR) without increasing measurement time, and acquires 4D light fields more efficiently and densely than the straightforward method.

We implemented the prototype setup, and confirmed through a number of experiments that the proposed method can increase the SNR of the acquired images from which the 4D light field of a self-luminous extended light source is computed. In other words, the proposed method can acquire the models of

self-luminous light sources in the real world more efficiently and densely than the straightforward method. The main contribution of this paper is to demonstrate that the proposed method using a programmable filter is effective for modeling self-luminous extended light sources in the real world.

The rest of this paper is organized as follows. We briefly summarize related work in Section 2. A method for acquiring 4D light fields of self-luminous extended light sources by using a programmable filter and a diffuse-reflection board is proposed in Section 3. We report the experimental results in Section 4 and present concluding remarks in Section 5.

2 Related Work

Existing techniques can be classified into 3 categories; (i) techniques for acquiring 2D radiant intensity distributions of self-luminous point light sources, (ii) techniques for acquiring 4D light fields of self-luminous extended light sources, and (iii) techniques for acquiring 4D light fields of general scenes. In this section, we briefly explain the existing techniques in each category, and then describe the relationship between those techniques and our proposed method.

2D radiant intensity distributions of self-luminous point light sources

Since the size of a point light source is negligible, the radiant intensity distribution of a self-luminous point light source is described by a 2D function, *i.e.* a function with respect to the direction seen from the center of the point light source. Verbeck and Greenberg [14] propose a basic method for measuring the 2D radiant intensity distributions of anisotropic point light sources by using a goniophotometer. Their method can directly sample the radiant intensity distribution of a light source by moving a sensor around the light source. However, their method requires a large amount of measurement time because it samples the radiant intensity distribution only at a single direction at a time.

To cope with this problem, image-based techniques, which can sample the radiant intensity distribution at a large number of directions simultaneously, are proposed. Rykowski and Kostal [9] propose an efficient method for measuring 2D radiant intensity distributions of LEDs by using the imaging sphere. They make use of the combination of a hemispherical chamber with diffuse coating and a hemispherical mirror, and capture the radiant intensity distribution with 2π steradian field of view at a time. Tan and Ng [12] use a diffuse translucent sheet and a flatbed scanner, and Moreno and Sun [5] use a diffuse translucent screen and a camera for efficiently capturing the 2D radiant intensity distributions of LEDs.

It is demonstrated that the above methods are useful for modeling real-world point light sources, in particular for inspecting LEDs. Unfortunately, however, we cannot use them for acquiring the 4D light fields of self-luminous extended light sources because they assume point light sources, *i.e.* light sources with negligible sizes.

4D light fields of self-luminous extended light sources

As mentioned in the introduction, the radiant intensity distributions of self-luminous extended light sources are described by 4D light fields. In a similar manner to Verbeck and Greenberg [14], Ashdown [3] proposes a basic method for measuring the 4D light field of a self-luminous light source by using a gonio-photometer. However, his method requires a huge amount of measurement time because it samples the 4D distribution only at a single point in the 4D space at a time.

To cope with this problem, image-based techniques are proposed also for measuring 4D light fields. Goesele *et al.* [4] propose a method for measuring 4D light fields of self-luminous light sources by using an optical filter and a diffuse-reflection board. Their method recovers the 4D light field from the images of the board illuminated by the light radiated from an extended light source and passing through the optical filter. Although their method is suitable for measuring a wide range of a 4D light field and works well with an optimally-designed optical filter, it is not easy to acquire a 4D light field efficiently and densely because the optical filter is static and one has to slide the position of the light source (or the optical filter) manually during the measurement.

Aoto *et al.* [2] propose a method for recovering the 4D light field of a self-luminous light source from the images of a diffuse-reflection board moving in front of the light source. Their method is unique in the sense that it does not require any static or dynamic filters but use only a diffuse-reflection board. However, it would be difficult to stably recover the high-frequency components of the 4D light field from the images of the diffuse-reflection board because diffuse reflectance behaves like a low-pass filter [8].

4D light fields of general scenes

Other than the above techniques specialized for measuring the 4D light fields of self-luminous extended light sources, there are a number of techniques for acquiring 4D light fields of general scenes. Since it is impossible to cover all the existing techniques due to limited space, we briefly mention the advantages and limitations of some of representative approaches when they are used for measuring the 4D light fields of self-luminous extended light sources.

One approach to general 4D light field acquisition is to use a spherical mirror array [13] and a camera array [16]. Those methods have the advantage that they can measure the wide range of the 4D light field of a self-luminous extended light source. However, they are not suited for densely measuring the 4D light field because it is not easy to place spherical mirrors and cameras densely.

Another approach to general 4D light field acquisition is to use a single camera with a micro-lens array [6] and a coded aperture [7]. Those methods have the advantage that they can measure the 4D light field of a self-luminous extended light source densely. However, they are not suited for measuring the wide range of the 4D light field because their measurement ranges are limited. Note that the objective of our study is not to acquire the incoming intensity distribution to a small area in a scene but to acquire the outgoing intensity distribution from an extended light source. In general, light field cameras are suited for the former purpose but are not suited for the latter purpose.

3 Proposed Method

3.1 Light Source Model

Fig. 1 shows the cross section of our proposed setup which consists of a pair of an LC panel and a diffuse-reflection board. Actually, we place the light source close to the LC panel as much as possible so that we can acquire a wider range of the light field. Our proposed method acquires the description of the light passing through a point $\boldsymbol{x}$ on the LC panel toward a direction $\boldsymbol{l}$ by using the images of the diffuse-reflection board illuminated by the transmitted light. For the reason described below, we move the diffuse-reflection board and observe the reflection of the transmitted light on the board twice at the positions 1 and 2. Note that we assume that a light source radiates unpolarized light since the transmittance of an LC panel depends on polarization state[1].

We assume that a self-luminous extended light source is approximately represented by a set of anisotropic point light sources, and therefore the light passing through $\boldsymbol{x}$ toward $\boldsymbol{l}$ comes from an unknown anisotropic point light source $\boldsymbol{c}$. We denote the surface normal and distance of the board at the first position by $\boldsymbol{n}_1$ and r_1, and those at the second position by $\boldsymbol{n}_2$ and r_2. We assume that the geometry of the setup is calibrated in advance, *i.e.* we assume that those surface normals and distances are known. On the other hand, there are two unknowns; one is the distance d between $\boldsymbol{x}$ and the point light source $\boldsymbol{c}$, and the other is the radiant intensity k of the light source toward the direction $\boldsymbol{l}$. Our proposed method estimates those two parameters for each $(\boldsymbol{x}, \boldsymbol{l})$ by using two radiances observed on the diffuse-reflection board at the positions 1 and 2.

When the diffuse-reflection board is placed at the first position, the radiance i_1 of the reflected light is given by[2]

$$i_1 = k\frac{(-\boldsymbol{l})^\top \boldsymbol{n}_1}{(d+r_1)^2}, \tag{1}$$

assuming the Lambertian model and the attenuation according to the inverse-square law[3]. Similarly, when the board is placed at the second position, the radiance i_2 is given by

$$i_2 = k\frac{(-\boldsymbol{l})^\top \boldsymbol{n}_2}{(d+r_2)^2}. \tag{2}$$

Taking the ratio of eq.(1) and eq.(2), we can derive

$$\frac{(d+r_1)^2}{(d+r_2)^2} = \frac{i_2}{i_1}\frac{(-\boldsymbol{l})^\top \boldsymbol{n}_1}{(-\boldsymbol{l})^\top \boldsymbol{n}_2} \equiv \alpha. \tag{3}$$

[1] One could use a depolarizing filter in front of the LC panel.

[2] In general, the transmittance of an LC panel depends on the direction of incident light. Since we used an LC display with a wide viewing angle of 165° in our experiments, we do not take the angle-dependency into consideration.

[3] This description is more complicated than that of the 4D light field because we take the distance from an anisotropic point light source into consideration.

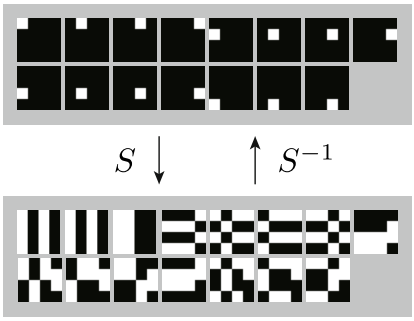


Fig. 2. The filters for the straightforward measurement (top) and the multiplexed measurement (bottom). Here, $n = 15$ for display purpose.

Thus, we can estimate one of the unknowns d as

$$d = \frac{\sqrt{\alpha} r_2 - r_1}{1 - \sqrt{\alpha}}. \tag{4}$$

Substituting eq.(4) into eq.(1) and/or eq.(2), we can estimate the other unknown k.

3.2 Straightforward Measurement

The straightforward method for measuring the light field of a self-luminous extended light source is to capture the images of the diffuse-reflection board at the first and second positions by using a set of *single filters* shown in the top of Fig. 2. Specifically, we divide an area of interest of the LC panel into n square patches, and then set the transmittance of a single patch to 1 and those of the other patches to 0 at a time in turn. The advantage of using a programmable filter, *i.e.* an LC panel in our case, is that we can control the transmittance both spatially and temporally without direct manual manipulation.

Unfortunately, however, such a straightforward measurement has limitations. In order to acquire light fields more densely, we need to make the size of each patch smaller. Since the transmittance of only a single patch is 1 in the straightforward measurement, the smaller the size of each patch is, the smaller the amount of light passing through the LC panel and reflected on the diffuse-reflection board is. Therefore, if we make the size of each patch smaller while keeping the measurement time constant, the SNRs of the captured images decrease and then the accuracy of the recovered light field is also degraded. On the other hand, if we make the size of each patch smaller while keeping the SNRs of the captured images constant, we need a longer exposure time for each image and then we need longer total measurement time. Hence, the straightforward measurement has a tradeoff between its accuracy and efficiency.

3.3 Multiplexed Measurement

To cope with the limitations of the straightforward measurement, our proposed method makes more use of the feature that the transmittance of the LC panel can be controlled both spatially and temporally. Our method utilizes multiplexed sensing [10,11,15], which is a well-known technique for increasing SNR without increasing measurement time, and acquires 4D light fields more efficiently and densely than the straightforward method.

Specifically, we use the *multiplexed filters* in which the transmittances of about half of the patches are 1 and those of the other patches are 0 as shown in the bottom of Fig. 2, and capture the images of the diffuse-reflection board illuminated by the transmitted light. We can obtain those n multiplexed filters by applying the so-called S-matrix, which is constructed on the basis of the Hadamard matrix of order $(n+1)$, for n individual filters. In an opposite manner, we can obtain the single filters by applying the inverse matrix S^{-1} to the multiplexed filters. Therefore, by applying S^{-1} to the captured images of the diffuse-reflection board under the multiplexed filters, we can obtain the decoded images under the single filters. It is known that S^{-1} can be computed analytically: $S^{-1} = 2(2S^{\top} - 1_n)/(n+1)$, where 1_n is an $n \times n$ matrix whose all elements are 1. See Sloane *et al.* [11] for more detail.

It is known that the ratio of the SNR of multiplexed sensing $\mathrm{SNR}_{\mathrm{multi}}$ and that of single sensing $\mathrm{SNR}_{\mathrm{single}}$ is at most

$$\frac{\mathrm{SNR}_{\mathrm{multi}}}{\mathrm{SNR}_{\mathrm{single}}} \simeq \frac{\sqrt{n}}{2}, \tag{5}$$

when n, *i.e.* the number of the patches in our case, is large enough. Therefore, the proposed method based on multiplexed sensing can acquire light fields more efficiently and densely than the straightforward method while keeping the SNR constant.

4 Experiments

4.1 Multiplexed Sensing

To demonstrate the effectiveness of multiplexed sensing, we compared the images of the diffuse-reflection board captured and decoded by multiplexed sensing with those captured by single sensing. We used a fluorescent light located nearby the LC panel and set the number of patches n to 63. Because a small amount of light passes through the LC panel even though the transmittance is set to 0, we captured an image when all the transmittances are set to 0 and then subtracted this image from all the images captured under the single and multiplexed filters.

Fig. 3 shows the example images of the diffuse-reflection board under a certain single filter taken with a fixed exposure time. We consider the average of 1000 images taken under the same condition as the ground truth (left). We can see that the image captured by single sensing (middle) is grained due to

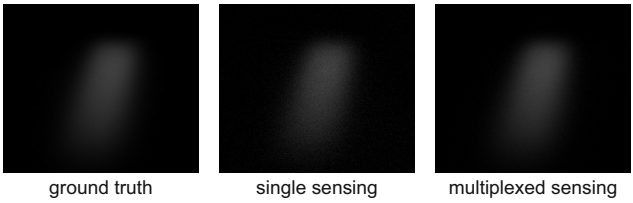


Fig. 3. The example images of the diffuse-reflection board under a single filter; the ground truth computed by averaging, the captured image by single sensing, and the captured and decoded image by multiplexed sensing from left to right. Pixel values are scaled for display purpose.

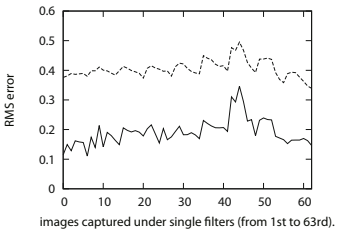


Fig. 4. The RMS errors of the images of the diffuse-reflection board under the single filters; captured by the straightforward measurement (dotted line) and captured and decoded by the multiplexed measurement (solid line)

noise. On the other hand, we can see that the image captured and decoded by multiplexed sensing is similar to the ground truth. This result qualitatively demonstrates that our proposed method based on multiplexed sensing works better than the straightforward method based on single sensing.

In addition, we conducted quantitative evaluation. Fig. 4 shows the RMS errors of the images of the diffuse-reflection board under all the single filters. We can see that the RMS errors of the captured and decoded images by multiplexed sensing (solid line) are always smaller than those of captured images by single sensing (dotted line) although the gain of multiplexed sensing, *i.e.* $\mathrm{SNR}_{\mathrm{multi}}/\mathrm{SNR}_{\mathrm{single}} \simeq 0.40/0.19 \simeq 2.1$ is smaller than the theoretical upper limit $\sqrt{n}/2 \simeq 4.0$. This result quantitatively demonstrates that the proposed method based on multiplexed sensing works better than the straightforward method based on single sensing.

4.2 Image Reconstruction

To demonstrate the effectiveness of the proposed method, we acquired the 4D light fields of three light sources and used them for image reconstruction. In this experiment, as described in Section 3, we acquired the light fields from the images of the diffuse-reflection board at the positions 1 and 2 by using the straightforward method based on single sensing and the proposed method based

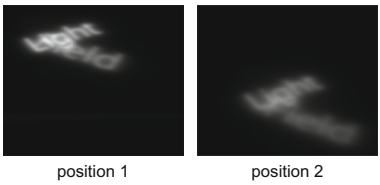


Fig. 5. The images of the diffuse-reflection board placed at the positions 1 (left) and 2 (right) for measurement under two projectors

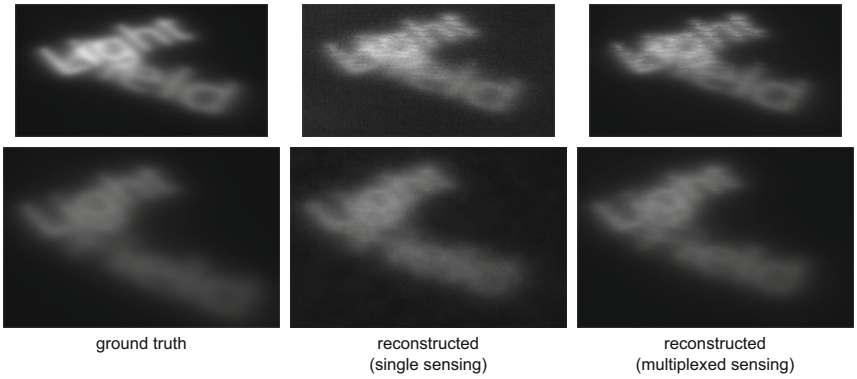


Fig. 6. The reconstruction results on the two projectors. The closeup images of the diffuse-reflection board at the positions 3 (top) and 4 (bottom); the ground truth image (left) and the reconstructed images by using single sensing (middle) and multiplexed sensing (right).

on multiplexed sensing. Then, we reconstructed the images of the board at two positions different from those for measurement, say positions 3 and 4, when the transmittances of all the patches are set to 1 by using the acquired light fields. Specifically, the intensity of each pixel in the reconstructed image is computed by assuming that the corresponding surface point is illuminated by n anisotropic point light sources whose intensities and distances are estimated as described in Section 3.1.

The first light source is two projectors. Fig. 5 shows the images of the diffuse-reflection board placed at the positions 1 (left) and 2 (right) for measurement. The transmittances of all the patches are set to 1 for display purpose. We can see that the characters of "Light" radiated from one projector cross the characters of "Field" radiated from another projector.

Fig. 6 shows the closeup images of the diffuse-reflection board at the positions 3 (top) and 4 (bottom); the ground truth images and the reconstructed images by using single sensing and multiplexed sensing from left to right. Here, the number of patches n is 255. We can see that both the straightforward method and the proposed method can capture how the characters radiated from the two projectors cross according to the distance from the projectors. Furthermore,

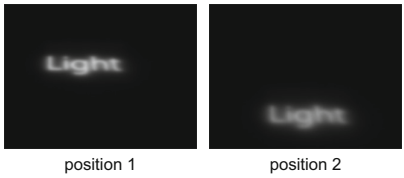


Fig. 7. The images of the diffuse-reflection board placed at the positions 1 (left) and 2 (right) for measurement under a single projector

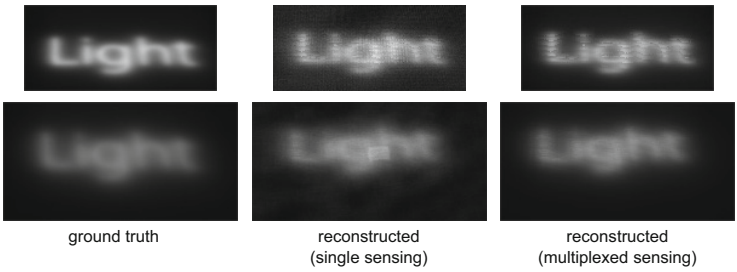


Fig. 8. The reconstruction results on the single projector. The closeup images of the diffuse-reflection board at the positions 3 (top) and 4 (bottom); the ground truth image (left) and the reconstructed images by using single sensing (middle) and multiplexed sensing (right).

we can see that the images reconstructed by using the proposed method based on multiplexed sensing are less noisy than the images reconstructed by using the straightforward method based on single sensing. Although some artifacts due to the discretization (the number of patches $n = 255$ is not necessarily large enough) and errors in geometric calibration are still visible, this result demonstrates that the proposed method works better than the straightforward method.

The second light source is a single projector. Fig. 7 shows the images of the diffuse-reflection board placed at the positions 1 (left) and 2 (right) for measurement. We can see that the characters of "Light" radiated from the projector is in focus and out of focus depending on the distance from the projector.

Fig. 8 shows the closeup images of the diffuse-reflection board at the positions 3 (top) and 4 (bottom); the ground truth images and the reconstructed images by using single sensing and multiplexed sensing from left to right. Here, the number of patches n is 255. We can see that both the straightforward method and the proposed method can capture how the characters radiated from the projector blur according to the distance from the projector. Similar to the above, we can see that the images reconstructed by using the proposed method are less noisy than the images reconstructed by using the straightforward method.

The third light source is an electric torch which consists of three LEDs. Fig. 9 shows the images of the diffuse-reflection board placed at the positions 1 (left)

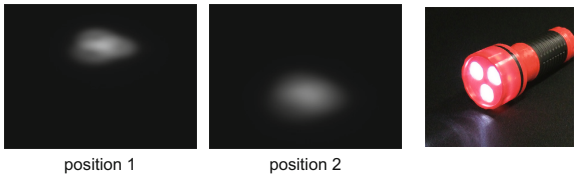


Fig. 9. The images of the diffuse-reflection board placed at the positions 1 (left) and 2 (right) for measurement under an electric torch

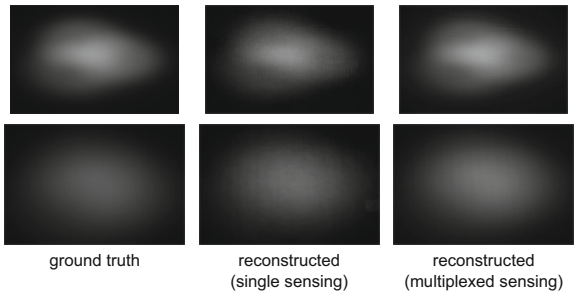


Fig. 10. The reconstruction results on the electric torch. The closeup images of the diffuse-reflection board at the positions 3 (top) and 4 (bottom); the ground truth image (left) and the reconstructed images by using single sensing (middle) and multiplexed sensing (right).

and 2 (right) for measurement. We can see that the lights radiated from the three LEDs cross each other depending on the distance from the torch.

Fig. 10 shows the closeup images of the diffuse-reflection board at the positions 3 (top) and 4 (bottom); the ground truth images and the reconstructed images by using single sensing and multiplexed sensing from left to right. Here, the number of patches n is 255. We can see that both the straightforward method and the proposed method can capture how the lights radiated from the three LEDs cross each other according to the distance from the torch. Similar to the above, we can see that the images reconstructed by using the proposed method are less noisy than the images reconstructed by using the straightforward method.

5 Conclusions and Future Work

In this paper, we proposed a method for measuring 4D light fields of self-luminous extended light sources by using an LC panel, *i.e.* a programmable filter and a diffuse-reflection board. Our proposed method recovers the 4D light field from the images of the board illuminated by the light radiated from an extended light source and passing through the LC panel. Our method makes use of the feature that the transmittance of the LC panel can be controlled both spatially

and temporally, and recovers 4D light fields efficiently and densely on the basis of multiplexed sensing. We implemented the prototype setup, and confirmed through a number of experiments that the proposed method works better than the straightforward measurement.

One direction of future study is to use more sophisticated filters, *e.g.* filters for adaptive sampling and compressive sensing. Another direction of future study is the applications of the acquired light fields to computer vision problems such as image-based modeling.

Acknowledgements. A part of this work was supported by JSPS KAKENHI Grant No. 24650077.

References

1. Adelson, E., Bergen, J.: The plenoptic function and the elements of early vision. Computational Models of Visual Processing, MIT Press, pp. 3–20 (1991)
2. Aoto, T., Sato, T., Mukaigawa, Y., Yokoya, N.: Linear estimation of 4-D illumination light field from diffuse reflections. In: Proc. ACPR 2013, pp. 495–500 (2013)
3. Ashdown, I.: Near-field photometry: a new approach. Journal of Illuminating Engineering Society **22**(1), 163–180 (1993)
4. Goesele, M., Granier, X., Heidrich, W., Seidel, H.: Accurate light source acquisition and rendering. In: Proc. ACM SIGGRAPH 2003, pp. 621–630 (2003)
5. Moreno, I., Sun, C.-C.: Three-dimensional measurement of light-emitting diode radiation pattern: a rapid estimation. Measurement Science and Technology **20**(7), 1–6 (2009)
6. Ng, R., Levoy, M., Bredif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Stanford Tech Report CTSR 2005–02 (2005)
7. Liang, C.-K., Lin, T.-H., Wong, B.-Y., Liu, C., Chen, H.: Programmable aperture photography: multiplexed light field acquisition. In: Proc. ACM SIGGRAPH 2008 (2008)
8. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: Proc. ACM SIGGRAPH 2001, pp. 117–128 (2001)
9. Rykowski, R., Kostal, H.: Novel approach for LED luminous intensity measurement. In: Proc. SPIE, vol. 6910 (2008)
10. Schechner, Y., Nayar, S., Belhumeur, P.: A theory of multiplexed illumination. In: Proc. ICCV 2003, pp. 808–815 (2003)
11. Sloane, N., Fine, T., Phillips, P., Harwit, M.: Codes for multiplex spectrometry. Applied Optics **8**(10), 2103–2106 (1969)
12. Tan, H., Ng, T.: Light-emitting-diode inspection using a flatbed scanner. Optical Engineering **47**(10) (2008)
13. Unger, J., Wenger, A., Hawkins, T., Gardner, A., Debevec, P.: Capturing and Rendering with Incident Light Fields. In: Proc. EGSR 2003, pp. 1–10 (2003)
14. Verbeck, C., Greenberg, D.: A comprehensive light-source description for computer graphics. IEEE CG&A **4**(7), 66–75 (1984)
15. Wetzstein, G., Ihrke, I., Heidrich, W.: On plenoptic multiplexing and reconstruction. IJCV **101**(2), 384–400 (2013)
16. Wilburn, B., Joshi, N., Vaish, V., Talvala, E.-V., Antunez, E., Barth, A., Adams, A., Horowitz, M., Levoy, M.: High performance imaging using large camera arrays. In: Proc. ACM SIGGRAPH 2005, pp. 765–776 (2005)

Light Field from Smartphone-Based Dual Video

Bernd Krolla[1](✉), Maximilian Diebold[2], and Didier Stricker[1]

[1] German Research Center for Artificial Intelligence, Kaiserslautern, Germany
{bernd.krolla,didier.stricker}@dfki.de
[2] Heidelberg Collaboratory for Image Processing, Heidelberg, Germany
maximilian.diebold@iwr.uni-heidelberg.de

Abstract. In this work, we introduce a light field acquisition approach for standard smartphones. The smartphone is manually translated along a horizontal rail, while recording synchronized video with front and rear camera. The front camera captures a control pattern, mounted parallel to the direction of translation to determine the smartphones current position. This information is used during a postprocessing step to identify an equally spaced subset of recorded frames from the rear camera, which captures the actual scene. From this data we assemble a light field representation of the scene. For subsequent disparity estimation, we apply a structure tensor approach in the epipolar plane images.

We evaluate our method by comparing the light fields resulting from manual translation of the smartphone against those recorded with a constantly moving translation stage.

Keywords: Computer vision · Light field imaging · Video processing

1 Introduction

While processing capabilities and hardware specifications of todays smartphones approach those of classical desktop computers, they are additionally equipped with a wide set of various sensors.

Besides multiple processing units and high amounts of memory, the latest smartphones are typically provided with GPS, IMUs, compass, (stereo) cameras, and other sensors.

Currently ongoing research within the domain of depth sensing technologies [6,10] including new camera systems such as [5,8,9,22] will most likely introduce an additional set of sensors for smartphones in the near future.

Besides this research and development of future devices, most of todays produced smartphones are typically equipped with at least one rear camera at the backside, as well as a front camera, which faces towards the user. Setting up on this hardware configuration, we aim to perform light field acquisition in an easy and end-user friendly manner.

We therefore introduce a new acquisition approach for light fields exploiting the availability of front and rear cameras of todays smartphones (Figure 1). In this context, different methods to precisely localize the smartphone during the light field acquisition are evaluated and discussed.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 600–610, 2015.
DOI: 10.1007/978-3-319-16181-5_46

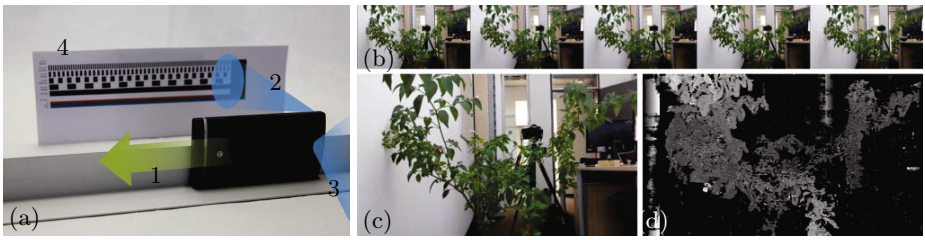


Fig. 1. (a) We manually translate a smartphone along a horizontal rail (1) while recording synchronized video with front (2) and rear camera (3). The front camera captures a control pattern (4) allowing the identification of the smartphones position while the main camera captures the actual light field data (b) of the scene (c). The resulting depth-map is shown in (d).

2 Related Work

A light field can be represented by the plenoptic function, introduced by Adelson and Bergen [1], Levoy and Hanrahan [12] and McMillan *et al.* [15].

The plenoptic function gives the fundamental understanding of representing and acquiring light fields e.g as 2D light field representation called Lumigraph, introduced by Gortler *et al.* [7]. Since then different methods have been established to exploit all information a light field provides.

Veeraraghavan *et al.* [21] introduce a light field acquisition camera using aperture masks. This masks attenuates the incident light rays without refracting them. Purpose of these masks is the modulation of the captured images. The light field is achieve by applying a Fourier transform based image demodulation.

An alternative approach also using aperture masks is called programmable aperture. Lian *et al.* [13] applying mask based multiplexing exploiting the fast multiple-exposures of cameras to generate the light field datasets.

In contrast to digital approaches, Levoy and Hanrahan [12] acquire light field data using a single moving camera. This is the simplest method and utilizes a computer controlled 3D translation stage called Gantry to capture suitable images for light field image processing.

A very similar approach is structure from motion introduced by Bolles *et al.* [2], having a straight-line camera motion system to capture a dense sequence of images. In their paper, they also introduce the exploitation of the epipolar plane images to obtain information about the three-dimensional position of objects and its usability.

Aside single moving cameras also large camera arrays as introduced by Wilburn *et al.* [25] are a possibility to capture light field datasets. While camera arrays for light field acquisition mostly have the constraint to be mounted on a planar grid with equidistant spacings between the cameras, Snavely *et al.* [19] introduces a method to reconstruct 3D object having a unstructured collection of images of the same object. The introduced system automatically computed

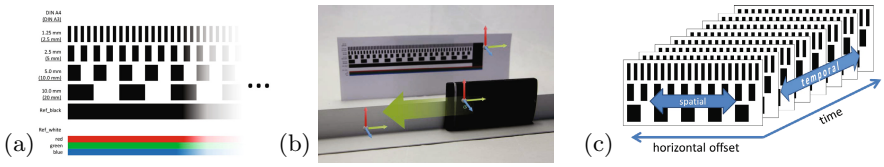


Fig. 2. (a) shows a part of the repetitive control pattern, which was used to locate the smartphones position during the horizontal translation (b). The captured video stream of this camera results in a set of video frames (c), which was used to localize the smartphones position.

the view point of each camera and generates a sparse 3D model of the scene and image, while the images can be captured in a random way.

Similar to that Davis *et al.* [3] present a system to interactively acquire light fields using a hand-held commodity camera. The system has real-time feedback to the photographer to obtain a dense light field of the captured object for the 3D reconstruction. An other possibility to capture wide angle light fields is been introduced by Taguchi *et al.* [20]. In this paper a spherical catadioptric cameras is modeled, using mirror balls mounted on a common plane. For the capturing, an aligned camera to the mirror set is been used to obtain the light field data.

While the above introduced methods are based on customary perspective cameras Adelson and Bergen [1] and Ng *et al.* [16,17] introduce so called plenoptic cameras having a micro lens array in front of the image sensor to obtain beside spacial information also angular information of the scene. Unfortunately, the obtained angular information is always combined with a reduction of spacial resolution. Thus Lumsdaine and Georgiev [4,14] and Perwass *et al.* [18] introduce focused plenoptic cameras. Difference to the already introduced plenoptic cameras is the changed focus position of the main lens. Thus a higher resolution in the resulting light field images is obtained, but also the computational effort is much higher.

The work of Levoy [11] provides a smartphone application, which allows the generation of computational images with a narrow depth of field. While the application is characterized by its good usability, a generation of disparity maps of the scene is not performed.

3 Method

Assuming to be provided with a smartphone, the required hardware setup was chosen to allow for a low-cost and end-user friendly light field acquisition. The presented approach is ready to be used with any *state-of-the-art* smartphone, which is able to capture *dual video* with its main (=*rear*) and sub (=*front*) camera as shown in Figure 1(a).

In this work, we used a smartphone, which records synchronized dual video with 24*fps* and a resolution of 640 × 360 pixel per video stream. Neglecting the additional setup for evaluation as detailed in Section 4, further necessary

equipment is limited to a rail allowing for horizontal sliding of the phone, as well as a control pattern provided as a simple printout on a paper sheet.

The control pattern as pictured in Figure 2(a) is horizontally subdivided into multiple binary patterns of different frequencies. This periodicity allows for an easy determination of the relative camera position, while excluding any absolute positioning of the camera towards the pattern. We achieved with the given layout an easy and fast processing leading to sufficient positioning results.

The actual capturing of the light field is shortly demonstrated in the supplementary video material and consists in the manual shift of the smartphone along the horizontally mounted rail.

The recording was done at a relatively low translation velocity ($\leq 3cm/s$) to allow for a dense sampling of the scene through the video frames and to avoid a degradation of the recorded data through motion blur or influence of the rolling shutter.

3.1 Key Frame Extraction

Being provided with the dual video stream of synchronized front and rear camera, we now aim to describe the captured light field information with a sparse subset of video frames to make it available for subsequent light field processing. To do so, we need to identify a set of equally spaced frames within the main video sequence.

Having the simultaneously recorded video stream of the front camera at hand, a wide variety of approaches is applicable to perform this task, which consists in the analysis of a two-dimensional space with a spatial and a temporal dimension as indicated in Figure 2(c).

In this work, we confine ourselves to evaluate the following methods:

Spacial Intensity Change Around a Fixed Key Position (SIC). When applying this approach, each frame of the front video stream is considered independently to retrieve information about the smartphones relative position towards the control pattern. The intensity gradient in the direction of translation is thereby computed at a preselected position as shown in Figure 3(a).

As soon as the calculated gradient exceeds a given threshold τ_g, indicating the passage of an intensity border within the binary control pattern, the corresponding video frame of the rear-camera is added to the light field representation. The value of τ_g was hereby identified as one third of the difference between the reference values for black and white intensity.

$$\tau_g = \frac{1}{3} \cdot \frac{i_{white} - i_{black}}{2} + i_{black}. \tag{1}$$

Being provided with those preselected keypoints, this approach allows an online detection of relevant frames while the video-capturing is still in progress.

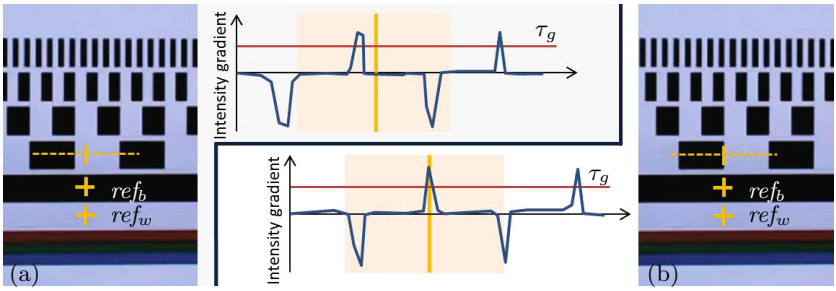


Fig. 3. (a) and (b) show two exemplary frames of the control pattern, captured by the front camera. A keypoint with accompanying evaluation window is indicated by the topmost orange mark. The two marks below are used to extract corresponding intensity values for white (ref_w) and black color (ref_b). Gradients of the intensity values for the two frames are assessed within the evaluation window (center) leading to a keyframe detection for frame (b).

Temporal Intensity Change on Fixed Key Position (TIC). For this approach, we choose keypoints within the video stream of the front camera in the same manner as for the SIC approach (See Figure 3). However, the introduced evaluation window as not used, but the intensity values at those keypoint positions were extracted for all frames of the video. To identify equidistant frames for light field parametrization, we then detected the edges of the binary control pattern at the chosen keypoints by comparing intensity values between current and proceeding frame. Whenever the difference of those intensities exceeded a given threshold τ the corresponding frame of the rear-camera was marked to be part of the light field.

The threshold τ was hereby computed in two different ways within this work: Assuming the overall intensity maximum i_{max} and minimum i_{min} to be given, we calculated a static threshold τ_s as average in a straight-forward manner by

$$\tau_s = \frac{(i_{max} + i_{min})}{2}. \tag{2}$$

While this threshold is easily determined during a postprocessing step, a temporally dynamic threshold τ_d is obtained by smoothing the intensity distribution with a Gaussian function. For frame i we obtain through discrete convolution of the intensity function f_{int} with the Gaussian distribution g:

$$\tau_d(i) = (f_{int} * g)(i) = \sum_m f_{int}[m] g[i - m] \tag{3}$$

Equidistant Frames in Time-Domain. To allow for a comparison of the presented methods for key frame extraction (SIC and TIC), we applied a further approach for keyframe extraction which is independent from the recorded video stream of the front camera (control pattern).

We used this method exclusively in conjunction with the translation stage and extracted a subset of frames by identifying every nth frame of the rear camera, while assuming constant translation velocity.

3.2 Light Field Processing

The captured keyframes of the main camera were rectified using the calibration approach of Vogiatzis *et al.* [23] and can be represented as three dimensional light field volume, utilizing the two plane parametrization as introduced by Gortler [7] called lumigraph. Thus we define the Π-plane containing the focal points $s \in \Pi$ of all cameras and the Ω-plane which denotes the image coordinates $(x, y) \in \Omega$. The resulting three dimensional light field volume becomes

$$L : \Omega \times \Pi \to \mathbb{R} \qquad (s, x, y) \mapsto L(s, x, y), \tag{4}$$

where $L(s, x, y)$ defines the color in each point.

In the resulting light field data is an epipolar plane image obtained by slicing through the light field volume. To achieve this the parameter y is set to a constant value $y*$. The resulting epipolar plane image is then defined by the function

$$S_{y^*} : \Sigma_{y^*} \to \mathbb{R} \tag{5}$$

$$(x, s) \mapsto S_{y^*}(x, s) := L(s, x, y^*). \tag{6}$$

An epipolar plane image contains information about the scene depth in terms of depth dependent orientations, see Figure 4.

To analyze these orientations, we use the structure tensor

$$J = \xi * \begin{pmatrix} \left(\frac{\partial \hat{S}}{\partial x}\right)^2 & \frac{\partial \hat{S}}{\partial x} \cdot \frac{\partial \hat{S}}{\partial s} \\ \frac{\partial \hat{S}}{\partial s} \cdot \frac{\partial \hat{S}}{\partial x} & \left(\frac{\partial \hat{S}}{\partial s}\right)^2 \end{pmatrix} =: \begin{pmatrix} J_{xx} & J_{xs} \\ J_{xs} & J_{ss} \end{pmatrix} \tag{7}$$

with the abbreviation

$$\hat{S} := \sigma * S_{y^*}, \tag{8}$$

where σ and ξ define a Gaussian smoothing. The resulting scene disparity information can now be computed as given in [24] using the equation

$$d = \tan\left(\frac{1}{2} \arctan\left(\frac{2J_{xs}}{J_{xx} - J_{ss}}\right)\right), \tag{9}$$

where only the structure tensor components are used to compute the underlying orientations.

4 Evaluation and Results

To evaluate the proposed approach we exploit the dual capturing mode of a *state-of-the-art* smartphone for parallel video acquisition of front and rear camera.

Fig. 4. Example of an Epipolar Plane Image (EPI) assembled from 31 images

Fig. 5. (a) On site capturing setup: A tripod-mounted and battery-powered translation stage allows for horizontal camera shifts with welldefined velocities as well as for manual operation

Besides a manual translation of the smartphone along a rail (Figure 1(a)), also a translation stage as shown in Figure 5 was used to capture light field data of different scenes. This stage provides a translation range of more than 25cm and operates highly accurate regarding the precision of velocity and positioning.

While the shifting velocity during manual operation could not be measured precisely, the velocity of the translation stage was set to a constant value of 7 mm/s during the acquisition process.

We evaluated besides the office scene (Figure 1) two outdoor scenes, while applying manual and automatic translation techniques. Figures 6 and 7 provide an overview of the obtained results, while the performance of the introduced keyframe detection approaches is discussed below.

Spacial Intensity Change Around a Fixed Key Position (SIC). Since all frames in this approach are evaluated independently from each other, it cannot be avoided that directly consecutive frames are detected as keyframes for the light field: While the gradient detection implies the evaluation of the keypoints neighborhood, it occurs, that consecutive frames are selected as key frames (e.g. for the 2.5mm pattern in Figure 7(c)), especially at low translation speeds of the camera and high recording frequencies.

Temporal Intensity Change on Fixed Key Position (TIC). This approach uses the two previously introduced thresholds τ_s and τ_d. For large parts of the evaluated scenarios, both approaches deliver very similar results (Figures 6 and 7)

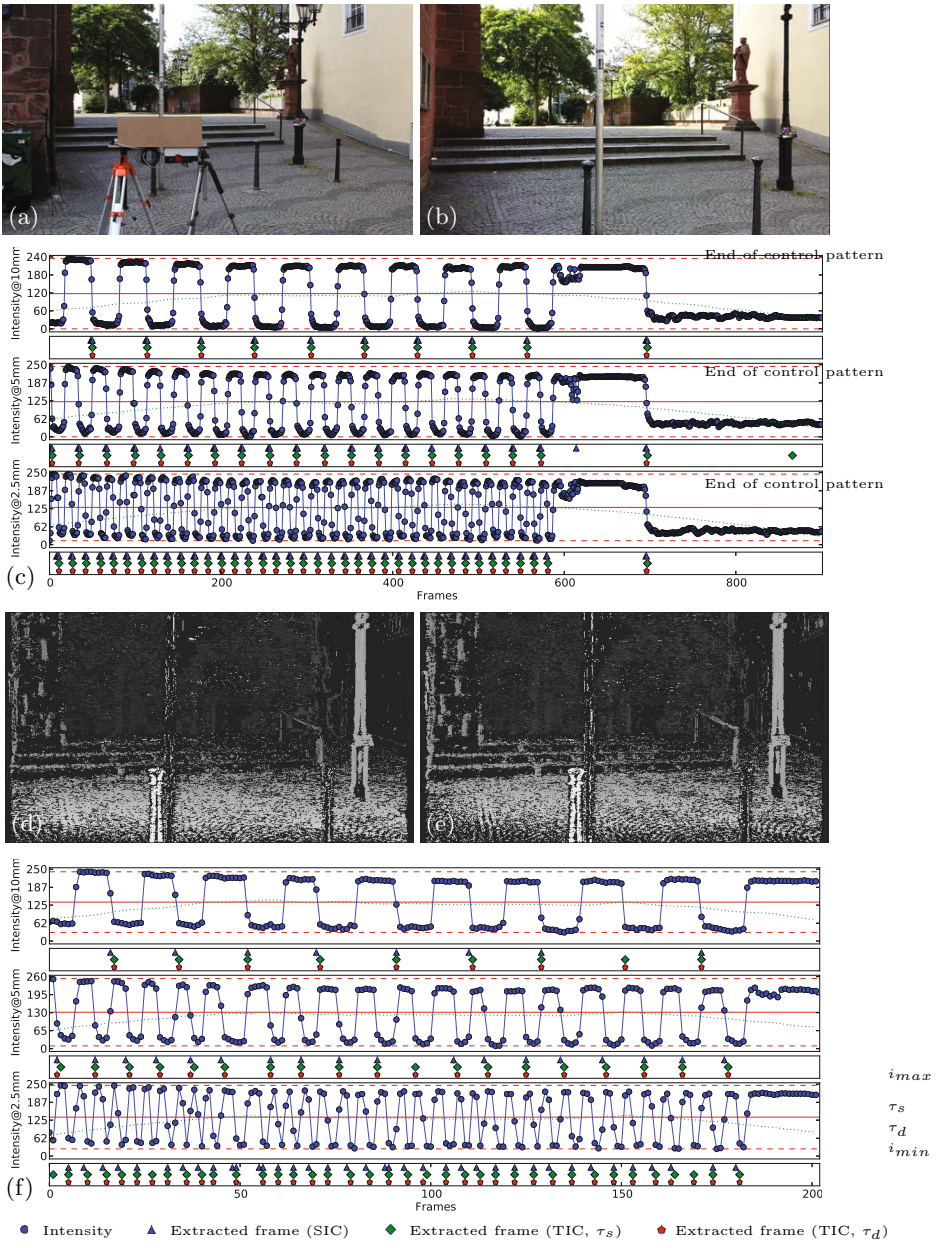


Fig. 6. (a) On site acquisition setup. (b) Exemplary frame captured by the smartphones rear camera. (c) Temporal intensity distribution for different keypositions in the front camera stream with an overview of extracted keyframes for different extraction methods (automatic translation). Resulting disparity maps for automatic translation (d), using the equidistant extraction approach and for the TIC extraction approach (e), using the $5mm$ control pattern. (f) Extracted keyframes using manual translation.

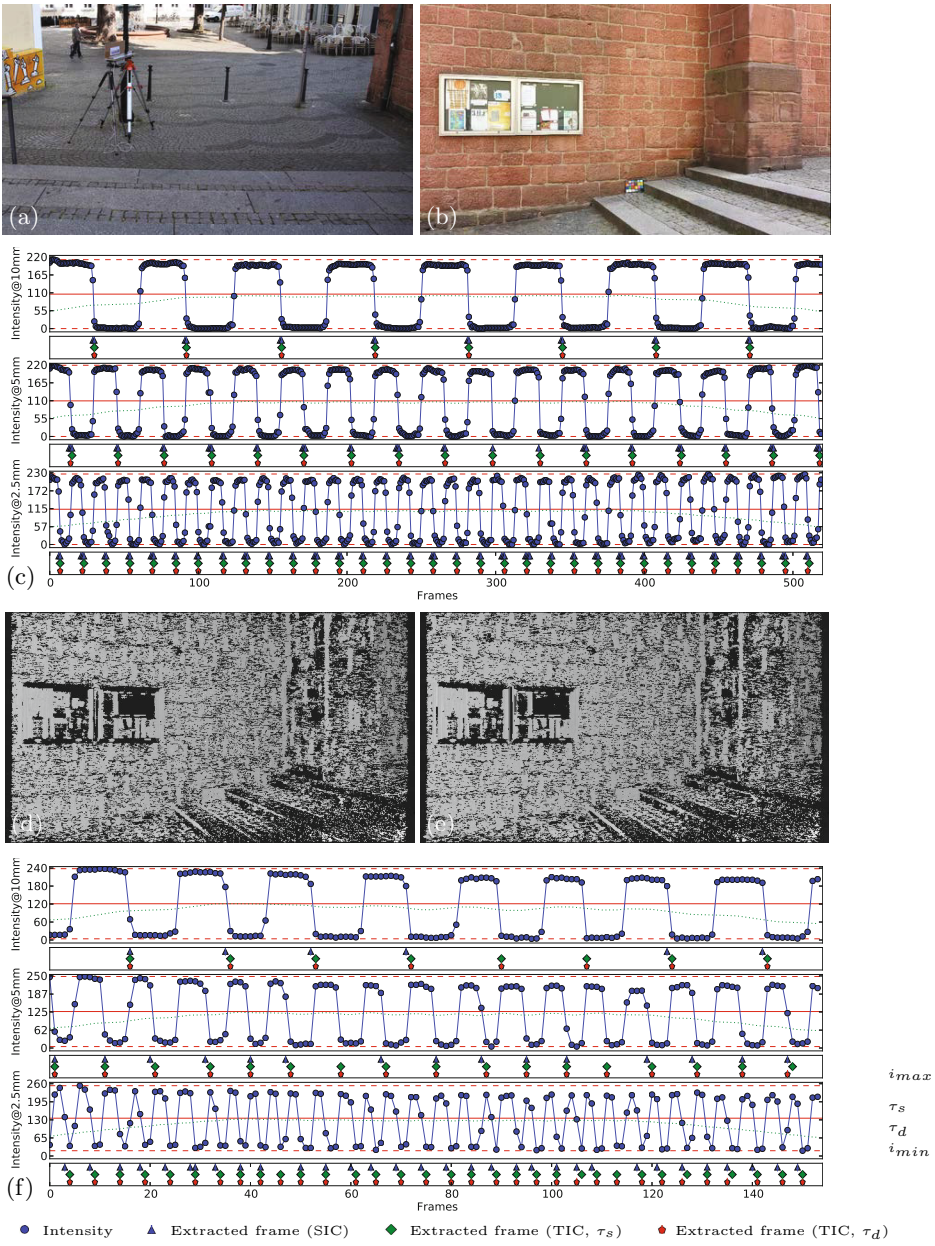


Fig. 7. (a) On site acquisition setup. (b) Exemplary frame captured by the smartphones rear camera. (c) Temporal intensity distribution for different keypositions in the front camera stream with an overview of extracted keyframes for different extraction methods (automatic translation). Resulting disparity maps for automatic translation (d), using the equidistant extraction approach and for the TIC extraction approach (e), using the 2.5mm control pattern. (f) Extracted keyframes using manual translation.

Since the acquisition of the light fields requires constant illumination conditions, the property of the dynamically calculated threshold τ_d to adopt to possible illumination changes within the scene is obsolete. Additionally tends this technique to deliver unreliable results at the start and end points of the recorded video streams. For the generation of disparity maps, resulting from the TIC approach, we therefore used exclusively the simpler static thresholding technique basing on τ_s.

Equidistant Frames in Time-Domain. This approach for keyframe extraction relies on constant translation velocity of the smartphone and was applied for evaluation purposes exclusively in conjunction with the translation stage. Since this method does not exploit any information from the control pattern, we used it to assess the results of the TIC and SIC approch (Figure 6 and 7).

5 Discussion

While providing in this work a conceptual overview over the proposed light field acquisition approach, we observed a variety of aspects, which currently prevent further improvement of results.

The recording with two independently managed (front and rear) camera systems complicates a full parameter control. Both cameras were checked to capture frames synchronously, while further camera parameters such as focus, white-balance or ISO-values remain uncorrelated. Establishing a strongly coupled camera pair, which assures the named parameters to be mutually controlled would allow to exploit especially prepared control pattern for global white-balancing.

During the evaluating of manually captured scenes, we furthermore noticed a high sensitivity of the light field processing methods against camera shakes, which require the user for careful acquisition. Image registration techniques as part of the postprocessing could possibly reduce this demand.

6 Conclusion

In this work, we introduced a light field acquisition approach for standard smartphones exploiting synchronized dual video capturing of front and rear camera. We evaluated the proposed method for different scenes and achieved comparable results for the proposed TIC keyframe extraction approach and the equidistant frame extraction method, relying on the capturings with the translation stage.

Acknowledgements. This work was funded by Sony Deutschland, Stuttgart Technology Center, EuTEC and is a result of the research cooperation between STC EuTEC, the Heidelberg Collaboratory for Image Processing (HCI) and the German Research Center for Artificial Intelligence (DFKI).

We would like to thank Thimo Emmerich, Yalcin Incesu and Oliver Erdler from STC EuTEC for their feedback to this work and all the fruitful discussions.

References

1. Adelson, E.H., Bergen, J.R.: The plenoptic function and the elements of early vision. Computational Models of Visual Processing **1**, 43–54 (1991)
2. Bolles, R.C., Baker, H.H., Marimont, D.H.: Epipolar-plane image analysis: An approach to determining structure from motion. International Journal of Computer Vision **1**(1), 7–55 (1987)
3. Davis, A., Levoy, M., Durand, F.: Unstructured Light Fields. Comp. Graph. Forum, 31, 305–314 (2012)
4. Georgiev, T., Lumsdaine, A.: Focused plenoptic camera and rendering. Journal of Electronic Imaging **19**, 021106 (2010)
5. Raytrix GmbH. Raytrix (2014). http://www.raytrix.de/
6. Google. Project tango (2014). https://www.google.com/atap/projecttango/
7. Gortler, S.J., Grzeszczuk, R., Szeliski, R., Cohen, M.F.: The Lumigraph. In: Siggraph (1996)
8. HTC. Htc one m8 (2014). http://www.htc.com/us/smartphones/htc-one-m8/
9. Lytro Inc., Lytro (2014). https://store.lytro.com/
10. Occipital Inc., Structure sensor (2014). http://structure.io/
11. Levoy, M.: Synthcam (2014). https://sites.google.com/site/marclevoy/
12. Levoy, M., Hanrahan, P.: Light field rendering. pp. 31–42 (1996)
13. Liang, C.-K., Lin, T.-H., Wong, B.-Y., Liu, C., Chen, H.: Programmable Aperture Photography: Multiplexed Light Field Acquisition 27(3), 1–10 (2008)
14. Lumsdaine, A.. Georgiev, T.: The Focused Plenoptic Camera. In: Proc. IEEE Int. Conference on Computational Photography, pp. 1–8 (2009)
15. McMillan, L., Bishop, G.: Plenoptic modeling: An image-based rendering system. pp. 39–46 (1995)
16. Ng, R.: Digital Light Field Photography. PhD thesis, Stanford University (2006). Note: thesis led to commercial light field camera, see also http://www.lytro.com
17. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenoptic camera. Technical Report CSTR 2005–02, Stanford University (2005)
18. Perwass, C., Wietzke, L.: The Next Generation of Photography (2010). http://www.raytrix.de
19. Snavely, N., Seitz, S., Szeliski, R.: Photo Tourism: Exploring image collections in 3D (2006). http://phototour.cs.washington.edu/bundler/
20. Taguchi, Y., Agrawal, A., Ramalingam, S., Veeraraghavan, A.: Axial Light Fields for Curved Mirrors: Reflect your Perspective, Widen your View (2010)
21. Veeraraghavan, A., Raskar, R., Agrawal, A., Mohan, A., Tumblin, J.: Dappled Photography: Mask Enhanced Cameras for Heterodyned Light Fields and Coded Aperture Refocussing 26(3), 1–69 (2007)
22. Venkataraman, K., Lelescu, D., Duparré, J., McMahon, A., Molina, G., Chatterjee, P., Mullis, R., Nayar, S.: Picam: an ultra-thin high performance monolithic camera array. ACM Transactions on Graphics (TOG) 32(6), 166 (2013). http://www.pelicanimaging.com/technology/
23. Vogiatzis, G., Hernández, C.: Video-based, real-time multi-view stereo. Image and Vision Computing **29**(7), 434–441 (2011)
24. Wanner, S., Goldluecke, B.: Variational Light Field Analysis for Disparity Estimation and Super-Resolution (2013)
25. Wilburn, B., Joshi, N., Vaish, V., Talvala, E.-V., Antunez, E., Barth, A., Adams, A., Horowitz, M., Levoy, M.: High performance imaging using large camera arrays. ACM Transactions on Graphics **24**, 765–776 (2005)

W15 - Computer Vision for Road Scene Understanding and Autonomous Driving

Ten Years of Pedestrian Detection, What Have We Learned?

Rodrigo Benenson(✉), Mohamed Omran, Jan Hosang, and Bernt Schiele

Max Planck Institute for Informatics, Saarbrücken, Germany
{rodrigo.benenson,mohamed.omran,jan.hosang,bernt.schiele}@mpi-inf.mpg.de

Abstract. Paper-by-paper results make it easy to miss the forest for the trees. We analyse the remarkable progress of the last decade by discussing the main ideas explored in the 40+ detectors currently present in the Caltech pedestrian detection benchmark. We observe that there exist three families of approaches, all currently reaching similar detection quality. Based on our analysis, we study the complementarity of the most promising ideas by combining multiple published strategies. This new decision forest detector achieves the current best known performance on the challenging Caltech-USA dataset.

1 Introduction

Pedestrian detection is a canonical instance of object detection. Because of its direct applications in car safety, surveillance, and robotics, it has attracted much attention in the last years. Importantly, it is a well defined problem with established benchmarks and evaluation metrics. As such, it has served as a playground to explore different ideas for object detection. The main paradigms for object detection "Viola&Jones variants", HOG+SVM rigid templates, deformable part detectors (DPM), and convolutional neural networks (ConvNets) have all been explored for this task.

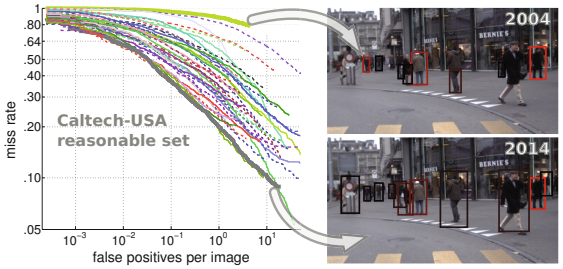


Fig. 1. The last decade has shown tremendous progress on pedestrian detection. What have we learned out of the 40+ proposed methods?

The aim of this paper is to review progress over the last decade of pedestrian detection (40+ methods), identify the main ideas explored, and try to quantify which ideas had the most impact on final detection quality. In the next sections we review existing datasets (section 2), provide a discussion of the different approaches (section 3), and experiments reproducing/quantifying the recent years' progress (section 4, presenting experiments over ∼ 20 newly trained

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 613–627, 2015.
DOI: 10.1007/978-3-319-16181-5_47

(a) INRIA test set **(b)** Caltech-USA test set **(c)** KITTI test set

Fig. 2. Example detections of a top performing method (`SquaresChnFtrs`)

detector models). Although we do not aim to introduce a novel technique, by putting together existing methods we report the best known detection results on the challenging Caltech-USA dataset.

2 Datasets

Multiple public pedestrian datasets have been collected over the years; INRIA [1], ETH [2], TUD-Brussels [3], Daimler [4] (Daimler stereo [5]), Caltech-USA [6], and KITTI [7] are the most commonly used ones. They all have different characteristics, weaknesses, and strengths.

INRIA is amongst the oldest and as such has comparatively few images. It benefits however from high quality annotations of pedestrians in diverse settings (city, beach, mountains, etc.), which is why it is commonly selected for training (see also §4.4). ETH and TUD-Brussels are mid-sized video datasets. Daimler is not considered by all methods because it lacks colour channels. Daimler stereo, ETH, and KITTI provide stereo information. All datasets but INRIA are obtained from video, and thus enable the use of optical flow as an additional cue.

Today, Caltech-USA and KITTI are the predominant benchmarks for pedestrian detection. Both are comparatively large and challenging. Caltech-USA stands out for the large number of methods that have been evaluated side-by-side. KITTI stands out because its test set is slightly more diverse, but is not yet used as frequently. For a more detailed discussion of the datasets please consult [7,8]. INRIA, ETH (monocular), TUD-Brussels, Daimler (monocular), and Caltech-USA are available under a unified evaluation toolbox; KITTI uses its own separate one with unpublished test data. Both toolboxes maintain an online ranking where published methods can be compared side by side.

In this paper we use primarily Caltech-USA for comparing methods, INRIA and KITTI secondarily. See figure 2 for example images. Caltech-USA and INRIA results are measured in log-average miss-rate (MR, lower is better), while KITTI uses area under the precision-recall curve (AUC, higher is better).

Value of Benchmarks. Individual papers usually only show a narrow view over the state of the art on a dataset. Having an official benchmark that collects detections from all methods greatly eases the author's effort to put their curve into context, and provides reviewers easy access to the state of the art results.

Table 1. Listing of methods considered on Caltech-USA, sorted by log-average miss-rate (lower is better). Consult sections 3.1 to 3.9 for details of each column. See also matching figure 3. "HOG" indicates HOG-like [1]. Ticks indicate salient aspects of each method.

Method	MR	*Family*	*Features*	*Classifier*	*Context*	*Deep*	*Parts*	*M-Scales*	*More data*	*Feat. type*	*Training*
VJ [9]	94.73%	DF	✓	✓						Haar	I
Shapelet [10]	91.37%	-	✓							Gradients	I
PoseInv [11]	86.32%	-					✓			HOG	I+
LatSvm-V1 [12]	79.78%	DPM					✓			HOG	P
ConvNet [13]	77.20%	DN				✓				Pixels	I
FtrMine [14]	74.42%	DF	✓							HOG+Color	I
HikSvm [15]	73.39%	-		✓						HOG	I
HOG [1]	68.46%	-	✓	✓						HOG	I
MultiFtr [16]	68.26%	DF	✓	✓						HOG+Haar	I
HogLbp [17]	67.77%	-	✓							HOG+LBP	I
AFS+Geo [18]	66.76%	-			✓					Custom	I
AFS [18]	65.38%	-								Custom	I
LatSvm-V2 [19]	63.26%	DPM		✓			✓			HOG	I
Pls [20]	62.10%	-	✓	✓						Custom	I
MLS [21]	61.03%	DF	✓							HOG	I
MultiFtr+CSS [22]	60.89%	DF	✓							Many	T
FeatSynth [23]	60.16%	-	✓	✓						Custom	I
pAUCBoost [24]	59.66%	DF	✓	✓						HOG+COV	I
FPDW [25]	57.40%	DF								HOG+LUV	I
ChnFtrs [26]	56.34%	DF	✓	✓						HOG+LUV	I
CrossTalk [27]	53.88%	DF			✓					HOG+LUV	I
DBN−Isol [28]	53.14%	DN					✓			HOG	I
ACF [29]	51.36%	DF	✓							HOG+LUV	I
RandForest [30]	51.17%	DF		✓						HOG+LBP	I&C
MultiFtr+Motion [22]	50.88%	DF	✓						✓	Many+Flow	T
SquaresChnFtrs [31]	50.17%	DF	✓							HOG+LUV	I
Franken [32]	48.68%	DF		✓						HOG+LUV	I
MultiResC [33]	48.45%	DPM			✓		✓	✓		HOG	C
Roerei [31]	48.35%	DF	✓					✓		HOG+LUV	I
DBN−Mut [34]	48.22%	DN			✓		✓			HOG	C
MF+Motion+2Ped [35]	46.44%	DF			✓				✓	Many+Flow	I+
MOCO [36]	45.53%	-	✓		✓					HOG+LBP	C
MultiSDP [37]	45.39%	DN	✓		✓	✓				HOG+CSS	C
ACF-Caltech [29]	44.22%	DF	✓							HOG+LUV	C
MultiResC+2Ped [35]	43.42%	DPM			✓		✓	✓		HOG	C+
WordChannels [38]	42.30%	DF	✓							Many	C
MT-DPM [39]	40.54%	DPM					✓	✓		HOG	C
JointDeep [40]	39.32%	DN			✓					Color+Gradient	C
SDN [41]	37.87%	DN				✓	✓			Pixels	C
MT-DPM+Context [39]	37.64%	DPM			✓		✓	✓		HOG	C+
ACF+SDt [42]	37.34%	DF	✓						✓	ACF+Flow	C+
SquaresChnFtrs [31]	34.81%	DF	✓							HOG+LUV	C
InformedHaar [43]	34.60%	DF	✓							HOG+LUV	C
Katamari-v1	22.49%	DF	✓		✓				✓	HOG+Flow	C+

The collection of results enable retrospective analyses such as the one presented in the next section.

3 Main Approaches to Improve Pedestrian Detection

Figure 3 and table 1 together provide a quantitative and qualitative overview over 40+ methods whose results are published on the Caltech pedestrian detection benchmark (July 2014). Methods marked in italic are our newly trained models (described in section 4). We refer to all methods using their Caltech benchmark shorthand. Instead of discussing the methods' individual particularities, we identify the key aspects that distinguish each method (ticks of table 1) and group them accordingly. We discuss these aspects in the next subsections.

Brief Chronology. In 2003, Viola and Jones applied their `VJ` detector [44] to the task of pedestrian detection. In 2005 Dalal and Triggs introduced the landmark `HOG` [1] detector, which later served in 2008 as a building block for the now classic deformable part model DPM (named `LatSvm` here) by Felzenswalb et al. [12]. In 2009 the Caltech pedestrian detection benchmark was introduced, comparing seven pedestrian detectors [6]. At this point in time, the evaluation metrics changed from per-window (FPPW) to per-image (FPPI), once the flaws of the per-window evaluation were identified [8]. Under this new evaluation metric some of the early detectors turned out to under-perform.

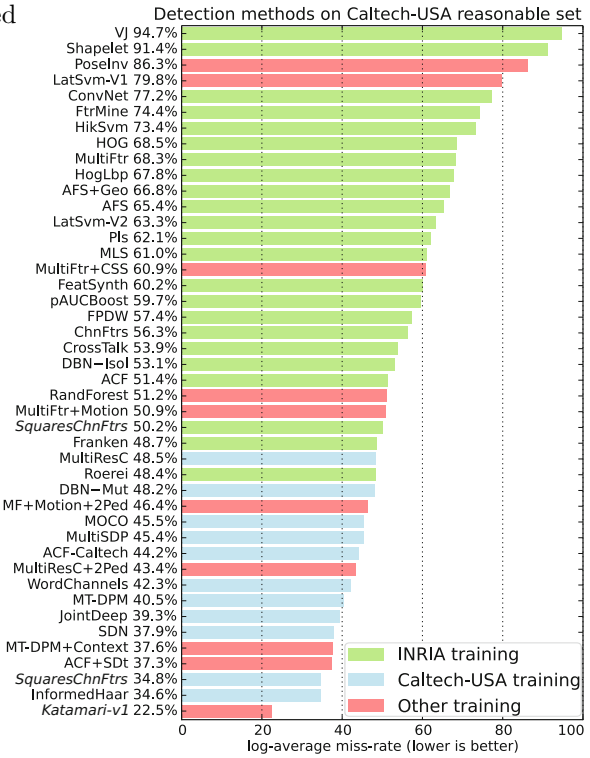


Fig. 3. Caltech-USA detection results

About one third of the methods considered here were published during 2013, reflecting a renewed interest on the problem. Similarly, half of the KITTI results for pedestrian detection were submitted in 2014.

3.1 Training Data

Figure 3 shows that differences in detection performance are, not surprisingly, dominated by the choice of training data. Methods trained on Caltech-USA systematically perform better than methods that generalise from INRIA. Table 1 gives additional details on the training data used[1]. High performing methods with "other training" use extended versions of Caltech-USA. For instance `MultiResC+2Ped` uses Caltech-USA plus an extended set of annotations over INRIA, `MT-DPM+Context` uses an external training set for cars, and `ACF+SDt` employs additional frames from the original Caltech-USA videos.

3.2 Solution Families

Overall we notice that out of the 40+ methods we can discern three families: 1) DPM variants (`MultiResC` [33], `MT-DPM` [39], etc.), 2) Deep networks (`JointDeep` [40], `ConvNet` [13], etc.), and 3) Decision forests (`ChnFtrs`, `Roerei`, etc.). On table 1 we identify these families as `DPM`, `DN`, and `DF` respectively.

Based on raw numbers alone boosted decision trees (`DF`) seem particularly suited for pedestrian detection, reaching top performance on both the "train on INRIA, test on Caltech", and "train on Caltech, test on Caltech" tasks. It is unclear however what gives them an edge. The deep networks explored also show interesting properties and fast progress in detection quality.

Conclusion Overall, at present, DPM variants, deep networks, and (boosted) decision forests all reach top performance in pedestrian detection (around 37 % MR on Caltech-USA, see figure 3).

3.3 Better Classifiers

Since the original proposal of `HOG+SVM` [1], linear and non-linear kernels have been considered. `HikSvm` [15] considered fast approximations of non-linear kernels. This method obtains improvements when using the flawed FPPW evaluation metric (see section 3), but fails to perform well under the proper evaluation (FPPI). In the work on `MultiFtrs` [16], it was argued that, given enough features, Adaboost and linear SVM perform roughly the same for pedestrian detection.

Recently, more and more components of the detector are optimized jointly with the "decision component" (e.g. pooling regions in `ChnFtrs` [26], filters in `JointDeep` [40]). As a result the distinction between features and classifiers is not clear-cut anymore (see also sections 3.8 and 3.9).

Conclusion There is no conclusive empirical evidence indicating that whether non-linear kernels provide meaningful gains over linear kernels (for pedestrian detection, when using non-trivial features). Similarly, it is unclear whether one particular type of classifier (e.g. SVM or decision forests) is better suited for pedestrian detection than another.

[1] "Training" data column: I→INRIA, C→Caltech, I+/C+ →INRIA/Caltech and additional data, P→Pascal, T→TUD-Motion, I&C→both INRIA and Caltech.

3.4 Additional Data

The core problem of pedestrian detection focuses on individual monocular colour image frames. Some methods explore leveraging additional information at training and test time to improve detections. They consider stereo images [45], optical flow (using previous frames, e.g. `MultiFtr+Motion` [22] and `ACF+SDt` [42]), tracking [46], or data from other sensors (such as lidar [47] or radar).

For monocular methods it is still unclear how much tracking can improve per-frame detection itself. As seen in figure 4 exploiting optical flow provides a non-trivial improvement over the baselines. Curiously, the current best results (`ACF-SDt` [42]) are obtained using coarse rather than high quality flow. In section 4.2 we inspect the complementarity of flow with other ingredients. Good success exploiting flow and stereo on the Daimler dataset has been reported [48], but similar results have yet to be seen on newer datasets such as KITTI.

Conclusion Using additional data provides meaningful improvements, albeit on modern dataset stereo and flow cues have yet to be fully exploited. As of now, methods based merely on single monocular image frames have been able to keep up with the performance improvement introduced by additional information.

3.5 Exploiting Context

Sliding window detectors score potential detection windows using the content inside that window. Drawing on the context of the detection window, i.e. image content surrounding the window, can improve detection performance. Strategies for exploiting context include: ground plane constraints (`MultiResC` [33], `RandForest` [30]), variants of auto-context [49] (`MOCO` [36]), other category detectors (`MT-DPM+Context` [39]), and person-to-person patterns (`DBN−Mut` [34], `+2Ped` [35], `JointDeep` [40]).

Figure 4 shows the performance improvement for methods incorporating context. Overall, we see improvements of 3 ∼ 7 MR percent points. (The negative impact of `AFS+Geo` is due to a change in evaluation, see section 3.) Interestingly, `+2Ped` [35] obtains a consistent 2 ∼ 5 MR percent point improvement over existing methods, even top performing ones (see section 4.2).

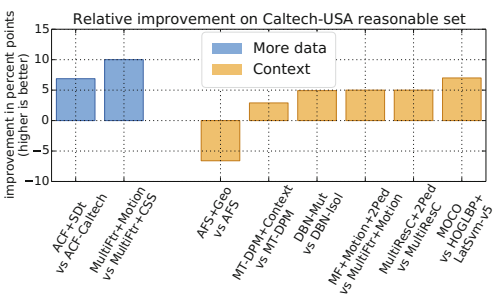


Fig. 4. Caltech-USA detection improvements for different method types. Improvement relative to each method's relevant baseline ("method vs baseline").

Conclusion Context provides consistent improvements for pedestrian detection, although the scale of improvement is lower compared to additional test data (§3.4) and deep architectures (§3.8). The bulk of detection quality must come from other sources.

3.6 Deformable Parts

The DPM detector [19] was originally motivated for pedestrian detection. It is an idea that has become very popular and dozens of variants have been explored.

For pedestrian detection the results are competitive, but not salient (`LatSvm` [12,50], `MultiResC` [33], `MT-DPM` [39]). More interesting results have been obtained when modelling parts and their deformations inside a deep architecture (e.g.`DBN--Mut` [34], `JointDeep` [40]).

`DPM` and its variants are systematically outmatched by methods using a single component and no parts (`Roerei` [31], `SquaresChnFtrs` see section 4.1), casting doubt on the need for parts. Recent work has explored ways to capture deformations entirely without parts [51,52].

Conclusion For pedestrian detection there is still no clear evidence for the necessity of components and parts, beyond the case of occlusion handling.

3.7 Multi-scale Models

Typically for detection, both high and low resolution candidate windows are resampled to a common size before extracting features. It has recently been noticed that training different models for different resolutions systematically improve performance by $1 \sim 2$ MR percent points [31,33,39], since the detector has access to the full information available at each window size. This technique does not impact computational cost at detection time [53], although training time increases.

Conclusion Multi-scale models provide a simple and generic extension to existing detectors. Despite consistent improvements, their contribution to the final quality is rather minor.

3.8 Deep Architectures

Large amounts of training data and increased computing power have lead to recent successes of deep architectures (typically convolutional neural networks) on diverse computer vision tasks (large scale classification and detection [54, 55], semantic labelling [56]). These results have inspired the application of deep architectures to the pedestrian task.

`ConvNet` [13] uses a mix of unsupervised and supervised training to create a convolutional neural network trained on INRIA. This method obtains fair results on INRIA, ETH, and TUD-Brussels, however fails to generalise to the Caltech setup. This method learns to extract features directly from raw pixel values.

Another line of work focuses on using deep architectures to jointly model parts and occlusions (`DBN-Isol` [28], `DBN-Mut` [34], `JointDeep` [40], and `SDN` [41]). The performance improvement such integration varies between 1.5 to 14 MR percent points. Note that these works use edge and colour features [28,34, 40], or initialise network weights to edge-sensitive filters, rather than discovering features from raw pixel values as usually done in deep architectures. No results have yet been reported using features pre-trained on ImageNet [54,57].

Conclusion Despite the common narrative there is still no clear evidence that deep networks are good at learning features for pedestrian detection (when using pedestrian detection training data). Most successful methods use such architectures to model higher level aspects of parts, occlusions, and context. The obtained results are on par with DPM and decision forest approaches, making the advantage of using such involved architectures yet unclear.

3.9 Better Features

The most popular approach (about 30 % of the considered methods) for improving detection quality is to increase/diversify the features computed over the input image. By having richer and higher dimensional representations, the classification task becomes somewhat easier, enabling improved results. A large set of feature types have been explored: edge information [1,26,41,58], colour information [22,26], texture information [17], local shape information [38], covariance features [24], amongst others. More and more diverse features have been shown to systematically improve performance.

While various decision forest methods use 10 feature channels (`ChnFtrs`, `ACF`, `Roerei`, `SquaresChnFtrs`, etc.), some papers have considered up to an order of magnitude more channels [16,24,30,38,58]. Despite the improvements by adding many channels, top performance is still reached with only 10 channels (6 gradient orientations, 1 gradient magnitude, and 3 colour channels, we name these HOG+LUV); see table 1 and figure 3. In section 4.1 we study in more detail different feature combinations.

From all what we see, from `VJ` (95% MR) to `ChnFtrs` (56.34% MR, by adding HOG and LUV channels), to `SquaresChnFtrs-Inria` (50.17% MR, by exhaustive search over pooling sizes, see section 4), improved features drive progress. Switching training sets (section 3.1) enables `SquaresChnFtrs-Caltech` to reach state of the art performance on the Caltech-USA dataset; improving over significantly more sophisticated methods. `InformedHaar` [43] obtains top results by using a set of Haar-like features manually designed for the pedestrian detection task. In contrast `SquaresChnFtrs-Caltech` obtains similar results without using such hand-crafted features and being data driven instead.

Upcoming studies show that using more (and better features) yields further improvements [59,60]. It should be noted that better features for pedestrian detection have not yet been obtained via deep learning approaches (see caveat on ImageNet features in section 3.8).

Conclusion In the last decade improved features have been a constant driver for detection quality improvement, and it seems that it will remain so in the years to come. Most of this improvement has been obtained by extensive trial and error. The next scientific step will be to develop a more profound understanding of the what makes good features good, and how to design even better ones[2].

[2] This question echoes with the current state of the art in deep learning, too.

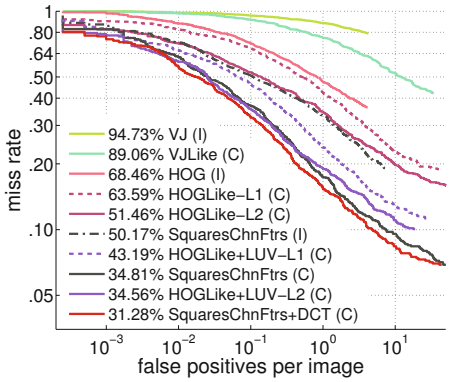


Fig. 5. Effect of features on detection performance. Caltech-USA reasonable test set.

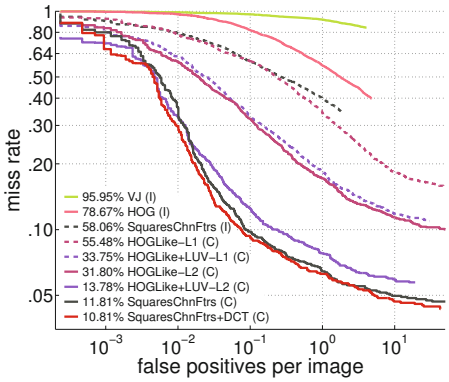


Fig. 6. Caltech-USA training set performance. (I)/(C) indicates using INRIA/Caltech-USA training set.

4 Experiments

Based on our analysis in the previous section, three aspects seem to be the most promising in terms of impact on detection quality: better features (§3.9), additional data (§3.4), and context information (§3.5). We thus conduct experiments on the complementarity of these aspects.

Among the three solution families discussed (section 3.2), we choose the Integral Channels Features framework [26] (a decision forest) for conducting our experiments. Methods from this family have shown good performance, train in minutes∼hours, and lend themselves to the analyses we aim.

In particular, we use the (open source) `SquaresChnFtrs` baseline described in [31]: 2048 level-2 decision trees (3 threshold comparisons per tree) over `HOG+LUV` channels (10 channels), composing one 64×128 pixels template learned via vanilla AdaBoost and few bootstrapping rounds of hard negative mining.

4.1 Reviewing the Effect of Features

In this section, we evaluate the impact of increasing feature complexity. We tune all methods on the INRIA test set, and demonstrate results on the Caltech-USA test set (see figure 5). Results on INRIA as well as implementation details can be found in the supplementary material.

The first series of experiments aims at mimicking landmark detection techniques, such as `VJ` [44], `HOG`+linear SVM [1], and `ChnFtrs` [26]. `VJLike` uses only the luminance colour channel, emulating the Haar wavelet like features from the original [44] using level 2 decision trees. `HOGLike-L1/L2` use 8×8 pixel pooling regions, 1 gradient magnitude and 6 oriented gradient channels, as well as level 1/2 decision trees. We also report results when adding the LUV colour channels `HOGLike+LUV` (10 feature channels total). `SquaresChnFtrs` is the baseline described in the beginning of section 4, which is similar to `HOGLike+LUV` to but with square pooling regions of any size.

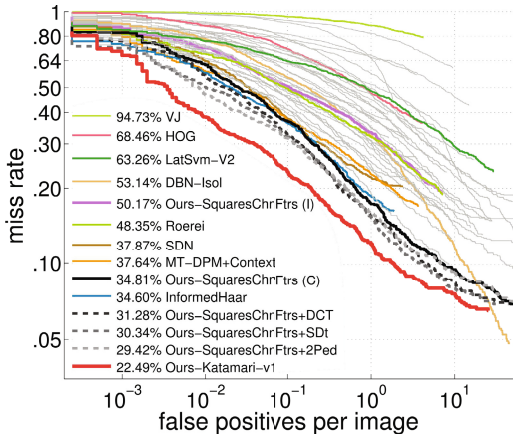


Fig. 7. Some of the top quality detection methods for Caltech-USA. See section 4.2.

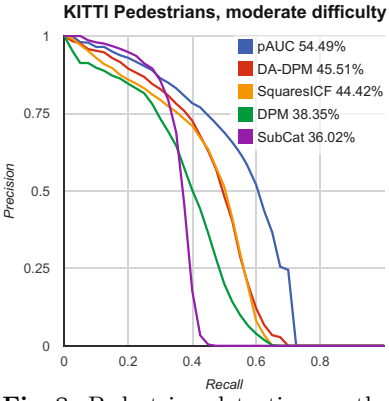


Fig. 8. Pedestrian detection on the KITTI dataset

Inspired by [60], we also expand the 10 HOG+LUV channels into 40 channels by convolving each channel with three DCT (discrete cosine transform) basis functions (of 7×7 pixels), and storing the absolute value of the filter responses as additional feature channels. We name this variant `SquaresChnFtrs+DCT`.

Conclusion Much of the progress since `VJ` can by explained by the use of better features, based on oriented gradients and colour information. Simple tweaks to these well known features (e.g. projection onto the DCT basis) can still yield noticeable improvements.

4.2 Complementarity of Approaches

After revisiting the effect of single frame features in section 4.1 we now consider the complementary of better features (HOG+LUV+DCT), additional data (via optical flow), and context (via person-to-person interactions).

We encode the optical flow using the same SDt features from `ACF+SDt` [33] (image difference between current frame T and coarsely aligned T-4 and T-8). The context information is injected using the `+2Ped` re-weighting strategy [35] (the detection scores are combined with the scores of a "2 person" DPM detector). In all experiments both DCT and SDt features are pooled over 8×8 regions (as in [33]), instead of "all square sizes" for the HOG+LUV features.

The combination `SquaresChnFtrs+DCT+SDt+2Ped` is called `Katamari-v1`. Unsurprisingly, `Katamari-v1` reaches the best known result on the Caltech-USA dataset. In figure 7 we show it together with the best performing method for each training set and solution family (see table 1). The supplementary material contains results of all combinations between the ingredients.

Conclusion Our experiments show that adding extra features, flow, and context information are largely complementary (12 % gain, instead of 3 + 7 + 5 %), even when starting from a strong detector.
It remains to be seen if future progress in detection quality will be obtained by further insights of the "core" algorithm (thus further diminishing the relative improvement of add-ons), or by extending the diversity of techniques employed inside a system.

4.3 How Much Model Capacity Is Needed?

The main task of detection is to generalise from training to test set. Before we analyse the generalisation capability (section 4.4), we consider a necessary condition for high quality detection: is the learned model performing well on the training set?

In figure 6 we see the detection quality of the models considered in section 4.1, when evaluated over their training set. None of these methods performs perfectly on the training set. In fact, the trend is very similar to performance on the test set (see figure 5) and we do not observe yet symptoms of over-fitting.

Conclusion Our results indicate that research on increasing the discriminative power of detectors is likely to further improve detection quality. More discriminative power can originate from more and better features or more complex classifiers.

4.4 Generalisation Across Datasets

For real world application beyond a specific benchmark, the generalisation capability of a model is key. In that sense results of models trained on INRIA and tested on Caltech-USA are more relevant than the ones trained (and tested) on Caltech-USA.

Table 2. Effect of training set over the detection quality. Bold indicates second best training set for each test set, except for ETH where bold indicates the best training set.

Test set \ Training set	INRIA	Caltech-USA	KITTI
INRIA	*17.42* %	60.50 %	**55.83** %
Caltech-USA	**50.17** %	*34.81* %	61.19 %
KITTI	**38.61** %	28.65 %	*44.42* %
ETH	**56.27** %	76.11%	61.19 %

Table 2 shows the performance of `SquaresChnFtrs` over Caltech-USA when using different training sets (MR for INRIA/Caltech/ETH, AUC for KITTI). These experiments indicate that training on Caltech or KITTI provides little generalisation capability towards INRIA, while the converse is not true. Surprisingly, despite the visual similarity between KITTI and Caltech, INRIA is the second best training set choice for KITTI and Caltech. This shows that Caltech-USA pedestrians are of "their own kind", and that the INRIA dataset is effective due to its diversity. In other words few diverse pedestrians (INRIA) is better than many similar ones (Caltech/KITTI).

The good news is that the best methods seem to perform well both across datasets and when trained on the respective training data. Figure 8 shows methods trained and tested on KITTI, we see that `SquaresChnFtrs` (named `SquaresICF` in KITTI) is better than vanilla DPM and on par with the best known DPM variant. The currently best method on KITTI, `pAUC` [59], is a variant of `ChnFtrs` using 250 feature channels (see the KITTI website for details on the methods). These two observations are consistent with our discussions in sections 3.9 and 4.1.

Conclusion While detectors learned on one dataset may not necessarily transfer well to others, their ranking is stable across datasets, suggesting that insights can be learned from well-performing methods regardless of the benchmark.

5 Conclusion

Our experiments show that most of the progress in the last decade of pedestrian detection can be attributed to the improvement in features alone. Evidence suggests that this trend will continue. Although some of these features might be driven by learning, they are mainly hand-crafted via trial and error.

Our experiment combining the detector ingredients that our retrospective analysis found to work well (better features, optical flow, and context) shows that these ingredients are mostly complementary. Their combination produces best published detection performance on Caltech-USA.

While the three big families of pedestrian detectors (deformable part models, decision forests, deep networks) are based on different learning techniques, their state-of-the-art results are surprisingly close.

The main challenge ahead seems to develop a deeper understanding of what makes good features good, so as to enable the design of even better ones.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
2. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR. IEEE Press, June 2008
3. Wojek, C., Walk, S., Schiele, B.: Multi-cue onboard pedestrian detection. In: CVPR (2009)
4. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: Survey and experiments. PAMI (2009)
5. Keller, C.G., Llorca, D.F., Gavrila, D.M.: Dense stereo-based roi generation for pedestrian detection. In: Denzler, J., Notni, G., Süße, H. (eds.) Pattern Recognition. LNCS, vol. 5748, pp. 81–90. Springer, Heidelberg (2009)
6. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: a benchmark. In: CVPR (2009)
7. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and PatternRecognition (CVPR) (2012)

8. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. TPAMI (2011)
9. Viola, P., Jones, M.: Robust real-time face detection. IJCV (2004)
10. Sabzmeydani, P., Mori, G.: Detecting pedestrians by learning shapelet features. In: CVPR (2007)
11. Lin, Z., Davis, L.S.: A pose-invariant descriptor for human detection and segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 423–436. Springer, Heidelberg (2008)
12. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR (2008)
13. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: CVPR (2013)
14. Dollár, P., Tu, Z., Tao, H., Belongie, S.: Feature mining for image classification. In: CVPR (2007)
15. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR (2008)
16. Wojek, C., Schiele, B.: A performance evaluation of single and multi-feature people detection. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 82–91. Springer, Heidelberg (2008)
17. Wang, X., Han, X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV (2009)
18. Levi, D., Silberstein, S., Bar-Hillel, A.: Fast multiple-part based object detection using kd-ferns. In: CVPR (2013)
19. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI (2010)
20. Schwartz, W., Kembhavi, A., Harwood, D., Davis, L.S.: Human detection using partial least squares analysis. In: ICCV (2009)
21. Nam, W., Han, B., Han, J.: Improving object localization using macrofeature layout selection. In: ICCV, Visual Surveillance Workshop (2011)
22. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: CVPR (2010)
23. Bar-Hillel, A., Levi, D., Krupka, E., Goldberg, C.: Part-based feature synthesis for human detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 127–142. Springer, Heidelberg (2010)
24. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Efficient pedestrian detection by directly optimize the partial area under the roc curve. In: ICCV (2013)
25. Dollár, P., Belongie, S., Perona, P.: The fastest pedestrian detector in the west. In: BMVC (2010)
26. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: BMVC (2009)
27. Dollár, P., Appel, R., Kienzle, W.: Crosstalk cascades for frame-rate pedestrian detection. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 645–659. Springer, Heidelberg (2012)
28. Ouyang, W., Wang, X.: A discriminative deep model for pedestrian detection with occlusion handling. In: CVPR (2012)
29. Dollár, P., Appel, R., Belongie, S., Perona, P.: Fast feature pyramids for object detection. PAMI (2014)
30. Marin, J., Vazquez, D., Lopez, A., Amores, J., Leibe, B.: Random forests of local experts for pedestrian detection. In: ICCV (2013)
31. Benenson, R., Mathias, M., Tuytelaars, T., Van Gool, L.: Seeking the strongest rigid detector. In: CVPR (2013)

32. Mathias, M., Benenson, R., Timofte, R., Van Gool, L.: Handling occlusions with franken-classifiers. In: ICCV (2013)
33. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 241–254. Springer, Heidelberg (2010)
34. Ouyang, W., Zeng, X., Wang, X.: Modeling mutual visibility relationship with a deep model in pedestrian detection. In: CVPR (2013)
35. Ouyang, W., Wang, X.: Single-pedestrian detection aided by multi-pedestrian detection. In: CVPR (2013)
36. Chen, G., Ding, Y., Xiao, J., Han, T.X.: Detection evolution with multi-order contextual co-occurrence. In: CVPR (2013)
37. Zeng, X., Ouyang, W., Wang, X.: Multi-stage contextual deep learning for pedestrian detection. In: ICCV (2013)
38. Costea, A.D., Nedevschi, S.: Word channel based multiscale pedestrian detection without image resizing and using only one classifier. In: CVPR, June 2014
39. Yan, J., Zhang, X., Lei, Z., Liao, S., Li, S.Z.: Robust multi-resolution pedestrian detection in traffic scenes. In: CVPR (2013)
40. Ouyang, W., Wang, X.: Joint deep learning for pedestrian detection. In: ICCV (2013)
41. Luo, P., Tian, Y., Wang, X., Tang, X.: Switchable deep network for pedestrian detection. In: CVPR (2014)
42. Park, D., Zitnick, C.L., Ramanan, D., Dollár, P.: Exploring weak stabilization for motion feature extraction. In: CVPR (2013)
43. Zhang, S., Bauckhage, C., Cremers, A.B.: Informed haar-like features improve pedestrian detection. In: CVPR (2014)
44. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: CVPR (2003)
45. Keller, C.G., Enzweiler, M., Rohrbach, M., Fernandez Llorca, D., Schnorr, C., Gavrila, D.M.: The benefits of dense stereo for pedestrian detection. IEEE Transactions on Intelligent Transportation Systems (2011)
46. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: Robust multi-person tracking from a mobile platform. PAMI (2009)
47. Premebida, C., Carreira, J., Batista, J., Nunes, U.: Pedestrian detection combining rgb and dense lidar data. In: IROS (2014)
48. Enzweiler, M., Gavrila, D.: A multilevel mixture-of-experts framework for pedestrian classification. IEEE Transactions on Image Processing (2011)
49. Tu, Z., Bai, X.: Auto-context and its application to high-level vision tasks and 3d brain image segmentation. PAMI (2010)
50. Yan, J., Lei, Z., Wen, L., Li, S.Z.: The fastest deformable part model for object detection. In: CVPR, June 2014
51. Hariharan, B., Zitnick, C.L., Dollár, P.: Detecting objects using deformation dictionaries. In: CVPR (2014)
52. Pedersoli, M., Tuytelaars, T., Gool, L.V.: Using a deformation field model for localizing faces and facial points under weak supervision. In: CVPR, June 2014
53. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: CVPR (2012)
54. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: arXiv (2014)
55. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: ICLR (2014)

56. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: JMLR (2014)
57. Azizpour, H., Razavian, A.S., Sullivan, J., Maki, A., Carlsson, S.: From generic to specific deep representations for visual recognition. CoRR (2014)
58. Lim, J., Zitnick, C.L., Dollár, P.: Sketch tokens: a learned mid-level representation for contour and object detection. In: CVPR (2013)
59. Paisitkriangkrai, S., Shen, C., van den Hengel, A.: Strengthening the effectiveness of pedestrian detection with spatially pooled features. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 546–561. Springer, Heidelberg (2014)
60. Nam, W., Dollár, P., Han, J.H.: Local decorrelation for improved detection. In: Nips (2014)

Fast 3-D Urban Object Detection on Streaming Point Clouds

Attila Börcs(✉), Balázs Nagy, and Csaba Benedek

Distributed Events Analysis Research Laboratory,
Institute for Computer Science and Control of the Hungarian Academy of Sciences,
Kende utca 13-17, Budapest H-1111, Hungary
{Attila.Borcs,Balazs.Nagy,Csaba.Benedek}@sztaki.mta.hu

Abstract. Efficient and fast object detection from continuously streamed 3-D point clouds has a major impact in many related research tasks, such as autonomous driving, self localization and mapping and understanding large scale environment. This paper presents a LIDAR-based framework, which provides fast detection of 3-D urban objects from point cloud sequences of a Velodyne HDL-64E terrestrial LIDAR scanner installed on a moving platform. The pipeline of our framework receives raw streams of 3-D data, and produces distinct groups of points which belong to different urban objects. In the proposed framework we present a simple, yet efficient hierarchical grid data structure and corresponding algorithms that significantly improve the processing speed of the object detection task. Furthermore, we show that this approach confidently handles streaming data, and provides a speedup of two orders of magnitude, with increased detection accuracy compared to a baseline connected component analysis algorithm.

Keywords: LIDAR · Urban object detection · 3-D point clouds · Dynamic processing

1 Introduction

1.1 Problem Statement

The reliable perception of the surrounding environment is an important task in outdoor robotics. Robustly detecting and identifying various urban objects are key problems for autonomous driving, and driving assistance systems. Future mobile vision systems promise a number of benefits for the society, including prevention of road accidents by constantly monitoring the surrounding vehicles or ensuring more comfort and convenience for the drivers. Vision systems with capability of handling continuously streamed sensor data have become important tools for robot perception [13]. Laser range sensors are particularly efficient

This work was partially funded by the Government of Hungary through a European Space Agency (ESA) Contract under the Plan for European Cooperating States (PECS), and by the Hungarian Research Fund (OTKA #101598).

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 628–639, 2015.
DOI: 10.1007/978-3-319-16181-5_48

for these tasks since in contrast to conventional camera systems they are highly robust against illumination changes or weather conditions, and they may provide a larger field of view. Moreover, LIDAR mapping systems are able to rapidly acquire large-scale 3-D point cloud data for real-time vision, with jointly providing accurate 3-D geometrical information of the scene, and additional features about the reflection properties and compactness of the surfaces. The detection of urban objects is a fundamental problem in any perception motivated point cloud processing task [15]. Although it is a challenging problem itself, it can be helpful for several robot vision tasks, such as object recognition, localization or feature extraction. We focus here on the object detection problem relying on large-scale terrestrial urban point clouds, more specifically, we use point set data obtained by a Velodyne HDL-64 S2 laser acquisition system. The problem of detecting objects on streaming point clouds is challenging for various reasons. First, the raw sensor measurements are noisy. Second, the point density is uneven: [2] typically in terrestrial LIDAR point clouds the point densities dominate from the direction the measurement is taken, causing strongly corrupted geometric properties of the objects such as missing object parts or deformed shapes. The object detection process is further complicated when the data is continuously streamed from a laser sensor on a moving platform or a mobile robot. In this case we are forced to complete a complex task within a very limited time frame.

1.2 Related Works

A number of approaches are available in the literature for solving 3-D object detection and recognition problems in outdoor laser scans. The used data structure are essential part all of the existing techniques, and they can be coarsely divided into *two categories*.

In the *first* category, traditional pre-computed tree-based data structures can be used, such as Kd-tree, Octree, range tree [3],[14]. These structures are efficient for performing range search, although there is a large processing overhead at initialization, and their performance rapidly degrades as updated data inserted after calling for reconstruction the tree structure [11]. Recent approaches apply different region growing techniques over tree-based structures to obtain coherent objects. The authors of [1] present an octree based occupancy grid representation to model the dynamic environment surrounding the vehicle and to detect moving objects based on inconsistencies between scans. However, the run-time and detection performance of the algorithm is not discussed here.

The *second* category of the methods focus on grid-based data structures and efficient dynamic processing techniques for fast detection or recognition of objects from streaming 3-D data. In [7] the authors propose a fast segmentation of point clouds into objects, which is accomplished by a standard connected component algorithm in a 2-D occupancy grid, and object classification is done on the raw point cloud segments with 3-D shape descriptors and a SVM classifier. Different voxel grid structures are also widely used to complete various scene understanding tasks, including segmentation, detection and recognition [11]. The data is stored here in cubic voxels for efficient retrieval of the 3-D points.

Efficient range search from streaming data is an essential component of any object detection problem, and can be used for retrieval of all points which fall within a certain distance of a given point. For this task, a scrolling voxel grid data structure was proposed by [11]. The data is quantized here into small voxels of a prespecified resolution, then the indices of the voxels are shifted using a circular buffer according to the robot motion. To handle querying a large subvolume of space in sparse data, a sparse global grid was proposed by [8], when all streamed measurements were stored in a voxel-based global map. All of the approaches mentioned above provide convincing object detection results in large scale 3-D environment but they have some important limitations. Firstly, standard connected component solutions over tree-based data structures give very precise detection results, but they are not fast enough to serve real-time vision systems. Although, there exist efficient data structures for modifying minimum spanning trees which have sublinear complexity for each online update [4], this solution is impractical with streaming 3-D data [8]. Secondly, recent studies which suggest voxel, 2-D, scrolling, octree -grid based data structures for detection or recognition tasks do not propose optimal grid parameter settings (*e.g.* grid resolution or grid cell size) in order to minimize execution time, and maximize detection accuracy. Instead, they choose one certain grid resolution heuristically, and evaluate the performance of their detection method on this predefined grid resolution.

2 Proposed Approach

We propose a new data structure and a corresponding algorithm which is a basis of an efficient range search technique and a connected component analysis approach for fast object detection. In addition an optimal parameter setting strategy is proposed for enhancing the accuracy, which leads to the same or better detection performance than the tree-based approaches. More specifically, the following four main improvements have been implemented:

◇ *Novel 2-D hierarchical grid structure for fast range search in 3-D:* a multi-level 2-D grid structure is presented with *two* different grid resolution levels (low and high). This structure is specifically designed for object detection *i.e.* connected component analysis tasks. We use these different grid levels to provide efficient and fast retrieval of 3-D point cloud features for the object detector module of our framework even in cases of strongly inhomogeneous point cloud density. We have experienced that standard 2-D grid structures [7] may give a decent result for region segmentation tasks *e.g.* ground detection, but they are not accurate enough near to the object boundaries, and they do not perform well in case of nearby urban objects. On one hand, using a large cell size multiple objects can occur within a given cell, resulting in several objects merged to the same extracted component. On the other hand a too dense grid structure (*i.e.* small cell size) may yield cells containing only a few points, which case does not enable us to calculate discriminative point cloud features for reliable classification. In Section 3 we introduce the proposed grid structure in details.

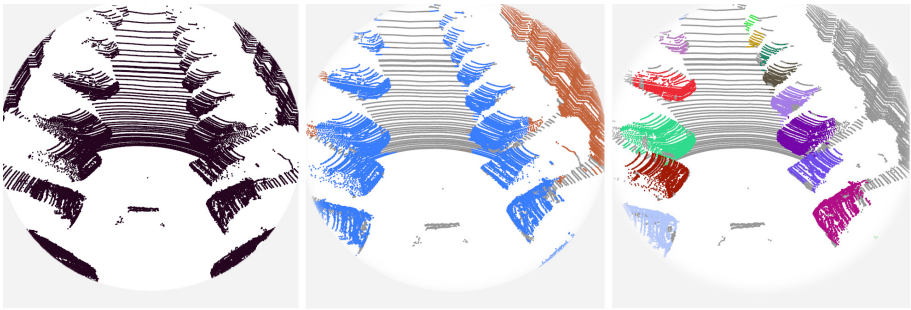


Fig. 1. Overview of the proposed object detection workflow [Note: color codes in Fig. (b): brown = walls, grey=ground, blue=other street objects]

◇ *Connected Component Algorithm for streaming data:* a simple, yet efficient connected component analysis method is proposed in the hierarchical grid data structure, which provides reliable detection results in urban environment with real-time performance. In contrast to previous works [6],[14] this module of our framework queries local 3-D point cloud features from the hierarchical grid, and decides which 3-D points belong to the same urban object. The algorithm relies on different merging criteria to fulfill this task. See Section 3.2 for the details.

◇ *Optimal grid resolution in urban environment:* In case of grid-based detection tasks, one of the biggest challenges is to find decent trade-off with respect to speed and accuracy. The major factor which can influence these properties is the grid resolution *i.e.* the size of a grid cell. It is crucial to select optimal grid resolution to keep the detection accuracy high, and the processing time low. In [7] the grid size has been selected manually without justification. Other approaches measure the entropies of the misclassification rate within the grid cell compared to different cell sizes. As a compromise to balance efficiency and accuracy they choose a certain grid resolution [8]. In contrast to above solutions, we propose a novel statistical metrics for approximation of the optimal grid resolution in terms of object detection.

◇ *Publishing a new large dataset of hand-labeled 3-D point clouds:* We implemented a 3-D point cloud annotation tool for two reasons: First, we intend to provide a free annotated dataset to the research community. Second, using the Ground Truth (GT) we can evaluate the performance of our algorithm quantitatively, and we can compare it to earlier solutions.

The detailed description of the proposed object detection framework is structured as follows. In Section 3 we present a data structure that will allow us to perform fast retrieval of 3-D point cloud features for segmentation and detection purposes. In Section 3.1 we describe our point cloud segmentation algorithm

(see Fig. 1 b)). The point cloud is classified into large semantic regions such as *ground, walls, street objects* to prepare the data for object detection, which is presented in Section 3.2 (see Fig. 1 c)). We discuss the parameter sensitivity and the performance evaluation of the proposed grid model in Section 4 and 5.

3 Data Structures

In this section, we introduce the grid based data structures used in the proposed system. First, we construct a *Simple Grid Model* [9] which will be used for initial point cloud segmentation, *i.e.* separating regions of roads, walls and short street objects. Second, we present a novel *Hierarchical Grid Model* which will be used for robust 3-D object detection from the strongly inhomogeneous density point clouds in challenging dense urban environments where several nearby object may be located close to each other.

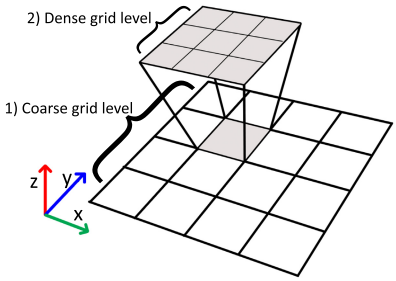


Fig. 2. Visualization of our *hierarchical grid model* data structure - *(bottom)* the coarse grid level: the 3-D space coarsely quantized into 2-D grid cells, *(top)* the dense grid level: each grid cell on the coarse level subdivided into smaller cells

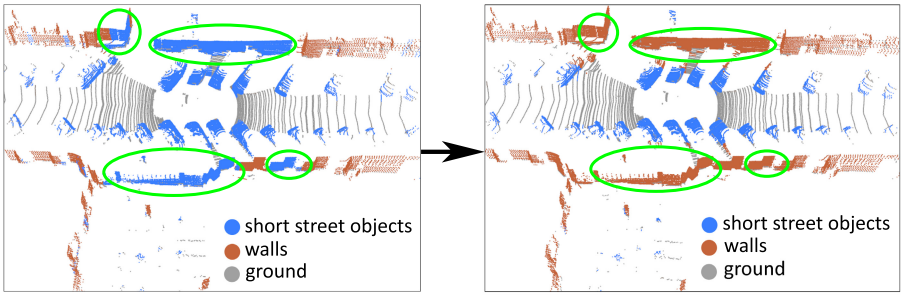
◇ *Simple Grid Model:* We fit a regular 2-D grid S with W_S rectangle side length onto the $P_{z=0}$ plane (using the Velodyne sensor's vertical axis as the z direction and the sensor height as a reference coordinate), where $s \in S$ denotes a single cell. We assign each $p \in \mathcal{P}$ point of the point cloud to the corresponding cell s_p, which contains the projection of p to $P_{z=0}$. Let us denote by $\mathcal{P}_s = \{p \in \mathcal{P} : s = s_p\}$ the point set projected to cell s. Moreover, we store the height coordinate and different height properties such as, maximum $z_{\max}(s)$, minimum $z_{\min}(s)$ and average $\hat{z}(s)$ of the elevation values within cell s, which quantities will be used later in point cloud segmentation.

◇ *Hierarchical Grid Model:* Our key idea is to create an extended grid based approach (see Fig. 2) called *hierarchical grid model* which uses a coarse and dense grid resolution. The cell s of the coarse grid level is subdivided into smaller cells $s'_d | d \in \{1, 2, \ldots, \xi^2\}$, with cell side length $W_{s'_d} = W_s/\xi$, where ξ is a scaling

factor (used $\xi = 3$). We store each 3-D point in the coarse and dense grid level as well. We use this data construction to perform object detection, as detailed in Section 3.2.

3.1 Point Cloud Segmentation Using a Simple Grid Model

In our system, point cloud segmentation is achieved by a *simple grid based* approach. Our goal is to discriminate regions of ground, walls and short street objects in the input cloud. For ground segmentation we apply a locally adaptive terrain modeling approach similarly to [9], which is able to accurately extract the road regions, even if their surfaces are not perfectly planar.



a) Initial point cloud segmentation b) Output of the Hough-based wall fitting step

Fig. 3. The refinement of the point cloud segmentation result with probabilistic Hough transformation - (*left*) the misclassified cloud regions denoted by green circles. *(right)* Point cloud segmentation after Hough-based wall fitting step.

We use point height information for assigning each grid cell to the corresponding cell class. Before that, we detect and remove grid cells that belong to clutter regions, thus we will not visit these cells later and save processing time. We classify any cell to *clutter*, which contains less points than a predefined threshold (typically 4-8 points). After clutter removal all the points in a cell are classified as *ground*, if the difference of the minimal and maximal point elevations in the cell is smaller than a threshold (used 25cm), moreover the average of the elevations in neighboring cells does not exceeds an allowed height range. A cell belongs to the class of *tall structure objects* (e.g. traffic signs, building walls, lamp post etc.), if either the maximal point height within the cell is larger than a predefined value (used 140cm), or the observed point height difference is larger than a threshold (used 310cm). The rest of the points in the cloud are assigned to class *short street object* belonging to vehicles, pedestrians, mail boxes, billboards etc. Due to the limited vertical view angle of the Velodyne LIDAR (+2° up to -24.8° down), the defined elevation criteria may fail near to the sensor position. In narrow streets where road sides located closely to the measurement position, several nearby grid cells can be misclassified regularly

e.g. some parts of the walls and the building facades are classified to *short street object* cell class instead of *tall object* cell class (see Fig. 3a)). Our aim is to filter out all of the tall objects, facades and wall structures from the scene, and use only the *short object* class labels for object detection. For this purpose we proposed a probabilistic Hough transformation based segmentation refinement. The grid cells with class labels *tall object* and *short street object* were projected into a pixel lattice (*i.e.* an image), and a probabilistic Hough transformation [12] was used to detect long - elongated structures, which belong to facades or walls in the original point cloud, thereafter the detected lines were back projected into a cloud. The class labels of the grid cells are updated from *short street object* to *tall object* if 1): the grid cell position fits the detected Hough-lines, and 2): the class label of the grid cell was *short street object* before the Hough based refinement step (see Fig. 3b)).

3.2 Urban Object Detection with a Hierarchical Grid Model

In this section we present the object detection step of our framework. Our aim is to find distinct groups of points which belong to different urban objects on the scene. We used the initial segmentation from Section 3.1, with considering the *short object* cell class as *foreground*, while we label the other classes as *background*. For object detection we use the *hierarchical grid model*: On one hand, the coarse grid resolution is appropriate for a rough estimation of the 3-D blobs in the scene, in this way we can also roughly estimate the size and the location of possible object candidates. In addition, we optimize the grid resolution parameter with a statistical approach (see Section 4), instead of setting the cell size parameters by hand similarly to [7], [8]. On the other hand, using a dense grid resolution beside a coarse grid level, is efficient for calculating point cloud features from a smaller subvolume of space, therefore we can refine the detection result derived from the coarse grid resolution. The proposed object detection algorithm consists of three main steps: *First*, we visit every cell of the coarse grid and for each cell s we consider the cells in its 3×3 neighborhood. We visit the neighbor cells one after the other in order to calculate two different point cloud features: (i) the maximal elevation value $Z_{max}(s)$ within a coarse grid cell and (ii) the point cloud density (*i.e.* point cardinality) of a dense grid cell. *Second* our intention is to find connected 3-D blobs within the foreground regions, by merging the coarse level grid cells together. We use an elevation-based cell merging criterion to perform this step. $\psi(s, s_r) = |Z_{max}(s) - Z_{max}(s_r)|$ is a merging indicator, which measures the difference between the maximal point elevation within cell s and its neighboring cell s_r. If the ψ indicator is smaller than a predefined value, we assume that s and s_r belong to the same 3-D object. *Third*, we perform a detection refinement step on the dense grid level. The elevation based cell merging criterion on the coarse grid level often yields that nearby and self-occluded objects are merged into a same blob. We handle this issue by measuring the point density in each sub-cell s'_d at the dense grid level. Our assumption is here that the nearby objects, which were erroneously merged at the coarse level, could be appropriately separated at the fine level, as the examples in Fig. 4

show. Note that using our Velodyne Lidar camera, the density of the recorded point cloud strongly decreases as a function of the distance from the sensor. We had to compensate this effect by a sensor distance based weighting of the cells during the density based merging step. After the weighting step, we expect by an order of magnitude similar point density in each sub-cell s'_d which belongs to the object candidates. On the other hand, if we observe several empty or low-density sub-cells on the border of two neighboring super-cells, or in the center line of a large cell we can assume that the blob extracted at the coarse level should be divided into two objects. Let us present three typical urban scenarios when the *simple* coarse grid model merges the close objects to the same extracted component, while using a *hierarchical* grid model with coarse and dense grid level, the objects can be appropriately separated. We consider two neighboring super-cell pairs -marked by red - in Fig. 4a) and Fig. 4b), respectively. In both cases the cells contain points from different objects, which fact cannot be justified at the coarse cell level. However, at the dense level, we can identify connected regions of near-empty sub-cells (denoted by gray), which separate the two objects. Fig. 4c) demonstrates a third configuration, where a super-cell intersects two objects, but at the sub-cell level, we can find a separator line.

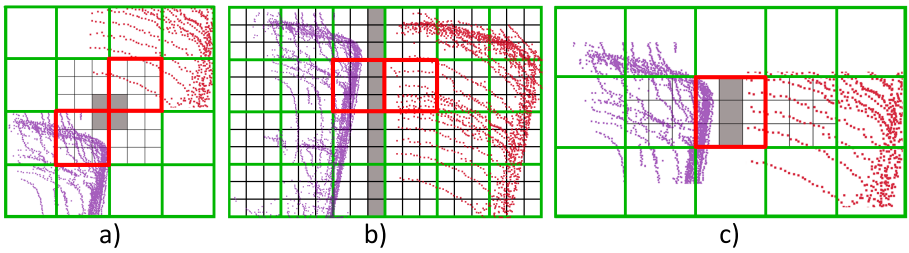


Fig. 4. Separation of close objects at the dense grid level [color codes: green lines =coarse grid level, black lines=dense grid level, grey cells= examined regions for object separation]

4 Data Charasteristic Analysis and Parameter Sensitivity

Data Characteristic Analysis:

By using a terrestrial laser scanner, such as the Velodyne LIDAR the data density decreases significantly as function of measurement distance from the sensor. This inhomogeneous point density makes the cell-merging based object detection task challenging. In order to compensate these artifacts for our sensor, we analyzed 1600 relevant frames from three different urban scenarios containing main roads, narrow streets and intersections. We create rings around the sensor position, thereafter we set the width of each ring to 1m, and we shift the disjunct rings from 1 to 80 meter from the sensor. Finally we measure the distribution of the point density in every ring normalized by the ring area as shown in Fig. 5a).

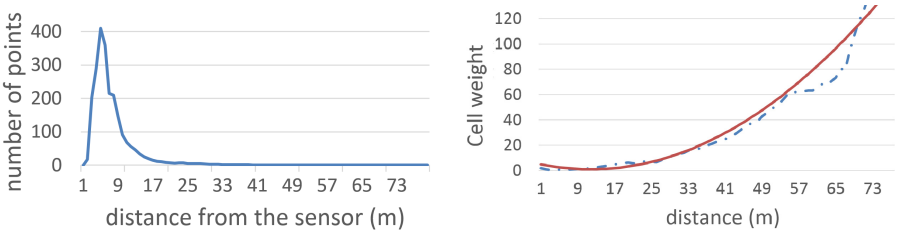


Fig. 5. (a) Point density vs. measurement distance from the sensor. (b) Grid cell weights vs. measurement distance from the sensor. [Note: color codes of Fig. (b): blue = derived weight function, red= sixth-degree polynomial fit of the weight function]

We derive a weight distribution by normalizing the point density function with the maximal point density, and use this function for create weights for the coarse and dense grid cells of the *hierarchical grid model*. Near to the sensor the weight distribution does not modify the point density of the cell, while far from the sensor where the grid cells might contain less points, we enrich the point density by the sixth-degree polynomial fit to the weight distribution, as shown in Fig. 5b).

Parameter Sensitivity:

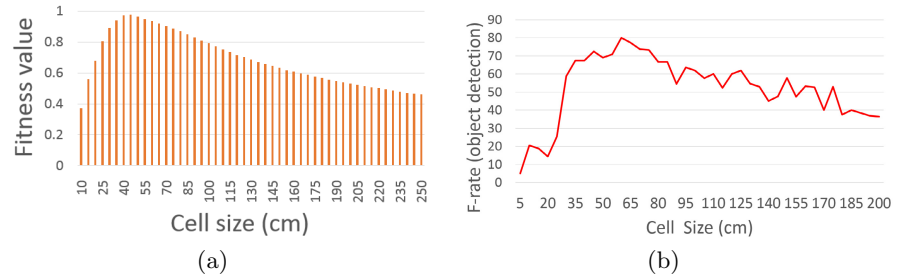


Fig. 6. (a) The distribution of the proposed *cell fitness value* for estimating optimal grid resolution. (b) The F-rate values (harmonic mean of precision and recall) of the detection step as a function of cell size.

In case of a grid based detection task one of the major factors, which affect the accuracy and the speed of the algorithm is the grid resolution (*i.e.* cell size). In order to approximate the optimal range of grid resolution, we propose a statistical metric called *cell fitness value*, which measures the ratio of *dense* (d), *sparse* (s) and *empty* (e) grid cells in different grid resolutions. We call a grid cell *dense* if it is contains more point than a minimal threshold $t_{(min)}$. We experienced that our initial point cloud segmentation method needs at least 20 points in a cell for appropriate results, therefore we choose $t_{(min)} = 20$. Finally we derived

the *cell fitness value* $f \in [0,1]$ as follows: $f = \frac{\#d}{(\#d+\#s)-\#e}$, where # denotes the number of the cells on the screen (see Fig. 6a)), in order to maximize the relative frequency of the *dense* grid cells. Moreover, the distribution of the *cell fitness value* f clearly has a maximum range as a function of grid resolution, therefore we choose an optimal grid resolution corresponds to this maximum range (used 60 cm).

Table 1. Numerical comparison of the detection results obtained by the Connected Component Analysis [14] and the proposed *Hierarchal Grid Model*. The number of objects (NO) are listed for each data set, also and in aggregate.

Point Cloud Dataset	NO	Conn. Comp. Analysis [14]		Prop. Hierarchical Grid	
		F-rate(%)	Avg. Processing Speed (fps)	F-rate(%)	Avg. Processing Speed (fps)
Budapest Dataset #1	669	77	0.38	89	29
Budapest Dataset #2	429	64	0.22	79	25
KITTI Dataset [5]	496	75	0.46	82	29
Overall	1594	72	0.35	83	28

5 Performance Evaluation and Conclusion

We evaluated our method in three urban LIDAR sequences, concerning different urban scenarios, such as main roads, narrow streets and intersections. Two scenarios recorded in the streets of Budapest, one scenario has been selected from the KITTI Vision Benchmark Suite [5]. The data flows have been recorded by a Velodyne HDL-64E S2 camera with a 10Hz rotation speed. We have compared our *hierarchical grid model* to a connected component analysis which uses kd-tree based solution for range search [14]. Qualitative results on four sample frames are shown in Fig. 7 and in Fig. 8.[1] For Ground Truth (GT) generation, we have developed a 3-D annotation tool, which enables labeling the urban objects manually as object or background. We manually annotated 1594 urban objects. To enable fully automated evaluation, we need to make first a non-ambiguous assignment between the detected objects and ground truth (GT) object samples. We use Hungarian algorithm [10] to find maximum matching. Thereafter, we counting the Missing Objects (MO), and the Falsely detected Objects (FO). These values are compared to the Number of real Objects (NO), and the F-rate of the detection (harmonic mean of precision and recall) is also calculated. We have also measured the processing speed in frames per seconds (fps). The numerical performance analysis is given in Table 1. The results confirms that proposed model surpasses the Connected Component Analysis technique in F-rate for all the scenes. Moreover, the proposed *Hierarchical Grid Model* significantly faster on streaming data, and less influenced by the inhomogeneous density of the point

[1] Demonstration videos and GT data are also available at the following url: http://web.eee.sztaki.hu/~borcs/demos.html

cloud. In urban point clouds we measure 0.35 fps average average processing with Connected Component Analysis [14] and 28 fps with the proposed *Hierarchical Grid Model*.

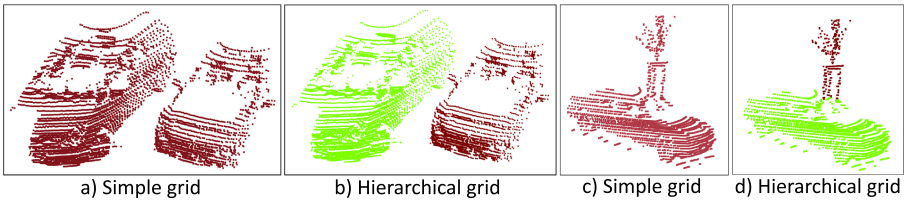


Fig. 7. Object separation for a case of nearby objects. Comparison of the *Simple Grid Model* Fig. a), c) and the *Hierarchical Grid Model* Fig. b), d).

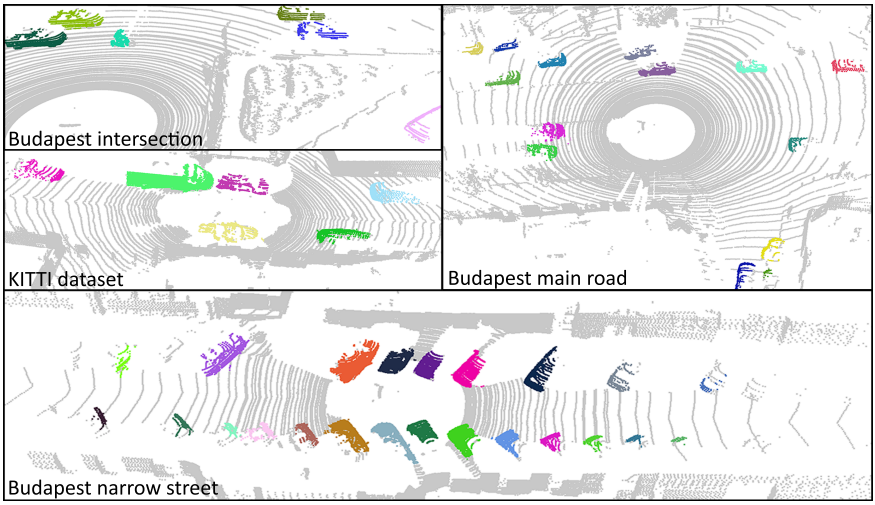


Fig. 8. Object detection results on different urban scenarios

As a conclusion, we have proposed a novel data structure, called *Hierarchical Grid Model* and corresponding connected component analysis algorithm to find distinct groups of 3-D points which belong to different urban objects in LIDAR point clouds. We propose a statistical metric for approximation of optimal grid resolution in terms of object detection. The model has been quantitatively validated based on Ground Truth data, and the advantages of the proposed solution versus a baseline technique have been demonstrated.

References

1. Azim, A., Aycard, O.: Detection, classification and tracking of moving objects in a 3D environment. In: Intelligent Vehicles Symposium, pp. 802–807 (2012)
2. Behley, J., Steinhage, V., Cremers, A.B.: Performance of histogram descriptors for the classification of 3D laser range data in urban environments. In: ICRA, pp. 4391–4398. IEEE
3. Benedek, C., Molnár, D., Szirányi, T.: A Dynamic MRF Model for Foreground Detection on Range Data Sequences of Rotating Multi-beam Lidar. In: Jiang, X., Bellon, O.R.P., Goldgof, D., Oishi, T. (eds.) WDIA 2012. LNCS, vol. 7854, pp. 87–96. Springer, Heidelberg (2013)
4. Frederickson, G.N.: Data structures for on-line updating of minimum spanning trees. In: Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing, STOC 1983, pp. 252–257. ACM, New York (1983). http://doi.acm.org/10.1145/800061.808754
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
6. Golovinskiy, A., Funkhouser, T.: Min-cut based segmentation of point clouds. In: IEEE Workshop on Search in 3D and Video (S3DV) at ICCV, September 2009
7. Himmelsbach, M., Müller, A., Lüttel, T., Wünsche, H.J.: LIDAR-based 3D Object Perception. In: Proceedings of 1st International Workshop on Cognition for Technical Systems. München, October 2008
8. Hu, H., Munoz, D., Bagnell, J.A., Hebert, M.: Efficient 3-D scene analysis from streaming data. In: IEEE International Conference on Robotics and Automation (ICRA) (2013)
9. Józsa, O., Börcs, A., Benedek, C.: Towards 4D virtual city reconstruction from Lidar point cloud sequences. In: ISPRS Workshop on 3D Virtual City Modeling, ISPRS Annals Photogram. Rem. Sens. and Spat. Inf. Sci., vol. II-3/W1, Regina, Canada, pp. 15–20 (2013)
10. Kuhn, H.: The Hungarian method for the assignment problem. Naval Research Logistic Quarterly **2**, 83–97 (1955)
11. Lalonde, J.F., Vandapel, N., Hebert, M.: Data structures for efficient dynamic processing in 3-D. The International Journal of Robotics Research **26**(8), 777–796 (2007)
12. Matas, J., Galambos, C., Kittler, J.: Robust detection of lines using the progressive probabilistic Hough transform. Comput. Vis. Image Underst. 78(1), 119–137 (2000). http://dx.doi.org/10.1006/cviu.1999.0831
13. McNaughton, M., Urmson, C., Dolan, J.M., Lee, J.W.: Motion planning for autonomous driving with a conformal spatiotemporal lattice. In: ICRA, pp. 4889–4895. IEEE (2011)
14. Rusu, R.B., Cousins, S.: 3D is here: Point cloud library (PCL). In: International Conference on Robotics and Automation, Shanghai, China (2011)
15. Thrun, S., et. al.: Stanley: The robot that won the darpa grand challenge. Journal of Field Robotics 23(9), 661–692 (2006)

Relative Pose Estimation and Fusion of Omnidirectional and Lidar Cameras

Levente Tamas[1], Robert Frohlich[2], and Zoltan Kato[2(✉)]

[1] Robotics Research Group, Technical University of Cluj-Napoca,
Dorobantilor st. 73, 400609 Cluj-Napoca, Romania
levente.tamas@aut.utcluj.ro
[2] Institute of Informatics, University of Szeged,
P.O. Box 652, Szeged H-6701, Hungary
{frohlich,kato}@inf.u-szeged.hu

Abstract. This paper presents a novel approach for the extrinsic parameter estimation of omnidirectional cameras with respect to a 3D Lidar coordinate frame. The method works without specific setup and calibration targets, using only a pair of 2D-3D data. Pose estimation is formulated as a 2D-3D nonlinear shape registration task which is solved without point correspondences or complex similarity metrics. It relies on a set of corresponding regions, and pose parameters are obtained by solving a small system of nonlinear equations. The efficiency and robustness of the proposed method was confirmed on both synthetic and real data in urban environment.

Keywords: Omnidirectional camera · Lidar · Pose estimation · Fusion

1 Introduction

There is a considerable research effort invested in autonomous car driving projects both at academic and industrial levels. While for special scenarios, such as highways, there are a number of successful applications, there is still no general solution for complex environments such as urban areas [5,11]. Recent developments in autonomous driving in urban environment rely on a great variety of close-to-market visual sensors, which requires the fusion of the visual information provided by these sensors [4].

One of the most challenging issues is the fusion of 2D RGB imagery with other 3D range sensing modalities (*e.g.* Lidar) which can also be formulated as a camera calibration task. Internal calibration refers to the self parameters of the camera, while external parameters describe the *pose* of the camera with respect to a world coordinate frame. The problem becomes more difficult, when the RGB image is recorded with a non-conventional camera, such as central catadioptric or dioptric (*e.g.* fish-eye) panoramic cameras. Although such lenses have a more complex geometric model, their calibration also involves internal parameters and external pose. Recently, the geometric formulation of omnidirectional

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 640–651, 2015.
DOI: 10.1007/978-3-319-16181-5_49

systems were extensively studied [1,6,14,17,24,25]. The internal calibration of such cameras depends on these geometric models. Although different calibration methods and toolboxes exist [10,12,24] this problem is by far not trivial and is still in focus [25].

While internal calibration can be solved in a controlled environment, using special calibration patterns, pose estimation must rely on the actual images taken in a real environment. There are popular methods dealing with point correspondence estimation such as [24] or other fiducial marker images suggested in [10], which may be cumbersome to use in real life situations. This is especially true in a multimodal setting, when omnidirectional images need to be combined with other non-conventional sensors like lidar scans providing only range data. The Lidar-omnidirectional camera calibration problem was analyzed from different perspectives: in [22], the calibration is performed in natural scenes, however the point correspondences between the 2D-3D images are selected in a semi-supervised manner. [15] tackles calibration as an observability problem using a (planar) fiducial marker as calibration pattern. In [19], a fully automatic method is proposed based on mutual information (MI) between the intensity information from the depth sensor and the omnidirectional camera. Also based on MI, [28] performs the calibration using particle filtering. However, these methods require a range data with recorded intensity values, which is not always possible and often challenged by real-life lighting conditions.

This paper introduces a novel region based calibration framework for non-conventional 2D cameras and 3D lidar. Instead of establishing point matches or relying on artificial markers or recorded intensity values, we propose a pose estimation algorithm which works on segmented planar patches. Since segmentation is required anyway in many real-life image analysis tasks, such regions may be available or straightforward to detect. The main advantage of the proposed method is the use of regions instead of point correspondence and a generic problem formulation which allows to treat several types of cameras in the same framework. We reformulate pose estimation as a shape alignment problem, which is accomplished by solving a system of nonlinear equations based on the idea of [2]. However, the equations are constructed in a different way here due to the different dimensionality of the lidar and camera coordinate frames as well as the different camera model used for omnidirectional sensors. The method has been quantitatively evaluated on a large synthetic dataset and it proved to be robust and efficient in real-life situations.

2 Omnidirectional Camera Model

A unified model for central omnidirectional cameras was proposed by Geyer and Daniilidis [6], which represents central panoramic cameras as a projection onto the surface of a unit sphere. This formalism has been adopted and models for the internal projection function have been proposed by Micusik [13,14] and subsequently by Scaramuzza [23] who derived a general polynomial form of the internal projection valid for any type of omnidirectional camera. In this work, we will use the latter representation.

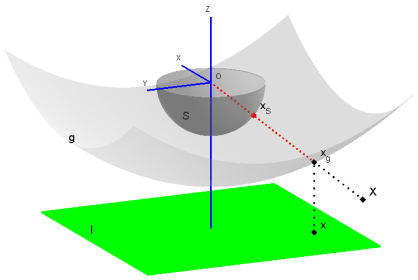


Fig. 1. Omnidirectional camera model

Let us first see the relationship between a point $\mathbf{x}$ in the omnidirectional image $\mathcal{I}$ and its representation on the unit sphere $\mathcal{S}$ (see Fig. 1). Note that only the half sphere on the image plane side is actually used, as the other half is not visible from image points. Following [23,24], we assume that the camera coordinate system is in $\mathcal{S}$, the origin (which is also the center of the sphere) is the projection center of the camera and the z axis is the optical axis of the camera which intersects the image plane in the *principal point*. To represent the nonlinear (but symmetric) distortion of central omnidirectional optics, [23,24] places a surface g between the image plane and the unit sphere $\mathcal{S}$, which is rotationally symmetric around z. The details of the derivation of g can be found in [23,24]. Herein, as suggested by [24], we will use a fourth order polynomial $g(\|\mathbf{x}\|) = a_0 + a_2\|\mathbf{x}\|^2 + a_3\|\mathbf{x}\|^3 + a_4\|\mathbf{x}\|^4$ which has 4 parameters representing the internal parameters (a_0, a_2, a_3, a_4) of the camera (only 4 parameters as a_1 is always 0 [24]). The bijective mapping $\Phi : \mathcal{I} \to \mathcal{S}$ is composed of 1) lifting the image point $\mathbf{x} \in \mathcal{I}$ onto the g surface by an orthographic projection

$$\mathbf{x}_g = \begin{bmatrix} \mathbf{x} \\ a_0 + a_2\|\mathbf{x}\|^2 + a_3\|\mathbf{x}\|^3 + a_4\|\mathbf{x}\|^4 \end{bmatrix} \tag{1}$$

and then 2) centrally projecting the lifted point $\mathbf{x}_g$ onto the surface of the unit sphere $\mathcal{S}$:

$$\mathbf{x}_\mathcal{S} = \Phi(\mathbf{x}) = \frac{\mathbf{x}_g}{\|\mathbf{x}_g\|} \tag{2}$$

Thus the omnidirectional camera projection is fully described by means of unit vectors $\mathbf{x}_\mathcal{S}$ in the half space of $\mathbb{R}^3$.

Let us see now how a 3D world point $\mathbf{X} \in \mathbb{R}^3$ is projected onto $\mathcal{S}$. This is basically a traditional central projection onto $\mathcal{S}$ taking into account the extrinsic pose parameters, rotation $\mathbf{R}$ and translation $\mathbf{t}$, acting between the camera (represented by $\mathcal{S}$) and world coordinate system. Thus for a world point $\mathbf{X}$ and its image $\mathbf{x}$ in the omnidirectional camera, the following holds on the surface of $\mathcal{S}$:

$$\Phi(\mathbf{x}) = \mathbf{x}_\mathcal{S} = \Psi(\mathbf{X}) = \frac{\mathbf{RX} + \mathbf{t}}{\|\mathbf{RX} + \mathbf{t}\|} \tag{3}$$

3 Pose Estimation

Consider a Lidar camera with a 3D coordinate system having its origin in the center of laser sensor rotation, x and y axes pointing to the right and down, respectively, while z is pointing away from the sensor. Setting the world coordinate system to the Lidar's coordinate system, we can relate a 3D Lidar point $\mathbf{X}$ with its image $\mathbf{x}$ in the omnidirectional camera using (3). In practical applications, like robot navigation or data fusion, the omnidirectional camera is usually calibrated (*i.e.* its intrinsic parameters (a_0, a_2, a_3, a_4) are known) and the relative pose $(\mathbf{R}, \mathbf{t})$ has to be estimated. Inspired by [3], [26] we will reformulate pose estimation as a 2D-3D shape alignment problem. Our solution is based on the correspondence-less 2D shape registration approach of Domokos *et al.* [2], where non-linear shape deformations are recovered via the solution of a nonlinear system of equations. This method was successfully applied for a number of registration problems in different domains such as volume [21] or medical [16] image registration. In our case, however, the registration has to be done on the spherical surface $\mathcal{S}$, which requires a completely different way to construct equations.

Any corresponding $(\mathbf{X}, \mathbf{x})$ Lidar-omni point pair satisfies (3). Thus a classical solution of the pose estimation problem is to establish a set of such point matches using *e.g.* a special calibration target or, if lidar points contain also the laser reflectivity value, by standard intensity-based point matching, and solve for $(\mathbf{R}, \mathbf{t})$. However, we are interested in a solution *without* a calibration target or correspondences because in many practical applications (*e.g.* infield mobile robot, autonomous driving systems), it is not possible to use a calibration target and most lidar sensors will only record depth information. Furthermore, lidar and camera images might be taken at different times and they need to be fused later based solely on the image content.

We will show that by identifying a single planar region both in the lidar and omni camera image, the extrinsic calibration can be solved. Since point correspondences are not available, (3) cannot be used directly. However, individual point matches can be integrated out yielding the following integral equation on the sphere $\mathcal{S}$:

$$\iint_{\mathcal{D}_\mathcal{S}} \mathbf{x}_\mathcal{S} \, \mathrm{d}\mathcal{D}_\mathcal{S} = \iint_{\mathcal{F}_\mathcal{S}} \mathbf{z}_\mathcal{S} \, \mathrm{d}\mathcal{F}_\mathcal{S} \tag{4}$$

$\mathcal{D}_\mathcal{S}$ and $\mathcal{F}_\mathcal{S}$ denote the surface patches on $\mathcal{S}$ corresponding to the omni and lidar planar regions $\mathcal{D}$ and $\mathcal{F}$, respectively. The above equation corresponds to a system of 2 equations, because a point on the surface $\mathcal{S}$ has only 2 independent components. However, we have 6 pose parameters (3 rotation angles and 3 translation components). To construct a new set of equations, we adopt the general mechanism from [2] and apply a function $\omega : \mathbb{R}^3 \to \mathbb{R}$ to both sides of the equation, yielding

$$\iint_{\mathcal{D}_\mathcal{S}} \omega(\mathbf{x}_\mathcal{S}) \, \mathrm{d}\mathcal{D}_\mathcal{S} = \iint_{\mathcal{F}_\mathcal{S}} \omega(\mathbf{z}_\mathcal{S}) \, \mathrm{d}\mathcal{F}_\mathcal{S} \tag{5}$$

To get an explicit formula for the above integrals, the surface patches $\mathcal{D}_\mathcal{S}$ and $\mathcal{F}_\mathcal{S}$ can be naturally parameterized via Φ and Ψ over the planar regions $\mathcal{D}$ and $\mathcal{F}$. Without loss of generality, we can assume that the third coordinate of $\mathbf{X} \in \mathcal{F}$ is 0, hence $\mathcal{D} \subset \mathbb{R}^2$, $\mathcal{F} \subset \mathbb{R}^2$; and $\forall \mathbf{x}_\mathcal{S} \in \mathcal{D}_\mathcal{S} : \mathbf{x}_\mathcal{S} = \Phi(\mathbf{x}), \mathbf{x} \in \mathcal{D}$ as well as $\forall \mathbf{z}_\mathcal{S} \in \mathcal{F}_\mathcal{S} : \mathbf{z}_\mathcal{S} = \Psi(\mathbf{X}), \mathbf{X} \in \mathcal{F}$ yielding the following form of (5):

$$\iint_{\mathcal{D}} \omega(\Phi(\mathbf{x})) \left\| \frac{\partial \Phi}{\partial x_1} \times \frac{\partial \Phi}{\partial x_2} \right\| \mathrm{d}x_1 \, \mathrm{d}x_2 = \iint_{\mathcal{F}} \omega(\Psi(\mathbf{X})) \left\| \frac{\partial \Psi}{\partial X_1} \times \frac{\partial \Psi}{\partial X_2} \right\| \mathrm{d}X_1 \, \mathrm{d}X_2 \quad (6)$$

where the magnitude of the cross product of the partial derivatives is known as the surface element. Adopting a set of nonlinear functions $\{\omega_i\}_{i=1}^{\ell}$, each ω_i generates a new equation yielding a system of ℓ independent equations. Although arbitrary ω_i functions could be used, power functions are computationally favorable [2] as these can be computed in a recursive manner:

$$\omega_i(\mathbf{x}_\mathcal{S}) = x_1^{l_i} x_2^{m_i} x_3^{n_i}, \text{ with } 0 \leq l_i, m_i, n_i \leq 2 \text{ and } l_i + m_i + n_i \leq 3 \quad (7)$$

Algorithm 1. The proposed calibration algorithm.

Input: 3D point cloud and 2D omnidirectional binary image representing the same region, and the g coefficients

Output: External Parameters of the camera as $\mathbf{R}$ and $\mathbf{t}$

1: Back-project the 2D image onto the unit sphere.
2: Back-project the 3D template onto the unit sphere.
3: Initialize the rotation matrix $\mathbf{R}$ from the centroids of the shapes on the sphere.
4: Initialize the translation $\mathbf{t}$ by translating $\mathcal{F}$ in the direction of its centroid until the area of $\mathcal{F}_\mathcal{S}$ and $\mathcal{D}_\mathcal{S}$ on the unit sphere are approximately equal.
5: Construct the system of equations of (4) with the polynomial ω_i functions.
6: Solve the set of nonlinear system of equations in (6) using the LM algorithm

Hence we are able to construct an overdetermined system of 15 equations, which can be solved in the *least squares sense* via a standard *Levenberg-Marquardt* algorithm. The solution directly provides the pose parameters of the omni camera. To guarantee an optimal solution, initialization is also important. In our case, a good initialization ensures that the surface patches $\mathcal{D}_\mathcal{S}$ and $\mathcal{F}_\mathcal{S}$ overlap as much as possible. This is achieved by computing the centroids of the surface patches $\mathcal{D}_\mathcal{S}$ and $\mathcal{F}_\mathcal{S}$ respectively, and initializing $\mathbf{R}$ as the rotation between them. Translation of the planar region $\mathcal{F}$ will cause a scaling of $\mathcal{F}_\mathcal{S}$ on the spherical surface. Hence an initial $\mathbf{t}$ is determined by translating $\mathcal{F}$ along the axis going through the centroid of $\mathcal{F}_\mathcal{S}$ such that the area of $\mathcal{F}_\mathcal{S}$ becomes approximately equal to that of $\mathcal{D}_\mathcal{S}$. The summary of the proposed algorithm with the projection on the unit sphere is presented in Algorithm 1.

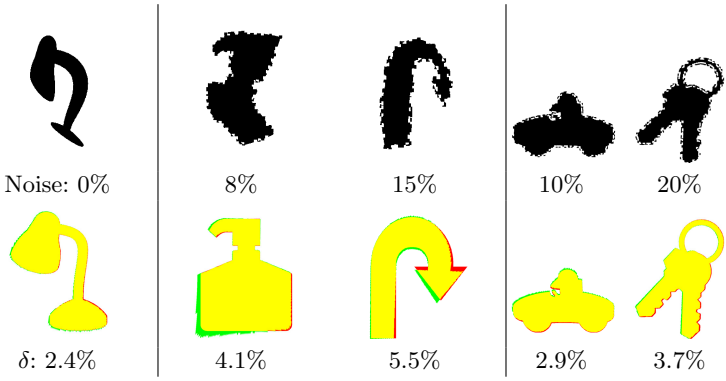


Fig. 2. Examples of various amount of simulated segmentation errors. First column contains a result without such errors, the next two show segmentation errors on the omnidirectional image, while the last two on the 3D planar region. The second row shows the δ error and the backprojected shapes overlayed in green and red colors(best viewed in color).

4 Evaluation on Synthetic Data

For the quantitative evaluation of the proposed method, we generated a benchmark set using 30 different shapes as 3D planar regions and their omnidirectional images taken by a virtual camera, a total of 150 2D-3D data pairs. The synthetic omni images were generated by a virtual camera being randomly rotated in the range of $(-40^\circ \cdots 40^\circ)$ and randomly translated in the range of $(0 \cdots 200)$. Assuming that the planar shape on the 800×800 template image represents a $5m \times 5m$ planar patch in 3D space, the $(0 \cdots 200)$ translation is equivalent to $(0 \cdots 1.25)$ meter in metric coordinates.

In practice, the planar regions used for calibration are segmented out from the lidar and omni images. In either case, we cannot produce perfect shapes, therefore robustness against segmentation errors was also evaluated on simulated data (see samples in Fig. 2): we randomly added or removed squares uniformly around the boundary of the shapes, both in the omni images and on the 3D planar regions, yielding a segmentation error of 5%–20% of the original shape.

The algorithm was implemented in Matlab and all experiments were run on a standard quad-core PC. Quantitative comparisons in terms of the various error plots are shown in Fig. 3, Fig. 4, and Fig. 5 (each test case is sorted independently in a best-to-worst sense). Calibration errors were characterized in terms of the percentage of non-overlapping area of the reference 3D shape and the backprojected omni image (denoted by δ in Fig. 5), as well as the error in each of the estimated pose parameters given in degrees in Fig. 4 and in cm in Fig. 3. Note that our method is quite robust against segmentation errors up to 15% error level.

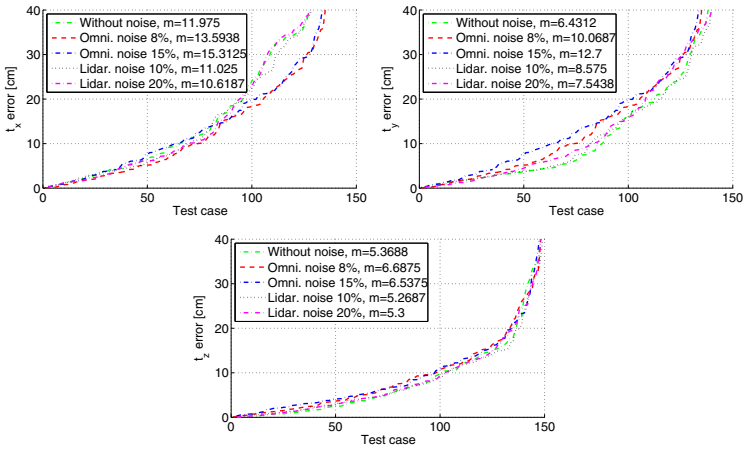


Fig. 3. Translation errors in cm along the x, y, and z axis. m denotes median error, *Omni noise* and *Lidar noise* stand for segmentation error on the omni and 3D regions, respectively (best viewed in color).

5 Experimental Validation

For the experimental validation of the proposed algorithm, two different omnidirectional camera setups are shown. In order to fuse the omnidirectional camera data and the lidar scan both the internal and external parameters are needed. The internal parameters of the omnidirectional camera were determined using the toolbox of [24]. Note that this internal calibration needs to be performed only once for a camera. The external parameters are then computed using the proposed algorithm.

5.1 Region Segmentation

In order to make the pose estimation user friendly, the region selection both in 2D and 3D was automated with efficient segmentation algorithms.

There are several automated or semi-automated 2D segmentation algorithms in the literature including clustering, histogram thresholding, energy based or region growing variants [29]. In this work we used a simple region growing algorithm which proved to be robust enough in urban environment [20].

For the 3D segmentation a number of point cloud segmentation methods are available, including robust segmentation [18] or difference of normals based segmentation[8]. Like in 2D, region growing gave stable results in our test cases thus it was suitable for the fusion algorithm as to extract planar input regions. This segmentation algorithm is based on a set composition principle, *i.e.* an initial starting point (seed) is selected from the original point cloud, and iteratively the set is completed with neighbor points which have similar normal

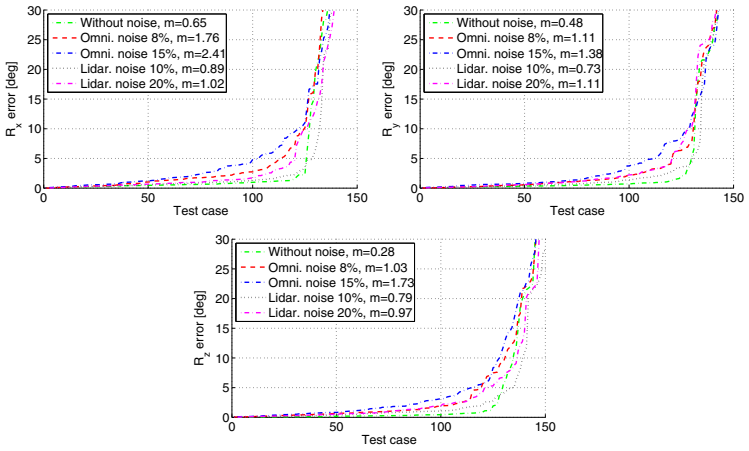


Fig. 4. Rotation errors in degrees along the x, y, and z axis. m denotes median error, *Omni noise* and *Lidar noise* stand for segmentation error on the omni and 3D regions, respectively (best viewed in color).

(within a certain threshold limit c_θ), or their curvature is less than a specified curvature threshold c_t.

Considering the correspondence establishment between the segmented 2D and 3D regions as minimal one-click user intervention, this aspect represents the only human interaction in the current procedure. After the first 2D-3D region pair establishment, further ones can easily be added by searching with a sample consensus approach for the neighbor plain patches. We remark, that a fully automatic region correspondence could be implemented by detecting and extracting windows [7] (see *e.g.* Fig. 7) which are typically planar and present in urban scenes.

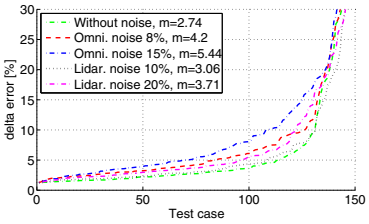


Fig. 5. Backprojection (δ) errors (best viewed in color)

5.2 Urban Data Fusion Results

The input data with the segmented regions as well as the results are shown in Fig. 6 and Fig. 7 for a catadioptric-lidar and dioptric-lidar camera pairs respectively. The omnidirectional images were captured with a commercial SLR

camera with a catadioptric lens (Fig. 6) and a fish-eye (Fig. 7) respectively. For the 3D range data, a custom lidar was used in Fig. 6 to acquire data similar to the one described in [27] with an angular resolution up to half degree and a depth accuracy of $1cm$. In Fig. 7, the 3D point cloud was recorded by a Velodyne Lidar mounted on a moving car [9].

After the raw data acquisition, the segmentation was performed in both domains. For the 3D data, segmentation yields a parametric equation and boundaries of the selected planar region. This was then uniformly sampled to get a dense homogeneous set of points (*i.e.* we do not rely on the Lidar resolution after segmentation), which was subsequently transformed with a rigid motion $(\mathbf{R_0}, \mathbf{t_0})$ into the $Z = 0$ plane yielding appropriate point coordinates $\mathbf{X}$ used in the right hand side of (6). The $\mathbf{x}$ points of the left hand side of (6) are fed with the pixel coordinates of the segmented omnidirectional image.

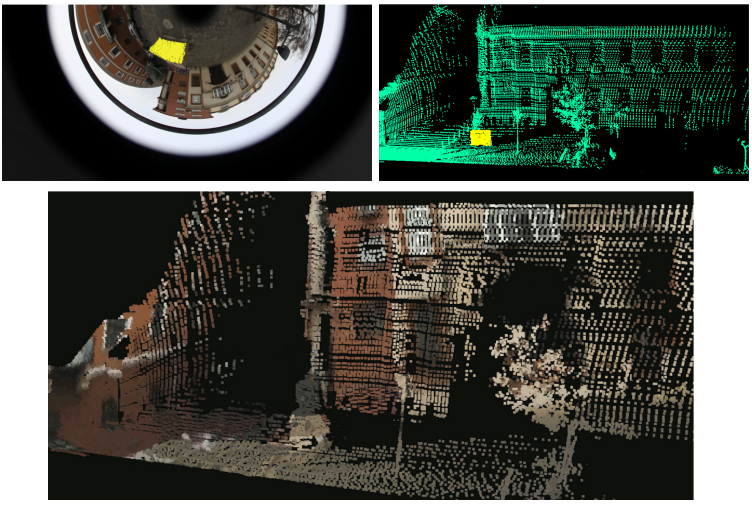


Fig. 6. Catadioptric and lidar images with segmented area marked in yellow, and the fused images after pose estimation (best viewed in color)

Once the output $(\mathbf{R}, \mathbf{t})$ is obtained from Algorithm 1, the final transformation acting between the lidar and omni camera can then be computed as a composite rigid transformation of $(\mathbf{R_0}, \mathbf{t_0})$ and $(\mathbf{R}, \mathbf{t})$. The final computed transformation was used to fuse the depth and RGB data by reprojecting the point cloud on the image plane using the internal and external camera parameters, and thus obtaining the color for each point of the 3D point cloud. The method proved to be robust against segmentation errors, but a sufficiently large overlap between the regions is required for better results.

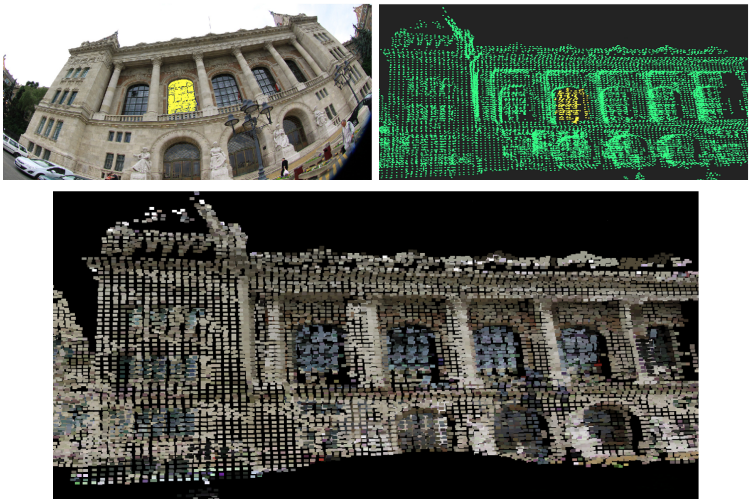


Fig. 7. Dioptric (fisheye) and lidar images with segmented area marked in yellow, and the fused images after pose estimation (best viewed in color)

6 Conclusions

In this paper a new method for pose estimation of non-conventional cameras is proposed. The method is based on a point correspondence-less registration technique, which allows reliable estimation of extrinsic camera parameters. The algorithm was quantitatively evaluated on a large synthetic data set and proved to be robust on real data fusion as well.

Acknowledgments. This research was partially supported by the European Union and the State of Hungary, co-financed by the European Social Fund through projects TAMOP-4.2.4.A/2-11-1-2012-0001 National Excellence Program and FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV-2012-0013); as well as by Domus MTA Hungary. The authors gratefully acknowledge the help of Csaba Benedek from DEVA Lab., SZTAKI in providing us with preprocessed Velodyne Lidar scans, as well as the discussions with Radu Orghidan and Cedric Demonceaux in the early stage of this work. The catadioptric camera was provided by the Multimedia Technologies and Telecommunications Research Center of UTCN with the help of Camelia Florea.

References

1. Baker, S., Nayar, S.K.: A Theory of Single-Viewpoint Catadioptric Image Formation. International Journal of Computer Vision **35**(2), 175–196 (1999)
2. Domokos, C., Nemeth, J., Kato, Z.: Nonlinear Shape Registration without Correspondences. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(5), 943–958 (2012)

3. Frohlich, R., Tamas, L., Kato, Z.: Homography Estimation between Omnidirectional Cameras without Point Correspondences. In: ICRA Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras. Hong Kong, June 2014
4. Furgale, P.T., Schwesinger, U., Rufli, M., Derendarz, W., Grimmett, H., Muehlfellner, P., Wonneberger, S., Timpner, J., Rottmann, S., Li, B., Schmidt, B., Nguyen, T.N., Cardarelli, E., Cattani, S., Bruning, S., Horstmann, S., Stellmacher, M., Mielenz, H., Köser, K., Beermann, M., Hane, C., Heng, L., Lee, G.H., Fraundorfer, F., Iser, R., Triebel, R., Posner, I., Newman, P., Wolf, L.C., Pollefeys, M., Brosig, S., Effertz, J., Pradalier, C., Siegwart, R.: Toward automated driving in cities using close-to-market sensors: An overview of the V-Charge Project. In: Intelligent Vehicles Symposium, Gold Coast City, Australia, pp. 809–816, June 2013
5. Geiger, A., Lauer, M., Wojek, C., Stiller, C., Urtasun, R.: 3D Traffic Scene Understanding From Movable Platforms. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(5), 1012–1025 (2014)
6. Geyer, C., Daniilidis, K.: A unifying theory for central panoramic systems. In: European Conference on Computer Vision, Dublin, Ireland, pp. 445–462, June 2000
7. Goron, L., Tamas, L., Reti, I., Lazea, G.: 3D laser scanning system and 3D segmentation of urban scenes. In: Automation Quality and Testing Robotics Conference, vol. 1, pp. 81–85. IEEE, Cluj-Napoca, May 2010
8. Ioannou, Y., Taati, B., Harrap, R., Greenspan, M.: Difference of Normals as a Multi-scale Operator in Unorganized Point Clouds. International Conference on 3D Imaging. Modeling, Processing, Visualization and Transmission, pp. 501–508. IEEE Computer Society, Los Alamitos, CA, USA (September (2012)
9. Józsa, O., Börcs, A., Benedek, C.: Towards 4D virtual city reconstruction from Lidar point cloud sequences. In: ISPRS Workshop on 3D Virtual City Modeling, ISPRS Annals of Photogrammetry, Remote Sensing and the Spatial Information Sciences, vol. II-3/W1, pp. 15–20. Regina, Canada (2013)
10. Kannala, J., Brandt, S.S.: A Generic Camera Model and Calibration Method for Conventional, Wide-Angle, and Fish-Eye Lenses. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(8), 1335–1340 (2006)
11. Lin, D., Fidler, S., Urtasun, R.: Holistic Scene Understanding for 3D Object Detection with RGBD Cameras. In: International Conference on Computer Vision. pp. 1417–1424. IEEE Computer Society, Sydney, December 2013
12. Mei, C., Rives, P.: Single View Point Omnidirectional Camera Calibration from Planar Grids. In: International Conference on Robotics and Automation, Roma, Italy, pp. 3945–3950, April 2007
13. Mičušík, B.: Two-View Geometry of Omnidirectional Cameras. Phd thesis, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic, June 2004
14. Mičušík, B., Pajdla, T.: Para-catadioptric Camera Auto-calibration from Epipolar Geometry. In: Asian Conference on Computer Vision, Seoul, Korea South, vol. 2, pp. 748–753, January 2004
15. Mirzaei, F.M., Kottas, D.G., Roumeliotis, S.I.: 3D LIDAR-camera intrinsic and extrinsic calibration: Identifiability and analytical least-squares-based initialization. International Journal of Robotics Research **31**(4), 452–467 (2012)

16. Mitra, J., Kato, Z., Marti, R., Oliver, A., Lladó, X., Sidibé, D., Ghose, S., Vilanova, J.C., Comet, J., Meriaudeau, F.: A Spline-based Non-linear Diffeomorphism for Multimodal Prostate Registration. Medical Image Analysis **16**(6), 1259–1279 (2012)
17. Nayar, S.K.: Catadioptric Omnidirectional Camera. In: Conference on Computer Vision and Pattern Recognition. pp. 482–488. Washington, USA, June 1997
18. Nurunnabi, A., Belton, D., West, G.: Robust Segmentation in Laser Scanning 3D Point Cloud Data. In: Digital Image Computing Techniques and Applications, Fremantle, Australia, pp. 1–8, December 2012
19. Pandey, G., McBride, J.R., Savarese, S., Eustice, R.M.: Automatic Targetless Extrinsic Calibration of a 3D Lidar and Camera by Maximizing Mutual Information. In: AAAI National Conference on Artificial Intelligence, Toronto, Canada, pp. 2053–2059, July 2012
20. Preetha, M., Suresh, L., Bosco, M.: Image Segmentation using Seeded Region Growing. Computing. In: Electronics and Electrical Technologies, Nagercoil, India, pp. 576–583, March 2012
21. Santa, Z., Kato, Z.: Correspondence-Less Non-rigid Registration of Triangular Surface Meshes. In: Conference on Computer Vision and Pattern Recognition, pp. 2275–2282. IEEE Computer Society, Portland, June 2013
22. Scaramuzza, D., Harati, A., Siegwart, R.: Extrinsic Self Calibration of a Camera and a 3D Laser Range Finder from Natural Scenes. In: International Conference on Intelligent Robots and Systems, San Diego, USA, pp. 4164–4169, October 2007
23. Scaramuzza, D., Martinelli, A., Siegwart, R.: A Flexible Technique for Accurate Omnidirectional Camera Calibration and Structure from Motion. In: International Conference on Computer Vision Systems, Washington, USA, pp. 45–51, January 2006
24. Scaramuzza, D., Martinelli, A., Siegwart, R.: A Toolbox for Easily Calibrating Omnidirectional Cameras. In: International Conference on Intelligent Robots and Systems, Bejing, China, pp. 5695–5701, October 2006
25. Schoenbein, M., Strauss, T., Geiger, A.: Calibrating and Centering Quasi-Central Catadioptric Cameras. In: International Conference on Robotics and Automation, Hong-Kong, China, pp. 1253–1256, June 2014
26. Tamas, L., Kato, Z.: Targetless Calibration of a Lidar - Perspective Camera Pair. International Conference on Computer Vision. Bigdata3dcv Workshops, Australia, Sydney, pp. 668–675, December (2013)
27. Tamas, L., Majdik, A.: Heterogeneous Feature Based Correspondence Estimation. In: Multisensor Fusion and Integration for Intelligent Systems, pp. 89–94. IEEE, Hamburg, June 2012
28. Taylor, Z., Nieto, J.: A Mutual Information Approach to Automatic Calibration of Camera and Lidar in Natural Environments. In: Australian Conference on Robotics and Automation, Wellington, Australia, pp. 3–8, December 2012
29. Vantaram, S.R., Saber, E.: Survey of Contemporary Trends in Color Image Segmentation. Journal of Electronic Imaging **21**(4), 1–28 (2012)

Good Edgels to Track: Beating the Aperture Problem with Epipolar Geometry

Tommaso Piccini[1(✉)], Mikael Persson[1], Klas Nordberg[1], Michael Felsberg[1], and Rudolf Mester[1,2]

[1] CVL, ISY, Linköping University, Linköping, Sweden
tommaso.piccini@liu.se
[2] VSI Lab, C.S. Department, Goethe University, Frankfurt, Germany

Abstract. An open issue in multiple view geometry and structure from motion, applied to real life scenarios, is the sparsity of the matched key-points and of the reconstructed point cloud. We present an approach that can significantly improve the density of measured displacement vectors in a sparse matching or tracking setting, exploiting the partial information of the motion field provided by linear oriented image patches (edgels). Our approach assumes that the epipolar geometry of an image pair already has been computed, either in an earlier feature-based matching step, or by a robustified differential tracker. We exploit key-points of a lower order, *edgels*, which cannot provide a unique 2D matching, but can be employed if a constraint on the motion is already given. We present a method to extract edgels, which can be effectively tracked given a known camera motion scenario, and show how a constrained version of the Lucas-Kanade tracking procedure can efficiently exploit epipolar geometry to reduce the classical KLT optimization to a 1D search problem. The potential of the proposed methods is shown by experiments performed on real driving sequences.

Keywords: Densification · Tracking · Epipolar geometry · Lucas-Kanade · Feature extraction · Edgels · Edges

1 Introduction

Most methods for finding image-to-image correspondences by tracking or matching are applied on selected *key-points*, i.e., image locations which are unique in appearance and can easily be re-identified. While providing a coarse outline of the motion structure in a scene, a drawback of this approach is that the feature points are distributed sparsely, and cannot produce a dense 3D point cloud.

A natural next step is to consider a scenario where the relative motion between the camera and the scene/object already has been determined with high accuracy, e.g., by means of sparse feature matching and windowed bundle adjustment, and is represented by the epipolar structure. After this step, it is possible to increase the density of the motion field, aiming for a dense 3D reconstruction. This can be done by stereo matching on rectified image pairs,

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 652–664, 2015.
DOI: 10.1007/978-3-319-16181-5_50

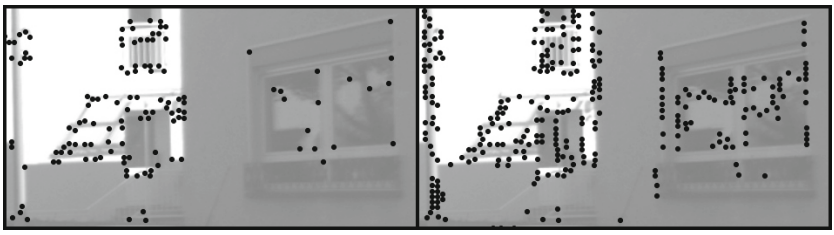


Fig. 1. Close up of an image patch showing a situation where the *GETT* keypoints strongly outnumber *GFTT*. **Left:** GFTT. **Right:** GETT.

and gives generally very good results in the case of sideways camera motion and relatively small rotations, but it is not a feasible approach for a general motion, such as forward/backward camera motion combined with significant rotations.

As an alternative, we instead extract edge pixels that are likely to track well given a known epipolar geometry, and perform a local search for the correspondence only on the physically plausible segment of the epipolar line. To do this, we modify the Kanade-Lucas-Tomasi (**KLT**) tracker to only operate along the epipolar line, thus allowing to also track several edge points (Fig. 1). Relevant applications of this method include:

- densification of the partial reconstruction for object detection in the context of autonomous robots.
- increase of the density of the reconstructed 3D point cloud as a final step of a multi-view structure from motion.

To summarize, this paper presents two novel contributions:

- We extract edgels that are likely to track well under the given relative motion. These edge points are called *Good Edgels To Track* (**GETT**).
- We densify the motion field by tracking these new edgel features, using a modified version of the KLT tracker that can deal with the aperture problem, given that the epipolar geometry is known.

2 Related Work

Many applications in computer vision, e.g., structure from motion and visual odometry, have as an initial step the computation of optical flow. This can be done in two ways: computing a global optical flow for the whole image plane [7,13] or by sparsely computing the displacement of a limited number of (feature) points in the image. The first approach provides a dense flow field, but has a high computational complexity. The sparse approach is more computationally efficient, but the sparsity of the motion field can be a limiting factor when a dense reconstruction is desired. The sparsity is a consequence of applying these

methods only on selected feature points, with a structure in their surroundings that allows them to be re-identified in subsequent images [1,10,22,24]. Such feature points can be found on a second image, depicting a slightly modified version of the same scene, either by matching or by tracking procedures. Matching procedures extract feature points in both images, and try to match the most similar pairs using a specifically designed descriptor [3,6,17,18,21]. Tracking is instead implemented generally as a search, in the second image, for a point that minimizes a cost function which measures the difference between small patches of the image surrounding the two points. The most common approach to tracking is based on the Lucas-Kanade tracker [19], which approximately minimizes a cost function calculated as the sum of the squared differences in the intensities of the pixels belonging to the patches. The KLT tracker has been shown to work reliably only on "corner points", known as *Good Features to Track* (GFTT), and although efforts have been made to improve the performance of the tracker on said points [2,8,23], not many papers in the literature deal with the sparsity of such features. GFTT-points can be identified by a structure tensor with large values for both its eigenvalues [22]. The 2D motion of these points can be identified, in both directions, with a gradient descent method on the matching criterion (e.g. the SSD). A natural next step would be to track points having only one large eigenvalue for their structure tensor. Intuitively, these points lie mainly on linear edges. At these points, however, a local analysis can only provide the motion component in a direction perpendicular to the edge, corresponding to the eigenvector for the largest eigenvalue of the structure tensor. This is the well-known *aperture problem*.

A framework for matching edgels is proposed in [16], and is shown to outperform matching based on, e.g., SIFT or SURF, when a planar object is tracked. The limitation to planar objects restricts, however, the application of this approach. Another attempt to solve the aperture problem is made in [4], by jointly tracking edge points and nearby corner points with a method that combines the KLT tracker and the global Horn-Schunck method. In this framework, the motion of corner points helps in solving for the second degree of freedom of nearby edge points thanks to a regularization term inserted in the optimization process.

A similar, but more generalized approach is taken in [26] where a semi-global matching method is adapted to work exclusively on fractions of the epipolar line. Also in this case, a regularization term is included in the optimization process to allow the matching of weaker points (edgels or even pixels lying on planar, textureless areas). The optimization problem is, however, NP-hard and non parallelizable so the dynamic programming approach used to solve it takes several seconds to compute.

In [15] the authors propose a method to obtain a dense optical flow which does not include a regularization term allowing each point to be treated individually. In this approach a Delaunay triangulation is performed over the motion field generated by a sparse matching framework to produce a prior on the motion field for all the points in the convex hull of the sparse features. This information is then used in combination with the estimated epipolar geometry and the trifocal

tensor to steer a maximum-a-posteriori estimate for the best matching candidate for each pixel while also cutting down the number of possible candidates significantly

In our work, a different approach is taken. Corner points are often sufficient in number to produce a useful inter-frame motion estimate, which can be made even more accurate by the means of a global or local bundle adjustment, but are just too sparse to produce a dense reconstruction. Hence, there is no need to include edgels at this early stage. In most cases, including them at this stage would just make the egomotion estimate more computationally complex without a significant gain in accuracy and robustness. Instead, if we are only interested in making the motion flow in an image pair denser, the corresponding epipolar geometry is already available to a high degree of accuracy, and it is possible to formulate the KLT tracker as a 1-dimensional problem by constraining its optimization procedure exclusively on an epipolar line. Thus, edgels that are not parallel to the corresponding epipolar line can be tracked. We refer to this modified version of the KLT tracker as *Epipolar KLT*.

The idea of epipolar KLT has recently been proposed by Trummer et al [25], applied to the scenario of a camera mounted on a robotic arm for which the motion parameters are known. That work, however, does not fully exploit this tool for the tracking of specifically extracted edgels and, in fact, they reject the idea of completely constraining the tracker to the epipolar lines due to possible uncertainty in the epipolar geometry. Instead, they formulate a biased 2-dimensional tracker that favors steps in the direction of the epipolar line and reduces the freedom in the perpendicular direction with empirically chosen weights. Not completely trusting the epipolar geometry can be reasonable in high precision applications, but there is no reason to ignore already established camera poses when just aiming at making the motion field denser.

3 Methods

3.1 Background

The KLT tracker The classical KLT tracker has been repeatedly derived in literature, and we will only present parts of the derivation which are of interest for our modifications. For a complete derivation, see [2,5,19] among others. Let $W'_{\mathcal{J}}$ be a small window in the second image around a feature point, $\mathcal{I}(\mathbf{x})$ and $\mathcal{J}(\mathbf{x})$, respectively intensity value of the first and second images at position $\mathbf{x}$. In the simplest case, the change in the position of a point from $\mathcal{I}$ to $\mathcal{J}$ consists in a 2D translation, represented by the 2 values of the translation vector $\mathbf{v}$. Also, in practice, the inverse warping function is often used since it makes the procedure computationally more efficient.

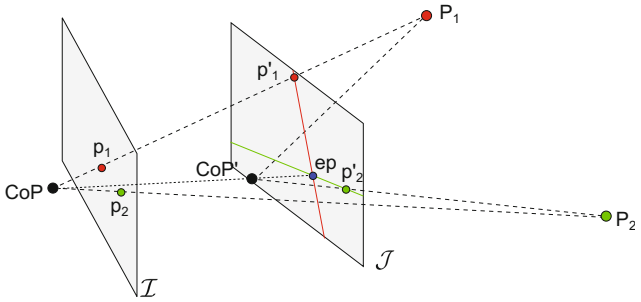


Fig. 2. Schematics of some geometric entities involved in two view geometry

The expression for the cost function is:

$$G(\Delta\mathbf{v}) := \sum_{\mathbf{x}'_i \in W'_{\mathcal{J}}} (\mathcal{I}(\mathbf{x}'_i - (\mathbf{v} + \Delta\mathbf{v}) - \mathcal{J}(\mathbf{x}'_i))^2 \tag{1}$$

to be minimized w.r.t. $\Delta\mathbf{v}$. To solve the equation, a first order Taylor expansion is needed. Let $x_i = x'_i - (\mathbf{v} + \Delta\mathbf{v})$, setting, then, the first order derivative of the cost function to zero to find the stationary point we obtain:

$$\sum_{\mathbf{x}'_i \in W'_{\mathcal{J}}} \nabla\mathcal{I}(\mathbf{x}_i)\, \nabla\mathcal{I}(\mathbf{x}_i)^T \Delta\mathbf{v} = \sum_{\mathbf{x}'_i \in W'_{\mathcal{J}}} (\mathcal{I}(\mathbf{x}_i) - \mathcal{J}(\mathbf{x}'_i))\nabla\mathcal{I}(\mathbf{x}_i) \tag{2}$$

In matrix form we can write it as $\mathbf{A}\,\Delta\mathbf{v} = \mathbf{b}$. Solving for $\Delta\mathbf{v}$, the update rule for the iteration scheme is $\mathbf{v} = \mathbf{v} + \Delta\mathbf{v}$, and $\mathbf{x}'_i = \mathbf{x}_i + \mathbf{v}$. The procedure is repeated until convergence or until some breaking criterion is met (for example, when the displacement is smaller than a threshold).

Epipolar Geometry This section contains a short review of the geometrical entities involved in epipolar geometry which are necessary to derive our algorithms. A more complete presentation is made, e.g., in [11]. With reference to Figure 2, for a given image pair $\mathcal{I}$ and $\mathcal{J}$ produced by two cameras, or the same camera in two different positions, having their center of projection in **CoP** and **CoP′** respectively, we can make the following observations (assuming lens distortion is absent or the images are rectified):

1. The projection of **CoP** in the image plane of $\mathcal{J}$ is the *epipole* **ep**.
2. For every 3D point $\mathbf{P}_k$ observed by both the cameras, the projection of the $\mathbf{P}_k$ in $\mathcal{J}$, $\mathbf{p}'_k$, must lie on a line in $\mathcal{J}$, given as the projection of the optic ray through $\mathbf{p}_k$ in $\mathcal{I}$, i.e., the 3D line connecting **CoP** and $\mathbf{p}_k$, onto $\mathcal{J}$. This line is the *epipolar line* generated by $\mathbf{p_k}$ in $\mathcal{J}$.
3. All the epipolar lines on an image plane, for a certain camera pair, meet in the corresponding epipole, **ep**.

Given the images $\mathcal{I}$ and $\mathcal{J}$, the corresponding epipolar geometry is represented by the fundamental matrix $\mathbf{F}$ The epipolar line in $\mathcal{J}$ corresponding to a point of pixel coordinates $\mathbf{x}$ in $\mathcal{I}$ is given by $\tilde{\mathbf{e}} = \mathbf{F}\tilde{\mathbf{x}} = \begin{bmatrix} e_1 & e_2 & e_3 \end{bmatrix}^\top$, where $\tilde{\mathbf{x}}$ is the column vector of the homogeneous pixel coordinates of the point $\mathbf{x}$ and $\tilde{\mathbf{e}}$ is the dual homogeneous coordinates of the epipolar line. The 2-dimensional unity vector representing the direction of the epipolar line is then $\hat{\mathbf{e}} = \begin{bmatrix} e_2 & -e_1 \end{bmatrix}^T / \sqrt{e_1^2 + e_2^2}$.

Epipolar KLT We now have all the tools to derive the Epipolar KLT. Given an initialization on the epipolar line for the translation vector (i.e. $\mathbf{v}$ such that $\mathbf{x}+\mathbf{v}$ lies on $\tilde{\mathbf{e}}$), we can substitute $\hat{\mathbf{e}}$ in (1) to obtain a 1-dimensional cost function

$$G(\alpha) := \sum_{\mathbf{x}'_i \in W'_{\mathcal{J}}} (\mathcal{I}(\mathbf{x}'_i - (\mathbf{v} + \alpha\hat{\mathbf{e}})) - \mathcal{J}(\mathbf{x}'_i))^2 \tag{3}$$

to be minimized w.r.t. the scalar value α.

After the first order Taylor expansion and solving for the stationary point, we obtain the following expression to be iteratively minimized to converge toward the solution:

$$\alpha \sum_{\mathbf{x}'_i \in W'_{\mathcal{J}}} \nabla\mathcal{I}(\mathbf{x}_i)\, \nabla\mathcal{I}(\mathbf{x}_i)^T \hat{\mathbf{e}} = \sum_{\mathbf{x}'_i \in W'_{\mathcal{J}}} (\mathcal{I}(\mathbf{x}_i) - \mathcal{J}(\mathbf{x}'_i))\nabla\mathcal{I}(\mathbf{x}_i). \tag{4}$$

In matrix form we can write the last expression as

$$\alpha \mathbf{A}\, \hat{\mathbf{e}} = \mathbf{b}, \tag{5}$$

which is an over-determined 2 equation system. By pre-multiplying both sides by $\hat{\mathbf{e}}^T$ (i.e. projecting the 2D problem on the direction of the epipolar line), we project the system into one dimension: $\alpha\hat{\mathbf{e}}^T \mathbf{A}\, \hat{\mathbf{e}} = \hat{\mathbf{e}}^T \mathbf{b}$. Solving for α, the update rule for the iteration scheme is $\mathbf{v} = \mathbf{v} + \alpha\hat{\mathbf{e}}$ and $\mathbf{x}'_i = \mathbf{x}_i + \mathbf{v}$. The procedure is repeated until convergence or until some breaking criterion is met (for example, when the displacement is smaller than a threshold).

3.2 Good Edgels to Track

As already mentioned, assuming a static scene and a reliable estimate of the relative camera motion between two frames, we can track two different kinds of features:

- Corner features
- Edge features non parallel to the motion

Similarly to the case of corner features, edge features are identified from the eigenvalues of their structure tensor: they are associated with a structure tensor with at least one eigenvalue over a certain threshold. However, not all edge features are useful for our method: edges parallel to the their own epipolar line cannot be tracked reliably.

Let $\nabla\mathcal{I}_u(\mathbf{x})$ and $\nabla\mathcal{I}_v(\mathbf{x})$ be the horizontal and vertical gradient of the image calculated in $\mathbf{x}$ respectively. The structure tensor $T(\mathbf{x})$ for the point $\mathbf{x}$ is:

$$T(\mathbf{x}) = \begin{bmatrix} \nabla\mathcal{I}_u(\mathbf{x})^2 & \nabla\mathcal{I}_u(\mathbf{x})\nabla\mathcal{I}_v(\mathbf{x}) \\ \nabla\mathcal{I}_u(\mathbf{x})\nabla\mathcal{I}_v(\mathbf{x}) & \nabla\mathcal{I}_v(\mathbf{x})^2 \end{bmatrix}. \tag{6}$$

Given a pixel $\mathbf{x}$, its structure tensor T and its corresponding epipolar line $\tilde{\mathbf{e}}$ in the matching image, its score in the GETT sense is

$$\text{score}(\mathbf{x}) = \hat{\mathbf{e}}^T T(\mathbf{x})\, \hat{\mathbf{e}}, \tag{7}$$

where $\hat{\mathbf{e}}$ is the unit vector representing the direction of the epipolar line in $\mathcal{J}$ for $\mathbf{x}$. Note that the value of the score is bounded by the larger and smaller eigenvalues of the structure tensor and corresponds, intuitively, to measuring the structure tensor along the direction of the optimization process.

3.3 Epipolar KLT Initialization

To assure convergence to a reasonable solution it is critical to initialize the tracker on the epipolar line. In our tests we assume that no prior information on the depth of the points is available. We therefore assume that every point has an infinite depth, in this case we can find a starting point on the epipolar line that is also a hard limit for our procedure as only one of the 2 segments of the epipolar line defined by this point is physically reasonable.

Let $\mathbf{K}$ be the intrinsic parameter matrix for the camera and $\mathbf{R}$ the inter-camera rotation between the 2 views. For a point $\tilde{\mathbf{x}} \in \mathcal{I}$ in homogeneous coordinates, its infinity projection on $\mathcal{J}$ is given by

$$\text{start}(\tilde{\mathbf{x}}) = \mathbf{K}\,\mathbf{R}\,\mathbf{K}^{-1}\tilde{\mathbf{x}}. \tag{8}$$

We refer to this initialization procedure as *initialization at infinity*.

Depending on the scene structure and the camera motion, this initialization can fail, in particular if the scene has large variation in depth and the translational motion component is significant relatively to the rotational component. This is the case for most driving sequences, such as those offered by the KITTI dataset [9], where this initialization is sensitive to repeating patterns and large displacements. This simple initialization, however, works in practice for the majority of the points. To increase the number of good matches, we do however perform an initialization in multiple steps as outlined in Algorithm 2. The details are given below.

4 Experiments

We tested our algorithm against the OpenCV [1] standard KLT implementation. For a fair comparison, the epipolar constraint has been imposed to the output

[1] http://opencv.org

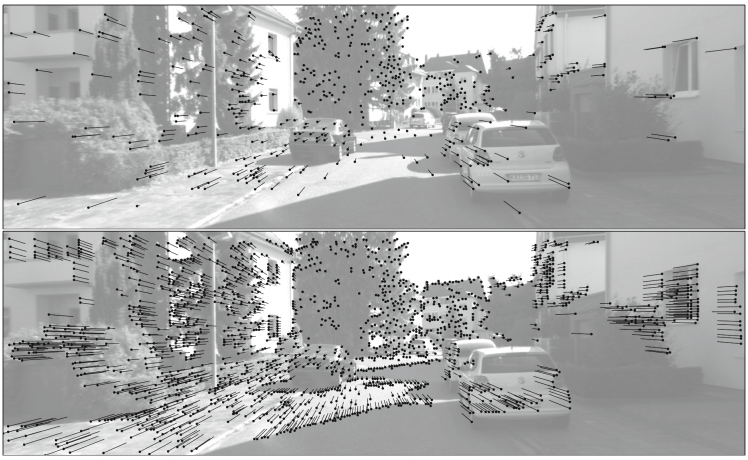


Fig. 3. Sample tracking output image from Sequence 1. **Top:** Tracking performed on GFTT with the standard KLT procedure. **Bottom: Tracking performed on GETT with the epipolar KLT.**

Algorithm 1. Testing pseudocode for standard KLT procedure

procedure TRACKGOODFEATURESTOTRACK(I, J, K, R, t)
  $p \leftarrow$ extractGoodFeaturesToTrack(I)
  $q \leftarrow$ standardKLT(I, J, p)
  $r \leftarrow$ standardKLT(J, I, q)
  $corr \leftarrow \{(p_i, q_i) | \|p_i - r_i\| < \tau\}$
  $corr \leftarrow$ epipolarFilter(corr, F)
  return *corr*
end procedure

of the standard tracker to reduce outliers. Both trackers have been fed with the same parameters in term of windows size (5×5), pyramid level (2nd level) and termination criteria (maximum 10 iterations, minimum displacement length of 0.1 pixels). The two testing schemes are outlined in Algorithms 1 and 2.

4.1 Implementation Details

In this subsection we go into the details of the testing scheme outlined in Algorithm 2.

Extraction of corners and edges We extract the corner and edge keypoints respectively according to the classical GFTT scheme and the hereby presented GETT scheme. The procedure works in the following way:

1. Extract the structure tensor for each pixel in the image.
2. Calculate the GFTT and GETT score for each pixel. The following steps are computed independently for GFTT and GETT points.

Algorithm 2. Testing pseudocode for our approach

procedure TRACKMIXEDFEATURES(I, J, K, R, t)
  $F \leftarrow K^{-T}[t]_x RK^{-1}$
  $p \leftarrow$ extractCornersEdges(I, F)
  $start_p \leftarrow KRK^{-1}p$
  $q \leftarrow$ epipolarKLT(I, J, p, start_p, F)
  $start_q \leftarrow KR^TK^{-1}q$
  $r \leftarrow$ epipolarKLT(J, I, q, start_q, F^T)
  $corr \leftarrow \{(p_i, q_i) | \|p_i - r_i\| < \tau\}$
  $clusters \leftarrow$ clusterize(p)
  $c_{MEAN} \leftarrow$ clustermeans(clusters, corr)
  $start_p \leftarrow$ moveStartingPoints(start_p, c_MEAN)
  $q \leftarrow$ epipolarKLT(I, J, p, start_p, F)
  $corr \leftarrow \{(p_i, q_i)\}$
  $c_{MEAN} \leftarrow$ clustermeans(clusters, corr)
  $corr \leftarrow$ filterOutliers(corr, c_MEAN)
  return *corr*
end procedure

3. An Non-Maximum Suppression (**NMS**) filter is applied to the score images to obtain a better distribution of the keypoints. A sliding window is applied to each score image suppressing all the points which are not a local maximum. The window size is set at the same size of the tracking window at the lowest pyramid level.
4. A threshold τ is computed as a fraction of the highest score found in the image and the points with a score $> \tau$ are used as keypoints.

Tracker initialization The first initialization is given by the *initialization at infinity* presented in Section 3.3.

Tracking The tracking is done with the Epipolar KLT in Section 3.1, including a pyramidal *coarse-to-fine* scheme similar to the one presented in [5]. At the end of the tracking procedure, features presenting a large error (see [8]) or that moved on the wrong segment of the epipolar line are rejected as outliers.

Backtracking As suggested in [12] , a backtracking step is performed to further remove outliers. Features are tracked backwards from image $\mathcal{J}$ to image $\mathcal{I}$ and when the procedure converges to a point different from the original feature, the feature is considered an outlier. Note that this step is proven particularly useful in the case of repeating patterns due to the naivety of the initialization.

Clusterization and tracker reinitialization The tracks surviving the first two steps are in general of good quality, but are not dense due to the naive initialization step performed. Therefore, we compute a new initialization based on region-wise local mean displacement for the available tracks. The image is subdivided in macro-regions (in our tests we use a 3×7 grid to determine the

Method	Sequence 1	Sequence 2	Sequence 3	Total
GETT + EpiKLT	3542 ± 464	4439 ± 848	4603 ± 484	4196 ± 779
GFTT + KLT	2537 ± 425	3593 ± 881	3908 ± 551	3347 ± 874

Table 1. Average number of features extracted over the 3 sequences ± standard deviation

Method	Sequence 1	Sequence 2	Sequence 3	Total
GETT + EpiKLT	1700 ± 324	1656 ± 470	2121 ± 373	1818 ± 448
GFTT + KLT	509 ± 250	412 ± 238	909 ± 337	610 ± 352

Table 2. Average number of features tracked over the 3 sequences ± standard deviation

regions) and the mean displacement along the epipolar line is robustly computed for each region, based on the available data of the surviving tracks belonging to the region. The displacement mean for each cluster of the image is robustly computed using the algorithm presented in [14]. This method works on 1-dimensional data, it orders the data and determines the mean giving the maximum amount of inliers, using a sliding window. The region-wise mean so computed is then used to reinitialize the tracker by moving the starting position of the tracker, away from the infinity projection of the feature, along the epipolar line.

Outlier removal The last tracking step has proven to provide a much higher inlier ratio than the initialization at infinity, making a new backtracking step unnecessary. To remove the remaining outliers, we perform a filtering based on a re-computed region based mean, and remove points that do not behave according to the majority of the points in the region (we assume a Gaussian distribution and set the filter threshold at twice the standard deviation for the region).

Note that this complicated procedure is only necessary when no prior information on the depth structure of the scene is known. For image sequences, features can in general be tracked more than once, thus providing a depth estimate and a better initialization of the tracker. For new features, even accounting for reasonable depth discontinuities, the initialization can be performed by using the already available depth information of similar nearby points.

4.2 Results

We tested our method on sequences extracted from the KITTI dataset [9], consisting of driving scenes recorded by a calibrated stereo setup. Each of the extracted sequences consists of 199 image pairs and present different settings:

- Sequence 1 shows an urban scenario: mainly buildings on the sides of the road.
- Sequence 2 is set in a suburban scenario. Mostly trees surround the road

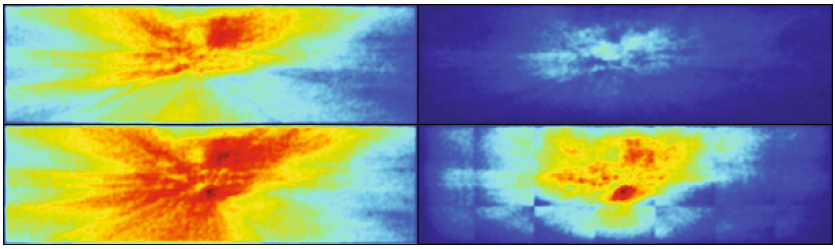


Fig. 4. Distribution of extracted GFTT (**top-left**) and GETT (**bottom-left:**). Distribution of successfully tracked GFTT (**top-right**) and GETT (**bottom-right**).

– Sequence 3 also has a suburban setting and shows a variety of buildings, vegetation and parked cars on the side of the road.

In our tests, we disregard the stereo information, focusing on a monocular scenario and making use of the given camera calibration. To show the robustness of the method, we do not use the ground truth egomotion provided with the KITTI dataset, we take instead the estimates produced by an implementation of [20]. The approach consists in a matching procedure performed on FAST keypoints [24] and BRIEF descriptors [6], the epipolar geometry estimation is made robust with RANSAC, and stabilized with a windowed bundle adjustment.

Tables 1 and 2 show, as expected, the higher density of GETT compared to GFTT which, combined with a better inlier ratio granted by the proposed tracking procedure, allows to sensibly increase density of the motion field. Notice that, as shown in Figure 4 and 3 the increase in density is particularly evident in areas of the image closer to the camera and on the road plane. This is a very desirable feature for many applications since closeby objects are generally of more interest and can be measured more precisely.

5 Conclusions

In this work we presented a complete framework for the extraction and tracking of edge elements in an image pair. Such edge elements would normally suffer from the aperture problem and cannot be tracked successfully. By exploiting a known estimate of the epipolar geometry of the scene, our algorithm allows to extract and successfully track specific edgels that are likely to behave well with the given egomotion. We have shown that such keypoints ($GETT$) consistently outnumber the standard $GFTT$ in different settings, have a higher inlier ratio when tracked with the procedure we presented and are also more uniformly spread on the image plane. These are all very desirable qualities that allow to produce a much denser point cloud in any structure from motion application.

Acknowledgments. This work has been supported by ELLIIT, Strategic Area for ICT research, and CADICS, funded by the Swedish Government, and by iQMatic, funded by VINNOVA.

References

1. Agrawal, M., Konolige, K., Blas, M.R.: CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 102–115. Springer, Heidelberg (2008)
2. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: A unifying framework. International Journal of Computer Vision **56**(3), 221–255 (2004)
3. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
4. Birchfield, S.T., Pundlik, S.J.: Joint tracking of features and edges. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2008) (2008)
5. Bouguet, J.Y.: Pyramidal implementation of the affine Lucas Kanade feature tracker: description of the algorithm. Intel Corporation, Tech. rep. (2001)
6. Calonder, M., Lepetit, V., Strecha, C., Fua, P.: BRIEF: Binary Robust Independent Elementary Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 778–792. Springer, Heidelberg (2010)
7. Farnebäck, G.: Two-Frame Motion Estimation Based on Polynomial Expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003)
8. Fusiello, A., Trucco, E.: Improving feature tracking with robust statistics. Pattern Analysis & Applications pp. 312–320 (1999)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2012) (2012)
10. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey Vision Conference, pp. 23.1–23.6 (1988)
11. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2004)
12. Hedborg, J., Forssén, P.-E., Felsberg, M.: Fast and Accurate Structure and Motion Estimation. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Wang, J.-X., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnação, M.L., Silva, C.T., Coming, D. (eds.) ISVC 2009, Part I. LNCS, vol. 5875, pp. 211–222. Springer, Heidelberg (2009)
13. Horn, B., Schunck, B.: Determining optical flow. In: SPIE 0281, Techniques and Applications of Image Understanding, vol. 319 (1981)
14. Jonsson, E., Felsberg, M.: Efficient robust mean value computation of 1D features. In: Proceedings of Svenska Sällskapet för Automatiserad Bildanalys. SSBA-2005 (2005)
15. Kitt, B., Lategahn, H.: Trinocular optical flow estimation for intelligent vehicle applications. In: International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 300–306. IEEE, September 2012
16. Lee, T., Soatto, S.: Fast planar object detection and tracking via edgel templates. In: IEEE Workshop on the Applications of Computer Vision (WACV), pp. 473–480. IEEE, January 2012
17. Leutenegger, S.: BRISK: Binary robust invariant scalable keypoints. IEEE Int. Conf. on Computer Vision (ICCV) **2011**, 2548–2555 (2011)
18. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)

19. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artificial Intellicence (IJCAI) (1981)
20. Persson, M.: Online Monocular SLAM. Master's thesis, Computer Vision Laboratory, Linköping University, Sweden (December 2013)
21. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: IEEE Int. Conf. on Computer Vision (ICCV), pp. 2564–2571 (2011)
22. Shi, J., Tomasi, C.: Good features to track. In: IEEE Conf. on Computer Vision and Pattern Recognition CVPR 1994, pp. 593–600 (1994)
23. Tommasini, T., Fusiello, A.: Making good features track better. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 178–183 (1998)
24. Trajković, M., Hedley, M.: Fast corner detection. Image and Vision Computing **16**(1998), 75–87 (1998)
25. Trummer, M., Denzler, J., Munkelt, C.: KLT tracking using intrinsic and extrinsic camera parameters in consideration of uncertainty. International Conference on Computer Vision Theory and Applications (VISAPP) (2008)
26. Yamaguchi, K., McAllester, D., Urtasun, R.: Robust monocular epipolar flow estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1862–1869. IEEE, June 2013

W16 - Soft Biometrics

Facial Age Estimation Through the Fusion of Texture and Local Appearance Descriptors

Ivan Huerta[1(✉)], Carles Fernández[2], and Andrea Prati[1]

[1] DPDCE, University IUAV, Santa Croce 1957, 30135 Venice, Italy
{huertacasado,aprati}@iuav.it
[2] Herta Security, Pau Claris 165 4-B, 08037 Barcelona, Spain
carles.fernandez@hertasecurity.com

Abstract. Automatic extraction of soft biometric characteristics from face images is a very prolific field of research. Among these soft biometrics, age estimation can be very useful for several applications, such as advanced video surveillance [5,12], demographic statistics collection, business intelligence and customer profiling, and search optimization in large databases. However, estimating age from uncontrollable environments, with insufficient and incomplete training data, dealing with strong person-specificity, and high within-range variance, can be very challenging. These difficulties have been addressed in the past with complex and strongly hand-crafted descriptors, which make it difficult to replicate and compare the validity of posterior classification schemes. This paper presents a simple yet effective approach which fuses and exploits texture- and local appearance-based descriptors to achieve faster and more accurate results. A series of local descriptors and their combinations have been evaluated under a diversity of settings, and the extensive experiments carried out on two large databases (MORPH and FRGC) demonstrate state-of-the-art results over previous work.

Keywords: Age estimation · CCA · HOG · LBP · SURF

1 Introduction

The problem of age estimation from images has historically been one of the most challenging within the field of facial analysis. Some of the reasons are the uncontrollable nature of the aging process, the strong specificity to the personal traits of each individual [24], high variance of observations within the same age range, and the fact that it is very hard to gather complete and sufficient data to train accurate models [7].

This process can be made easier by having available large and representative collections of age-annotated images. However, in the past the available databases were often very limited and strongly skewed. This is especially disadvantageous for applications like video surveillance and forensics, which need to work correctly when facing unknown subjects and a lack of any additional cues. Fortunately, the recent availability of large databases like MORPH [21] and FRGC [20] offers

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 667–681, 2015.
DOI: 10.1007/978-3-319-16181-5_51

a great opportunity to make advances in the field. Keeping in mind that any training data set which is representative of the whole population cannot exist, the only viable option is to develop methods that are able to exploit large databases in order to gain substantial generalization capabilities.

The inherent difficulties in the facial age estimation problem, such as limited imagery, challenging subject variability, and subtle visual age patterns, have derived research in the field into building particularly complex feature extraction schemes. The most typical ones consist of either hand-tuned multi-level filter banks, that intend to emulate the behavior of primary visual cortex cells, or fine-grained facial meshes to accomplish precise alignment through dozens of facial landmarks. In any case, the resulting extraction schemes are difficult to replicate, and the high-dimensional visual descriptors in many cases take considerable time to be extracted and processed.

On the other hand, during the last decade, several fields within image classification and object recognition have proposed different families of very fast and descriptive feature extraction schemes, which have become well-known for being especially invariant to rotation, scale, illumination, and alignment. Such histogram-based descriptors, which typically capture local intensity variations or local neighborhood patterns from spatial grids, are nowadays a fundamental tool to deal with highly adverse and unconstrained environments for a variety of applications.

In this paper we conduct a thorough evaluation of a series of common local visual descriptors, in order to investigate their utility towards the automatic facial age estimation problem. The contributions are as follows:

- We review some of the most efficient and effective local visual descriptors from image classification, and explore their suitability to extract age-related discriminative patterns.
- We demonstrate that the fusion of textural and local appearance-based descriptors achieves state-of-the-art results, improving over complex feature extraction schemes that were previously proposed.
- Candidate descriptors are exhaustively evaluated regarding optimal parameters and regularization, in terms of mean average errors and cumulative score curves over two large databases.

The paper is structured as follows: next section gathers and comments on previous related work on facial age estimation. The candidate descriptors to be evaluated are reviewed in Section 3, along with the selected classification scheme. Evaluation is presented out in Section 4, by first describing available large databases with age annotations, and subsequently analyzing the extensive experiments carried out over the combinations of local descriptors. Finally, Section 5 summarizes the results and draws some conclusions.

2 Related Work

After an initial interest on automatic age estimation from images dated back in the early 2000s [13–15], research in the field has experienced a renewed interest from 2006 on, since the availability of large databases like MORPH-Album 2 [21], which increased by 55× the amount of real age-annotated data with respect to traditional age databases. Therefore, this database has deeply been employed in recent works by applying over it different descriptors and classification schemes.

Feature extraction scheme. Regarding visual features, flexible shape and appearance models such as ASM (Active Shape Model) and AAM (Active Appearance Model) have been some of the primary cues used to model aging patterns [2,7,8,13]. Such statistical models capture the main modes of variation in shape and intensity observed in a set of faces, and allow face signatures based on such characterizations to be encoded.

Bio-Inspired Features (BIF) [22] and its derivations have consistently been used for age estimation in the last years [7,12]. These feed-forward models consist of a number of layers intertwining convolutionally and pooling processes. First, an input image is mapped to a higher-dimensional space by convolving it with a bank of multi-scale and multi-orientation Gabor filters. Later, a pooling step downscales the results with a non-linear reduction, typically a MAX or STD operation, progressively encoding the results into a vector signature. In [17], the authors carefully design a two-layer simplification of this model for age estimation by manually setting the number of bands and orientations for convolution and pooling. Such features are also used in their posterior works [9–11].

Features extracted from local neighborhoods have very rarely been used for the purpose of age estimation. In [24], LBP histogram features are combined with principal components of BIF, shape and textural features of AAM, and PCA projection of the original image pixels. HOG features have independently been used for age estimation in [4].

Classification scheme. With regards to the learning algorithm, several approaches have been proposed, including, among others, Support Vector Machines (SVM) / Support Vector Regressors (SVR) [2,12,17,24], neural networks [13] and their variant of Conditional Probability Neural Network (CPNN) [7], Random Forests (RF) [16], and projection techniques such as Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA), along with their regularized and kernelized versions [9–11]. An extensive comparison of these classification schemes for age estimation has been reported in our previous paper [4], and in particular the advantageousness of CCA was demonstrated over the others, both regarding accuracy and efficiency.

Specific attention must be given to the CCA technique, which is the main focus of this paper from the classification perspective. The PLS and CCA subspace learning algorithms were originally conceived to model the compatibility between two multidimensional variables. PLS uses latent variables to learn a new space in which such variables have maximum correlation, whereas CCA

finds basis vectors such that the projections of the two variables using these vectors are maximally correlated to each other. Both techniques have been adapted for label regression. To the best of our knowledge, the best current result over MORPH is achieved by combining BIF features with kernel CCA [10], although in that case the size of training folds is limited to 10K samples due to computational limitations.

The main contribution of this paper is the proposal of a novel combination of well-known local descriptors capturing texture and contour cues for the purpose of facial age estimation. The orthogonal nature of these features allows the exploitation of the benefits of each of them, bringing to performance which are superior than in the case of them applied separately. To the best of our knowledge, this approach has never been employed before for age estimation, and our experiments demonstrate comparable performance with respect to state-of-the-art results provided by complex and fine-tuned feature extraction schemes such as BIF [11]. Moreover, for the sake of simplicity and efficiency, a simple eye alignment operation is carried out through similarity transformation, as opposed to precise alignment approaches typically fitting active shape and appearance models with tens of facial landmarks.

3 Methodology

Preprocessing. In general, existing works tackle the problem of age estimation with visual features that are either complex and fine-tuned (e.g., BIF), or require precise statistical models involving tens of facial landmarks for accurate alignment (e.g., ASM and AAM). As opposed to them, we do not rely on precisely aligned appearance models; instead, our experiments will be evaluated using a simple alignment through the fiducial landmarks of the detected eye regions.

The facial region of each image has been detected with the face detector described in [19]. The relative alignment invariance of local descriptors based on concatenated cell histograms allows us to work with simple eye-aligned images. The fiducial markers corresponding to the eye centers have been obtained using the convolutional neural network for face alignment presented in [23]. The aligned version of each detected face is obtained by a non-reflective similarity image transformation that yields an optimal least-square correspondence between the eye centers and the target locations, that have been symmetrically placed at 25% and 75% of the alignment template. Unlike previous works like [10], which use input images of 60×60 pixels, our aligned image are resized to only 50×50 pixels.

Descriptors. The choice of visual features to be extracted from aligned images and sent to the classification scheme plays a fundamental role on the resulting estimation accuracy. In this paper, we have selected a number of significant local invariant descriptors that have been useful for image matching and object recognition in the past due to their expressiveness, fast computation, compactness, and invariance to misalignment and monotonic illumination changes. They include local appearance descriptors as HOG and texture descriptors as LBP and SURF.

Histograms of Oriented Gradients (HOG) [3] have largely been used as robust visual descriptors in many computer vision applications related to object detection and recognition. The horizontal and vertical gradients of the input image are computed, and the image region is divided into $C_x \times C_y$ grid cells. A histogram of orientations is assigned to each cell, in which every bin accounts for an evenly split sector of either the $[0, \pi]$ or $[-\pi, \pi]$ domain (for unsigned and signed versions, respectively). At each pixel location, the gradient magnitude and orientation is computed, and that pixel increments the assigned orientation bin of its correspondent cell by its gradient magnitude. Cell histograms are concatenated to provide the final descriptor. We use $\text{HOG}_{C,B}$ to denote $C \times C$ square grids and B orientation bins.

Local Binary Patterns (LBP) [18] have been long used as a textural descriptor for image classification, and more recently, variations of the original proposal have provided state-of-the-art results in fields like face and object recognition. The original operator describes every pixel in the image by thresholding its surrounding 3×3-neighborhood with its intensity value, and concatenating the 8 boolean tests as a binary number. A common extension considers generic pixel neighborhoods formed by P sampled pixel values at radius R from the central pixel. To build an LBP compact descriptor, a histogram is computed over the filtered result, in which each bin corresponds to a LBP code. Another typical extension reduces the dimensionality of the descriptor by assigning all *non-uniform* codes to a single bin, whereas uniform codes are defined as those having not more than 2 bitwise transitions from 0 to 1 or vice versa (e.g., 00111000, versus non-uniform 01001101). An LBP descriptor of generic neighborhood size and radius using uniform patterns is referred as $LBP^{u2}_{P,R}$, e.g. $LBP^{u2}_{8,2}$.

Speeded-Up Robust Features (SURF) [1] is an interest point detector and descriptor that is particularly invariant to scale and rotation. It has commonly been used in image matching and object recognition as a faster and comparable alternative to SIFT. In our case, we concentrate on the descriptor component of the upright version of the technique (U-SURF). The square image region to describe is partitioned into 4×4 subregions. Horizontal and vertical wavelet responses d_x and d_y are computed and weighted with a Gaussian. The sum of these responses and their absolute values are stored, generating a 4-dimensional vector $(\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|)$ for each subregion, and these are concatenated to form the final 64-dimensional descriptor of the image region, SURF_{64}. A common extension consists of doubling the number of features, by separately computing the sums of d_x and $|d_x|$ for $d_y < 0$ and $d_y \geq 0$, and equally for d_y given the sign of d_x, thus yielding SURF_{128}.

As gradient information is typically a very relevant cue to describe image content for many image descriptors, we have included raw magnitude gradient images (GRAD) as a baseline in our experiments for the evaluation of the proposed descriptors.

Classification. From the wide variety of learning schemes presented in the literature on facial age estimation, **Canonical Correlation Analysis (CCA)** and its derivations have recently obtained state-of-the-art results in challenging large databases such as MORPH [11]. This projection technique involves low computational effort and unprecedented accuracy in the field, for which we use it as our chosen regression learning algorithm. CCA is posed as the problem of relating data $\mathbf{X}$ to labels $\mathbf{Y}$ by finding basis vectors w_x and w_y, such that the projections of the two variables on their respective basis vectors maximize the correlation coefficient

$$\rho = \frac{{w_x}^T \mathbf{X}\mathbf{Y}^T w_y}{\sqrt{({w_x}^T \mathbf{X}\mathbf{X}^T w_x)({w_y}^T \mathbf{Y}\mathbf{Y}^T w_y)}}, \tag{1}$$

or, equivalently, finding $\max_{w_x, w_y} {w_x}^T \mathbf{X}\mathbf{Y}^T w_y$ subject to the scaling ${w_x}^T \mathbf{X}\mathbf{X}^T w_x$=1 and ${w_y}^T \mathbf{Y}\mathbf{Y}^T w_y$=1. For age estimation, labels in $\mathbf{Y}$ are unidimensional, so a least squares fitting suffices to relate these labels to the projected data features. Thus, only w_x is computed, by solving the following generalized eigenvalue problem:

$$\mathbf{X}\mathbf{Y}^T \left(\mathbf{Y}\mathbf{Y}^T +_y I\right)^{-1} \mathbf{Y}\mathbf{X}^T w_x = \lambda \left(\mathbf{X}\mathbf{X}^T + I\right) w_x \tag{2}$$

When projecting through the solution w_x, the dimensionality of data features is reduced to one dimension per output (a single numerical value in our case), so the aforementioned label fitting simply consists on finding the scalar value that optimally adapts the projected values to the ground truth age, in the least-squares sense. The described procedure can be stabilized through regularization, by modifying the eigenvalue problem in the following manner:

$$\mathbf{X}\mathbf{Y}^T \left((1-\gamma_y)\mathbf{Y}\mathbf{Y}^T + \gamma_y I\right)^{-1} \mathbf{Y}\mathbf{X}^T w_x = \lambda \left((1-\gamma_x)\mathbf{X}\mathbf{X}^T + \gamma_x I\right) w_x \tag{3}$$

Regularization terms $\gamma_x, \gamma_y \in [0, 1]$ have been included in Eq. 3 to prevent overfitting. Although CCA also admits extension to a kernelized version, in that case covariance matrices become computationally intractable with over 10K samples. In practice, regularized CCA works comparably to KCCA [10], it is much less computationally demanding, and will allow us to reproduce the same exact validation schemes over large databases.

4 Experimental Results

Age databases. Due to the nature of the age estimation problem, there is a restricted number of publicly available databases providing a substantial number of face images labeled with accurate age information. Table 1 shows the summary of the existing databases with main reference, number of samples, number of subjects, and comments.

From the information in Table 1, we see that PAL and FG-NET are comparatively negligible to the rest in terms of number of samples. Additionally, age

Table 1. Description of the existing databases for age estimation

Database	Samples	Subjects	Comments
PAL [15]	580	580	Limited number of samples
FG-NET [14]	1,002	82	Limited number of samples and subjects
GROUPS [6]	28,231	28,231	Ages discretized into seven age intervals
FRGC v2.0 [20]	44,278	568	Large database; many samples per subjects
MORPH II [21]	55,134	13,618	Large database; high diversity

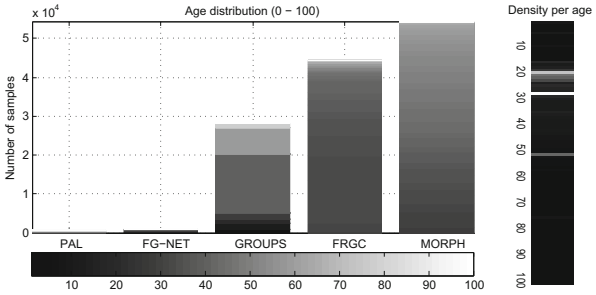


Fig. 1. Age distribution and density per database. In the left graphic (Age distribution) different ages are represented by the intensity. In the right graphic (Desity per age) the intensity represent the density (white color more density). PAL and FG-NET are relatively negligible compared to others, and GROUPS only provides age intervals, so we focus on MORPH II and FGRC. Age samples are mainly skewed towards 20–30 and 50 year old.

annotations in GROUPS are discretized into seven age intervals, which makes it unsuitable for training accurate age estimation models. Moreover, FG-NET contains only 82 subjects, so a *leave-one-person-out* validation scheme is employed by convention, to avoid optimistic biasing by identity replication. Given such limitations, and the recent tendency to use MORPH as a standard for age estimation, we concentrate on this database and on FRGC to provide experimental evaluations. Although the FRGC database is comparable to MORPH regarding number of samples, image quality and age range coverage, we have only found one previous publication on age estimation including FRGC as part of their experiments [4]. Figure 1 offers a graphical visualization and comparison of the analyzed databases, by number of samples and density of age ranges.

Metrics. To evaluate the accuracy of the age estimators, the conventional metrics are the Mean Average Error (MAE) and the Cumulative Score (CS). MAE computes the average age deviation error in absolute terms, $MAE = \sum_{i=1}^{M} |\hat{a}_i - a_i|/M$, with $\hat{a}_i$ the estimated age of the i-th sample, a_i its real age and M the total of samples. CS is defined as the percentage of images for which the error e is no higher than a given number of years l, as $CS(l) = M_{e \leq l}/M$ [2,12,24] .

Cx	Cy	B																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	3	7.84	8.16	7.32	7.06	7.11	6.97	6.88	6.86	6.73	6.77	6.66	6.58	6.60	6.56	6.55	6.48	6.48	6.49
4	4	7.47	7.17	6.84	6.82	6.62	6.56	6.35	6.42	6.28	6.28	6.17	6.18	6.16	6.16	6.08	6.06	6.04	6.06
5	5	6.68	6.45	6.02	6.05	5.76	5.75	5.55	5.53	5.47	5.44	5.39	5.37	5.38	5.35	5.33	5.31	5.31	5.29
6	6	6.15	6.07	5.66	5.67	5.53	5.43	5.30	5.32	5.26	5.23	5.17	5.18	5.16	5.14	5.13	5.11	5.12	5.10
7	7	5.90	5.70	5.47	5.36	5.13	5.10	4.98	4.99	4.93	4.93	4.89	4.89	4.88	4.85	4.85	4.85	4.85	4.84
8	8	5.58	5.44	5.19	5.13	4.97	4.94	4.84	4.86	4.80	4.80	4.76	4.77	4.75	4.75	4.74	4.73	4.74	4.73
9	9	5.36	5.25	5.02	4.98	4.86	4.81	4.73	4.75	4.71	4.69	4.66	4.67	4.64	4.64	4.65	4.64	4.63	4.64
10	10	5.28	5.13	4.98	4.91	4.77	4.73	4.68	4.69	4.64	4.61	4.61	4.60	4.59	4.58	4.59	4.59	4.59	4.59
11	11	5.10	5.01	4.83	4.76	4.66	4.62	4.55	4.57	4.54	4.50	4.50	4.50	4.49	4.47	4.50	4.49	4.49	4.50
12	12	5.33	5.21	5.03	4.97	4.84	4.82	4.77	4.78	4.72	4.71	4.70	4.72	4.70	4.69	4.70	4.70	4.71	4.71
13	13	5.11	5.00	4.82	4.80	4.66	4.65	4.60	4.61	4.57	4.56	4.54	4.56	4.55	4.54	4.55	4.56	4.56	4.57
14	14	4.97	4.87	4.70	4.68	4.57	4.55	4.50	4.51	4.47	4.46	4.45	4.48	4.46	4.46	4.47	4.47	4.48	4.49
15	15	4.83	4.78	4.59	4.56	4.47	4.45	4.41	4.42	4.39	4.38	4.38	4.40	4.39	4.39	4.40	4.41	4.41	4.43
16	16	5.56	5.44	5.29	5.24	5.14	5.09	5.08	5.10	5.04	5.06	5.05	5.06	5.07	5.04	5.09	5.10	5.11	5.11
17	17	5.39	5.29	5.13	5.08	5.00	4.95	4.96	4.97	4.92	4.92	4.92	4.94	4.95	4.92	4.96	4.98	4.99	5.01
18	18	5.21	5.11	4.96	4.92	4.84	4.81	4.81	4.82	4.77	4.77	4.79	4.80	4.81	4.80	4.83	4.85	4.88	4.88
19	19	5.03	4.89	4.76	4.74	4.65	4.64	4.62	4.63	4.61	4.60	4.62	4.63	4.63	4.63	4.67	4.69	4.71	4.72
20	20	4.90	4.78	4.67	4.63	4.55	4.55	4.54	4.54	4.54	4.53	4.54	4.56	4.57	4.57	4.60	4.63	4.65	4.67
21	21	4.82	4.71	4.61	4.59	4.50	4.50	4.49	4.51	4.50	4.50	4.51	4.53	4.54	4.55	4.58	4.61	4.63	4.66
22	22	4.73	4.64	4.54	4.52	4.45	4.44	4.44	4.46	4.46	4.46	4.47	4.50	4.50	4.52	4.55	4.59	4.61	4.64
23	23	4.68	4.61	4.50	4.48	4.42	4.41	4.41	4.44	4.44	4.45	4.45	4.49	4.50	4.51	4.56	4.60	4.61	4.65
24	24	4.64	4.57	4.48	4.47	4.41	4.40	4.41	4.43	4.44	4.44	4.45	4.50	4.51	4.53	4.57	4.62	4.64	4.67
25	25	6.14	6.07	6.02	5.90	5.89	5.86	5.84	5.88	5.89	5.90	5.93	5.96	5.99	6.03	6.12	6.12	6.18	6.24

Fig. 2. Results for $\mathrm{HOG}_{C,B}$ feature for a single scale with image size 50×50 at varying grid size C (rows) and number of bins B (columns). The bordered cell shows the best value.

Related publications typically supply either an eleven-point curve for age deviations $[0-10]$, or simply the value $CS(5)$.

All through the rest of this paper, the optimal parameters are searched so as to minimize the MAE score over MORPH, using 5-fold cross-validation in all cases. In particular, the division into training and validation sets is made so that all the instances of the same subject are contained in one single fold at a time; this applies to all the presented experiments. Descriptors are always directly extracted from the aligned version of detected faces.

Parameter analysis. In order to evaluate in depth the performance of the analyzed features for age estimation, we have conducted an analysis of the different parameters for the compared feature detectors. In the case of $\mathrm{HOG}_{C,B}$, the optimal parameters for grid size $C{\times}C$ and number of bins B have been obtained through exhaustive logarithmic grid search and 5-fold cross-validation, for single and multiple scales. Our implementation of HOG incorporates 50% cell overlapping for smoothness and global L2 normalization, instead of per-cell. Multiscale variations are achieved by concatenating the feature vectors obtained by the descriptor at different scales. In order to have a fair comparison with the results reported in [11], images have been processed at 50×50 (similar to the 60×60 size used in that paper). However, we also evaluate the effect of different image sizes on the final performance in Fig. 4, where images of size 100×100 were used. In summary, Figs. 2, 3 and 4 report the individual analysis of HOG descriptors

Cx	Cy	B													
		5	6	7	8	9	10	11	12	13	14	15	16	17	18
8	8					4.62	4.63	4.58	4.59	4.58	4.58				
9	9					4.50	4.51	4.48	4.48	4.47	4.48				
10	10	4.72	4.67	4.56	4.55	4.51	4.52	4.49	4.49	4.48	4.50	4.50	4.49	4.64	4.52
11	11	4.61	4.56	4.48	4.47	4.43	4.44	4.42	4.43	4.43	4.44	4.45	4.44	4.47	4.48
12	12	4.72	4.68	4.60	4.61	4.57	4.59	4.57	4.58	4.58	4.61	4.60	4.63	4.64	4.66
13	13	4.74	4.73	4.61	4.62	4.58	4.60	4.57	4.57	4.57	4.59	4.58	4.59	4.59	4.62
14	14	4.63	4.62	4.53	4.53	4.48	4.52	4.49	4.49	4.49	4.53	4.52	4.54	4.55	4.57
15	15	4.52	4.51	4.45	4.45	**4.41**	4.45	4.42	4.43	4.44	4.47	4.46	4.49	4.51	4.54
16	16				5.04	5.04	5.08	5.04	5.07	5.07	5.09	5.12			

Fig. 3. Results for $\text{HOG}_{C,B}^{\times 3}$ feature for 3 scales concatenating descriptors over 50×50, 25×25, and 13×13 images, at varying grid size C (rows) and number of bins B (columns). The bordered cell shows the best value.

Cx	Cy	B											
		6	7	8	9	10	11	12	13	14	15	16	17
7	7	5.39	5.13	5.09	4.97	4.95	4.87	4.88	4.85	4.82	4.82	4.81	4.80
8	8	5.15	4.93	4.91	4.80	4.79	4.73	4.72	4.70	4.67	4.66	4.65	4.66
9	9	4.85	4.70	4.65	4.59	4.59	4.53	4.51	4.49	4.48	4.48	4.47	4.48
10	10	4.87	4.67	4.62	4.54	4.55	4.49	4.49	4.46	4.44	4.44	4.44	4.43
11	11	4.64	4.50	4.48	4.41	4.42	4.37	4.37	4.36	4.34	4.35	4.34	4.34
12	12	4.63	4.51	4.47	4.41	4.42	4.38	4.38	4.37	4.36	4.36	4.35	4.36
13	13	4.52	4.41	4.38	4.33	4.33	4.30	4.29	4.28	4.28	4.28	4.28	4.28
14	14	4.47	4.36	4.33	4.31	4.30	4.28	4.29	4.27	4.26	4.28	4.27	4.29
15	15	4.37	4.28	4.26	4.23	4.23	4.21	4.22	4.20	4.20	4.21	4.22	4.24
16	16	4.44	4.35	4.33	4.30	4.31	4.30	4.28	4.29	4.27	4.29	4.29	4.30
17	17	4.36	4.28	4.26	4.24	4.25	4.23	4.23	4.23	4.22	4.24	4.24	4.25
18	18	4.30	4.23	4.21	4.20	4.20	4.19	4.18	4.19	4.19	4.20	4.21	4.22
19	19	4.26	4.20	4.18	4.17	4.18	4.17	**4.16**	4.17	4.17	4.19	4.19	4.22
20	20	4.41	4.34	4.24	4.33	4.33	4.32	4.34	4.34	4.35	4.38	4.37	4.40

Fig. 4. Results for $\text{HOG}_{C,B}$ feature for a single scale with size image 100×100 at varying grid size C (rows) and number of bins B (columns). The bordered cell shows the best value.

for a single scale at 50×50 pixels; for 3-scales at {50×50, 25×25, 13×13}; and for a single scale at 100×100, respectively. Fig. 4 shows that 100×100 images provide even better scores than the traditional sizes in the literature, although we conduct the rest of experiments for 50×50 pixels for fair comparison. Single scale HOG performed better than multiscale.

A similar grid search procedure has been chosen to optimize the parameters of LBP and SURF descriptors. In the case of $\text{LBP}_{P,R}^{u2}$ the analysis has been carried out by searching the optimal number of sampled neighbors P and radius R, for one and three scales, constraining the number of neighbors to either 8 or 16, see Table 2. In the case of SURF, multiple scales have been tested for both

Table 2. MAE for the single-scale descriptor $\text{LBP}^{u2}_{P,R}$ at 50×50 pixels, and for the 3-scale $\text{LBP}^{u2\times 3}_{P,R}$ concatenating 50×50, 25×25, and 13×13. Neighborhoods of 8 and 16 are shown.

	(Size)	Radius R								
		2	3	4	5	6	7	8	9	10
$\text{LBP}^{u2}_{8,R}$	(59)	7.17	7.12	7.15	7.30	7.55	7.82	8.04	8.11	8.08
$\text{LBP}^{u2}_{16,R}$	(243)	6.88	6.70	**6.66**	6.76	7.06	7.25	7.40	7.51	7.81
$\text{LBP}^{u2\times 3}_{8,R}$	(177)	6.48	6.49	6.66	6.82	10.75	-	-	-	-
$\text{LBP}^{u2\times 3}_{16,R}$	(729)	6.18	**6.13**	12.41	11.32	12.26	-	-	-	-

Table 3. MAE results for SURF at one and multiple scale combinations. Size in brackets.

Scale	SURF_{64}	SURF_{128}	Multiscale	$\text{SURF}^{\times S}_{64}$	$\text{SURF}^{\times S}_{128}$
1.6	**6.09** (320)	**5.72** (640)	{1.6, 2}	5.73 (640)	5.39 (1280)
1.8	6.21 (320)	5.77 (640)	{1.6, 2.4}	5.71 (640)	5.41 (1280)
2.0	6.24 (320)	5.81 (640)	{2, 3}	5.95 (640)	5.60 (1280)
2.4	6.65 (320)	6.24 (640)	{1.6, 1.8, 2}	5.67 (960)	5.34 (1920)
3.0	6.93 (320)	6.59 (640)	{1.6, 2, 2.4}	**5.59** (960)	**5.30** (1920)
4.0	7.46 (320)	7.12 (640)	{1.6, 2.4, 3}	5.60 (960)	5.33 (1920)
5.0	7.52 (320)	7.26 (640)	{2, 2.4, 3}	5.84 (960)	5.53 (1920)

the original and extended descriptor (SURF_{64} and SURF_{128}) over the keypoints, which are the five fiducial points previously used for the alignment, as shown in Table 3.

The optimal regularization cost γ^*, as defined in Section 3, differs for each computed feature and parameter. For this reason, initially the above-mentioned grid search has been performed without regularization ($\gamma = 0$). Once the best parameters for the feature detectors have been identified, the optimal regularization cost has been searched by looking for the optimal (minimum) MAE. Additionally, we impose $\gamma_x = \gamma_y$. However, our experiments suggest that no significant changes can be noticed when incorporating regularization because of the relative size of the database to the descriptor, as shown in Fig. 5. As the number of database examples M increases well over the dimensionality of the feature N, i.e. $M \gg N$, the optimal regularization cost γ^* tends to zero.

In order to improve the accuracy of the estimation, and taking advantage of the orthogonal nature of different descriptors, a thorough analysis of fusion combinations among feature candidates has been carried out. Although more combinations have been tested, Table 4 shows the most significant ones: single-scale $\text{HOG}_{8,9}$ and $\text{HOG}_{15,13}$; 3-scale $\text{LBP}^{u2\times 3}_{16,3}$; the raw gradient magnitude GRAD; and the 3-scale $\text{SURF}^{\times 3}_{64}$ and $\text{SURF}^{\times 3}_{128}$ with scales 1.6, 2, and 2.4. Feature combinations have been obtained by concatenating the descriptors and exploiting the best parameters obtained previously.

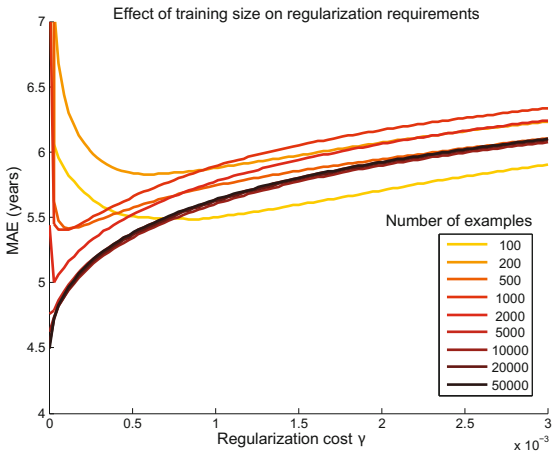


Fig. 5. The need for regularization depends strongly on the ratio between training examples M and feature dimensionality N. This figure shows 5-fold cross-validation results using 576-dimensional $\text{HOG}_{8,9}$ and CCA, through different values of γ and increasing examples from 100 to 50K. As M increases the optimal γ^* decays, dropping to zero for $M \gg N$.

As observed from the results summarized in Table 4, $\text{SURF}_{128}^{\times 3}$ reduces its MAE when fused with other features (from 5.30 years down to 4.33 when combined with $\text{HOG}_{15,13}$ and $\text{LBP}_{16,3}^{u2\times 3}$), and performs worse than SURF_{64} under the same combination. The best result is obtained when combining $\text{HOG}_{15,13}$, $\text{LBP}_{16,3}^{u2\times 3}$ and $\text{SURF}_{64}^{\times 3}$. This combination has the advantage of fusing texture and local appearance-based descriptors. Another noticeable remark is the so-called curse of dimensionality: the addition of further descriptors into higher dimensional features not always enhances the result.

The specific size of the most accurate descriptors does not seem to be correlated to their accuracy either, at least not after proper regularization has been applied. The HOG family of descriptors behaves particularly well for the different granularities that were tested, $\text{HOG}_{8,9}$ and $\text{HOG}_{15,13}$, of 576 and 2925 dimensions respectively. This suggests that local appearance information is particularly useful and quite sufficient for capturing age patterns. The size of the descriptor deserves important consideration in the case of CCA, as it strongly affects the computational efficiency of the training process, and plays an important role in the stability of the solution: higher $\frac{M}{N}$ ratios result in more stable pseudo-inverse matrices when searching for the CCA projection matrix.

Table 5 shows the effect of regularization on the features that yielded best MAE scores in our experiments, over the MORPH database and using the regularized CCA regression technique. The optimal regularization costs are provided. We have also included the best results (to the best of our knowledge) achieved using the BIF descriptor, which is very commonly used in age estima-

Table 4. MAE results for the fusion of different descriptors that yielded best results. $HOG_{8,9}$ and $HOG_{15,13}$ have a single scale. $LBP^{u2\times3}_{16,3}$ is computed at the original, half and quarter image size. GRAD is formed concatenating all gradient magnitude values. $SURF^{\times3}_{64}$ and $SURF^{\times3}_{128}$ are aggregated SURF descriptors with scales $\{1.6, 2, 2.4\}$. The best result is achieved by combining $HOG_{15,13}$, $LBP^{u2\times3}_{16,3}$, and $SURF^{\times3}_{64}$.

$HOG_{8,9}$	$HOG_{15,13}$	$LBP^{u2\times3}_{16,3}$	GRAD	$SURF^{\times3}_{64}$	$SURF^{\times3}_{128}$	(Size)	MAE
•						(576)	4.84
	•					(2925)	4.38
		•				(729)	6.13
			•			(2500)	5.58
				•		(960)	5.59
					•	(1920)	5.30
$HOG_{8,9}$	$HOG_{15,13}$	$LBP^{u2\times3}_{16,3}$	GRAD	$SURF^{\times3}_{64}$	$SURF^{\times3}_{128}$	(Size)	MAE
•		•				(1305)	4.66
•		•	•			(3805)	4.53
•		•		•		(2265)	4.42
•		•			•	(3225)	4.61
•		•	•	•		(4765)	4.51
•		•	•		•	(5725)	4.72
	•	•				(3654)	4.33
	•		•			(5420)	4.33
	•	•	•			(6154)	4.30
	•			•		(3885)	4.30
	•				•	(4845)	4.33
	•	•		•		(4614)	**4.27**
	•	•			•	(5574)	4.33
	•	•	•	•		(7114)	4.31
	•	•	•		•	(8074)	4.34
		•	•			(3229)	5.07
		•		•		(1689)	5.31
		•			•	(2649)	6.45

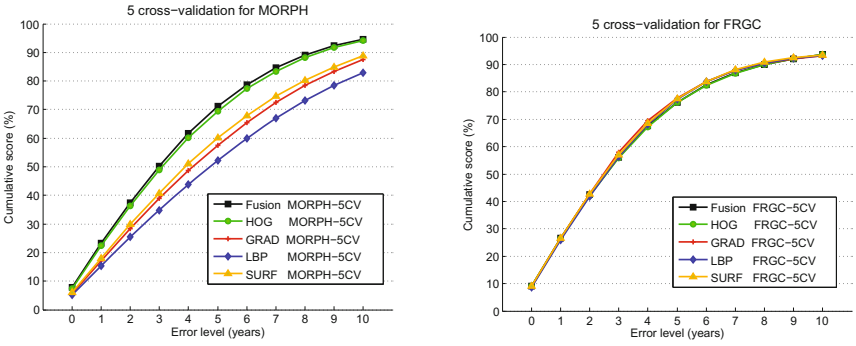
Table 5. Results for non-regularized CCA ($\gamma = 0$) and for CCA with the regularization cost γ^* yielding the best MAE, for each descriptor

	$HOG_{15,13}$	GRAD	$LBP^{u2\times3}_{16,3}$	$SURF^{\times3}_{128}$	BIF [11]	Fusion
(Size)	(2925)	(2500)	(729)	(1920)	(4376)	(4614)
MAE ($\gamma = 0$)	4.38	5.58	6.13	5.30	5.37	**4.27**
MAE (best γ^*)	4.34	5.49	6.13	5.29	4.42	**4.25**
	(γ^*=0.001)	(γ^*=0.002)	($\gamma^*\to0$)	($\gamma^*\to0$)	(γ^*=0.05)	($\gamma^*\to0$)

tion and provides the lowest MAE for MORPH in the literature [11]. The size of BIF after dimensionality reduction (4376) is very similar to the proposed fusion without any further processing (4614). Nonetheless, our proposed fusion of local

Table 6. MAE and CS(5) scores for MORPH and FRGC. Each descriptor has optimal parameters.

	MAE					CS(5)				
	HOG	GRAD	LBP	SURF	Fusion	HOG	GRAD	LBP	SURF	Fusion
MORPH–5CV	4.34	5.49	6.13	5.29	**4.25**	69.5%	57.6%	52.1%	60.2%	**71.2%**
FRGC–5CV	4.19	4.38	4.45	4.44	**4.17**	76.0%	**77.9%**	77.4%	77.5%	76.2%



(a) (b)

Fig. 6. 5-fold cross-validation (5CV) Cumulative Score curves of the Feature descriptor techniques evaluated in: (a) MORPH and (b) FRGC databases

descriptors improves over the best registered result in this database, reducing it from 4.42 down to 4.25. It is noteworthy to see how differently regularization contributes to each descriptor. For instance, it does not affect LBP, but it improves BIF by 18%.

Finally, these results have been obtained for FRGC as well. Table 6 contains global MAE errors and CS(5) values for MORPH and FRGC, whereas Figure 6 shows the complete cumulative score curves for error levels between 0 and 10. From Figure 6(a) it can be seen that for the MORPH database, the fusion of descriptors consistently improves over individual features, even for their optimal configuration of parameters and regularization. On the other hand, the FRGC curves are practically identical. As stated at the beginning of this section, this may be due to the lack of variability in the images of this database, in which every individual averages 80 images, and all very alike. In terms of MAE, the fusion of descriptors always obtains the best score.

5 Conclusions

We have provided a thorough evaluation on the effectiveness of local invariant descriptors, both individually and combined, towards the automatic estimation of apparent age from facial images, using a standard classification technique.

In our experiments, the early fusion of HOG, LBP and SURF descriptors over eye-aligned images provides state-of-the-art results over two large databases, MORPH and FGRC. Concretely, the proposed fusion of descriptors at 50×50 pixel images improves over the best MAE score reported using the CCA technique, resulting in 4.25 years compared to the 4.38 of BIF at 60×60 pixels. Our experiments also show that this distance can be further increased when using larger images and a single HOG descriptor (MAE 4.16).

Our approach requires few feature tuning; it does not involve statistical face models requiring precise annotation of tens of facial landmarks; and it does not require additional cues. We have explored the robustness of the descriptors in terms of parameter settings and in the presence and lack of regularization. Finally, we have demonstrated that local appearance information is sufficient for capturing age information from faces, although it is further improved with textural cues. Canonical Correlation Analysis has proved to be a very effective and efficient technique for age estimation, working consistently for an ample variety of descriptors.

Acknowledgments. This work has been partially supported by the Spanish Ministry of Science and Innovation (MICINN) through the Torres-Quevedo funding program (PTQ-11-04401).

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
2. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: CVPR, pp. 585–592. IEEE (2011)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR, vol. 1, pp. 886–893. IEEE (2005)
4. Fernández, C., Huerta, I., Prati, A.: A comparative evaluation of regression learning algorithms for facial age estimation. In: FFER in Conjunction with ICPR. IEEE (in press, 2014)
5. Fu, Y., Guo, G., Huang, T.: Age synthesis and estimation via faces: A survey. TPAMI **32**(11), 1955–1976 (2010)
6. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 256–263. IEEE (2009)
7. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. TPAMI **35**, 2401–2412 (2013)
8. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. TPAMI **29**(12), 2234–2240 (2007)
9. Guo, G., Mu, G.: Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In: CVPR, pp. 657–664. IEEE (2011)
10. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: CCA vs. PLS. In: 10th Int. Conf. on Automatic Face and Gesture Recognition. IEEE (2013)

11. Guo, G., Mu, G.: A framework for joint estimation of age, gender and ethnicity on a large database. Image and Vision Computing (2014)
12. Han, H., Otto, C., Jain, A.K.: Age estimation from face images: human vs. machine performance. In: International Conference on Biometrics (ICB). IEEE (2013)
13. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. TSMC-B **34**(1), 621–628 (2004)
14. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. TPAMI **24**(4), 442–455 (2002)
15. Minear, M., Park, D.C.: A lifespan database of adult facial stimuli. Behavior Research Methods, Instruments, & Computers **36**(4), 630–633 (2004)
16. Montillo, A., Ling, H.: Age regression from faces using random forests. In: ICIP, pp. 2465–2468. IEEE (2009)
17. Mu, G., Guo, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: CVPR, pp. 112–119. IEEE (2009)
18. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (2002)
19. Oro, D., Fernández, C., Saeta, J.R., Martorell, X., Hernando, J.: Real-time GPU-based face detection in HD video sequences. In: ICCV Workshops, pp. 530–537 (2011)
20. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: CVPR, pp. 947–954. IEEE (2005)
21. Ricanek, K., Tesafaye, T.: MORPH: a longitudinal image database of normal adult age-progression. In: Automatic Face and Gesture Recognition, pp. 341–345 (2006)
22. Riesenhuber, M., Poggio, T.: Hierarchical models of object recognition in cortex. Nature Neuroscience **2**(11), 1019–1025 (1999)
23. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: CVPR, pp. 3476–3483. IEEE (2013)
24. Weng, R., Lu, J., Yang, G., Tan, Y.P.: Multi-feature ordinal ranking for facial age estimation. In: AFGR. IEEE (2013)

Privacy of Facial Soft Biometrics: Suppressing Gender But Retaining Identity

Asem Othman(✉) and Arun Ross

Michigan State University, East Lansing, MI, USA
{aothman,rossarun}@cse.msu.edu

Abstract. We consider the problem of perturbing a face image in such a way that it cannot be used to ascertain soft biometric attributes such as age, gender and race, but can be used for automatic face recognition. Such an exercise is useful for extending different levels of privacy to a face image in a central database. In this work, we focus on masking the gender information in a face image with respect to an automated gender estimation scheme, while retaining its ability to be used by a face matcher. To facilitate this privacy-enhancing technique, the input face image is combined with another face image via a morphing scheme resulting in a mixed image. The mixing process can be used to progressively modify the input image such that its gender information is progressively suppressed; however, the modified images can still be used for recognition purposes if necessary. Preliminary experiments on the MUCT database suggest the potential of the scheme in imparting "differential privacy" to face images.

1 Introduction

Most operational face recognition systems store the original face image of a subject in the database along with the extracted feature set (template). Storing the original image would allow the system to extract new feature sets and recompute templates if the feature extractor and matching modules are changed. However, face images offer additional information about an individual which can be automatically deduced. For instance, it has been shown that *automated* schemes can be used to extract soft biometric attributes such as age [4], gender [10], and race [8] from a face image. This can be viewed as privacy leakage since an entity can learn additional information about a person (or population) from the stored data, without receiving authorization from the person for such a disclosure. Therefore, it is necessary to ensure that face images stored in a system are used *only* for the intended purpose and not for purposes that may result in a "function creep" [21].

In this work, we investigate the possibility of suppressing the soft biometric attribute of a face (e.g., gender) while simultaneously preserving the ability of the face matcher to recognize the individual (see Figure 1). Such a capability will ensure that the stored biometric data is not used for purposes beyond what was expressed during the time of data collection. However, at the same time, it is necessary that any such perturbation does not drastically impact the recognition accuracy of the automated face matcher.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 682–696, 2015.
DOI: 10.1007/978-3-319-16181-5_52

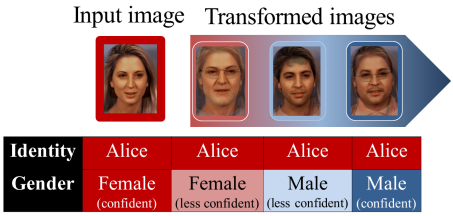


Fig. 1. An illustration of progressively suppressing the gender of an input image while retaining identity with respect to an automated face matcher.

In this paper, a mixing approach is used to transform an input face image into a look-alike face image (i.e., similar facial features and may be appearance) while suppressing a specific soft biometric attribute, viz., gender. The **degree of suppression** is assessed by an *automated* gender classifier since the goal of this work is to disallow automated algorithms from extracting information beyond what was intended at the time of data collection[1]. A typical gender classifier outputs a label indicating the gender of a face image along with a confidence value of this determination. If the proposed method is successful, then either the confidence values associated with the transformed images will decrease (i.e., gender suppression) or their output label will change (i.e., gender conversion).

1.1 Related Work

Biometrics Privacy: In the context of biometrics, privacy refers to the assurance that the biometric data collected from an individual is not used to deduce any type of information about the individual. i.e., it should be used only for matching purposes.

Extensive work on preserving personal information has been done in the data mining community [1]. Their goal is to enable researchers and organizations to learn statistical properties of an underlying population[2] as a whole, while protecting sensitive information of the individuals in the population against "linkage attacks" [12]. Approaches such as k-anonymity [19], l-diversity [9], t-closeness [7], and differential privacy [2] have been proposed to preserve privacy of the personal data in statistical databases. These approaches employ techniques such as data perturbation and sub-sampling [19] (i.e., non-interactive privacy model), or provide an interface through which users may query about the data and get possibly noisy answers (i.e., interactive privacy model) [2]. In the context of biometrics, Newton et al. [13] and Gross et al. [5] introduced a face anonymization algorithm that minimized the chances of performing automatic face recognition on surveillance images while preserving details of the face such as expression,

[1] It is also possible to suppress soft biometric information from a human vision perspective - however, the work here does not explore the cognitive-psychological aspects of the transformed image.

[2] This population can be represented as statistical database.

gender and age. However, the identities of original face images are irrevocably lost, thereby undermining the use of such techniques in biometric databases.

In related literature [16], to protect the privacy of stored biometric data (e.g., face images) in a central database, template protection approaches have been proposed. Most of these template protection approaches replace the stored feature set in the central database with a transformed feature set or a cryptographic key that has been generated from the feature set or bound with it. These approaches, such as fuzzy vault cryptosystems, invariably result in loss of matching accuracy as demonstrated in the literature [16]. Further, when the feature extraction scheme is changed, the cryptosystem has to be changed. Some researchers have addressed the challenge of protecting biometric data at the image-level [15][22][6][3][14], *but their goal was to perturb identity.* Our methodology, on the other hand, only perturbs soft biometric attributes at the image level, while retaining identity, to prevent any gender profiling on an individual[3].

Face fusion for gender conversion: Fusing face images in order to change a perceived soft biometric (such as age, gender, and/or race) has been researched in both computer vision and graphics literature due to its many interesting applications. Regarding gender conversion while preserving face identity, there are two methods: a prototype-based approach [17] and a component-based approach [18]. In the prototype-based approach [17], prototypes for the two gender groups (male and female) are computed to describe the typical characteristics of males and females, respectively, and the difference between these two prototypes is used to modify the gender appearance of an input face image. A component-based approach was proposed by Suo et al. [18] as an alternative approach to gender conversion. Their approach starts by decomposing a source face image into several facial components. Next, these facial components are replaced with templates taken from the opposite gender group and the resulting mosaic is assembled using seamless image editing techniques. The identity of the source image is preserved by selecting replacement templates that are similar to that of the source components and penalizing large alterations in the image editing step. However, the goal of the gender conversion approaches described above is to generate a *single* face image that preserves the identity but modifies the gender. In this paper, our main objective is different. The input face image has to be transformed to *multiple* images that are similar to it but with the gender information suppressed at different levels. In other words, some of the generated images will be perceived to be of the same gender but with less confidence values, while other images will be perceived to be of the opposite gender with different confidence values.

1.2 Proposed Method

To generate a face image with aforementioned properties, the principle of face morphing is used. Consider two face images F_1 and F_2. The morphing algorithm generates intermediate images along the continuum from F_1 to F_2, and

[3] Population privacy is enhanced because an adversary cannot draw any conclusion about the gender of face database users.

their positions on this continuum are specified by the morphing parameters. The parameters, described later, are used to determine the rate of warping and color blending. So, as the morphing proceeds along the continuum from F_1 to F_2, the first image (F_1) is gradually distorted and is faded out, while the second image (F_2) is faded in (see Figure 4).

The key *contributions* of this paper are summarized as follows.

- Progressively suppressing the gender attribute of a face while preserving its identity from the face matcher's perspective. To the best of our knowledge, this paper is the first to present the *potential* of imparting differential privacy to face images via a simple face morphing technique.
- The degree of suppression has been systematically quantified by utilizing an *automated* gender classifier. The proposed method is expected to be applicable across different gender classifiers since it has not been particularly tuned to a specific one.

The rest of the paper is organized as follows. Section 2 discusses the face morphing technique to perturb gender attributes. Section 3 reports the experimental results and Section 4 concludes the paper.

2 Face Morphing

Figure 2 shows the three distinct phases in the generation of a mixed face image (MF) : facial feature extraction, image warping and cross-dissolving.

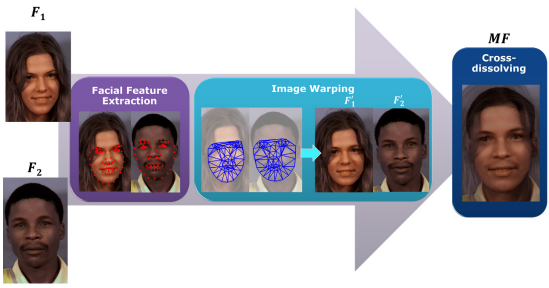


Fig. 2. Proposed approach for suppressing gender while retaining identity. Here, the gender of F_1 is perturbed by mixing it with F_2 resulting in image MF. However, an automated face matcher can successfully match MF with F_1.

2.1 Facial Feature Extraction

Morphing two face images to generate an intermediate face image involves the nontrivial task of locating facial features. For both face images, F_1 and F_2, the prominent facial features are characterized by a pre-defined set of control points.

Both sets of control points, X_1 and X_2, associated with the two face images (see Figure 2), are stored in a vector format. This representation does not include any information about the connection between the control points:

$$X_j = [x_{1j}, x_{2j}, x_{3j}, \ldots, x_{nj}, y_{1j}, y_{2j}, y_{3j}, \ldots y_{nj}]^T, \tag{1}$$

where $j \in \{1, 2\}$ and $n = 76$ is the number of control points. Errors in landmark annotation can cause a ghost-like effect on the subsequently generated image. Since extracting control points automatically is not the focus of this work, a pre-annotated face image database was used (see Section 3). This minimizes the ghost-like effect.

2.2 Image Warping

Once the corresponding control points between the two face images are known, the next step is to perform image warping by mapping each facial feature (e.g., mouth, nose and eyes) in the individual face images to its corresponding feature in the mixed image. A triangulation-based warping scheme is used to deform the face images [20]. First, the intermediate control points set (which defines the shape of the facial features of the mixed face image) is determined. From the control point sets X_1 and X_2 of the face images F_1 and F_2, respectively, the intermediate control point set (X_m) is obtained by linear interpolation as follows:

$$X_m = (1 - \alpha) \cdot X_1 + \alpha \cdot X_2, \tag{2}$$

where $\alpha \in [0, 1]$ is the **warping factor** that determines how the individual shapes of the two face images are integrated into the shape of the mixed face. Next, the face region of each face image is dissected into a suitable set of triangles by utilizing the control points as the vertices of the triangles. Generating an optimal triangulation has to be guaranteed in order to avoid skinny triangles and, therefore, Delaunay triangulation was utilized to construct a triangular mesh for each face image. An example of face images tessellated into triangular regions according to the annotated control points is shown in Figure 2.

Finally, the affine transformation that relates each triangular region in the original face image (F_1 or F_2) to the corresponding triangle in the intermediate image is computed. Suppose that $T_1 = [P_1, P_2, P_3]^T$ ($T_2 = [R_1, R_2, R_3]^T$) is a triangular region in X_1 (X_2) and $T_m = [Q_1, Q_2, Q_3]^T$ is the corresponding triangular region in X_m (see Figure 3). A_1 (A_2) is the affine transformation that maps all points in T_1 (T_2) onto T_m.

$$T_m = A_j T_j, \tag{3}$$

where $j \in \{1, 2\}$. Together, T_1's (T_2's) vertices and T_m's vertices are used in (3) to compute the parameters of the affine transformation A_1 (A_2).

As shown in Figure 2, this results in two warped face images F_1' and F_2' such that F_1' and F_2' have similar shapes.

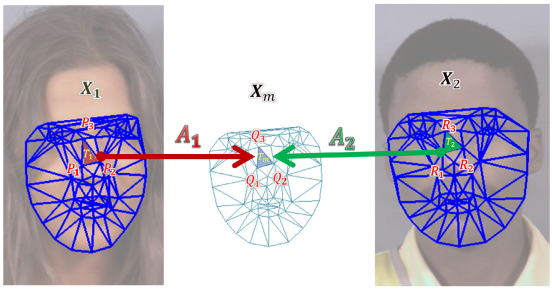


Fig. 3. Generating the corresponding triangle, T_m, in the intermediate image based on the triangles T_1 and T_2 in the original images

2.3 Image Cross-dissolving

The final step to obtain the mixed face image, is simply a cross-dissolving process of the two warped images. If F_1' and F_2' are the warped images, the mixed face image is obtained by linearly interpolating their pixel intensities, such that

$$MF = (1 - \beta) \cdot F_1' + \beta \cdot F_2', \tag{4}$$

where $\beta \in [0, 1]$ is the **color-dissolving factor** that determines the relative influence of the appearance of the two face images on the mixed face image MF. Figure 4 shows different examples of mixed face images along the continuum from F_1 to F_2 by varying the warping factor (α) and the cross-dissolving factor (β).

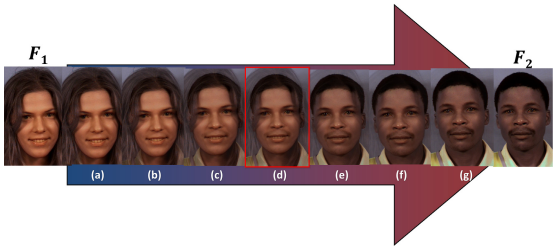


Fig. 4. Mixed face images along the continuum from F_1 to F_2 where $\alpha = \beta =$ (a) 0.1, (b) 0.2, (c) 0.3, (d) 0.5, (e) 0.7, (f) 0.8 and (g) 0.9

3 Experiments and Discussion

The purpose of the following experiments is to systematically investigate if mixing an input face image with different face images from the opposite gender group will (1) suppress the gender attribute of the input face image to different degrees, and (2) preserve the identity of the input image with respect to a face

matcher. To generate a mixed face image from two face images; i.e., a male face image F_m and a female face image F_f, the morphing technique described earlier is utilized and the mixed face image can be anywhere along the continuum from F_m to F_f. But where on this continuous continuum should the mixed face image be? The position of the mixed face on this continuum is specified by the morphing parameters, i.e., α and β. Although the two parameters can be different, the best visually appealing mixed face image along this continuum is observed when $\alpha = \beta$. Also, if $\alpha = \beta < 0.5$, the gender of the source image will not be suppressed effectively. Contrarily, if $\alpha = \beta > 0.5$, the identity of the source will be suppressed and the mixed image will not be similar to it. Thus, we select $\alpha = \beta = 0.5$.

3.1 Performance Metrics

The notion of similarity/dissimilarity between face images is assessed using the match scores generated by a Verilook[4] face matcher. In the context of identification, a higher rank-1 accuracy would imply a higher similarity; in the context of verification, a lower Equal Error Rate (EER) would imply higher similarity. So we use rank-1 accuracy and EER to characterize notions of similarity and dissimilarity.

The gender of a face image (male or female) is assessed using a VeriLook gender classifier [5], which also outputs classification confidence values ($C^{'}$) along with the gender label. These confidence values are in the [0, 100] interval. A confidence value of 0 indicates that the image is in the boundary of the male and female class[6]. However, the software labels the image as female when the confidence value is 0. Here, we mapped the resultant confidence values as follows:

$$C = \begin{cases} C^{'}/100 & \text{if class} = \text{male;} \\ -C^{'}/100 & \text{if class} = \text{female,} \end{cases}$$

where male and female are the labels computed by the gender classifier. This mapping results in a gender axis with two ends: 1 (i.e., male with a confidence value = 100%) and -1 (i.e., female with a confidence value = 100%). This gender axis will be used to quantify as well as visualize the degree to which gender is suppressed in the forthcoming experiments.

3.2 Database and Baseline Performance

The performance of the proposed approach was tested using a dataset from the MUCT database [11]. MUCT database consists of 3755 face images of 276 subjects. We selected the first 2 samples captured by camera "a" (usually the

[4] http://www.neurotechnology.com

[5] Since the proposed method is not particularly tuned to the specific gender classifier used, it is expected to work as well on other types of gender classifiers.

[6] This assertion has been confirmed by consulting the technical support at Neurotechnology.

frontal face) of each subject [7]. For each subject, one sample was added to the probe set and the other sample was added to the gallery set resulting in a probe set P and gallery set G each containing 276 face images. The images in P were matched against those in G. This resulted in a rank-1 accuracy of 95% and an Equal Error Rate (EER) of 3.5%. This dataset was used since the facial landmarks (control points) of individual images were annotated and available online, and also because it contains a comparable number of males and females (i.e., 131 males and 145 females). The ground truth for gender was obtained from the filename ("m" for male and "f" for female). The Verilook gender classifier was used to classify the face images in the gallery set G. Figure 5 shows examples of face images from G along the gender axis based on the predicted gender and confidence values. There are only 5 images from G that were misclassified (see Figure 5). Therefore, from the perspective of this automated gender classifier, the gallery set will be divided into a male dataset G_m consisting of 132 males and a female dataset G_f consisting of 144 females.

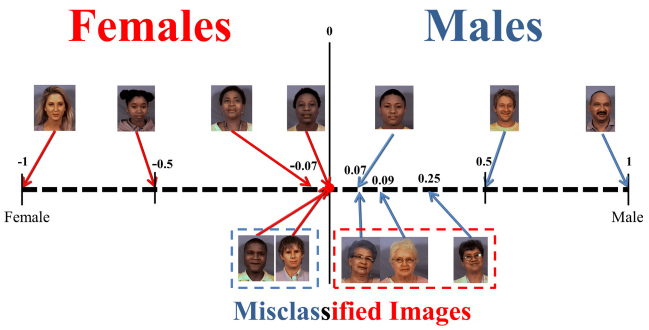


Fig. 5. Examples of frontal face images from the gallery set G shown on the gender axis. The gender axis is based on the gender as estimated by the classifier along with the confidence value, C. Misclassified images are depicted below the axis.

3.3 Degrees of Gender Suppression

In this experiment, the possibility of generating images with different gender suppression levels is tested. Every face image in the male gallery set G_m is mixed with every face image in the female gallery set G_f. This results in 19,008 mixed face images. For every male face image, there are 144 corresponding mixed face images. For every female face image, there are 132 corresponding mixed face images. Figure 6 shows the distribution of confidence values (C) of the mixed

[7] Camera "a" was the only camera that was directly in front of the subject's face. Images captured by other cameras exhibited some pose variations. In this paper, we used only frontal images to examine viability of the proposed approach. There are two or three images per subject captured by camera "a" and we selected two samples in order to have the same number of samples for all subjects.

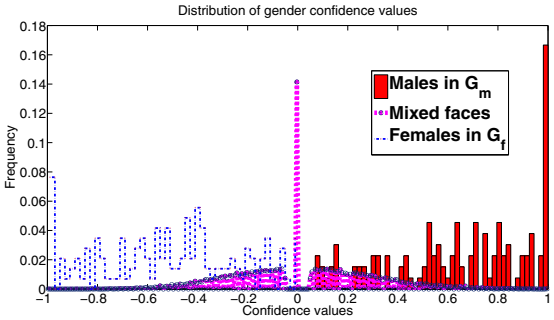


Fig. 6. Distribution of confidence values of the male, female and mixed images.

images as well as the original male and female images. This graph clearly suggests the potential of suppressing gender attributes using the proposed method. However, the objective is not just to suppress the gender attribute. We are looking to generate images that reveal different levels of gender. Another way of looking at this is as follows: if a male face image F_m is fused with different female face images F_fs, the resulting face images (MFs) should have confidence values (C_{mf}s) that vary from -1 to the confidence value of the original male image, C_m. To determine the *degree* of gender suppression, the **male suppression-level** of a mixed image is computed as follows:

$$S_m = \frac{C_m - C_{mf}}{C_m + 1}. \tag{5}$$

If $S_m = 0$, this indicates that the mixed image has the same confidence value as the original male image (F_m) and the gender is not suppressed. If $S_m = 1$, this indicates that the mixed image has been classified as female with $C_{mf} = -1$ and the gender of the original male image is completely suppressed. On the other hand, if the source is a female image which has been mixed with a set of male images, the goal would be to generate mixed images with suppression-levels that start from $C_{mf} = C_f$ and end at $C_{mf} = 1$. Therefore, if the input image is female, the **female suppression-level** can be computed as follows:

$$S_f = \frac{|C_f| + C_{mf}}{|C_f| + 1}. \tag{6}$$

Note that in this particular database S_m and $S_f \in [0, 1]$ because, as shown in Figure 6, the confidence value of a mixed face image (C_{mf}) is always less than the confidence values of the original male subjects (C_m) and greater than the confidence values of the original female subjects (C_f), i.e., $C_f \leq C_{mf} \leq C_m$. Figures 7 and 8 show S_m and S_f, respectively, for all mixed images (i.e., the 19,008 face images). These graphs suggest the possibility of having different levels of gender suppression. This can be observed by viewing the range of colors in each row or each column.

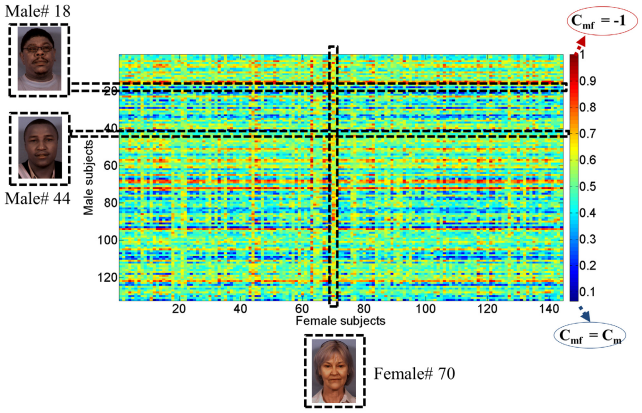


Fig. 7. Plot of male suppression-levels of the mixed images where points along the horizontal axis is the id # of female subjects and points along the vertical axis is the id # of male subjects. Male suppression-levels after mixing male subjects #18 and #44 are highlighted. Additionally, the male-suppression level when female subject #70 is used is also highlighted. See text for explanation as to why these three subjects are highlighted.

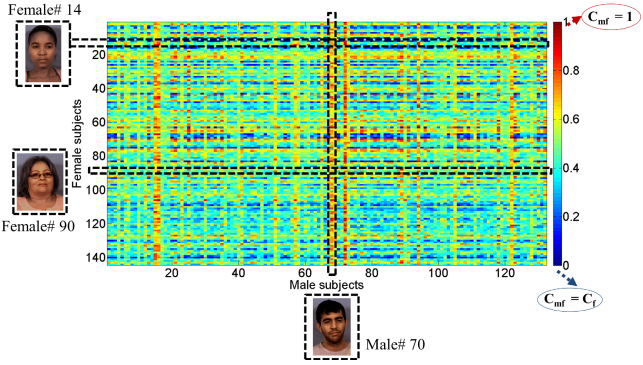


Fig. 8. Plot of female suppression-levels of the mixed images where points along the horizontal axis is the id # of male subjects and points along the vertical axis is the id # of female subjects. Female suppression-levels after mixing female subjects #14 and #90 are highlighted. Additionally, the female-suppression level when male subject #70 is used is also highlighted. See text for explanation as to why these three subjects are highlighted.

Figure 9 show two male images and the results of mixing them with different females images along with gender confidence values. Figure 11(a) show S_m for all mixed images generated by these two male subjects. Note that, the male suppression-levels (S_m) of mixed images generated by male subject #18 tend to be closer to the original gender confidence value (i.e., S_m tends to be closer to 0). On the other hand, the mixed images of male subject #44 tend to be

classified as females and are closer to the target (i.e., $C_{mf} \simeq -1$). A similar effect on female subject #18 and #44, can be seen in Figure 10. Figure 11(b) shows the female suppression-levels (S_f) for all mixed images generated by these two female subjects. We also observed that some female images when mixed with input male images cause most of the mixed images to have male suppression-levels that are closer to 1 (i.e., C_{mf} are closer to -1). For example, as shown in Figures 7 and 9, female subject #70 strongly suppressed the gender of most male images. Similarly, as shown in Figures 8 and 10, male subject #70 has strongly affected most of the female suppression-levels of the mixed face images.

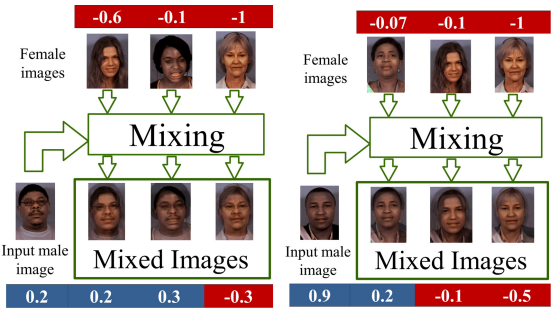


Fig. 9. Examples of mixed face images after mixing the face images of male subjects (a)# 18 and (b)#44 with different female face images, along with the confidence value (C) of each image. The blue (red) color indicates that the image is labeled as a male (female).

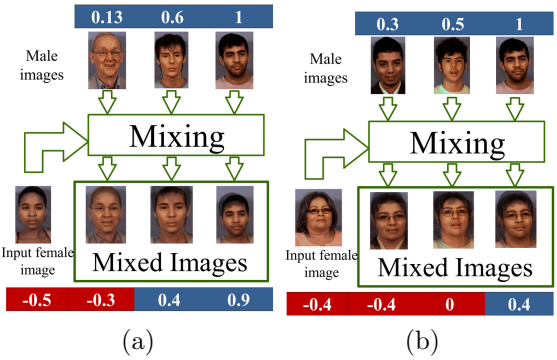
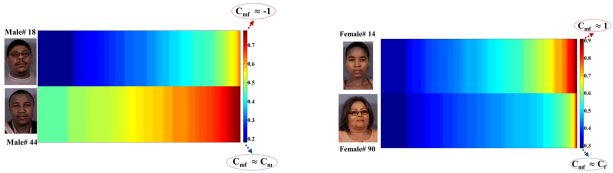


Fig. 10. Examples of mixed face images after mixing the face images of female subjects (a) # 14 and (b) #90 with different male face images, along with the confidence value (C) of each image

These results prove that we can generate different versions of an input face image with different levels of gender suppression. Note that the confidence values associated with the original images play an important role. As shown in



(a) Male suppression-levels (b) Female suppression-levels

Fig. 11. Plot of sorted suppression-levels corresponding to the mixed images generated for (a) male subjects #18 & #44, and (b) female subjects #14 & #90

Figure 9(a), the facial appearance of the input face image has also a role. Hence, the facial hair of the male subject #18, i.e., the mustache, results in mixed images with more maleness, although the original male image has a low gender confidence value.

3.4 Similarity to the Original Face Images

After suppressing or modifying the gender of the face image by mixing, the identity information should be preserved effectively in the resultant images. Therefore, in this experiment, the similarity between the mixed face image and original face images (i.e., male and female face images) was evaluated. To this end, the mixed face images generated in Experiment 1 (i.e., 19,008 face images) were matched against the original images in the probe set P (see Figure 12). Here, a genuine score is generated when the mixed face image is matched with either of the original face images (i.e., the images that were mixed) and the rest are impostor scores.

The resultant rank-1 accuracy of matching mixed images against original images in P was 95% (and the EER was 5%). These results indicate that the original images are reasonably similar to the mixed images. Hence, the identities of the originals have been preserved in the mixed faces, which is our objective.

4 Summary and Future Work

In this work, we explored the possibility of generating mixed face images that suppress the gender of a face image to *different degrees*. In this regard, we mixed an input face image with different face images from the opposite gender and determined if the mixed images suppress the gender while bearing sufficient similarity to the input face image in terms of a face matcher. We utilized a gender classifier along with the resultant confidence value to assess the gender information. To mix two face images, a face morphing technique was adopted in this work. Experiments on the MUCT dataset indicate that (a) the mixed face suppresses the gender of original face images to different degrees, and (b) the mixed face exhibits similarity with the original image and so identity with respect to a face matcher is retained. Figure 13 shows that the distribution of male and female suppression-levels are not uniform distributions. While it is

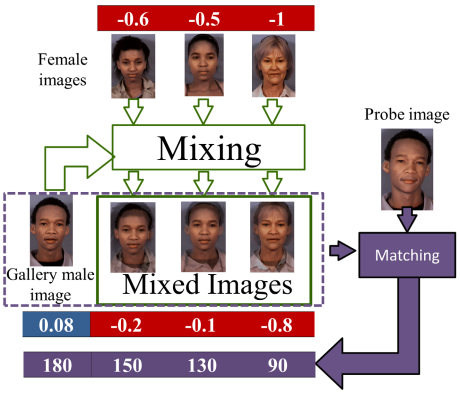


Fig. 12. Examples of mixing the face images of a male subject with different female face images, along with the confidence value (C) of each image. Match scores generated by matching the input probe against the mixed images and the gallery image of the male subject are shown below the confidence values of the gallery and mixed images. These scores are similarity scores and in the [0,180] interval.

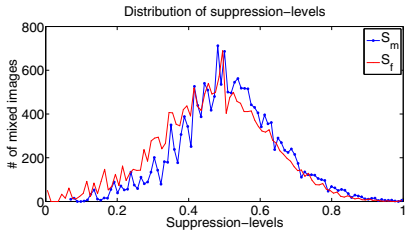


Fig. 13. Distributions of male (S_m) and female (S_f) suppression-levels of all mixed images (i.e., 19,800 face images). These distributions suggest that it is possible to suppress gender to different degrees.

possible to suppress a face image to different degrees, these degrees may not form a continuous or complete continuum[8]. Therefore, further work is needed to test this approach on a larger database having subjects with large variation in their gender confidence values. Other morphing approaches based on radial basis functions and multi-level free-form deformation [20] could be explored. The technique could potentially be extended to suppress different soft biometric attributes simultaneously (to different degrees) thereby supporting a differential privacy framework. Note that mixing more than two images is possible, but this may suppress individual identities and the mixed image is likely to be less similar to the originals. Future work will also investigate the possibility of utilizing the proposed approach as a privacy-enhancing technique by mixing faces of different subjects in order to *hide* the original identities.

[8] This continuum should start from the gender confidence value of the input face image and end at the maximum confidence value of the opposite gender (i.e., +1 or -1).

Acknowledgments. The authors are grateful to Cunjian Chen for his assistance with the gender prediction experiments.

References

1. Agrawal, R., Srikant, R.: Privacy-preserving data mining. ACM Sigmod. Record **29**(2), 439–450 (2000)
2. Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
3. Färberböck, P., Hämmerle-Uhl, J., Kaaser, D., Pschernig, E., Uhl, A.: Transforming rectangular and polar iris images to enable cancelable biometrics. In: Image Analysis and Recognition, pp. 276–286. Springer (2010)
4. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(11), 1955–1976 (2010)
5. Gross, R., Sweeney, L., De la Torre, F., Baker, S.: Model-based face de-identification. In: Computer Vision and Pattern Recognition Workshop (CVPRW), pp. 161–168. IEEE Computer Society, Los Alamitos (2006)
6. Hämmerle-Uhl, J., Pschernig, E., Uhl, A.: Cancelable Iris Biometrics Using Block Re-mapping and Image Warping. In: Samarati, P., Yung, M., Martinelli, F., Ardagna, C.A. (eds.) ISC 2009. LNCS, vol. 5735, pp. 135–142. Springer, Heidelberg (2009)
7. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: IEEE 23rd International Conference on Data Engineering (ICDE), vol. 7, pp. 106–115 (2007)
8. Lu, X., Jain, A.K.: Ethnicity identification from face images. In: SPIE Defense and Security Symposium, pp. 114–123 (2004)
9. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. In: ACM Transactions on Knowledge Discovery from Data (TKDD) (2007)
10. Makinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(3), 541–547 (2008)
11. Milborrow, S., Morkel, J., Nicolls, F.: The MUCT Landmarked Face Database. Pattern Recognition Association of South Africa (2010). http://www.milbo.org/muct
12. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy, pp. 111–125 (2008)
13. Newton, E., Sweeney, L., Malin, B.: Preserving privacy by de-identifying face images. IEEE Transactions on Knowledge and Data Engineering **17**, 232–243 (2005)
14. Othman, A., Ross, A.: On mixing fingerprints. IEEE Transactions on Information Forensics and Security **8**(1), 260–267 (2013)
15. Ratha, N., Connell, J., Bolle, R.: Enhancing security and privacy in biometrics-based authentication systems. IBM Systems Journal **40**(3), 614–634 (2001)
16. Rathgeb, C., Uhl, A.: A survey on biometric cryptosystems and cancelable biometrics. EURASIP Journal on Information Security **1**, 1–25 (2011)
17. Rowland, D., Perrett, D.: Manipulating facial appearance through shape and color. Computer Graphics and Applications **15**(5), 70–76 (1995)

18. Suo, J., Lin, L., Shan, S., Chen, X., Gao, W.: High-resolution face fusion for gender conversion. In: IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, pp. 1–12 (2011)
19. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **10**(05), 557–570 (2002)
20. Wolberg, G.: Image morphing: a survey. The Visual Computer **14**(8), 360–372 (1998)
21. Woodward, J., Orlans, N., Higgins, P.: Biometrics:identity assurance in the information age. McGraw-Hill/Osborne, New York (2003)
22. Zuo, J., Ratha, N., Connell, J.: Cancelable iris biometric. In: IEEE 19th International Conference on Pattern Recognition (ICPR), pp. 1–4 (2008)

Exploring the Magnitude of Human Sexual Dimorphism in 3D Face Gender Classification

Baiqiang Xia[2,3(✉)], Boulbaba Ben Amor[1,3], and Mohamed Daoudi[1,3]

[1] Institut Mines-Télécom/Télécom Lille, Lille, France
[2] University of Lille1, Villeneuve-d'ascq, France
[3] CRIStAL (UMR CNRS 9189), Lille, France
xia.baiqiang@telecom-lille.fr

Abstract. Human faces demonstrate clear Sexual Dimorphism (SD) for recognizing the gender. Different faces, even of the same gender, convey different magnitude of sexual dimorphism. However, in gender classification, gender has been interpreted discretely as either male or female. The exact magnitude of the sexual dimorphism in each gender is ignored. In this paper, we propose to evaluate the SD magnitude, using the ratio of votes from the Random Forest algorithm performed on 3D geometric features related to the face morphology. Then, faces are separated into a *Low-SD* group and a *High-SD* group. In the *Intra-group* experiments, when the training is performed with scans of similar SD magnitude than the testing scan, the classification accuracy improves. In *Inter-group* experiments, the scans with low magnitude of SD demonstrate higher gender discrimination power than the ones with high SD magnitude. With a *decision-level* fusion method, our method achieves 97.46 % gender classification rate on the 466 earliest 3D scans of FRGCv2 (mainly neutral), and 97.18 % on the whole FRGCv2 dataset (with expressions).

Keywords: 3D face · Gender classification · Sexual dimorphism · Random forest

1 Introduction

Human faces exhibit clear sexual dimorphism (SD), in terms of masculinity and femininity [29], for recognizing their gender. In several anthropometry studies, researchers have concluded that male faces usually possess more prominent features than female faces [6,18,36,41]. Automated gender classification has gradually developed as an active research area, sine 1990s. Abundant works have been published, concerning (i) different face modalities (2D texture images or 3D scans), (ii) different face descriptions (2D pixels, 3D point cloud, or more complex features like LBP, AAM, wavelets, etc.), and (iii) different classifiers (Random Forest, SVM, Adaboost, etc.). Earlier gender classification works mainly focused on 2D texture of faces. Recently, 3D face-based gender classification has been investigated in several studies [11,12,25,33,37,38] and has demonstrated its benefits compared to 2D face-based approaches. Compared to 2D face images,

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 697–710, 2015.
DOI: 10.1007/978-3-319-16181-5_53

the 3D scans have better robustness to illumination and pose changes. Also, the 3D face scans are able to capture complete information of the facial shape.

The first work of 3D-based gender classification is proposed by *Liu and Palmer* in [25]. Considering the human faces are approximately symmetric, they extract features from the height and orientation differences on symmetric facial points from 101 full 3D face models. Using LDA outputs, they achieve 91.16% and 96.22% gender recognition rate considering the height and orientation differences, respectively. In [12], *Vignali et al.* use the 3D coordinates of 436 face landmark points as features, and achieve 95% classification rate on 120 3D scans with LDA classifier. Considering the statistical differences shown in facial features between male and female, such as in the hairline, forehead, eyebrows, eyes, cheeks, nose, mouth, chin, jaw, neck, skin, beard regions [10], *Han et al.* extract geometric features with the volume and area information of faces regions [13]. They achieve 82.56% classification rate on 61 frontal 3D face meshes of GavabDB database, with the RBF-SVM classifier in 5-fold cross validation. In [15], *Hu et al.* divide each face into four regions in feature extraction, and find that the upper face is the most discriminating for gender. The classification rate reported is 94.3% with SVM classifier, on 945 3D neutral face scans. In [2], *Ballihi et al.* extract radial and iso-level curves and use a Riemannian shape analysis approach to compute lengths of geodesics between facial curves from a given face to the Male and Female templates computed using the Karcher Mean algorithm. With a selected subset of facial curves, they obtain 86.05% gender classification rate with Adaboost, in 10-fold cross-validation on the 466 earliest scans of FRGCv2 dataset. In [33], *Toderici et al.* obtain features with the wavelets and the MDS (Multi Dimensional Scaling). Experiments are carried out on the FRGCv2 dataset in 10-fold subject-independent cross validation. With polynomial kernel SVM, they achieved 93% gender classification rate with the unsupervised MDS approach, and 94% classification rate with the wavelets-based approach. In [11], *Gilani et al.* automatically detect the biologically significant facial landmarks and calculate the euclidean and geodesic distances between them as facial features. The minimal-Redundancy-Maximal-Relevance (mRMR) algorithm is performed for feature selection. In a 10-fold cross-validation with a LDA classifier, they achieve 96.12% gender classification rate on the FRGCv2 dataset, and 97.05% on the 466 earliest scans. Combining shape and texture, in [26], *Lu et al.* fuse the posterior probabilities generated from range and intensity images using SVM. In [19], *Huynh et al.* fuse the Gradient-LBP from range image and the Uniform LBP features from the gray image. In [34], *Wang et al.* fuse the results of gender classification from 8 facial regions, and achieve 93.7% gender classification rate on the FRGCv2 dataset, using both the range and texture data. In [35], *Wu et al.* combine shape and texture implicitly with needle maps recovered from intensity images. More recently, in [17], *Huang et al.* propose to use both boosted local texture and shape features for gender and ethnicity classification. The local circular patterns (LCP) are used to compute the facial features. Then a boosting strategy is employed to highlight the most discriminating gender- and race-related features. Their method achieve 95.5% correct gender classification rate on a subset of FRGCv2 dataset using 10-fold cross validation.

In subjective experiments conducted in [16], the authors reported that human observers perform better on gender recognition with 3D scans than with 2D images. The study presented in [14] also confirms that, for gender classification, the usage of 2D images is limited to frontal views, while the 3D scans are adaptable to non-frontal facial poses. Despite the achievements of 3D face-based gender classification, the majority of related works have interpreted gender discretely as either male or female. To our knowledge, no work gives consideration to the fact that, even faces with the same gender can have different magnitude of sexual dimorphism. Thus, other than viewing faces equally as definitely male or female, we propose to evaluate the magnitude of sexual dimorphism first, and then explore the usage of this magnitude for gender classification. The rest of this paper is organized as following: in section 2, we present the adopted feature extraction procedure for 3D faces and emphasize our contribution; in section 3, we detail our gender classification method, including the Random Forest classifier and the evaluation protocols; experimental results and discussions are in section 4; section 5 makes the conclusions.

2 Methodology and Contributions

In face perception, researchers have revealed that facial sexual dimorphism relates closely with the anthropometric cues, such as the facial distinctiveness (the converse to **averageness**) [3], and the bilateral **asymmetry** [23]. The averageness of the face and its symmetry serve as covariants in judging the perceived health of potential mates in sexual selection [24,28,30], and also the attractiveness of face [20,22]. As stated earlier, the male faces usuallye Statistics on head and face of American and Chinese adults reported in [9,39,40] have also confirmed this point. Concerning the face symmetry, in [23], *Little et al.* reveal that the symmetry and sexual dimorphism from faces are related in humans, and suggest that they are biologically linked during face development. In [32], *Steven et al.* find that the masculinization of the face significantly covaries with the fluctuating asymmetry in men's face and body. In addition to facial averageness and symmetry, the global **spatiality** and local **gradient** relates closely with sexual dimorphism in the face. As demonstrated in [39,40], the sexual dimorphism exhibits inequally in magnitude in different spatial parts of the face. Also, sexual dimprphism demonstrates the developmental stability [24] in faces. It relates closely to the shape gradient which represents the local face consistency. Thus, considering sexual dimorphism is closely related to these morphological cues of the face, we explore the use of four descriptions based on the recently-developed Dense Scalar Field (DSF) [4,8]. Our facial descriptions quantify and reflect the averageness (*3D-avg.*), the bilateral symmetry (*3D-sym.*), the local changes (*3D-grad.*), and the global structural changes (*3D-spat.*) of the 3D face. Recall that the extraction of the DSFs features is based on a Riemannian shape analysis of elastic radial facial curves. They are designed to capture densely the shape deformation, in a vertex-level. They were first proposed in [4,8] for facial expression recognition from dynamic facial sequences. The extraction of the proposed descriptors is detailed in the following.

2.1 Feature Extraction Methodology

At first, the 3D scans are pre-processed to define the facial surface (remove noise, fill holes, etc.) and limit the facial region (remove hair, etc.). Then, a set of radial curves stemming from the automatically detected nose-tip of each scan are extracted with equal angular interpolation, $\alpha \in [0, 2\pi]$. The radial curve that makes an clockwise angle of α with the middle-up radial curve which passes through the center of the nose and the forehead is denoted as β_α. Given a facial surface S, it results in $S \approx \cup_\alpha \beta_\alpha$. Then, with the *Square-Root Velocity Function* (SRVF) representation introduced in [31], an elastic shape analysis of these curves is performed.

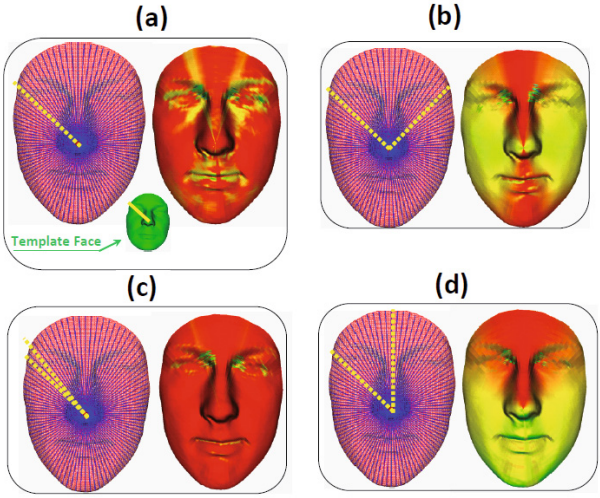


Fig. 1: Illustrations of different features on 3D shape of the preprocessed face. (a) the ***3D-avg.*** description: the DSF from radial curve in a preprocessed face to the radial curve in face template with the same index; (b) the ***3D-sym.*** description: the DSF from symmetrical radial curves. (c) the ***3D-grad.*** description: DSF from a pair of neighboring radial curves. (d) the ***3D-spat.*** description: DSF from each radial curve to the middle-up radial curve which passes through the middle of the nose and the forehead.

More formally, considering a given facial curve as a continuous parametrized function $\beta(t) \in \mathbb{R}^3$, $t \in [0, 1]$. β is represented by its *Square-Root Velocity Function* SRVF, q, according to :

$$q(t) = \dot{\beta}(t)/\sqrt{\|\dot{\beta}(t)\|}, t \in [0, 1]. \quad (1)$$

Then with its $\mathbb{L}^2$-norm $\|q(t)\|$ scaled to 1, the space of such functions: $\mathcal{C} = \{q : [0, 1] \to \mathbb{R}^3, \|q\| = 1\}$ becomes a Riemannian manifold and has a spherical

structure in the Hilbert space $\mathbb{L}^2([0,1],\mathbb{R}^3)$. Given two curves β_1 and β_2 represented as q_1 and q_2 on the manifold, the geodesic path between q_1, q_2 is given by the minor arc of the great circle connecting them on the Hypersphere [7]. To capture densely the shape deformation between the curves q_1 and q_2, the shooting vector $V_{q_1 \mapsto q_2}$ (tangent vector to $\mathcal{C}$ on q_1, and also an element of the tangent space on q_1 to the manifold $\mathcal{C}$, $T_{q_1}(\mathcal{C})$), is used. This vector represent the shooting direction along the geodesic connecting q_1 and q_2. Knowing that, the covariant derivative of the tangent vector field on geodesic path is equal to 0 (i.e. a geodesic corresponds to a constant velocity path on the manifold), this shooting vector characterizes the geodesic path. Here again, due to the spherical structure of $\mathcal{C}$, the shooting vector $V_{q_1 \mapsto q_2}$ is given by:

$$V_{q_1 \mapsto q_2} = \frac{\theta}{\sin(\theta)}(q_2^* - \cos(\theta)q_1), V_{q_1 \mapsto q_2} \in T_{q_1}(\mathcal{C}). \tag{2}$$

where $q_2^* \in [q_2]$ (element of the orbit of the shape q_2) denote the closest shape in $[q_2]$ to q_1 with respect to the metric $d_{\mathcal{C}}$. The shape q_2^* is given by $q_2^* = \sqrt{\dot{\gamma^*}}O^* q_2(\gamma^*)$, where γ^* is the optimal reparametrization that achieved the best matching between q_1 and q_2 and O^* gives the optimal rotation to align them. The angle $\theta = \cos^{-1}\langle q_1, q_2^* \rangle$, denotes the angle between q_1 and q_2^* (we refer the reader to [7] for further details on elastic shape analysis of facial radial curves). Recall that, shapes correspondence is an essential ingredient in shape analysis to find the efficient way to deform one shape into another. The elastic shape analysis framework used here achieves accurate dense correspondence and returns the optimal deformation from one shape into another.

With the magnitude of $V_{q_1 \mapsto q_2}$ computed at N indexed points of q_1 and q_2, the *Dense Scalar Field* (DSF), $DSF = \{\|V_\alpha^{(k)}\|, k = 1, 2, 3, .., N\})$ is built. It quantifies the shape difference between two curves at each point. Using this geometric deformation between pairwise curves on the face, we derive four different facial descriptions which reflect different morphological cues, as described earlier, which are the face averageness, symmetry, local shape changes (termed gradient) and the global spatial changes (termed spatial). The extracted DSF features are illustrated in Fig. 1. In each sub-figure of Fig. 1, the left part illustrates the extracted radial curves and the curve comparison strategy, the right part shows the DSF features as color-map on the face. On each face point, the warmer the color, the lower the deformation magnitude. The ***3D-avg.*** description shown in Fig. 1 (a) compares a pair of curves with the same angle from a preprocessed face and an template face. The average face template (presented in Fig. 1 (a)) is defined as the middle point of geodesic from a representative male face to a representative female face. The ***3D-sym.*** description shown in Fig. 1 (b) captures densely the deformation between bilateral symmetrical curves. This description allows to study the facial changes in terms of bilateral symmetry. The ***3D-grad.*** description shown in Fig. 1 (c) captures the deformation between pairwise neighboring curves. The idea behind is to approximate a local derivation or the gradient. The ***3D-spat.*** description shown in Fig. 1 (d) captures the deformation of a curve to the middle-up curve, emanating from the

nose tip and passing vertically through the nose and the forehead. As the middle-up curve locates at the most rigid part of the face, this description captures the spatial changes from the most rigid facial part in the face. We emphasize that, although all these descriptions are based on the same mathematical background, they relate to significantly different morphology cues of face shape.

2.2 Contributions

With the designed facial descriptions, we perform gender classification experiments with the Random Forest classifier. We propose to first evaluate the magnitude of sexual dimorphism of a face, then explore this magnitude in gender classification. In summary, we have made the following contributions:

- First, rather than taking a face as definitely male or female, we propose to evaluate its magnitude of sexual dimorphism using the ratio of votes from effective random forest based gender classification approach. To our knowledge, this is the first time in the literature of gender classification that gives consideration to the sexual dimorphism difference.
- Second, according to the magnitude of sexual dimorphism, we separate the instances into *High-SD* and *Low-SD* groups. With the *Intra-group* gender classification experiments performed within each group, we find out that the gender of instances are more accurately classified with the training instances which have similar sexual dimorphism.
- Third, in the *Inter-group* experiments, we demonstrate that the gender classification algorithm trained on *Low-SD* instances has good generalization ability on the *High-SD* instances, while the inverse is not true. It means that the instances with low magnitude of sexual dimorphism tell more accurately the discriminating cues of gender. When training only on the *Low-SD* instances, the classification results are even better than training on all instances.
- Last, with a decision-level fusion method performed on the results from four descriptions, we achieve 97.46% gender classification rate on the 466 earliest 3D scans of FRGCv2, and 97.18% on the whole FRGCv2 dataset.

In the next, we detail the experiments conducted for gender classification.

3 Gender Classification Experiments

We use Random Forest classifier on the DSFs features for Gender classification. The Random Forest is an ensemble learning method that grows many decision trees $t \in \{t_1, .., t_T\}$ for an attribute [5]. In growing of each tree, a number of $\boldsymbol{N}$ instances are sampled randomly with replacement from the data pool. Then a constant number of variables are randomly selected at each node of the tree, with which the node makes further splitting. This process goes on until the new nodes are totally purified in label. To classify a new instance, each tree

gives a prediction and the forest makes the overall decision by majority voting. Thus, associated to each forest decision, there is always a score in the range of [50%,100%], which indicates the ratio of trees voting for this decision. This ratio shows the confidence of the decision.

Our experiments are carried out on the Face Recognition Grand Challenge 2.0 (FRGCv2) dataset [27], which contains 4007 3D near-frontal face scans of 466 subjects. There are 1848 scans of 203 female subjects, and 2159 scans of 265 male subjects. Following the literature [2,11,34,37,38], we conducted two experiments: the *Expression-Dependent* experiment uses the 466 earliest scans of each subject in FRGCv2, for which the majority have neutral expression. The *Expression-Independent* experiments use the whole scans of FRGCv2 dataset, for which about 40% are expressive. These settings are designed to test the effectiveness of gender recognition algorithm, and its robustness against facial expressions. Under the Expression-Dependent setting, we use directly our DSFs features. Whereas, under the Expression-Independent setting, we first reduce the original DSF features into a salient subset using the *Correlation-based Feature Selection* (CFS)algorithm [1], to make feasible the evaluation on the whole dataset. The *CFS* belongs to the *the filters*, which select features with heuristics based on general characteristics of features. They are independent of learning algorithm and are generally much faster [21], compared with another school of feature selection methods, *the wrappers*. After Feature selection, we retain only 200-400 features for each description.

3.1 Leave-One-Person-Out (LOPO) Gender Classification

First, for both the *Expression-Dependent* and the *Expression-Independent* settings, we follow the *Leave-One-Person-Out (LOPO)* cross-validation with 100-tree Random Forest with each type of DSFs descriptions. Under the LOPO protocol, each time one subject is used for testing, with the remaining subjects used for training. No subject appears both in traning and in testing in the same round. The experimental results are reported in Table 1. For each DSF description, under the *Expression-Dependent* setting, the gender classification rate is always $> 84\%$. With the selected DSFs features in the *Expression-Independent* setting, the gender classification rate is always $> 85\%$. Generally speaking, these results demonstrate the relevance of using the 3D geometry (shape) of the face, in particular the morphological cues, for the gender classification task. The best classification rates is achieved by the **3D-avg.** description, which quantifies the shape divergence to a face template.

Despite the effectiveness of the previous descriptions, the gender of the training and testing instances had been interpreted discretely as either male or female during the experiments. Labeling a male or female only represents the dominant tendency of facial sexual dimorphism, which is masculinity or femininity. The varying magnitude of facial sexual dimorphism, especially in the same gender class, is ignored. Different male faces can have different magnitude of masculinity. Also, different female faces can have different magnitude of femininity. Thus, in our work, we propose to evaluate the magnitude of facial sexual dimorphism in

Table 1: Expression-Dependent Gender classification Results

Experiment/Description	**3D-avg.**	**3D-sym.**	**3D-grad.**	**3D-spat.**	# of Scans
Expression-Dependent	89.06%	87.77%	85.62%	84.12%	466
Expression-Independent	91.21%	90.49%	85.29%	87.79%	4005

gender classification, using the ratio of votes from Random Forest in the LOPO experiments. Recall that, in the Random Forest, the overall decision is made by majority voting of its decision trees. The ratio of votes signifies the confidence of the forest decision. In our case, the ratio of votes actually is interpreted as the magnitude of sexual dimorphism, in a statistical way. As in the forest, each tree is acting as an evaluator of the sexual dimorphism in the testing instance. The ratio of votes thus represents the statistical evaluation of the testing instance of the whole forest. Moreover, since the results shown in Table. 1 have demonstrated the effectiveness of the forests in gender classification, which means that the ratio of votes signifies well the sexual dimorphism. This is similar to human based gender evaluation. When a group of people are evaluating the same face, the more people voting for a gender (male or female), the higher the magnitude of sexual dimorphism exhibiting in the evaluated face (masculinity or femininity). The only underlying pre-condition is that, human observers have good accuracy in gender classification. Similarly, the only requirement of using the ratio of votes in random forest as evaluation of sexual dimorphism is that, the trees should be relevant to gender discrimination. This has already been justified by the effective results in Table. 1. Thus, we use the ratio of votes to indicate the magnitude of the dominant sexual dimorphism in a face. For example, if a scan is classified as male (or female) with 70% trees voting for it, we note it as having a magnitude of 0.7 in masculinity (or femininity). In our case, the ratio of votes, which signifies the magnitude of sexual dimorphism, is always in the range of $[0.5, 1.0]$.

In Fig. 2, we show the relationship between gender classification accuracy and the magnitude of sexual dimorphism in testing scans for both *Expression-Dependent* and *Expression-Independent* experiments. The magnitude of sexual dimorphism is divided into five equally spaced groups, and shown in the x-axis of each subplot of Fig. 2. The corresponding recognition rates in each description are shown as color-bars in y-axis. As shown in Fig. 2 (A), under the *Expression-Dependent* setting, the higher the magnitude of sexual dimorphism in testing scans, the higher the gender classification accuracy in each description. When the SD magnitude is as low as [0.5,0.6], the classification rate is only about 60%. The rate increases to 72%-75% when SD magnitude is within (0.6,0.7]. The classification rate reaches 79%-92% when the SD magnitude is (0.7,0.8]. Finally, when the SD magnitude is > 0.8, the accuracy reaches 97%-100%. The results under the *Expression-Independent* setting, shown in Fig. 2 (B), confirms these observations that the higher the SD magnitude in scans, the higher the gender classification accuracy in each description. The gender classification rate is very low when the SD magnitude is < 0.6. When the SD magnitude increases to > 0.8, the classification rate reaches as high as $> 90\%$. The observation from

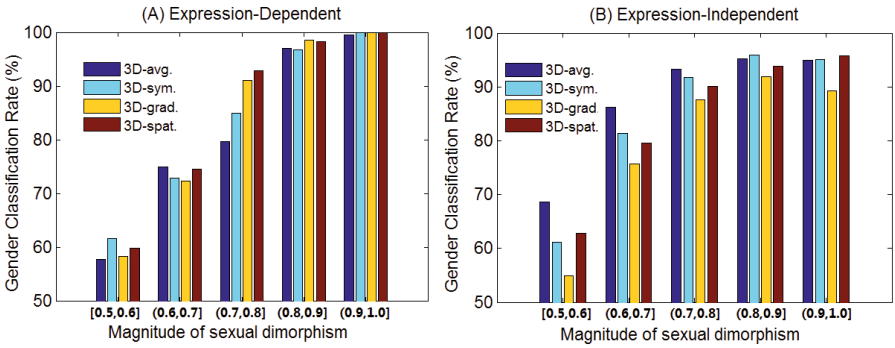


Fig. 2: Gender classification rates with different magnitude of sexual dimorphism

Fig. 2 under both settings matches the simple intuition that the lower the sexual dimorphism in a face, the harder for classifying its gender.

3.2 *Intra-group* and *Inter-group* Experiments

Results from Fig. 2 also show that the gender classification algorithm trained on general scans performs poorly when the testing scans have very low magnitude of sexual dimorphism. Thus, for both the experimental settings, we propose to separate the scans into a *High-SD* group and a *Low-SD* group in each description, according to the magnitude of sexual dimorphism evaluated in the corresponding LOPO experiments with the concerning description. The *High-SD* group comprises of instances with the magnitude of sexual dimorphism evaluated higher than 0.8. The *Low-SD* group is formed by instances with the magnitude of sexual dimorphism evaluated $\leq$ 0.8. With Fig. 2, it is clear that the accuracy differs in the two groups of each description. After this, we design two types of experiments, the *Intra-group* experiments and the *Inter-group* experiments. In the *Intra-group* experiments, we perform LOPO experiments on each group of each description. This type of experiments are designed to reveal that, for a testing scan, whether using scans with similar sexual dimprphism magnitude in training can determine more accurately its gender. In the *Inter-group* experiments, we train and test with different groups. Each group is used twice, once as training and once as testing. This type of experiments are aiming at testing the cross group ability of the gender classification algorithm trained on a specific group.

In Table 2 - 5, we show the results under the *Intra-group* and the *Inter-group* experiments with the *Expression-Dependent* settings, for each face description. In each table, the first row shows the testing data, and the first column indicates the experimental setting. For each group, the number of scans is shown in the last row of the table. Results from the previous LOPO experiments on the 466 scans are also presented in each table, labeled as *LOPO-ED*. In Table 2 - 5, we observe that, **First**, the results from the *Intra-group* experiments always outperform the

results from LOPO experiments, further it outperforms the results from the *Inter-group* experiments. It reveals that faces with similar magnitude of sexual dimorphism can determine more accurately the gender of each other. It is a better gender classification strategy to train and test within scans of similar sexual dimorphism magnitude. **Second**, when training with the *Low-SD* group, the testing results on the *High-SD* group are always effective ($> 90\%$), but the inverse is not true. The algorithm trained on the *Low-SD* instances has good generalization ability on the *High-SD* instances. It means that the *Low-SD* instances contains more discriminating cues of the gender. Taking the first two diagonal elements of each table, we can generate also the results for each description when training with only the *Low-SD* instances. Following this, our approach achieved 86.48% with the *3D-avg.* description, 89.06% using the *3D-sym.* description, 88.41% based the *3D-grad.* description, and 87.18% with the *3D-spat.* description. Except the *3D-avg.* description, these results outperform significantly the corresponding LOPO experiments and are comparable to the *Intra-group* experiment results shown in Table 3 - 5. It means that with only the *Low-SD* instances in training, we can achieve better results than with all the scans in training. The *Low-SD* instances should have higher priority to be selected for training, than the *High-SD* ones.

Table 2: Results of 3D-avg

	Low-SD	*High-SD*	All
Intra-Group	75.32%	98.70%	90.77%
Inter-Group	60.76%	92.21%	81.54%
LOPO-ED	70.25%	98.70%	89.06%
# of scans	158	308	466

Table 3: Results of 3D-sym

	Low-SD	*High-SD*	All
Intra-Group	77.78%	98.13%	89.48%
Inter-Group	70.71%	97.39%	86.05%
LOPO-ED	73.23%	98.51%	87.77%
# of scans	198	268	466

Table 4: Results of 3D-grad

	Low-SD	*High-SD*	All
Intra-Group	80.60%	99.49%	88.62%
Inter-Group	69.03%	98.99%	81.76%
LOPO-ED	75.75%	98.99%	85.62%
# of scans	268	198	466

Table 5: Results of 3D-spat

	Low-SD	*High-SD*	All
Intra-Group	81.96%	97.95%	86.70%
Inter-Group	71.56%	98.63%	80.04%
LOPO-ED	77.50%	98.63%	84.12%
# of scans	320	146	466

In parallel, we also performed the *Intra-group* experiments and the *Inter-group* experiments under the *Expression-Independent* setting on the whole FRGCv2 dataset. The results are shown in Tables 6 - 9. Similarly, the results from LOPO experiments on the 4005 scans are also shown in each table, labeled as *LOPO-EI*. They show, again, that the *Intra-group* experiments always outperform the results from LOPO experiments on the whole FRGCv2, further outperform the results from the *Inter-group* experiments. It confirms that faces with similar magnitude of sexual dimorphism can determine more accurately the gender of each other, and it is better gender classification strategy to train and test within scans of similar sexual dimorphism magnitude. When training with the *Low-SD* group, the testing results on the *High-SD* group are again always

effective (> 95%) in each description, but the inverse is not. It confirms the previous finding that the algorithm trained on the *Low-SD* instances has very generalization ability on the *High-SD* instances, and the *Low-SD* instances contain more accurately the discriminating cues of gender. When taking only the *Low-SD* instances in training, we achieve 92.78% in the *3D-avg.* description, 90.29% in the *3D-sym.* description, 88.07% in the *3D-grad.* description, and 87.32% in the *3D-spat.* description. These results are very close to the *Intra-group* experiment results, and most of them outperform the corresponding LOPO experiments, as shown in Table 6 - 9. It confirms that, with the *Low-SD* instances in training, we can achieve better results than with all the scans in training.

Table 6: Results of 3D-avg

	Low-SD	*High-SD*	All
Intra-Group	79.03%	98.31%	92.96%
Inter-Group	68.05%	98.06%	89.74%
LOPO-EI	72.55%	98.38%	91.21%
# of scans	1111	2894	4005

Table 7: Results of 3D-sym

	Low-SD	*High-SD*	All
Intra-Group	75.89%	97.43%	90.76%
Inter-Group	70.16%	96.75%	88.51%
LOPO-EI	75.00%	97.43%	90.49%
# of scans	1240	2765	4005

Table 8: Results of 3D-grad

	Low-SD	*High-SD*	All
Intra-Group	79.04%	97.38%	86.72%
Inter-Group	69.07%	97.20%	80.85%
LOPO-EI	76.50%	97.50%	85.29%
# of scans	2014	1991	4005

Table 9: Results of 3D-spat

	Low-SD	*High-SD*	All
Intra-Group	79.39%	98.24%	88.76%
Inter-Group	73.63%	97.74%	85.62%
LOPO-EI	77.46%	98.24%	87.79%
# of scans	2328	1687	4005

3.3 Decision-level Fusion for Gender Classification

As noted before, the four face descriptions reflect different perspectives for sexual dimorphism. All of them have demonstrated good competence in face gender classification. Thus, we explore in this section their fusion in gender classification. Following the idea that the higher the magnitude of sexual dimorphism, the higher the accuracy in gender classification, we propose to take the predicted label with the highest magnitude of sexual dimorphism given by the four descriptions in the *Intra-Group* experiments. In practice, this is equal to take the label which is associated with the highest ratio of votes among the predicted labeled given by each of the four description with the Random Forest classifier. With this fusion strategy, our method achieved 97.42% gender classification rate on the 466 earliest scans, and 97.18% gender classification rate on the whole FRGCv2 dataset. The details of these fusion results are shown in Fig. 3.

In each subplot of Fig. 3, the x-axis shows magnitude of sexual dimorphism. The blue bars show the gender classification rate (corresponds to the left-side y-axis), and the red line shows the number of instances (corresponds to the right-side y-axis). As shown in Fig. 3 (A), under the *Expression-Dependent* setting, the magnitude of sexual dimorphism for more than 60% instances (296 in 466) is greater than 0.9 in the fusion. The corresponding gender classification rate

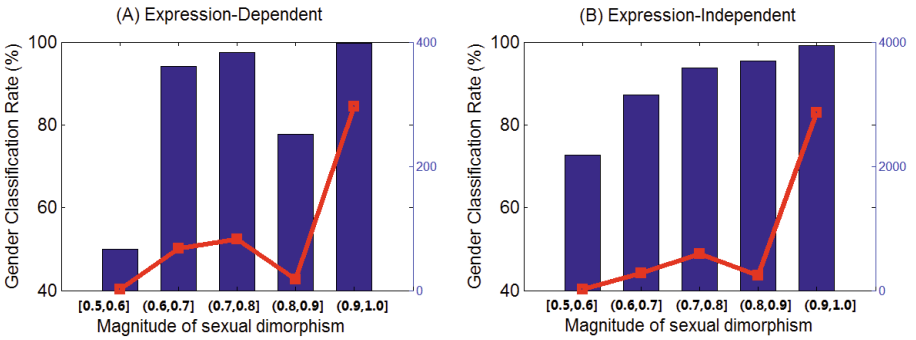


Fig. 3: Fusion details concerning different magnitude of sexual dimorphism

reaches 99.67%. Under the *Expression-Independent* setting, as shown in Fig. 3 (B), the magnitude of sexual dimorphism for more than 70% instances (2861 in 4005) is greater than 0.9 in the fusion. The corresponding gender classification rate reaches 99.2%. These results explain largely for the improvements of results in fusion. Also, the fusion significantly improves the classification accuracy in the instances with low sexual dimorphism magnitude under both experimental settings, especially when the instances are evaluated as (0.6,0.8] in sexual dimorphism magnitude in fusion.

4 Conlusion

In this work, we have proposed to use the ratio of votes from Random Forest to evaluate the sexual dimorphism of face instances. Four facial description designed to capture different perspectives of the facial morphology are built based on an elastic shape analysis of facial curves. We have discovered that instances with similar sexual dimorphism magnitude can determine more accurately the gender of each other in gender classification. Moreover, we reveal that the face instances with low magnitude of sexual dimorphism can tell more accurately the discriminating cues of gender. The gender classification algorithm trained with these instances have good generalization ability for gender classification of instances with high magnitude of sexual dimorphism, while the inverse is not. When training only on the instances with low magnitude of sexual dimorphism, better results can be achieved than training generally on all the instances. We also propose a decision-level fusion method, with which we achieve 97.46% gender classification rate on the 466 earliest scans of FRGCv2, and 97.18% on the whole FRGCv2 dataset.

References

1. Hall, M.A.: Correlation-based feature subset selection for machine learning. Ph.D thesis, Department of Computer Science, University of Waikato (1999)

2. Ballihi, L., Ben Amor, B., Daoudi, M., Srivastava, A., Aboutajdine, D.: Boosting 3D-geometric features for efficient face recognition and gender classification. IEEE Transactions on Information Forensics and Security **7**, 1766–1779 (2012)
3. Baudouin, J.Y., Gallay, M.: Is face distinctiveness gender based? Journal of Experimental Psychology: Human Perception and Performance **32**(4), 789 (2006)
4. Ben Amor, B., Drira, H., Berretti, S., Daoudi, M., Srivastava, A.: 4d facial expression recognition by learning geometric deformations. IEEE Transactions on Cybernetics, February 2014
5. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
6. Bruce, V., Burton, A.M., Hanna, E., Healey, P., Mason, O., Coombes, A., Fright, R., Linney, A.: Sex discrimination: How do we tell the difference between male and female faces? Perception. **22**(2), 131–152 (1993)
7. Drira, H., Ben Amor, B., Srivastava, A., Daoudi, M., Slama, R.: 3d face recognition under expressions, occlusions, and pose variations. IEEE Transactions on Pattern Analysis and Machine Intelligence. **35**, 2270–2283 (2013)
8. Drira, H., Ben Amor, B., Daoudi, M., Srivastava, A., Berretti, S.: 3d dynamic expression recognition based on a novel deformation vector field and random forest. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 1104–1107. IEEE (2012)
9. Du, L., Zhuang, Z., Guan, H., Xing, J., Tang, X., Wang, L., Wang, Z., Wang, H., Liu, Y., Su, W., et al.: Head-and-face anthropometric survey of chinese workers. Annals of occupational hygiene **52**(8), 773–782 (2008)
10. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**, 2234–2240 (2007)
11. Gilani, S.Z., Shafait, F., Ajmal, M.: Biologically significant facial landmarks: how significant are they for gender classification?. In: DICTA, pp. 1–8 (2013)
12. Guillaume, V., Harold, H., Eric, V.B.: Linking the structure and perception of 3d faces: gender, ethnicity, and expressive posture. In: International Conference on Audio-Visual Speech Processing (AVSP) (2003)
13. Han, X., Ugail, H., Palmer, I.: Gender classification based on 3D face geometry features using svm. In: CyberWorlds, pp. 114–118 (2009)
14. Hill, H., Bruce, V., Akamatsu, S.: Perceiving the sex and race of faces: the role of shape and colour. Proceedings of the Royal Society of London Series B Biological Sciences **261**(1362), 367–373 (1995)
15. Hu, Y., Yan, J., Shi, P.: A fusion-based method for 3D facial gender classification. Computer and Automation Engineering (ICCAE) **5**, 369–372 (2010)
16. Hu, Y., Fu, Y., Tariq, U., Huang, T.S.: Subjective experiments on gender and ethnicity recognition from different face representations. In: Boll, S., Tian, Q., Zhang, L., Zhang, Z., Chen, Y.-P.P. (eds.) MMM 2010. LNCS, vol. 5916, pp. 66–75. Springer, Heidelberg (2010)
17. Huang, D., Ding, H., Wang, C., Wang, Y., Zhang, G., Chen, L.: Local circular patterns for multi-modal facial gender and ethnicity classification. Image and Vision Computing (0) (2014)
18. Hunter, W.S., Garn, S.M.: Disproportionate sexual dimorphism in the human face. American Journal of Physical Anthropology **36**(1), 133–138 (1972)
19. Huynh, T., Min, R., Dugelay, J.: An efficient lbp-based descriptor for facial depth images applied to gender recognition using rgb-d face data. In: ACCV 2012, Workshop on Computer Vision with Local Binary Pattern Variants (2012)
20. Jones, B.C., DeBruine, L.M., Little, A.C.: The role of symmetry in attraction to average faces. Perception & Psychophysics **69**(8), 1273–1277 (2007)

21. Kohavi, R.: Wrappers for performance enhancement and oblivious decision graphs. Ph.D thesis, Stanford University (1995)
22. Komori, M., Kawamura, S., Ishihara, S.: Effect of averageness and sexual dimorphism on the judgment of facial attractiveness. Vision Research **49**(8), 862–869 (2009)
23. Little, A., Jones, B., Waitt, C., Tiddeman, B., Feinberg, D., Perrett, D., Apicella, C., Marlowe, F.: Symmetry is related to sexual dimorphism in faces: data across culture and species. PLoS One **3**(5), e2106 (2008)
24. Little, A.C., Jones, B.C., DeBruine, L.M., Feinberg, D.R.: Symmetry and sexual dimorphism in human faces: interrelated preferences suggest both signal quality. Behavioral Ecology **19**(4), 902–908 (2008)
25. Liu, Y., Palmer, J.: A quantified study of facial asymmetry in 3D faces. In: Analysis and Modeling of Faces and Gestures, pp. 222–229 (2003)
26. Lu, X., Chen, H., Jain, A.K.: Multimodal facial gender and ethnicity identification. In: International Conference on Advances in Biometrics, pp. 554–561 (2006)
27. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. Computer Vision and Pattern Recognition **1**, 947–954 (2005)
28. Rhodes, G., Zebrowitz, L.A., Clark, A., Kalick, S.M., Hightower, A., McKay, R.: Do facial averageness and symmetry signal health? Evolution and Human Behavior **22**(1), 31–46 (2001)
29. Shuler, J.T.: Facial sexual dimorphism and judgments of personality: a literature review. Issues **6**(1) (2012)
30. Smith, F.G., Jones, B.C., DeBruine, L.M., Little, A.C.: Interactions between masculinity-femininity and apparent health in face preferences. Behavioral Ecology **20**(2), 441–445 (2009)
31. Srivastava, A., Klassen, E., Joshi, S., Jermyn, I.: Shape analysis of elastic curves in euclidean spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**, 1415–1428 (2011)
32. Steven, W., Randy, T.: Facial masculinity and fluctuating asymmetry. Evolution and Human Behavior **24**(4), 231–241 (2003)
33. Toderici, G., O'Malley, S., Passalis, G., Theoharis, T., Kakadiaris, I.: Ethnicity- and gender-based subject retrieval using 3-D face-recognition techniques. International Journal of Computer Vision **89**, 382–391 (2010)
34. Wang, X., Kambhamettu, C.: Gender classification of depth images based on shape and texture analysis. In: Global Conference on Signal and Information Processing (GlobalSIP), pp. 1077–1080. IEEE (2013)
35. Wu, J., Smith, W., Hancock, E.: Gender classification using shape from shading. In: International Conference on Image Analysis and Recognition, pp. 499–508 (2007)
36. www.virtualffs.co.uk
37. Xia, B., Ben Amor, B., Drira, H., Daoudi, M., Ballihi, L.: Gender and 3D facial symmetry: what's the relationship?. In: IEEE Conference on Automatic Face and Gesture Recognition (2013)
38. Xia, B., Ben Amor, B., Huang, D., Daoudi, M., Wang, Y., Drira, H.: Enhancing gender classification by combining 3d and 2d face modalities. In: European Signal Processing Conference (EUSIPCO) (2013)
39. Young, J.: Head and face anthropometry of adult us civilians (1993)
40. Zhuang, Z., Bradtmiller, B.: Head and face anthropometric survey of us respirator users. Journal of Occupational and Environmental Hygiene **2**(11), 567–576 (2005)
41. Zhuang, Z., Landsittel, D., Benson, S., Roberge, R., Shaffer, R.: Facial anthropometric differences among gender, ethnicity, and age groups **54**(4), 391–402 (2010)

Towards Predicting Good Users for Biometric Recognition Based on Keystroke Dynamics

Aythami Morales(✉), Julian Fierrez, and Javier Ortega-Garcia

Biometric Recognition Group ATVS, EPS, Universidad Autonoma de Madrid,
C/Francisco Tomas y Valiente 11, 28049 Madrid, Spain
{aythami.morales,julian.fierrez,javier.ortega}@uam.es

Abstract. This paper studies ways to detect good users for biometric recognition based on keystroke dynamics. Keystroke dynamics is an active research field for the biometric scientific community. Despite the great efforts made during the last decades, the performance of keystroke dynamics recognition systems is far from the performance achieved by traditional hard biometrics. This is very pronounced for some users, who generate many recognition errors even with the most sophisticate recognition algorithms. On the other hand, previous works have demonstrated that some other users behave particularly well even with the simplest recognition algorithms. Our purpose here is to study ways to distinguish such classes of users using only the genuine enrollment data. The experiments comprise a public database and two popular recognition algorithms. The results show the effectiveness of the Kullback-Leibler divergence as a quality measure to categorize users in comparison with other four statistical measures.

Keywords: Keystroke · Typing patterns · Biometric · Authentication · Quality · Performance prediction

1 Introduction

Keystroke dynamics is a well-known biometric recognition technology which has attracted the interest of industry and researchers during the last decade [1][2]. The proliferation of web applications (e. g. e-banking or e-commerce) and the necessity of accurate and secure recognition methods has increased the interest in biometrics related with the user activity with the computer. Keystroke dynamics plays an important role in this context and its complementarity with other biometric modalities such as mouse dynamics has renovated the interest in these approaches [3]. The identification of people using their typing patterns can be applied to several scenarios including high security password authentication [1], text-independent authentication [4] and continuous authentication [5]. In summary, keystroke dynamics is an active research area with both scientific and industrial possibilities (e. g. DARPA, Active Authentication Program).

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 711–724, 2015.
DOI: 10.1007/978-3-319-16181-5_54

However, the accuracy of keystroke dynamics recognition systems is far from the performance achieved by the most competitive biometric traits and the error rates requested by international biometric standards (e. g. EN-50133-1). In terms of performance, keystroke dynamics can be considered halfway between hard and soft biometrics. As a behavioral biometric, it is highly user-dependent and it is difficult to generalize the performance among all population and scenarios. Previous works demonstrate the large user-variability of the error rates even with the most competitive recognition algorithms [6]. There are users with performances twenty times worst than others and therefore it is difficult to ascertain the overall accuracy. Predicting the performance of the users during the enrollment is a key to improve further recognition steps or the enrollment itself.

The performance of biometric recognition systems is strongly related with the quality of the samples [7]. Quality assessments have been studied for different biometrics traits such as fingerprint [8] or face [9] among others. Despite its well-known utility, the quality of keystroke dynamics has been scarcely studied [10]. The main reasons for this apparent disinterest could be found in the difficulties to establish a quality assessment of a behavioral biometric based only in timing between key events. It is not trivial defining the meaning of quality in keystroke dynamics technologies.

The term quality in the biometric literature have several meanings and applications. It is possible to distinguish between quality of biometric samples, quality of sensors and quality of the users among others. This paper focuses on quality of the users as a measure of their individual performance in terms of recognition accuracy. Low quality users imply users with low performances or high error rates while high quality users will be those users with high performances or low errors.

This paper studies different statistical measures for a reliable prediction of the quality of users for keystroke dynamics authentication. The purpose here is to analyze different ways to distinguish between users with well marked differences in terms of performance. The study assumes a scenario in which only the genuine enrollment data is used for both predicting the quality and enrolling the user. The experiments include a public benchmark dataset and two popular matchers for keystroke dynamics. The results suggest that it is possible to define a quality measure to categorize users correlated with their recognition performance.

The paper is organized as follows. Section 2 introduces the quality framework and proposes the Kullback-Leibler divergence as a feasible measure to establish the quality of the users of keystroke dynamics authentication systems. Section 3 presents the experimental protocol and results while Section 4 summarizes the main conclusions.

2 Quality Assessment for Keystroke Dynamics

Quality of biometric samples has become an important concern for the biometric community [7][11]. It is well known that the degradation of quality strongly affects the performance of biometric recognition systems and dealing with such

degradation is still an open challenge in many biometric traits. The quality of biometric samples is affected by many factors and it is difficult to generalize among all biometric technologies and sensors. The standard ISO/IEC 29794-1 has established normalizations and three main concepts related with the quality on biometric systems:

- **Character:** indicates the distinctiveness capability of the source.
- **Fidelity:** indicates the degree of similarity between the sample and its source.
- **Utility:** indicates the impact of a sample on the overall performance of the biometric system.

The quality measure of a biometric sample can be used for different purposes including: image enhancement [12], improving the matching algorithms [8] or optimized fusion strategies [13][11][14], see Fig. 1. Noteworthy, the quality is not exclusively related to a standalone sample and it is possible to measure the quality of a user or its enrollment set [14]. This quality evaluation of the users can be employed to improve the enrollment, the combination with other systems and the confidence on the authentication. The performance of the biometric recognition system is strongly influenced by the quality of the enrollment data and the evaluation of its utility is crucial in real applications [15].

Concerning keystroke dynamics, among the several factors that affect the quality of the biometric sample it is important to highlight:

- **Behavioral factors:** related with human emotional states, cooperativity or distractions. These factors also comprise the intrinsic characteristics of each user which include users particularly vulnerable to impersonation or users difficult to match, among others. The literature refer to this as biometric zoo [16] or menagerie [17].
- **Sensor factors:** related with the sensor, human-machine interactions, ease of use or maintenance. The proliferation of new portable devices and the necessity of interoperable schemes are important factors which affect the quality of the samples.

While the factors related with the sensor can be mitigated with hardware maintenance, the factors related with human behavior have more unpredictable consequences. How can we detect behavioral factors such as cooperativity or distraction from keystroke dynamic features? The features employed in keystroke dynamics are generally based on timing between key events [1][2] and the quality evaluation of these features arises several problems. In [10] the researchers analyzed six factors to explain different keystroke dynamics error rates (in order of relevance): algorithms, training amount, updating, typist-to-typist variation, feature set and impostor practice.

The quality evaluation of the keystroke dynamics has attracted scarce attention in the literature. The related works are focused mostly on outliers removal [18][19] and features improvement [20][21]. The outliers can be defined as samples with an unusual pattern in comparison with the available data from a

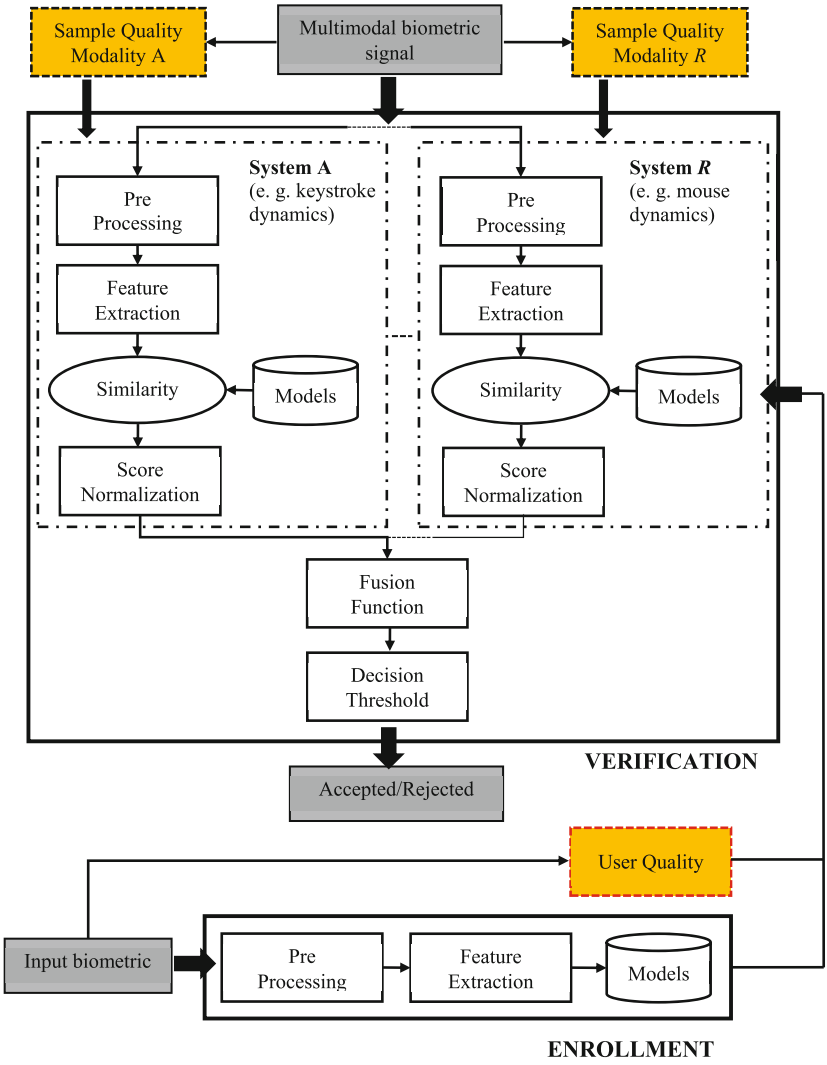


Fig. 1. Block diagram of multimodal biometric recognition/identification systems

specific user. The methodology used to discard these samples is traditionally based on statistical features (mean, variance, standard deviation) related with the distribution of the genuine data [18][19]. The inclusion of artificial rhythms and cues was proposed in [20] to improve the quality of data in terms of distinctive ability. In [21], the researchers established a quality classification in terms of uniqueness, inconsistency, and discriminability of the keystroke patterns. The main disadvantage of this classification is that all three measures need both gen-

uine and impostor data. Depending of the application, the impostor data may not be available (e. g. applications where the password is chosen by the user).

2.1 Measuring the Quality with the Kullback-Leibler Divergence

The entropy is a measure of the uncertainty in a random variable and it is related with the information present in any signal. Some researchers have studied the relationship between the performance of biometric recognition algorithms based on online signature and the entropy of the dynamic signals [22]. The researchers observed that high values of entropy implied higher error rates and low entropy values implied lower error rates. The reason of such behavior was explained with the stability of the genuine samples, which is greater for low entropy samples.

The Kullback-Leibler divergence (also called relative entropy or K-L divergence) is another information measure which have been proposed for biometric quality assessment [9]. The K-L divergence measures the difference between two probability distributions A and B in terms of the information needed to approximate A to B. In this paper we measure the K-L divergence from a feature vector $\mathbf{v}^Q = [v_1^Q, v_2^Q, ..., v_N^Q]$ (Query sample with N features) and the enrollment data mean $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_N]$ (generated with the enrollment data). The K-L divergence $D_{KL}(\mathbf{v}^Q||\boldsymbol{\mu})$ can then be defined as:

$$D_{KL}(\mathbf{v}^Q||\boldsymbol{\mu}) = \sum_{n=1}^{N} v_n^Q \log \frac{v_n^Q}{\mu_n} \tag{1}$$

where $\boldsymbol{\mu}$ is the enrollment data mean of the user obtained as:

$$\boldsymbol{\mu} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{v}_m^E \tag{2}$$

where each $\mathbf{v}_m^E$ is one out of the M enrollment samples (with N features each). Assuming $N = 31$ features and $M = 200$ enrollment samples, the Fig. 2 shows some examples of mean vectors (from the CMU benchmark dataset detailed in Section 3) as well as the $D_{KL}(\mathbf{v}^Q||\boldsymbol{\mu})$ obtained for each of the 50 query samples $\mathbf{v}^Q$ of the same users.

It can be seen that there are slight differences between the user enrollment data mean (note that the password was unique for all users). However, the K-L divergence between query samples and user background models shows different behavior and it is possible to find users with low stability (Fig. 2-Right black line) or users with very stable K-L divergence values (Fig. 2-Right grey lines). Next sections will analyze the correlation between K-L divergence and the performance of keystroke dynamics systems.

3 Experiments

The experiments are conducted to analyze: i) the quality dependence of keystroke dynamics and; ii) the utility of the K-L divergence for predicting the

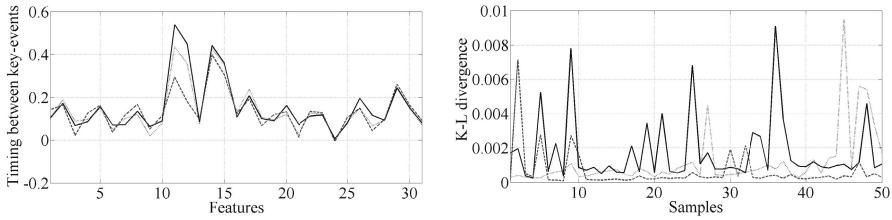


Fig. 2. The mean vectors of 3 different users (left) and the K-L divergence from 50 samples (feature vectors) of the same 3 users (right). $N = 31$ features and $M = 200$ enrollment samples.

performance of individual users. The experiments assume a scenario in which only genuine enrollment data is available (imposter data are not employed to model the users).

3.1 Database: the CMU Benchmark Dataset

The CMU benchmark dataset [19] comprises 51 subjects and 8 sessions with 50 repetitions per session. The time lapse between sessions is more than one day and the 400 typing samples were collected with an accuracy of 200 microseconds. The password was the same for all users and it consists of a ten characters typical strong password which includes uppercase, lowercase and symbols: **.tie5Roanl**. The feature data for each sample includes: hold time for each key (i.e. time between press and release); the keydown-keydown time between two keys (i.e. time between the press of the key 1 and the press of key 2); the keyup-keydown time between two keys (i.e. time between the release of key 1 and the press of key 2); the Enter key is included as a part of the password. The total number of features per samples is 31 (11 hold times, 10 keydown-keydown times and 10 keyup-keydown times).

The most attractive characteristics of the CMU benchmark dataset for this work can be summarized as: i) large number of samples per subject which allows an accurate modeling of the individual behavior; ii) publicly available benchmarks with several feature extraction and classification techniques [19].

3.2 Baseline Systems

The experimental protocol is the same as employed in popular benchmarks [18][19]. The 200 samples from the first 4 sessions are used as gallery/enrollment set. The genuine scores are obtained from the 200 samples corresponding to the last 4 sessions while the impostors are obtained from the first 50 samples of each subject in the database. The performance is evaluated in terms of Equal Error Rate (EER) for each of the 51 subjects in the database.

In this paper we evaluate two popular recognition algorithms for keystroke dynamics [18][19]. Both approaches have achieved the most competitive performances reported for the CMU benchmark dataset among more than 14 different systems. Both systems include training/modeling stages based exclusively on genuine data and other promising systems were discarded because they include impostor data during the training phase [23]. The approaches used in the experimental evaluation made in this paper are:

- **System A** - Modified Manhattan distance with Nearest Neighbor classifier [18]: this system is based on a combination of the Manhattan and the Mahalanobis distances. The method can be summarized as (see [18] for details): i) the feature vectors are normalized according principles inspired by the Mahalanobis distance (using the covariance matrix) and; ii) the normalized feature vectors are matched with the enrollment data using the Manhattan distance and a Nearest Neighbor classifier.
- **System B** - Scaled Manhattan distance [19]: this system is based on the simplicity of the Manhattan distance and its usefulness for decomposing into contributions made by each feature (see [19] for details). The distance is normalized by the average absolute deviation from the enrollment data.

Outlier removal is common in keystroke dynamics and it is a feasible method to improve the enrollment set. An outlier is a sample beyond the typical user variability and its inclusion in the enrollment set to model the user usually have a negative impact in the performance. The K-L divergence can be used to evaluate the stability of the enrollment set. Fig. 3 shows the mean K-L divergence for all the subjects in the CMU database for the different sessions available. The K-L divergence is estimated separately for each subject (there is one $\boldsymbol{\mu}$ for each subject which is calculated using all the samples available from the same user as it is described in section 2.1) and the results are averaged. Note that the users were not habituated to type the password (.tie5Roanl) and they needed a learning period in which they stabilized their typing patterns.

Fig. 3 clearly shows large differences between K-L divergences values from first and last sessions. The samples from the first session can be considered outliers. Table 1 shows how excluding such samples can improve the overall performance of the baseline systems. However, the improvement is slight (around 10% improvement of the average EER) and it is important to note that 150-200 enrollment samples could be considered excessive depending of the application.

3.3 Performance Evaluation

The standard ISO/IEC 29794 includes in the definition of the purpose of quality algorithms the next requirement: *"Quality algorithms shall produce quality scores that predict performance metrics such as either false match or false non-match".*

The quality is related with several factors including the acquisition, the features and the personal characteristics of the subjects among others. The performance of keystroke dynamics is highly user-dependent and it means that there

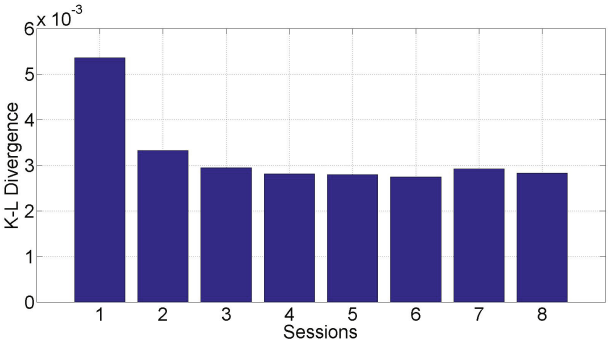


Fig. 3. Mean Kullback-Leibler divergence for each session on the CMU benchmark dataset

Table 1. Performance (EER in %) for different enrollment sets employed

	Enrollment Set	
	Sessions 1 to 4	**Sessions 2 to 4**
System A	8.89	8.20
System B	9.60	8.55

are users who exhibit an EER below 1% and others with EER greater than 20%, see Fig. 4.

Fig. 4 shows that different users present large variations in terms of performance. The reasons of such different performances vary with users who are easy to be recognized (Fig. 4c and Fig. 4d) and others are difficult to be recognized (Fig. 4a and Fig. 4b). The researchers analyzed and defined these classes of users as the biometric menagerie [17] or biometric zoo [16].

These performance evaluations are obtained a posteriori when the test data is compared with the enrollment data. Is it possible to predict the performance of each user based exclusively in her enrollment data? To answer this question it is necessary to determine if the enrollment data provided by the user contains enough information to ascertain the performance during the subsequent test phase.

Inspired by the methodology employed in [17], we divided the population in three groups according to their performance (obtained following the experimental protocol explained in Section 3.2) as Good (33% of users with lowest EER), Ugly (33% of users with highest EER) and Bad (the remaining 33% of users). This is not an ideal classification but allows us to group users with similar performances. See Table 2 for the resulting performance of the baseline systems in these three groups.

Fig. 5 shows the ROC curves for each of the performance groups of the two keystroke dynamics recognition algorithms employed in this work. The curves evidence the different performances for both recognition algorithms and the three groups considered.

Table 2. Performance (EER in %) according to the performance groups

System	All	Good		Bad		Ugly	
		Mean	Min/Max	Mean	Min/Max	Mean	Min/Max
A	8.20	4.31	0.45/6.65	8.58	6.90/9.75	15.31	9.80/24.50
B	8.55	4.27	0.90/6.90	8.65	7.10/11.10	16.66	11.10/27.55

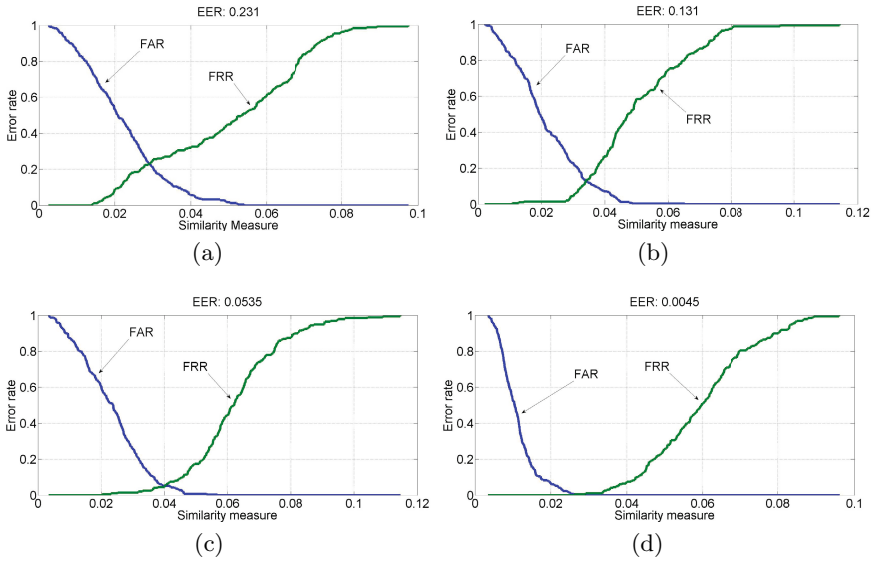


Fig. 4. False Acceptance and False Rejection curves for user 1 (a), user 2 (b), user 3 (c) and user 31 (d) from CMU benchmark dataset using the System A

3.4 Predicting the Quality

This paper analyzes five different measures for estimating the performance of keystroke dynamics users based exclusively on the enrollment data. Assuming that $\mathbf{v}^Q = [v_1^Q, v_2^Q, ..., v_N^Q]$ is the feature vector of the Query sample, N is the number of features ($N = 31$ for the CMU benchmark database) and $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_N]$ the enrollment data mean of the user (detailed in Section 2.1), the measures evaluated in this paper are defined as:

– **Variance:** the variance measures the stability of the data available for each user. A small variance indicates small differences between the query sample and the enrollment mean. A high variance indicates that the feature distribution is spread out around the mean. The variance is a valuable measure to characterize the stability of the data provided by the user. The variance is defined as:

$$\text{Variance} = \frac{1}{N} \sum_{n=1}^{N} \left(v_n^Q - \mu_n\right)^2 \tag{3}$$

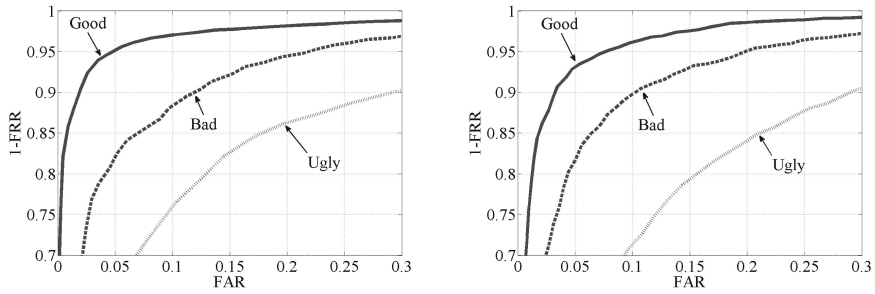


Fig. 5. ROC curves for different performance groups obtained using the System A (left) and System B (right)

– **Structural Content:** this is a popular quality measure for image analysis [24]. For a one-dimensional vector it is defined as the relative difference between the information of the query sample and the enrollment mean. The structural content is defined as:

$$\text{Structural Content} = \frac{\sum_{n=1}^{N} \left(v_n^Q\right)^2}{\sum_{n=1}^{N} \left(\mu_n\right)^2} \tag{4}$$

– **Entropy:** the entropy quantifies the expected value of the information contained in a sequence. The entropy value is defined as:

$$\text{Entropy} = -\sum_{n=1}^{N} v_n^Q \log v_n^Q \tag{5}$$

– **Kullback-Leibler Divergence:** as described in Eq. (1). The K-L divergence measures the amount of information needed to approximate two distributions.
– **Genuine scores:** it is possible to estimate the genuine score distribution of the enrollment data. From the 4 sessions available as enrollment, we used three sessions for training and the other one for validation. The protocol is repeated for all 4 sessions for a total number of genuine scores equal to 200.

The experiments conducted try to ascertain the capability of the proposed measures to predict the performance of each user in a keystroke dynamics system (using only its enrollment data). The protocol used to ascertain the prediction capability can be summarized in the following steps:

– Based on the performance obtained with each of the systems (performance reported in Table 2), we assign a quality value between 2 and 0, Q_i^A and Q_i^B, for each subject i (Good=2, Bad=1, Ugly=0, $i = 1, ..., 51$). Therefore there are two different quality values assigned for each user (one for each system).

Table 3. Distances between a posteriori user quality estimations $\hat{Q}_i^M$ and a priori quality prediction Q_i^A based on individual statistical measures from the enrollment data (Baseline System A)

Quality Feature	$\frac{1}{51}\sum \lvert Q_i^A - \hat{Q}_i^M \rvert$	# Large errors (out of 51 subjects)
Genuine scores	1.05	16
Structural content	0.58	5
Entropy	0.63	7
Variance	0.62	4
K-L divergence	**0.52**	**3**

Table 4. Distances between a posteriori user quality estimations $\hat{Q}_i^M$ and a priori quality prediction Q_i^B based on individual statistical measures from the enrollment data (Baseline System B)

Quality Feature	$\frac{1}{51}\sum \lvert Q_i^B - \hat{Q}_i^M \rvert$	# Large errors (out of 51 subjects)
Genuine scores	0.78	9
Structural content	0.66	6
Entropy	0.67	6
Variance	0.58	6
K-L divergence	**0.54**	**3**

- The five statistical measures are computed using the enrollment data. The average values across the 200 enrollment samples are computed for each subject i.
- For each average measure M ={Variance, Structural Content, Entropy, K-L Divergence, Genuine Scores}, the estimated quality $\hat{Q}_i^M$ of each user i is assigned as Good (33% of best M values), Ugly (33% of worst M values) and Bad (the remaining 33% of M values). The term best depends of the measure employed being the lowest values in case of {Variance, Entropy, Structural Content and K-L Divergence} and highest values in case of {Genuine Scores}.
- The mean distance between the real quality groups obtained with the test data, Q_i^A and Q_i^B, and the different estimated qualities $\hat{Q}_i^M$ is evaluated.

Tables 3 and 4 show the distances between real quality groups obtained with test samples and the predictions obtained with the enrollment data. Note that the quality, as employed in this section, depends of the performance of a specific matcher. The tables also show the number of large prediction errors, *i. e.* the number of good users estimated as ugly or vice versa.

As can be seen in Tables 3 and 4, the genuine scores obtained from the enrollment data are less competitive than other measures. The reason is that genuine scores are very sensitive to those users especially vulnerable to impersonation. The K-L divergence shows the most competitive performance with only 3 large errors (out of 51 subjects) and a mean estimation error around 0.5. Fig. 6 shows

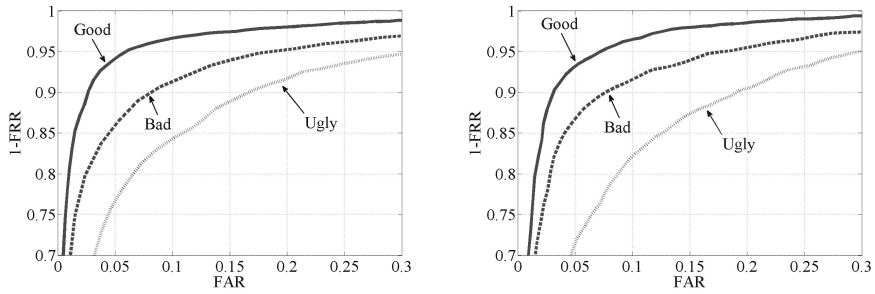


Fig. 6. ROC curves for different predicted qualities using the K-L divergence with Systems A (left) and B (right)

the ROC curves (averaged) of users classified by their predicted quality (using the K-L divergence).

The results show how the K-L divergence can be used to classify users apriori which will result in different performances groups in testing. The difference between classes is evident and the results suggest that the proposed measure is useful to predict the performance of keystroke dynamics using only the enrollment data.

4 Conclusions

This paper studied the feasibility of user quality prediction for biometric recognition based on keystroke dynamics. The usefulness of quality measures in biometrics is well-known and the scarce study on keystroke dynamics represents an open challenge for the scientific community. The performance of keystroke dynamics is highly user-dependent and it is usual to find large performance deviations among users even with the most competitive recognition algorithms. This paper analyzed five statistical measures for predicting the quality of users and the K-L divergence showed the most accurate results. The results showed that it is possible to ascertain the performance of users using exclusively the genuine enrollment data and encourage to further research in this area.

The work presented in this paper is focused on a limited dataset (i. e. same password and large amount of data per user) and future work includes other scenarios and databases. The prediction of performances when the password is different for each subject as well as text-independent keystroke dynamics are challenging scenarios to be studied.

Acknowledgments. Aythami Morales Moreno is supported by a Juan de la Cierva Fellowship from Spanish MINECO. This work has been partially supported by projects Bio-Shield (TEC2012-34881) from Spanish MINECO and BEAT (FP7-SEC-284989) from EU.

References

1. Peacock, A., Ke, X., Wilkerson, M.: Typing patterns: A key to user identication. IEEE Security and Privacy **2**(5), 40–47 (2004)
2. Banerjee, S.P., Woodard, D.L.: Biometric authentication and identification using keystroke dynamics: a survey. Journal of Pattern Recognition Research **7**, 116–139 (2012)
3. Bailey, K.O., Okolica, J.S., Peterson, G.L.: User identification and authentication using multimodal behavioral biometrics. Computers and Security **43**, 77–89 (2014)
4. Gunetti, D., Picardi, C.: Keystroke analysis of free text. ACM Transactions on Information and System Security **8**(3), 312–347 (2005)
5. Sim, T., Zhang, S., Janakiraman, R., Kumar, S.: Continuous verification using multimodal biometrics. IEEE Transactions on Pattern Analysis and Machine Intelligent **29**(4), 687–700 (2007)
6. Hocquet, S., Ramel, J.Y., Cardot, H.: User classication for keystroke dynamics authentication. In: International Conference on Biometrics, Seoul, Korea, pp. 531–539 (2007)
7. Alonso-Fernandez, F., Fierrez, J., Ortega-Garcia, J.: Quality measures in biometric systems. IEEE Security and Privacy **10**(9), 52–62 (2012)
8. Chen, Y., Dass, S.C., Jain, A.K.: Fingerprint quality indices for predicting authentication performance. In: International Conference on Audio and Video-Based Biometric Person Authentication, Hilton Rye Town, NY, USA, pp. 160–170 (2005)
9. Youmaran, R., Adler, A.: Measuring biometric sample quality in terms of biometric information. In: Biometrics Symposium, Baltimore, USA (2006)
10. Killourhy, K., Maxion, R.: Why did my detector do *That*?! predicting keystroke-dynamics error rates. In: Jha, S., Sommer, R., Kreibich, C. (eds.) RAID 2010. LNCS, vol. 6307, pp. 256–276. Springer, Heidelberg (2010)
11. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J., Bigun, J.: Discriminative multimodal biometric authentication based on quality measures. Pattern Recognition **38**(5), 777–779 (2005)
12. Hong, L., Wan, Y., Jain, A.K.: Fingerprint image enhancement: Algorithm and performance evaluation. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**, 777–789 (1998)
13. Fiérrez-Aguilar, J., Chen, Y., Ortega-Garcia, J., K.Jain, A.: Incorporating image quality in multi-algorithm fingerprint verification. In: Zhang, D., Jain, A.K. (eds.) ICB 2005. LNCS, vol. 3832, pp. 213–220. Springer, Heidelberg (2005)
14. Kumar, A., Zhang, D.: Improving biometric authentication performance from the user quality. IEEE Transactions on Instrumentation and Measurement **59**(3), 730–735 (2010)
15. Prabhakar, S., Pankanti, S., Jain, A.K.: Biometric recognition: Security and privacy concerns. IEEE Security Privacy Magacine **1**(2), 33–42 (2003)
16. Doddington, G., Liggett, W., Martin, A., Przybocki, M., Reynolds, D.: Sheep, goats, lambs and wolves: a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In: International Confenrence on Spoken Language Processing, Sydney, Austrailia (1998)
17. Yager, N., Dunstone, T.: The biometric menagerie. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(2), 220–230 (2010)
18. Zhong, Y., Deng, Y., Jain, A.K.: Keystroke dynamics for user authentication. In: IEEE Computer Society Workshop on Biometrics, Providence, USA (2012)

19. Killourhy, K.S., Maxion, R.A.: Comparing anomaly detectors for keystroke dynamics. In: International Conference on Dependable Systems and Networks, Estoril, Portugal, vol. 32, pp. 125–134 (2009)
20. Kang, P., Park, S., Hwang, S., Lee, H., Cho, S.: Improvement of keystroke data quality through artificial rhythms and cues. Computers and Security **27**, 3–11 (2008)
21. Cho, S., Hwang, S.: Artificial rhythms and cues for keystroke dynamics based authentication. In: Zhang, D., Jain, A.K. (eds.) ICB 2005. LNCS, vol. 3832, pp. 626–632. Springer, Heidelberg (2005)
22. Houmani, N., Garcia-Salicetti, S., Dorizzi, B.: A novel personal entropy measure confronted with online signature verification systems performance. In: IEEE Conference on Biometrics: Theory, Applications and Systems, Washington, USA, pp. 1–6 (2008)
23. Deng, Y., Zhong, Y.: Keystroke dynamics user authentication based on gaussian mixture model and deep belief nets. ISRN Signal Processing **2013**, 1–7 (2013)
24. Killourhy, K.S., Maxion, R.A.: Image quality measures and their performance. IEEE Transactions on Communications **43**(12), 2959–2965 (1995)

How Much Information Kinect Facial Depth Data Can Reveal About Identity, Gender and Ethnicity?

Elhocine Boutellaa[1,2], Messaoud Bengherabi[1], Samy Ait-Aoudia[2], and Abdenour Hadid[3(✉)]

[1] Centre de Développement des Technologies Avancées, Baba Hassen, Algeria
[2] Ecole Nationale Supèrieure d'Informatique, El Harrach, Algeria
[3] University of Oulu, Oulu, Finland
hadid@ee.oulu.fi

Abstract. Human face images acquired using conventional 2D cameras may have inherent restrictions that hinder the inference of some specific information in the face. The low-cost depth sensors such as Microsoft Kinect introduced in late 2010 allow extracting directly 3D information, together with RGB color images. This provides new opportunities for computer vision and face analysis research. Although more accurate sensors for detailed facial image analysis are expected to be available soon (e.g. Kinect 2), this paper investigates the usefulness of the depth images provided by the current Microsoft Kinect sensors in different face analysis tasks. We conduct an in-depth study comparing the performance of the depth images provided by Microsoft Kinect sensors against RGB counterpart images in three face analysis tasks, namely identity, gender and ethnicity. Four local feature extraction methods are investigated for both face texture and shape description. Moreover, the two modalities (i.e. depth and RGB) are fused to gain insight into their complementarity. The experimental analysis conducted on two publicly available kinect face databases, EurecomKinect and Curtinfaces, yields into interesting results.

1 Introduction

Human face is involved in an impressive variety of different activities. It houses the majority of our sensory apparatus - eyes, ears, mouth, and nose - allowing the bearer to see, hear, taste, and smell. Apart from these biological functions, it also provides a number of signals about our health, emotional state, identity, age, gender etc. Machine analysis of faces (i.e. automatic face analysis) plays also a key role in many emerging applications of computer vision, including biometric recognition systems, human-computer interfaces, smart environments, visual surveillance, and content-based retrieval of images from multimedia databases. Due to its many potential applications, automatic face analysis which includes, e.g., face detection, face recognition, gender classification, age estimation and facial expression recognition, has become one of the most active topics in computer vision research [1].

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 725–736, 2015.
DOI: 10.1007/978-3-319-16181-5_55

Face analysis problems have been mainly extensively studied using conventional RGB cameras at visible light. However, this makes some face analysis tasks a challenging problem. Furthermore, face images acquired using such conventional sensors may have inherent restrictions that hinder the inference of some specific information in the face. For instance, illumination changes are still challenges in face recognition while near infrared imaging is shown to be less prone to this problem; face spoofing (e.g. detecting sign of liveness) is a threat in face recognition using RGB images while thermal cameras can easily solve this problem; analysing faces under pose variations from 2D images is a complex task which can be better handled in 3D. So, face sensing using new technologies and beyond the visible light is needed.

The recent introduction of low-cost depth cameras (such as Microsoft Kinect) provides exciting new opportunities for computer vision and face analysis research. Kinect sensors allow extracting directly depth information, together with RGB color images. This is a potential alternative to classical 3D scanners which are usually slow, expensive and large-sized, making them inconvenient for many practical applications. Consequently, low-cost depth sensing has recently attracted a significant attention in the research community [2,3].

Among the major drawbacks of the facial depth images provided by Microsoft Kinect are the low-resolution and noisy nature of the images. This can be due, for instance, to missing data (holes) in some parts of the face, inaccurate depth value computation and limited distance coverage from the sensor (2 to 4 meters). More accurate sensors (e.g. Kinect 2) for more detailed facial image analysis are expected to be available soon. Despite of the aforementioned limitations, Kinect depth images (after efficient pre-processing) have already been successfully used in some facial analysis tasks such as head pose estimation [4] and gender classification [5]. Of significant importance when dealing with Kinect depth images is also the use of effective face descriptions. Local features are usually shown to perform better than global features due to their ability to cope with local changes.

The intriguing question is how much information Kinect depth data can reveal about faces? To answer this question and to gain insights into the usefulness of the depth images in different face analysis tasks, this work provides the first comprehensive analysis comparing the performance of the depth images versus RGB counterparts in three face analysis tasks, namely identity, gender and ethnicity. Four local feature extraction methods are considered for encoding face texture and shape: Local Binary Patterns (LBP) [6], Local Phase Quantization (LPQ) [7], Histogram of Oriented Gradients (HOG) [8] and Binarized Statistical Image Features (BSIF) [9]. Moreover, the complementarity of the two sources of information (i.e. depth and RGB) is also studied through experiments fusing the two modalities. Extensive experiments are conducted on two recent publicly available benchmark databases namely EurecomKinect [5] and Curtinfaces [10] face databases. The obtained results point out interesting findings.

The remainder of this paper is organized as follows. Section 2 reviews some works related to the use of Kinect depth images. Then, Section 3 presents our

methodology for studying the usefulness of Kinect depth images in different face analysis tasks. In Section 4, we describe the extensive experiments and discuss the obtained results. Section 5 draws some conclusions and highlights future perspectives.

2 Related Work

It is well-known that illumination and pose variations can be better tackled using 3D scans of faces than 2D images. However, 3D scanners are usually expensive, bulky and slow, and this limits their use in practical applications. The recent introduction of low-cost depth cameras (such as Microsoft Kinect) provides exciting new opportunities for computer vision and face analysis research. Kinect sensor allows extracting directly depth information, together with RGB color images of the scene at video rates. The sensor is based on time-of-flight technology and is initially introduced as a peripheral of Microsoft Xbox games console. Since its introduction in late 2010, it is widely adopted by the computer vision research community in various applications [2].

Among the attempts to use Kinect sensors for face analysis is the work of Li et al. [10] who aimed at tackling the problem of face recognition under pose, illumination, expression and disguise using Kinect. The authors proposed a preprocessing chain that generates canonical frontal views for both depth map and texture of the face regardless of its initial position. To this end, Iterative Closest Point (ICP) is used for registering a given face to a reference model. Then, facial symmetry is employed to recover missing face parts, fill holes and smooth the face depth data. Finally, sparse representation classifier (SRC) is used for both depth and texture separately. Experimental results on the CurtinFaces dataset [10] yields in a recognition rate of 88.7% using depth data only and 96.7% when face texture and depth are fused.

Similarly, Goswami et al. [11] used images obtained from Kinect for face recognition. The proposed method computes the HOG descriptor on the entropy of RGB-D faces and the saliency features from a 2D face. The probe RGB-D descriptor is used as input to a random decision forest classifier to establish the identity. Experimental results on a private database comprising 106 subjects with multiple RGB-D images of each subject indicated that the RGB-D information obtained by Kinect can be used to enhance face recognition performance compared to 2D and 3D approaches.

In another work, Min el al. [12] explored the use of Kinect sensor for real-time 3D face identification. Instead of registering a probe to all instances in the database, the authors proposed to only register it with several intermediate references (i.e. canonical faces) randomly selected from the gallery, thus reducing the processing time without significantly affecting the recognition performance. Moreover, ICP was implemented on a GPU. Good identification results were reported on a dataset of 20 subjects with an average speed ranging from 0.04 seconds to 0.38 seconds, depending on the number of canonical faces. It is worth noting, however, that the proposed approach was tested only under limited variations of head pose, expression and illumination.

More recently, Pamplona Segundo et al. [13] addressed continuous face authentication problem using 3D faces acquired with Kinect. Faces are first detected and normalized using ICP. Each face is then registered according to its pose and classified into frontal, left profile or right profile. HOG features are then extracted from the region of interest and matched to the corresponding region. The approach was evaluated on four 40 minutes long videos with variations in facial expression, occlusion and pose. An equal error rate (EER) of 0.8% was reported.

Inspired by 3DLBP [14], (a variant of LBP based on the statistics of range image differences), Huynh et al. [5] proposed a novel descriptor, called Gradient-LBP, and applied it to the problem of gender classification from Kinect depth images. Gradient-LBP encodes the facial depth difference as well as its sign. The depth differences are computed from different orientations yielding in a separate depth difference image per each orientation. Hence, the depth difference at each pixel for all the orientations is encoded. Experiments were carried out on both high quality 3D range images (obtained by a 3D scanner) and images of lower quality obtained from Kinect (EURECOM Kinect Face Dataset). The reported results pointed out the usefulness of facial depth information when used together with RGB images for gender classification.

It appears that most of the few attempts on using Kinect in face analysis are mainly devoted to the face recognition problem hence overlooking and ignoring other face analysis tasks such as gender recognition, age estimation and ethnicity classification. Moreover, most of the proposed works focused on the fusion of Kinect depth information and RGB images but did not explicitly explore how much information Kinect facial depth data alone can reveal about the faces. Some of the results are also reported on size-limited and/or private Kinect databases. Finally, most of the existing works used only basic features with the depth data and RGB images.

To tackle these drawbacks, this present work provides the first comprehensive analysis comparing the performance of the Kinect depth images versus RGB counterparts in three different face analysis tasks (identity, gender and ethnicity) and using four local feature extraction methods (LBP [6], LPQ [7], BSIF [9] and HOG [8]). Extensive experiments are carried out on two recent publicly available Kinect benchmark databases (EurecomKinect [5] and Curtinfaces [10]).

3 Methodology

3.1 Preprocessing

Since depth images provided by Kinect sensor are usually noisy and of low quality, a preprocessing is needed and crucial before further analysis. The noise in the depth images can be originated from the unknown distance between the sensor and the face. The depth maps usually contain many holes that should be filled. On the other side, 3D information is useful for assisting 2D analysis under severe pose variation by registration to a common face model using ICP algorithm.

Firstly, we transform the depth maps provided by Kinect into real world 3D coordinates. Thus, each pixel is represented by six values: x, y and z coordinates

Fig. 1. Examples of 2D (left) and 3D (right) cropped face images obtained with the Microsoft Kinect sensor after preprocessing

and the three RGB values. We translate the resulting cloud of points so that the nose tip is located at the origin by subtracting the nose coordinates. The nose has indeed been shown to be the most reliable point to crop the face region from the depth images [15]. Thus, we extract the face region using an ellipsoid centered at the nose tip. Therefore, the points located outside the ellipsoid are discarded. Then, we smooth and re-sample the face point cloud to a grid of 128×96. Examples of cropped 2D and 3D face images are shown in Fig. 1. Finally, for the 3D face, we drop the x and y coordinates hence keeping only the z coordinates describing the face shape.

3.2 Feature Extraction

After preprocessing, facial descriptors are computed from the depth and RGB images. The function of the descriptors is to convert the pixel-level information into a form, which captures the most important facial properties but is insensitive to irrelevant aspects caused by e.g. blur, noise and illumination changes. In contrast to global face descriptors which compute features directly from the entire face image, local face descriptors representing the features in small local image patches have proved to be more effective in real world conditions. Hence, we adopted four state-of-the-art local descriptors which are briefly described below.

Local Binary Patterns (LBP) [6] is defined as a gray-scale invariant texture measure, derived from a general definition of texture in a local neighborhood. The discriminative power, computational simplicity and tolerance against monotonic gray-scale changes are behind the great success of LBP in many computer vision problems. In LBP, a pixel code is computed by thresholding its value with the neighborhood. The signs of the differences are coded as a binary string which is converted to a decimal number representing the pixel code. The occurrences of the LBP codes in a given face image can be collected into a histogram. The classification can then be performed by computing histogram similarities.

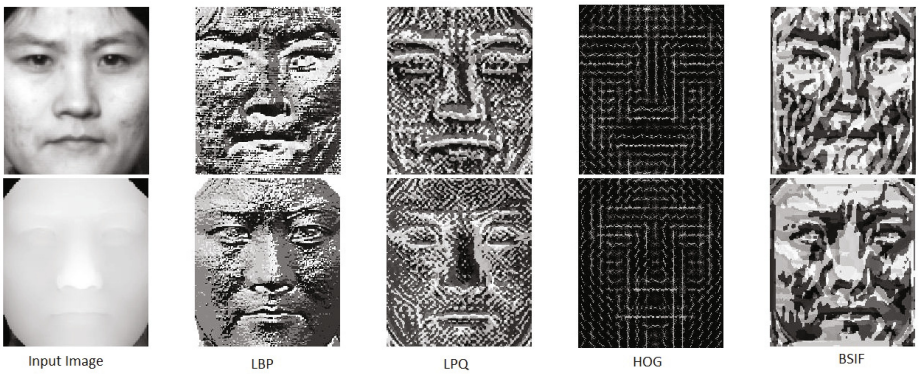


Fig. 2. Applying the four descriptors on a face texture and depth images. From left to right : the original face image (top: texture image and bottom: its corresponding depth image) and the resulting images after the application of LBP, LPQ, HOG and BSIF descriptors, respectively.

Local Phase Quantization (LPQ) [7] is shown to be robust for texture analysis [16] and face recognition [7] from blurred images. In LPQ, an image is described using the phase information of short-term Fourier transform (STFT) locally computed on a rectangular window at each pixel. The phase information of four Fourier coefficients are coded by examining the signs of the real and imaginary parts of each component. For a given image, each pixel is labeled with a blur invariant LPQ code. Similarly to LBP, the occurrences of the LPQ codes are collected into a histogram for classification.

Histogram of Oriented Gradients (HOG) [8] describes local object appearance and shape within an image by the distribution of intensity gradients or edge directions. The magnitudes of the gradient at each pixel are accumulated into a histogram according to the gradient direction. The image is first divided into small connected regions from which histograms of gradient directions or edge orientations of the pixels are extracted. The combination of these histograms yields in the HOG descriptor. The method was initially developed for human detection but later extended and applied to other computer vision problems including face analysis.

Binarized Statistical Image Features (BSIF) approach [9] was recently proposed for face recognition and texture classification. Inspired by LBP and LPQ, the idea behind BSIF is to automatically learn a fixed set of filters from a small set of natural images, instead of using hand-crafted filters such as in LBP and LPQ. The set of filters are learnt based on statistics of training images [9]. The training images, which are normalized to zero mean and unit variance, are randomly sampled into small patches. The mean of each patch is subtracted and PCA is applied to reduce the dimension and whiten the data. Finally, the filters

are estimated as the independent components obtained by ICA algorithm. An image is represented by the quantization of the filters responses.

Figure 2 shows the results of applying the four descriptors on a face texture and depth images of a subject from the FRGC database [17]. In the case of the BSIF descriptor, we extended the method to handle depth images as follows. We learnt the filters using facial depth images from the FRGC database. These filters are then used to compute BSIF features on Kinect depth images. We found this new learning to perform better than the original filters. To the best of our knowledge this is the first work that uses BSIF features for describing depth images.

For each descriptor (LBP, LPQ, HOG and BSIF), the RGB and depth images are first divided into several local regions from which local histograms are extracted and then concatenated into an enhanced feature histogram used for classification.

3.3 Classification

Once the face descriptors are extracted from both RGB and depth face images, we use the well-known support vector machine classifier (SVM) with an RBF kernel for the three face analysis problems. The libsvm[1] implementation is employed in our experiments. The face features are fed to the SVM classifier and the average classification accuracy is reported.

4 Experimental Analysis

For extensive experimental evaluation, we analyzed the performance of the four local descriptors (LBP [6], LPQ [7], BSIF [9] and HOG [8]) presented in Section 3.2 on two publicly available Kinect face databases, EurecomKinect [5] and Curtinfaces [10], containing both RGB and depth facial images. We report the results in three different face classification problems: face identification, gender recognition and ethnicity classification. We describe below the experimental data, the setup and the obtained results.

4.1 Experimental Data

The EurecomKinect face database [5] contains both RGB and depth facial images of 52 subjects acquired using Kinect sensor. There are 14 females and 38 males in the database. The people in the database belong to six different ethnicity groups (Asian , Black, Hispanics, Indian, Middle East and White). The data is captured in two sessions separated by two weeks. In each session, the facial images of each person are captured under 9 different facial variations (neutral, smile, open mouth, strong light, eyes occlusion, mouth occlusion, paper occlusion, left profile and right profile. Face image samples from this database are shown in Fig. 3.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm.

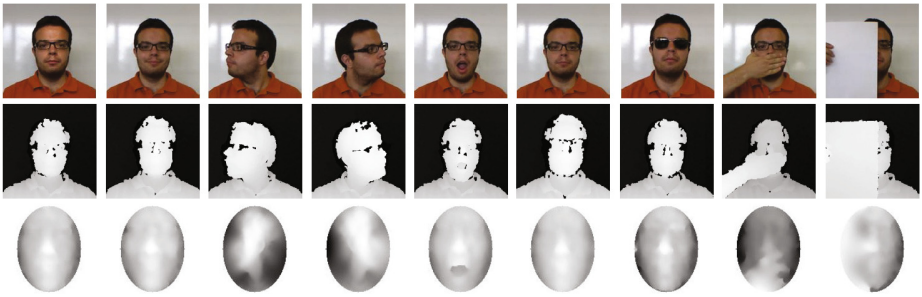


Fig. 3. Face samples from the EurecomKinect database. Top: RGB faces, middle: the corresponding raw depth maps and bottom: depth cropped face.

The CurtinFaces Kinect database [10] contains over 5000 images of 52 subjects in both RGB and depth maps obtained by Kinect sensor. The participants consist of 10 females and 42 males. Three ethnic groups (Caucasians, Chinese and Indians) are included. The facial images have various variations in pose, illumination, facial expression as well as sunglasses and hand disguise. The faces of each subject are provided with many combinations of these challenges. For each subject, there are 49 images under 7 poses and 7 facial expressions, 35 images under 5 illuminations and 7 expressions, and 5 images under disguise (sunglasses and hand). The full set for each person consists of 97 images. Face samples from this database are shown in Fig. 4.

4.2 Experimental Protocol and Setup

We considered experimental scenarios including face images under pose, illumination and expression variations. For fair evaluation, we divided each of the two databases into two subsets: development (*Dev*) and evaluation (*Eval*). In the case of EurecomKinect database, we used the images of the first session for training and those of the second session for tests. For CurtinFaces, we selected 18 and 69 images per person for training and test, respectively, so that both parts include pose, illumination and expression variations. In all the experiments, we tuned the optimal parameters of the methods on the development subset, and utilized these parameters to report the classification accuracy on the evaluation subset.

4.3 Experimental Results

Tables 1 and 2 summarize the classification performance of the four local descriptors on the two Kinect face databases for face identification, gender recognition and ethnicity classification. These results point out several findings:

– As expected, the overall classification rates indicate better performance on the EurecomKinect database (Table 1) compared to the CurtinFaces data-

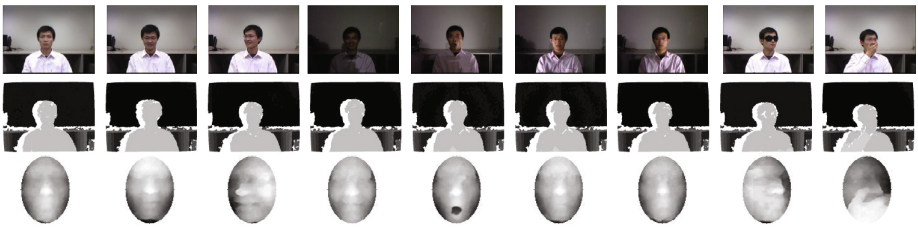


Fig. 4. Samples from the CurtinFaces database face images. Top: RGB faces, middle: their corresponding raw depth maps and bottom: depth cropped face.

Table 1. Classification rates (%) using texture (RGB), depth for facial identity, gender and ethnicity classification on EurecomKinect database

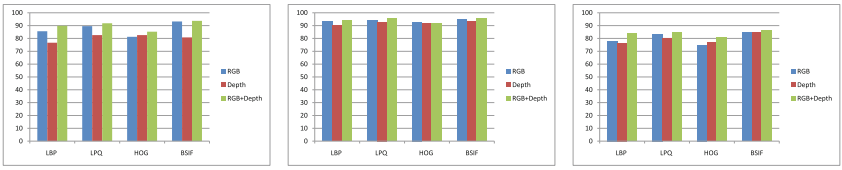
Method	Classification Rates (%)					
	Identity		Gender		Ethnicity	
	RGB	Depth	RGB	Depth	RGB	Depth
LBP [6]	100	94.2	94.2	96.1	97.1	81.7
LPQ [7]	99.0	91.3	99.0	88.4	98.0	78.8
HOG [8]	100	95.1	98.0	95.1	97.1	85.5
BSIF [9]	98.0	92.3	96.1	93.2	98.0	81.7

Table 2. Classification rates (%) using texture (RGB) and depth for facial identity, gender and ethnicity classification on CurtinFaces database

Method	Classification Rates (%)					
	Identity		Gender		Ethnicity	
	RGB	Depth	RGB	Depth	RGB	Depth
LBP [6]	85.6	76.7	93.2	90.0	78.0	76.5
LPQ [7]	89.6	82.6	94.2	92.7	83.5	79.8
HOG [8]	81.3	82.6	92.7	91.5	74.9	76.7
BSIF [9]	93.2	80.8	95.0	92.9	84.9	84.7

base (Table 2). CurtinFaces database is indeed more challenging in terms of variations of pose, expression and illumination.

- In overall, the RGB images yield in better performances compared to the depth images. Nevertheless, the results of the depth images alone are still good and actually much better than our expectations based on the human perception. It is indeed quite hard to visually distinguish the subjects using only the depth images.
- Regarding the best performing methods, the four different descriptors perform comparably on the RGB images under controlled conditions. On the depth images, HOG yields in the best classification rates under controlled environments followed by LBP and BSIF while LPQ seems to suffer the most. Under pose, expression and illumination variations, BSIF shows the highest robustness for both RGB and depth images followed by LPQ.
- A close look at the results in Table 1 and 2 indicates that gender classification is the least challenging task compared to face identification and



(a) Identity (b) Gender (c) Ethnicity

Fig. 5. Results (%) for (a) identity , (b) gender and (c) ethnicity classification using LBP, LPQ, HOG and BSIF methods on texture (RGB) and depth images and their feature level fusion (RGB+Depth) on CurtinFaces dataset. The results indicate a slight but a clear performance improvement in most cases when combining the two modalities, hence pointing out the usefulness of depth data when used together with RGB images.

ethnicity classification. This holds for both RGB and depth images and is in concordance with the findings of previous studies.

– The depth images provided by Kinect sensor are usually of low quality and noisy thus requiring a crucial preprocessing before analysis. The outcomes on such images are highly depending on the preprocessing step and hence cannot be easily generalized or compared to previously reported results if a different preprocessing is applied.

In another set of experiments, we analyzed the results of combining the RGB images with the depth information. We considered a simple feature level fusion strategy by concatenating the features extracted from RGB and depth images. As shown in figure 5, the obtained results indicated a slight but a clear performance improvement in all cases when combining the two modalities. This is also in agreement with the results of previous studies pointing out the usefulness of depth data when used together with RGB images [11].

5 Conclusion

We presented the first comprehensive study in the literature exploring the usefulness of the depth information acquired by the low-cost depth sensor, Microsoft Kinect, in different face analysis tasks including face identification, gender recognition and ethnicity classification. We experimented with four state-of-the-art local face descriptors on two publicly available Kinect face databases.

While it is difficult to visually distinguish the subjects using only the depth images, the obtained results showed that the depth information alone provides promising classification results beyond the expectations based on the human perception. This is a very interesting finding. With more accurate low-cost depth sensors for detailed facial image analysis which are expected to be available soon (e.g. Kinect 2), many face analysis problems will be much more feasible to solve.

The experiments also confirmed some findings of previous studies showing that (1) gender classification is the least challenging task compared to face identification and ethnicity classification, (2) combining the RGB images with the

depth information does provide performance enhancement and (3) the performance of Kinect depth images highly depends on the preprocessing step which is a very crucial step before further analysis.

Regarding the best performing methods, the introduced BSIF features derived from a new set of filters provide promising results for both RGB and depth images under different variations of pose, expression and illumination. This should be further investigated especially with the expected Kinect 2.

As a future work, it is of interest to extend the work to other face analysis related tasks including age estimation and kinship verification combining RGB and depth facial information.

Finally, it is worth mentioning that the findings of our work should be further confirmed with larger Kinect databases and under more challenging settings in terms of illumination and pose variations. Toward this goal, we plan to record a large Kinect face database and make it publicly available to the research community along with well-defined evaluation protocol and baseline results in order to follow the progress on using low-cost depth data in face analysis.

Acknowledgments. This work is funded by the CDTA-FNR FUSMBIO project. Authors are thankful for DGRSDT and MESRS for their support. We would like also to thank the authors of EurecomKinect [5] and Curtinfaces [10] for making these Kinect face databases publicly available for research purposes.

References

1. Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition, 2nd edn. Springer, New York (2011)
2. Han, J., Shao, L., Xu, D., Shotton, J.: Enhanced computer vision with Microsoft Kinect sensor: A review. IEEE Transactions on Cybernetics **43**(5), 1318–1334 (2013)
3. Andersen, M., Jensen, T., Lisouski, P., Hansen, A., Gregersen, T., Ahrendt, P.: Kinect depth sensor evaluation for computer vision applications. Technical report, Department of Engineering, Aarhus University, Denmark (2012)
4. Fanelli, G., Weise, T., Gall, J., Van Gool, L.: Real time head pose estimation from consumer depth cameras. In: Mester, R., Felsberg, M. (eds.) DAGM 2011. LNCS, vol. 6835, pp. 101–110. Springer, Heidelberg (2011)
5. Huynh, T., Min, R., Dugelay, J.-L.: An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In: Park, J.-I., Kim, J. (eds.) ACCV Workshops 2012, Part I. LNCS, vol. 7728, pp. 133–145. Springer, Heidelberg (2013)
6. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
7. Ahonen, T., Rahtu, E., Ojansivu, V., Heikkila, J.: Recognition of blurred faces using local phase quantization. In: 19th International Conference on Pattern Recognition, ICPR 2008, pp. 1–4, December 2008
8. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893, June 2005

9. Kannala, J., Rahtu, E.: BSIF: Binarized statistical image features. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 1363–1366 (2012)
10. Li, B., Mian, A., Liu, W., Krishna, A.: Using Kinect for face recognition under varying poses, expressions, illumination and disguise. In: 2013 IEEE Workshop on Applications of Computer Vision (WACV), pp. 186–192, January 2013
11. Goswami, G., Bharadwaj, S., Vatsa, M., Singh, R.: On RGB-D face recognition using kinect. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6, September 2013
12. Min, R., Choi, J., Medioni, G., Dugelay, J.: Real-time 3D face identification from a depth camera. In: 2012 21st International Conference on Pattern Recognition (ICPR), pp. 1739–1742, November 2012
13. Pamplona Segundo, M., Sarkar, S., Goldgof, D., Silva, L., Bellon, O.: Continuous 3D face authentication using RGB-D cameras. In: IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 64–69 (2013)
14. Huang, Y., Wang, Y., Tan, T.: Combining statistics of geometrical and correlative features for 3D face recognition. In: Proceedings of the British Machine Vision Conference, pp. 879–888, September 2006
15. Mian, A., Bennamoun, M., Owens, R.: An efficient multimodal 2D–3D hybrid approach to automatic face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(11), 1927–1943 (2007)
16. Ojansivu, V., Heikkilä, J.: Blur insensitive texture classification using local phase quantization. In: Elmoataz, A., Lezoray, O., Nouboud, F., Mammass, D. (eds.) ICISP 2008. LNCS, vol. 5099, pp. 236–243. Springer, Heidelberg (2008)
17. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 947–954 (2005)

An Overview of Research Activities in Facial Age Estimation Using the FG-NET Aging Database

Gabriel Panis and Andreas Lanitis(✉)

Visual Media Computing Lab, Department of Multimedia and Graphic Arts,
Cyprus University of Technology, Limassol, Cyprus
gpanis@gmail.com, andreas.lanitis@cut.ac.cy

Abstract. The FG-NET aging database was released in 2004 in an attempt to support research activities related to facial aging. Since then a number of researchers used the database for carrying out research in various disciplines related to facial aging. Based on the analysis of published work where the FG-NET aging database was used, conclusions related to the type of research carried out in relation to the impact of the dataset in shaping up the research topic of facial aging, are presented. In particular we focus our attention on the topic of age estimation that proved to be the most popular among users of the FG-NET aging database. Through the review of key papers in age estimation and the presentation of benchmark results the main approaches/directions in facial aging are outlined and future trends, requirements and research directions are drafted.

Keywords: Facial age estimation · Aging databases · FG-NET aging database

1 Introduction

The availability of public databases can play a crucial role in the development of a research field as it enables researchers to get engaged in research activities quickly and at the same time it promotes the idea of comparative evaluation. Especially in cases where the data collection process demands a lengthy procedure, the availability of public datasets can have a substantial impact on a field. In the research area of soft biometrics a typical example where the generation of suitable databases is, by nature, a lengthy process involves face aging datasets displaying age-separated face images of the same individual. Due to the non-availability of face aging databases, up to 2004 only a small number of researchers considered the problem of facial aging, mainly based on small in-house face datasets containing age-separated face images [37] [34] [22] [27] [26]. Back in 2004 two face aging datasets were made publicly available: The MORPH [36] and the FG-NET Aging Dataset (FG-NET-AD) [23] . When MORPH was first released it contained a large number of images but only about three instances

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 737–750, 2015.
DOI: 10.1007/978-3-319-16181-5_56

of the same person. On the other hand the FG-NET-AD contained a small number of images and subjects, but included about 12 age-separated images per subject. Despite the fact that none of the two datasets was ideal, both datasets played an instrumental role in initiating research activities in the area of facial aging. The availability of public aging databases promoted the topic of facial age estimation among the most extensively researched topics in soft biometrics.

Detailed coverage of the topics of facial aging can be found in related survey papers [35], [11] and books [9]. In [32] a technical report on the performance of age estimation algorithms is presented with emphasis given on the analysis of the results obtained rather than the adopted methodologies. Unlike the aforementioned survey papers, in this paper we concentrate our attention on the use of the FG-NET-AD. As part of this effort an analysis related to research publications reporting work where the FG-NET-AD was utilized is presented. In particular an analysis related to thematic areas of published papers over the years, an overview of the most representative papers reported in the literature and a collection of benchmark results are presented. The analysis of research outcomes related to the FG-NET-AD can be used for formulating trends and directions adopted in facial aging analysis and most importantly for shaping future directions in this area. In particular we focus our attention on research related to facial age estimation that attracted the interest of the majority of researchers working in facial aging.

The remainder of the paper is structured as follows. In section 2 information about the FG-NET-AD and statistics related to the dataset usage are presented. An overview of key papers in facial age estimation that report experimental results using the FG-NET-AD and a summary of comparative results are presented in sections 3 and 4. In Section 5 a discussion related to future research directions and needs for additional aging datasets are described, followed by concluding comments.

2 Database Description

2.1 FG-NET Project

The FG-NET-AD was generated as part of the Project FG-NET (Face and Gesture Recognition Network) [10]. FG-NET was funded by the European Union as part of the 5th Framework Programme, Information Society Technologies in the category of initiative Support Measures Networks of Excellence and Working Groups. The project consortium was comprised of the University of Manchester (UK) (project coordinator), Technological University of Munich (Germany), INRIA (France), Aalborg University (Denmark), Cyprus College (Cyprus) and IDIAP (Switzerland). One of the major aims of the project was to encourage research technology development in the area of face and gesture recognition by specifying and supplying image sets to support activities in face and gesture recognition. Within this context, among other datasets, the FG-NET-AD was generated.

2.2 Database Contents

The FG-NET-AD contains 1002 images from 82 different subjects with ages ranging between newborns to 69 years old subjects. However, ages between zero to 40 years are the most populated in the database. With the exception of images showing individuals at more recent ages, for which digital images were available, in most cases FG-NET-AD images were collected by scanning photographs of subjects found in personal collections. As a consequence the quality of images depends on the photographic skills of the photographer, the quality of the imaging equipment used, the quality of photographic paper and printing and the condition of photographs. As a result face images in the FG-NET-AD display considerable variability in resolution, quality, illumination, viewpoint and expression. Occlusions in the form of spectacles, facial hair and hats are also present in a number of images. Each image in the dataset was annotated with 68 landmark points located at key positions and also a semantic description of each image was recorded. In particular, information about the age, gender, expression, pose, image quality and appearance of occlusions (i.e. moustaches, beards, hats or spectacles) was recorded.

2.3 Analysis of Database Usage

So far the FG-NET-AD has been distributed to more than 4000 researchers, supporting in that way wide-spread aging-related research activities. In this section we present key facts related to the FG-NET-AD usage, based on a survey of the literature referencing the FG-NET-AD in articles listed at Google Scholar. Within this context the Google Scholar engine was used to search and identify the academic literature from 2005 until early 2014 that references the FG-NET-AD. A total number of 358 publications originated from 167 different institutions from 37 countries from all six continents were located. The distribution of the publications over the past 10 years is shown in Figure 1. Between 2005 to 2011 a steady increase in articles referencing the FG-NET-AD is observed, reflecting the gradual but steadily increasing interest of the research community in topics related to facial aging. The decrease from 2012 till 2014 is mainly attributed to scientific publications of the corresponding years, not indexed yet by the Google Scholar Engine.

Published papers describing research work using the FG-NET-AD were classified into the main research thematic areas of age estimation, face recognition, age progression/modeling, feature extraction/location, gender classification, biometrics, face modeling, face detection and pose estimation. Publications that were found to cover work extending across more than one thematic area were placed in all appropriate thematic areas. A significant number of other diverse thematic areas such as race classification, makeup detection, sketch matching, psychology related and perception related papers have been found in smaller numbers and have been grouped under the Other category. The distribution of papers into the main thematic areas over the years is shown in Figure 2. It is worth pointing out that although the primary scope of the dataset was to support

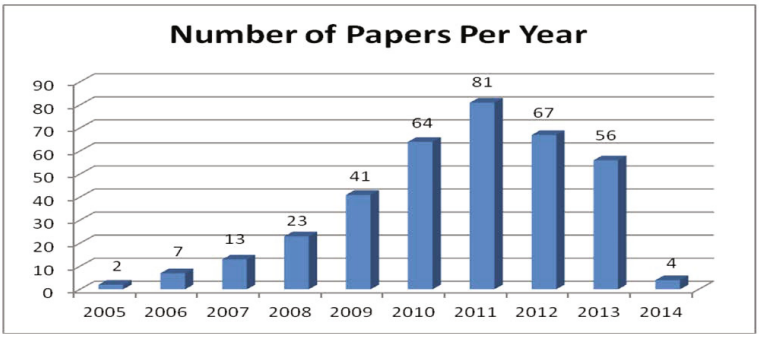


Fig. 1. Number of papers where images from the FG-NET-AD were used, per year

research in age progression, according to the findings in Figure 2, the highest share of papers are related with facial age estimation indicating the increased interest of the research community into soft biometrics.

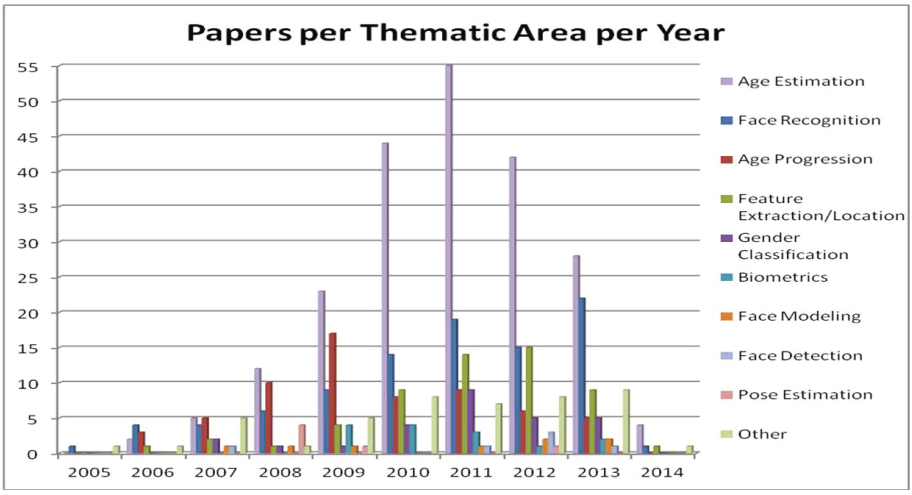


Fig. 2. Number of papers in different thematic areas per year

3 Research on Age Estimation Using the FG-NET-AD

According to the data presented in Figure 2, the topic of age estimation dominates research efforts in facial aging. The main reasons for this trend are:

i) Potential applications: Machine-based age estimation methods could figure in a wide range of applications involving man-machine interfaces such as age-adaptive interfaces and the enforcement of age-based access restrictions both to physical and electronic sites.

ii) Humans are not perfect in the task of age estimation; hence automated age estimates could complement/aid the task of human operators.

iii) The problem of age estimation, bears similarities with other standard face interpretation/pattern recognition tasks (i.e. face recognition, expression recognition etc.) hence the overall problem domain is more accessible to researchers.

iv) Accurate age estimates are usually required for other facial aging related applications (i.e. age invariant face recognition and age progression) hence the starting point in dealing with facial aging is usually the task of age estimation.

v) For age estimation there are concrete ways to test the performance of different algorithms allowing in that way the efficient comparative evaluation of different algorithms.

The output of an age estimation algorithm can be an estimate of the exact age of a person or the age group of a person. For exact age estimation the performance of an age estimation algorithm is usually based on the mean average error (MAE) between real and estimated ages over a test set and plots of Cumulative Score (CS) that shows the number of test cases which have an absolute error smaller than a given threshold. In the case of age-group age estimation errors usually refer to the percentage of correct classifications.

Researchers who carried out research in facial age estimation investigated the use of both standard pattern recognition/regression approaches and techniques adapted to the facial aging problem. In general most researchers conclude that the aging variability encountered in face images requires the use of dedicated techniques. The main trends of research activities are focused on determining suitable feature vectors that better reflect aging information in conjunction with efforts in customizing classification algorithms to take into account certain characteristics of the problem of age classification such as the problem of data sparseness i.e. the fact that for a given individual it is impossible to have training samples covering all the ages in the range of interest. A number of researchers deal with this problem by capitalizing on the observation that samples belonging to neighboring age groups display aging-related similarity even though they belong to different subjects.

Geng et al [13] generate aging patterns for each person in a dataset consisting of face images showing each subject at different ages. In this case the problem of data sparseness is addressed by filling in missing samples using the Expectation Maximization algorithm. Given a previously unseen face, the face is substituted at different positions in a pattern and the position that minimizes the reconstruction error indicates the age of the subject. Experimental results prove that this method outperformed previous approaches reported in the literature and also performed better than widely used classification methods.

A common trend in age estimation is the use of regression based on face subspace representations. Along this line Guo et al [16] propose a discriminative

subspace learning based on manifold criterion for low-dimensional representations of the aging manifold. Regression is applied on the aging manifold patterns in order to learn the relationship between coded face representations and age. A key aspect of the work described in [16] is the use of a global SVR for obtaining a rough age estimate, followed by refined age estimation using a local SVR trained using only ages within a small interval around the initial age estimate. Luu et al [31] project faces in an AAM [7] subspace and then adopt a two-stage hierarchical age estimation approach. The first stage involves the initial classification of faces into young and old followed by the use of an SVM regressor trained using images from the chosen age range, in order to get the final age estimate.

Instead of projecting faces in low dimensional subspaces, a number of researchers experimented with the use of different types of Biologically Inspired (bio-inspired) features derived from the facial area. For example Guo et al [18] propose a model that contains alternating layers called Simple and Complex cell units that resemble object recognition models of the human visual system. Features for the simple layer (S) are extracted based on Gabor filters with different scales, standard deviations and orientations. The C layer involves the use of a standard deviation function for pooling S-layer features at different bands, scales and orientations. The dimensionality of the feature vector is reduced using Principal Component Analysis and a support vector regressor is used for obtaining age estimates. The overall framework of using bio-inspired features [18], has been studied extensively both in the area of age estimation and age invariant face recognition [40]. In a more recent approach, El Dib et al [8] extract bio-inspired facial features at a fine level and information from the forehead is also utilized resulting in an error rate of 3.17 years. The trend of dealing with facial features at different levels was also adopted by Suo et al [38] who propose an age estimation algorithm based on a hierarchical face model. The model represents human faces at three levels that include the global appearance, facial components and skin zones. An age estimator is trained from the feature vectors and their corresponding age labels. Han et al [19] adopt a hierarchical approach where bio-inspired features are extracted from individual facial components. Facial components are then classified into one of four age groups and then within an age group an SVM regressor is trained to predict the age. It was found that the best performance was attained from a fusion of the best performing features, i.e. holistic bio-inspired features, shape and eye region bio-inspired features. Han et al [19] also ran an experiment involving human-based age estimation of images from the FG-NET-AD, using crowd-sourcing and the results were compared to the proposed automated method. For the FG-NET-AD, the human age estimation experiment generated a MAE of 4.7. Hong et al [20] introduce the so called biologically inspired active appearance model where instead of using pixel intensities, shape-free faces are represented by bio-inspired features Guo et al [18] during the process of AAM training. A regression-based age estimator is then used for estimating the age of samples based on the coded representations of faces.

As part of the efforts of using features related to the aging process Zhou et al [51] describe an age classification method based on the Radon transform.

Difference of Gaussians filtering is applied on the face image to extract perceptual features, which are processed using the Radon transform. An entropy-based SVM classification algorithm is then used to select features. The algorithm is tested regarding the accuracy of classifying a face as over twenty or under twenty years old. Choi et al [5] propose an age estimation method based on extracting features directly related to aging. Within this context authors propose the extraction of wrinkles using a set of region specific Gabor filters, each of which is designed based on the regional direction of wrinkles. Li et al [28] also attempt to provide a generalized framework for selecting Gabor features that preserve both global and local aging information and at the same time minimize the redundancy between features. The method was tested both on age group classification and exact age estimation.

In order to deal with the problem of data sparseness a number of researchers focused their attention on assigning age labels to different ages in a way that optimizes the training process. A method based on the relative ranking of age labels is proposed in [1]. The proposed ordinary hyperplane ranking algorithm is based on using relative ranking information and a cost-sensitive property to optimize the age estimation process. Within this context the age estimation problem is decomposed into a number of binary decisions that classify a given face into a class of faces with age greater or smaller than a given age. The combination of the results of all individual classifiers yields the final age estimation result. Chao et al [2] propose the label-sensitive concept in an attempt to take advantage of correlations that exist between different classes in age estimation. As part of this effort the learning process of samples belonging to a certain age, takes also into account weighted samples belonging to neighboring ages. The proposed formulation is used in conjunction with a customized age-oriented local regression algorithm that performs the age classification task in a hierarchical fashion. The problem of class similarity between adjacent ages is also addressed in [12] where the concept of using label distributions is introduced. Along these lines during the training process samples belonging to a certain age category contribute to the training process of the class they belong to and also to the training of adjacent classes. The proposed label distribution method was used in conjunction with the proposed IIS-LLD and CPNN label distribution learning algorithms.

The use of Neural Network-based techniques for age estimation was also investigated. Zheng et al [50] use a back propagation neural network, where the inputs are geometrical features and local binary patterns, in order to classify faces into juveniles and adults. Yin and Geng [47] use a Conditional Probability Neural Network where the inputs are a facial descriptor and an age estimate and the output is the probability that the face descriptor is extracted from a face showing the given age. Based on this methodology the training process for a certain age takes into account faces showing the exact age and also samples with other ages enlarging in that way the training set. As a result the learning process is more efficient.

The majority of age estimation methods reported in the literature are based on texture-based features. In contrast Thukral et al [39] use face shape landmarks

information in a hierarchical approach where the test image is first classified into an age group using several classifiers fused using the majority rule. Then the Relevance Vector Machine regression model of that age group is used to estimate the age. Wu et al [41] rely on facial shapes represented by point-coordinates on a Grassmann manifold. Based on this framework, the so called aging signature is extracted for each sample, by considering the tangent vectors of the deformation needed to deform a given face shape to the average face shape. A regressor-based age estimator that relates aging signatures to age is used during the age estimation process. This framework was also tested in the task of face verification.

4 Age Estimation FG-NET-AD Benchmark Results

In Table 1 we present a summary of age estimation results reported in the literature, in relation to experiments using images from the FG-NET-AD. The summary of the results presented can act as a benchmark for new age estimation experiments involving the FG-NET-AD. Most researchers reporting results using the FG-NET-AD adopted the Leave One Person Out (LOPO) approach where for each of the 82 subjects in the database, an age estimator is trained using images of the remaining 81 subjects and the results are averaged over the 82 trials. Given the small number of images available in the FG-NET-AD this is the optimum and recommended approach. The current benchmark for age estimation is the work of El Dib et al [8] where a mean average error of 3.17 is recorded when the LOPO approach is used. It is worth quoting that within three years of the publication of the first standardized age estimation results based on the LOPO method [13] reported MAE were almost halved [8], indicating in this way the benefits of standardised comparative evaluation. In the case of human-age estimation the recorded benchmark is 4.7 when all images from the FG-NET-AD were processed through crowd-sourcing by 10 volunteers [19]. Geng et al [12] also report a human age estimation MAE of 6.23 derived based on the observations of 29 volunteers using a sub-sample of 51 FG-NET-AD images. Clearly a number of reported algorithms match and even better the indicative performance of humans as recorded in [19] and [12].

In general two main trends seem to form the current directions in age estimation: The first is the use of bio-inspired features [18], [8], [20] and the second is the exploitation of age label distributions and ranking [2], [1], [12]. In addition efforts in investigating feature extraction from facial areas that contain increased age related information [19] also show promise.

5 Discussion

The availability of two publicly available aging databases (MORPH and FG-NET-AD) played an important role in initiating an increased interest in research related to facial aging among the computer vision community. According to the analysis of published work, the topic of facial age estimation has been the most

Table 1. A summary of Age Estimation Results using Images from the FG-NET-AD

Reference	**Method**	**Train-Test Images**	**Result (MAE)**
Ni2009 [33]	Multi-Instance Regression	600 train 402 test	9.49
Zhou2005 [52]	Regression using Boosting	800 train 202 test	5.81
Xiao2009 [42]	Regression with distance metric	300 train 702 test	5.04
Luu2009 [31]	2 stage SVR in AAM subspace	802 train 200 test	4.37
Han2013 [19]	**Human age estimation**	Entire FG-NET-AD	**4.7**
Geng2013 [12]	**Human age estimation**	51 images	**6.23**
Geng2007 [13]	Aging Pattern Subspace	LOPO	6.22
Thukral2012 [39]	Fused classifiers using shape	LOPO	6.2
Gunay2013[15]	Radon Features	LOPO	6.18
Wu2012 [41]	Grassmann manifold	LOPO	5.89
Yan2007 [45]	Regressor with uncertain labels	LOPO	5.78
Yan2007 [44]	Ranking with uncertain labels	LOPO	5.33
Yan2009 [43]	Submanifold Embedding	LOPO	5.21
Ylioinas2013[48]	Binary Pattern Density Estimate	LOPO	5.09
Guo2008 [16]	Manifold Learning and Regressor	LOPO	5.07
Kilinc2013[21]	Geometric and Gabor Binary Pattern	LOPO	5.05
Guo2008 [17]	Probabilistic Fusion Approach	LOPO	4.97
Liang2014[29]	Multi-feature ordinal ranking	LOPO	4.97
Yan2008 [46]	Regression from patch kernel	LOPO	4.95
Zhang2013 [49]	Hierarchical Model	LOPO	4.89
Li2012 [28]	Ordinal Discriminative Features	LOPO	4.82
Guo2009 [18]	Bio-inspired features (BIF)	LOPO	4.77
Yin2012 [47]	Probability Neural Network	LOPO	4.76
Geng2013 [12]	Learning Label Distribution	LOPO	4.76
Chen2013 [3]	Cumulative Attribute SVR	LOPO	4.67
Han2013 [19]	Component and holistic BIF	LOPO	4.6
Chen2013[4]	Pairwise Age Ranking	LOPO	4.56
Chang2011 [1]	Ordinal hyperplanes ranker	LOPO	4.48
Chao2013 [2]	Label-sensitive regression	LOPO	4.38
Hong2013 [20]	Bio-Inspired AAM	LOPO	4.18
El Dib2010 [8]	**Enhanced Bio-Inspired features**	LOPO	**3.17**

popular among researchers active in the area of facial aging. The increased interest in this area resulted in the development of robust age estimation algorithms capable of providing age estimates that can be used in most applications requiring user age information. It is anticipated that future research directions in age estimation will focus on the following issues:

i) Dealing with unconstraint face images: Developing age estimation algorithms that can deal with images captured under completely unconstraint conditions such as the ones captured by surveillance cameras. The ability of performing age estimation using unconstraint images will extend the range of possible applications where age estimation systems can be used.

ii) Age estimation based on video sequences: Currently almost all research efforts in age estimation deal with static images. However, temporal information that includes both face movements and expressions can also provide important age-related clues. A similar scenario was encountered in expression recognition that gradually moved away from dealing with static images as it became obvious that facial movements are also important for interpreting expressions [6].

iii) Multi-modal age estimation: Apart from the face, aging also affects other parts of the body [25] hence information fusion from different modalities could lead to more accurate age estimation systems. Although Some attempts of developing age estimation based on individual biometric modalities, such as gait [30], head movements [24] and fingerprints [14] were reported in the literature. It is anticipated that the topic of multi-modal biometric age estimation will attract substantial research interest in the near future.

iv) Age estimation results have reached error levels that make them suitable for several applications. However, there is still room for further improvements in age estimation tasks involving ages traditionally used as age thresholds (i.e age of 12, 15 and 18). For these particular ages additional research is required in order to further minimize age estimation errors.

In order to support the scenarios stated above, there is a clear need for developing new aging datasets that contain unconstraint images, video sequences and multi-modal biometric samples.

6 Conclusions

The FG-NET-AD was released back in 2004 in an attempt to encourage and promote research in the new (at that time) research topic of facial aging. It was fortunate that the release of the FG-NET-AD coincided with the release of the MORPH aging database [36] that provided different type of data and as a result supported complementary experimental investigations. Based on the analysis of published work, review of several key age estimation papers and analysis of the results reported, it is evident that the scopes that lead to the generation and distribution of the FG-NET-AD are fulfilled. A large number of researchers have benefited from using the database and as a result the topic of facial age estimation is now an established and well studied research area in computer

vision and the emerging field of soft biometrics in particular. Although the FG-NET-AD can still be used for supporting research related to facial aging, it is imperative that new aging databases are made available in order to support new types of experimentations that will further advance research in age estimation.

Acknowledgments. The generation of the FG-NET-AD was funded by the EU Project FG-NET, 5th FP. We would like to thank Mr E. Elefteriou who collected and annotated the images for the FG-NET-AD and Mrs A. Parmaxi for collecting data regarding the usage of the FG-NET-AD.

References

1. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 585–592. IEEE (2011)
2. Chao, W.L., Liu, J.Z., Ding, J.J.: Facial age estimation based on label-sensitive learning and age-oriented regression. Pattern Recognition **46**(3), 628–641 (2013)
3. Chen, K., Gong, S., Xiang, T., Loy, C.C.: Cumulative attribute space for age and crowd density estimation. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2467–2474. IEEE (2013)
4. Chen, Y., Hsu, C.: Subspace learning for facial age estimation via pairwise age ranking. IEEE Transactions on Information Forensics and Security **8**(12), 2164–2176 (2013)
5. Choi, S.E., Lee, Y.J., Lee, S.J., Park, K.R., Kim, J.: Age estimation using a hierarchical classifier based on global and local facial features. Pattern Recognition **44**(6), 1262–1281 (2011)
6. Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Computer Vision and Image Understanding **91**(1), 160–187 (2003)
7. Cootes, T.F., Edwards, G.J., Taylor, C.J., et al.: Active appearance models. IEEE Transactions on pattern analysis and machine intelligence **23**(6), 681–685 (2001)
8. El Dib, M.Y., El-Saban, M.: Human age estimation using enhanced bio-inspired features (ebif). In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 1589–1592. IEEE (2010)
9. Fairhurst, M.: Age factors in biometric processing. The Institution of Engineering and Technology (2013)
10. FG-NET (Face and Gesture Recognition Network) (2014). http://www-prima.inrialpes.fr/FGnet/ Accessed 10 June 2014
11. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(11), 1955–1976 (2010)
12. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(10), 2401–2412 (2013)
13. Geng, X., Zhou, Z.H., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2234–2240 (2007)
14. Gnanasivam, P., Muttan, D.S.: Estimation of age through fingerprints using wavelet transform and singular value decomposition. International Journal of Biometrics and Bioinformatics (IJBB) **6**(2), 58 (2012)

15. Günay, A., Nabiyev, V.V.: Age estimation based on local radon features of facial images. In: Computer and Information Sciences III, pp. 183–190. Springer (2013)
16. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: Image-based human age estimation by manifold learning and locally adjusted robust regression. IEEE Transactions on Image Processing **17**(7), 1178–1188 (2008)
17. Guo, G., Fu, Y., Dyer, C.R., Huang, T.S.: A probabilistic fusion approach to human age prediction. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. CVPRW 2008, pp. 1–6. IEEE (2008)
18. Guo, G., Mu, G., Fu, Y., Huang, T.S.: Human age estimation using bio-inspired features. In: IEEE Conference on Computer Vision and Pattern Recognition CVPR 2009, pp. 112–119. IEEE (2009)
19. Han, H., Otto, C., Jain, A.K.: Age estimation from face images: Human vs. machine performance. In: 2013 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2013)
20. Hong, L., Wen, D., Fang, C., Ding, X.: A new biologically inspired active appearance model for face age estimation by using local ordinal ranking. In: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, pp. 327–330. ACM (2013)
21. Kilinc, M., Akgul, Y.S.: Automatic human age estimation using overlapped age groups. In: Csurka, G., Kraus, M., Laramee, R.S., Richard, P., Braz, J. (eds.) VISIGRAPP 2012. CCIS, vol. 359, pp. 313–325. Springer, Heidelberg (2013)
22. Kwon, Y.H., da Vitoria Lobo, N.: Age classification from facial images. In: 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Proceedings CVPR 1994, pp. 762–767. IEEE (1994)
23. Lanitis, A.: Comparative evaluation of automatic age-progression methodologies. EURASIP Journal on Advances in Signal Processing 2008 (2008)
24. Lanitis, A.: Age estimation based on head movements: A feasibility study. In: 2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP), pp. 1–6. IEEE (2010)
25. Lanitis, A.: A survey of the effects of aging on biometric identity verification. International Journal of Biometrics **2**(1), 34–52 (2010)
26. Lanitis, A., Draganova, C., Christodoulou, C.: Comparing different classifiers for automatic age estimation. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **34**(1), 621–628 (2004)
27. Lanitis, A., Taylor, C.J., Cootes, T.F.: Toward automatic simulation of aging effects on face images. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(4), 442–455 (2002)
28. Li, C., Liu, Q., Liu, J., Lu, H.: Learning ordinal discriminative features for age estimation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2570–2577. IEEE (2012)
29. Liang, Y., Wang, X., Zhang, L., Wang, Z.: A hierarchical framework for facial age estimation. Mathematical Problems in Engineering 2014 (2014)
30. Lu, J., Tan, Y.P.: Gait-based human age estimation. IEEE Transactions on Information Forensics and Security **5**(4), 761–770 (2010)
31. Luu, K., Ricanek, K., Bui, T.D., Suen, C.Y.: Age estimation using active appearance models and support vector machine regression. In: IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems, 2009. BTAS 2009, pp. 1–5. IEEE (2009)
32. Ngan, M., Grotherh, P.: Face recognition vendor test (frvt) - performance of automated age estimation algorithms. NIST Interagency Report 7995 (2014)

33. Ni, B., Song, Z., Yan, S.: Web image mining towards universal age estimator. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 85–94. ACM (2009)
34. O'Toole, A.J., Price, T., Vetter, T., Bartlett, J., Blanz, V.: 3d shape and 2d surface textures of human faces: the role of averages in attractiveness and age. Image and Vision Computing **18**(1), 9–19 (1999)
35. Ramanathan, N., Chellappa, R., Biswas, S.: Computational methods for modeling facial aging: A survey. Journal of Visual Languages & Computing **20**(3), 131–144 (2009)
36. Ricanek, K., Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression. In: 7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006, pp. 341–345. IEEE (2006)
37. Rowland, D.A., Perrett, D.I.: Manipulating facial appearance through shape and color. IEEE Computer Graphics and Applications **15**(5), 70–76 (1995)
38. Suo, J., Wu, T., Zhu, S., Shan, S., Chen, X., Gao, W.: Design sparse features for age estimation using hierarchical face model. In: 8th IEEE International Conference on Automatic Face & Gesture Recognition, 2008. FG 2008, pp. 1–6. IEEE (2008)
39. Thukral, P., Mitra, K., Chellappa, R.: A hierarchical approach for human age estimation. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1529–1532. IEEE (2012)
40. Wang, S., Xia, X., Qing, Z., Wang, H., Le, J.: Aging face identification using biologically inspired features. In: 2013 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC), pp. 1–5. IEEE (2013)
41. Wu, T., Chellappa, R.: Age invariant face verification with relative craniofacial growth model. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VI. LNCS, vol. 7577, pp. 58–71. Springer, Heidelberg (2012)
42. Xiao, B., Yang, X., Xu, Y., Zha, H.: Learning distance metric for regression by semidefinite programming with application to human age estimation. In: Proceedings of the 17th ACM international conference on Multimedia, pp. 451–460. ACM (2009)
43. Yan, S., Wang, H., Fu, Y., Yan, J., Tang, X., Huang, T.S.: Synchronized submanifold embedding for person-independent pose estimation and beyond. IEEE Transactions on Image Processing **18**(1), 202–210 (2009)
44. Yan, S., Wang, H., Huang, T.S., Yang, Q., Tang, X.: Ranking with uncertain labels. In: 2007 IEEE International Conference on Multimedia and Expo, pp. 96–99. IEEE (2007)
45. Yan, S., Wang, H., Tang, X., Huang, T.S.: Learning auto-structured regressor from uncertain nonnegative labels. In: IEEE 11th International Conference on Computer Vision. ICCV 2007, pp. 1–8. IEEE (2007)
46. Yan, S., Zhou, X., Liu, M., Hasegawa-Johnson, M., Huang, T.S.: Regression from patch-kernel. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2008, pp. 1–8. IEEE (2008)
47. Yin, C., Geng, X.: Facial age estimation by conditional probability neural network. In: Liu, C.-L., Zhang, C., Wang, L. (eds.) CCPR 2012. CCIS, vol. 321, pp. 243–250. Springer, Heidelberg (2012)
48. Ylioinas, J., Hadid, A., Hong, X., Pietikäinen, M.: Age estimation using local binary pattern kernel density estimate. In: Petrosino, A. (ed.) ICIAP 2013, Part I. LNCS, vol. 8156, pp. 141–150. Springer, Heidelberg (2013)

49. Zhang, L., Wang, X., Liang, Y., Xie, L.: A new method for age estimation from facial images by hierarchical model. In: Proceedings of the Second International Conference on Innovative Computing and Cloud Computing, p. 88. ACM (2013)
50. Zheng, Y., Yao, H., Zhang, Y., Xu, P.: Age classification based on back-propagation network. In: Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service, pp. 319–322. ACM (2013)
51. Zhou, H., Miller, P., Zhang, J.: Age classification using radon transform and entropy based scaling svm. In: BMVC, pp. 1–12 (2011)
52. Zhou, S.K., Georgescu, B., Zhou, X.S., Comaniciu, D.: Image based regression using boosting method. In: Tenth IEEE International Conference on Computer Vision. ICCV 2005. vol. 1, pp. 541–548. IEEE (2005)

Gender Classification from Iris Images Using Fusion of Uniform Local Binary Patterns

Juan E. Tapia[1(✉)], Claudio A. Perez[1(✉)], and Kevin W. Bowyer[2(✉)]

[1] Department of Electrical Engineering and Advanced Mining Technology Center, Universidad de Chile, Santiago, Chile
{jtapiafarias,clperez}@ing.uchile.cl

[2] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, USA
kwb@nd.edu

Abstract. This paper is concerned in analyzing iris texture in order to determine "soft biometric", attributes of a person, rather than identity. In particular, this paper is concerned with predicting the gender of a person based on analysis of features of the iris texture. Previous researchers have explored various approaches for predicting the gender of a person based on iris texture. We explore using different implementations of Local Binary Patterns from the iris image using the masked information. Uniform LBP with concatenated histograms significantly improves accuracy of gender prediction relative to using the whole iris image. Using a subject-disjoint test set, we are able to achieve over 91 % correct gender prediction using the texture of the iris. To our knowledge, this is the highest accuracy yet achieved for predicting gender from iris texture.

Keywords: Biometrics · Iris · LBP · Gender classification

1 Introduction

Whenever people log onto computers, access an ATM, pass through airport security, use credit cards, or enter high-security areas, they need to verify their identities [1]. Thus, there is tremendous interest in improved methods for reliable and secure identification of people. Gender classification based on iris images is currently one of the most challenging problems in image analysis research [2,3]. In a biometric recognition framework, gender classification can help by requiring a search of only half of the subjects in the database [4].

One active area of "soft biometric" research involves classifying the gender of the person from the biometric sample. Most work done on gender classification has involved the analysis of face images and uses Local Binary Patterns (LBP) to increase the accuracy of the identification task [5]. Various types of classifiers have been used in gender classification after feature extraction and selection. Gender recognition is a fundamental task for human beings, as many social

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 751–763, 2015.
DOI: 10.1007/978-3-319-16181-5_57

functions critically depend on the correct gender perception. Automatic gender classification has many important applications, for example, intelligent user interface, visual surveillance, collecting demographic statistics for marketing, etc. Human faces provides important visual information for gender classification.

Gender classification from face images has received much research interest in the last two decades. Moghaddam and Yang [6] were the first to report the SVM with the Radial Basic Function kernel (SVM+RBF) as the best gender classifier. More recently, Makinen and Raisano [7] compared the performance of SVM with other classifiers including neural networks and Adaboost. According to their published results, SVM achieved the highest performance. In [4,8] was reported the extension of the use of feature selection based on mutual information and features fusion to improve gender classification of face images. The authors compare the results of fusing 3 groups of features, 3 spatial scales and 4 different mutual information measures to select features. They also showed improved results by fusion of LBP features with different radii and spatial scales, and the selection of features using mutual information.

Gender classification using iris information is a rather new topic, with only a few papers published [2,3,9]. Most gender classification methods reported in the literature use all iris texture features for classification or periocular images [10,11] and using LBP for identification. As a result, gender-irrelevant information might be fed into the classifier which may result in poor generalization, especially when the training set is small. It has been shown both theoretically and empirically that reducing the number of irrelevant or redundant features increases the learning efficiency of the classifier [12].

Thomas et al. [3] were the first to explore gender-from-iris, using images acquired with an LG 2200 sensor. They segmented the iris region, created a normalized iris image, and then a log-Gabor filtered version of the normalized image. In addition to the log-Gabor texture features, they used seven geometric features of the pupil and iris, and were able to reach a gender-prediction accuracy close to 80%.

Lagree et al. [2] experimented with iris images acquired using an LG 4000 sensor. Their work differs from Thomas [3] in several ways. They computed texture features separately for eight five-pixel horizontal bands, running from the pupil-iris boundary out to the iris sclera boundary, and ten twenty-four-pixel vertical bands from a 40x240 image. The normalized image is not processed by the log-Gabor filters that are used by IrisBEE software [13] to create the "iris code" for recognition purpose and no geometrics features are used. This approach reached an accuracy close to 62% for gender and close to 80% for ethnicity.

Bansal et al. [9] experimented with iris images acquired with a Cross Match SCAN-2 dual-iris camera. A statistical feature extraction technique based on correlation between adjacent pixels was combined with a 2D wavelet tree based on feature extraction techniques to extract significant features from the iris image. This approach reached an accuracy of 83.06% for gender classification. Nevertheless, the database used in this experiment was very small (300 images) compared to other studies published in the literature.

Actually numerous variants of LBP descriptors have been proposed in the last years [14–17]. Several works only utilized the uniform patterns but combining uniform patterns with a few non-uniform patterns was shown to improve performance [18,19].

In this paper we propose a new method to extract information from the iris image to improve gender classification. We first extract texture information in details using small windows and then concatenate the histogram information. Results indicate that each window contains useful information for gender classification. We also consider using overlapping windows, in order to obtaining a more representative histogram. Results indicate that using a subset of the iris region gives greater accuracy than using only the whole iris region. We then explore different implementations using traditional LBP, uniform histogram and concatenated histogram of overlapped windows. We are able to achieve over 91% correct gender classification with the Uniform LBP(8,1).

2 Methods

The iris feature extraction process involves the following steps. First, a camera acquires an image of the eye. All commercial iris recognition systems use near-infrared illumination, to be able to image iris texture of both "dark" and "light" eyes. Next, the iris region is located within the image. The annular region of the iris is transformed from raw image coordinates to normalized polar coordinates. This results in what is sometimes called an "unwrapped" or "rectangular" iris image. A texture filter is applied at a grid of locations on this unwrapped iris image, and the filter responses are quantized to yield a binary iris code [1]. Iris recognition systems operating on these principles are widely used in a variety of applications around the world.

The radial resolution (r) and angular resolution (θ) used during the normalization or "unwrapping" stage determine the size of the rectangular iris image, and can significantly influence the iris recognition rate. This unwrapping is referred to as using Daugman's rubber sheet model [20]. In this work we use a rectangular image of 20 (r) x 240 (θ), created using IrisBEE implementation, as illustrated in Figure 1.

The implementation also creates a segmentation mask of the same size as the rectangular image, masked by default 25% of fragile bits [21]. When using fragile bit masking, we mask a significant amount of information because it is not "stable". Rather than completely ignoring all of the fragile bits of information, we would like to find a different way of use those bits. We know that the values (zero/one) of those bits are not stable. However, the physical locations of those bits should be stable and might be used to improve our gender classification performance.

The segmentation mask indicates the portions of the normalized iris image that are not valid due to occlusion by eyelids, eyelashes or specular reflections (See, Figure 2.)

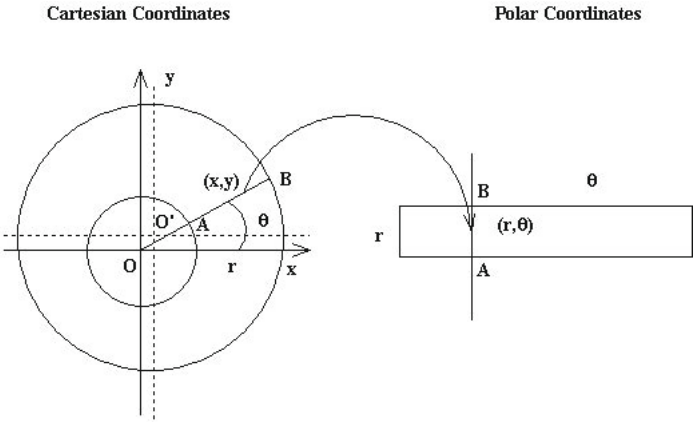


Fig. 1. Transformation of Cartesian coordinates (x, y) to Polar coordinate (r, θ) for generating the Unwrapper image

For the encoding stage, the output of the Gabor filters is transformed into the binary iris code by quantizing the phase information into four levels, for each possible quadrant in the complex plane. In coding only the phase information, the iris code keep only the most stable information of the iris, while discarding redundant or noisy information, which is represented by the amplitude component [20].

The points at which the filter is applied can be viewed as sampling at increments along the radial distance between the pupil-iris boundary and the iris-sclera boundary and at increments of angular distance around the iris. At each point that the filter is applied, a complex-valued result is obtained. The real part and the imaginary part of each result are each quantized to 0/1, giving two bits of iris code for each texture filter result.

Liu et al. [13] have collected a large data set of iris images, intentionally sampling a range of quality broader than that used by current commercial iris recognition systems. The author re- implemented the Daugman-like iris recognition algorithm developed by Masek [22] and also developed and implemented an improved iris segmentation and eyelid detection stage of the algorithm called

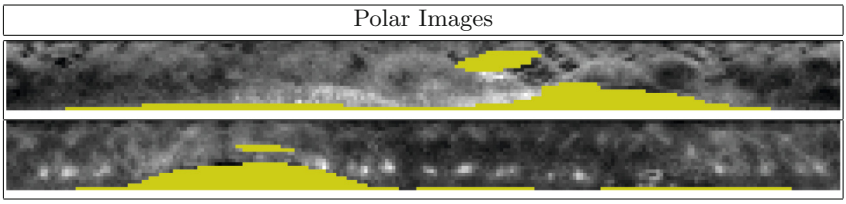


Fig. 2. Representation of polar image from the segmented iris region. The iris region is "unwrapped" to a rectangular image. The segmented areas of iris occlusion are shown in yellow.

IrisBEE, and experimentally verified the improvement in recognition performance using the collected dataset. Compared to Masek' s original segmentation approach, this improved segmentation algorithm leads to an increase of over 6% in the rank-one recognition rate.

Figure 3 shows examples of the original image for a female eye with the corresponding segmentation and unwrapped image.

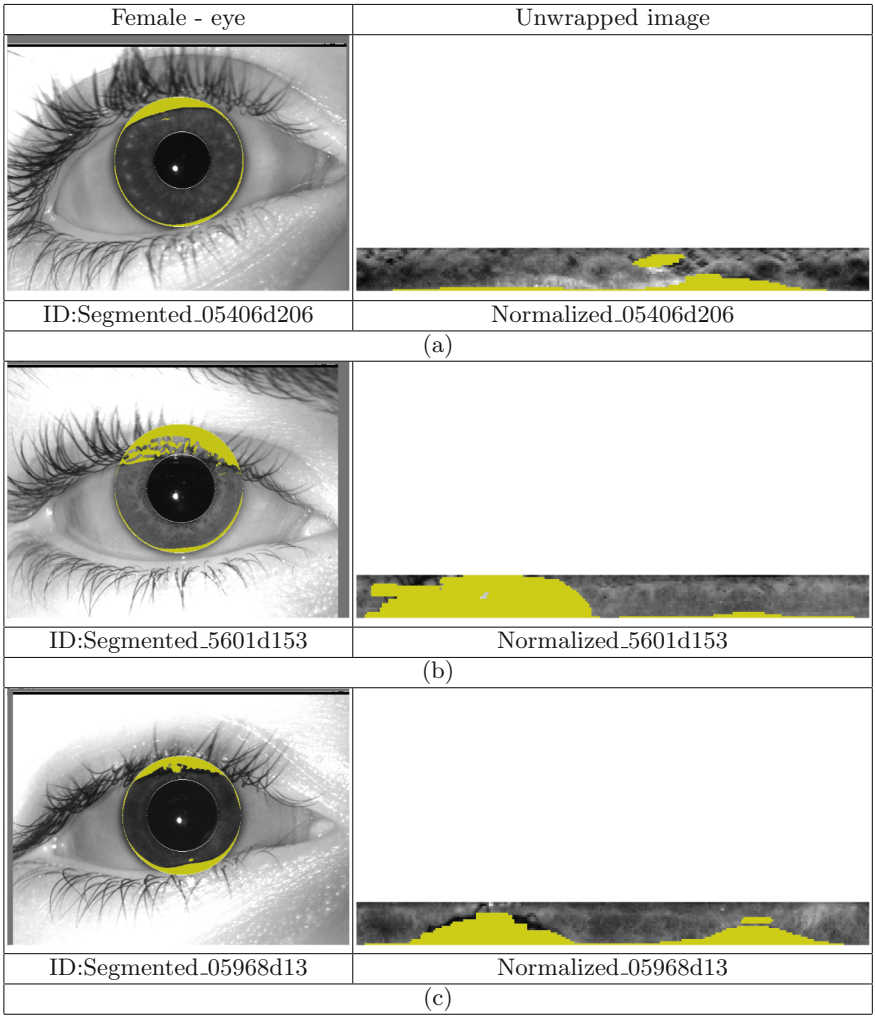


Fig. 3. Original images from a female subject with eyelids and eyelashes detection using IrisBEE implementation. The Images (a), (b) and (c) represent segmented and normalized image.

In this research, iris images were divided into 48 sub-regions, using windows size of 10x10 without overlapping and 59 bins for the LBP histogram. The $LBP(8, 1.u2)$ operator was adopted to extract LBP features.

The aim of this work is to find the best way for describing a given texture using a local binary pattern (LBP). First, several different approaches are compared, then the best fusion approach is tested and compared with several approaches proposed in the literature.

An SVM classifier with Gaussian kernel was trained using a LIBSVM implementation [23] with ten fold cross-validation procedure.

2.1 Local Binary Patterns (LBP)

LBP is a gray-scale texture operator which characterizes the spatial structure of the local image texture. Given a central pixel in the image, a binary pattern number is computed by comparing its value with those of its neighbors. The original operator used a 3x3 windows size. LBP features were computed from relative pixels intensities in a neighborhood.

$$LBP_{P,R}(x, y) = \bigcup_{(x',y')\in N(x,y)} h(I(x, y), I(x', y')) \tag{1}$$

where $N(x, y)$ is vicinity around (x, y),$\cup$ is the concatenation operator, P is number of neighbors and R is the radius of the neighborhood.

LBP was first introduced in [14] showing high discriminative power in distinguishing texture features, and is widely used for face analysis. As the neighborhood consists of 8 pixels, a total of $2^8 = 256$ different labels can be obtained depending on the relative gray values of the center and the pixels in the neighborhood (See, Figure 4.)

Later, in [17] the uniform local binary pattern (ULBP) was introduced, extending the original LBP operator to circular neighborhood with a different radius size and a small subset of LBP patterns selected. A uniformity measure of a pattern is used: U ("pattern") is the number of bitwise transitions from 0 to 1 or vice versa when the bit pattern is considered circular. A local binary pattern is called uniform if its uniformity measure is at most 2. For example, the patterns 00000000 (0 transitions), 01110000 (2 transitions) and 11001111 (2 transitions) are uniform whereas the patterns 11001001 (4 transitions) and 01010011 (5 transitions) are not. In uniform LBP mapping there is a separate output label for each uniform pattern and all the non-uniform patterns are assigned to a single label. Thus, the number of different output labels for mapping for patterns of P *bits* is $P(P - 1) + 3$. For instance, the uniform mapping produces 59 output labels for neighborhoods of 8 sampling points, and 243 labels for neighborhoods of 16 sampling points.

The reasons for omitting the non-uniform patterns are twofold. First, most of the local binary patterns in natural images are uniform. It was noticed experimentally in [14] that uniform patterns account for a bit less than 90% of all patterns when using the (8, 1) neighborhood. In experiments with facial images,

it was found that 90.6% of the patterns in the (8, 1) neighborhood and 85.2% of the patterns in the (8, 2) neighborhood are uniform. The second reason for considering uniform patterns is the statistical robustness. Using uniform patterns instead of all the possible patterns has produced better recognition results in many applications [4,19]. On one hand, there are indications that uniform patterns themselves are more stable, i.e. less prone to noise and on the other hand, considering only uniform patterns makes the number of possible LBP labels significantly lower and reliable estimation of their distribution requires fewer samples.

Rotation invariant patterns have been explored in [16], where patterns that represent 80% of all the patterns in training data are used. The uniform patterns allows to see the LBP method as a unifying approach to the traditionally divergent statistical and structural models of texture analysis.

In [17], was proposed CLBP using both the sign and magnitude information in the difference d between the central pixel, q_c, and some pixel in its neighborhood q_p.

In conventional LBP operator only the sign component of d is utilized. If $d_p = q_p - q_c$ its sign h is as we see above in Eq. (1), $h(d_p) = 1$ if $d_p \geq 0$, otherwise 0. CLBP utilizes the magnitude m_p of d_p, where $m_p = \|d_p\|$, for additional discriminant power. CLBP also considers the intensity of the central pixel, q_c. Thus, three operators are defined in CLBP:

CLBP_S, which considers the sign component of the difference, CLBP_M, which considers the magnitude component of the difference, and CLBP_C, which considers the intensity of the central pixel.

CLBP_S is the conventional LBP sign operator $h(x)$.

CLBP_M is defined as follows:

$$CLBP_{M_{P,R}} = \sum_{p=0}^{P-1} t(m_p, c)2^P \tag{2}$$

Where $t(x) = 1$ if $x \geq 0$, otherwise 0, and c is the mean value of absolute value of the differences between a pixel and one neighbor.

CLB_C is defined as follow:

$$CLBP_{C_{P,R}} = t(q_p - \tau_1) \tag{3}$$

where $t(x)$ is defined as in Eq.(2) and $\tau_1$is the average gray level of entire image. These three codes are then combined to form CLBP feature map of the original image.

In [15] was proposed Local Binary Pattern Histogram Fourier features (LBP-HF), a novel rotation invariant image descriptor computed from discrete Fourier transforms of local binary pattern (LBP) histograms. Unlike most other histogram based invariant texture descriptors which normalize rotation locally, the proposed invariants are constructed globally for the whole region to be described. In addition to being rotation invariant, the LBP-HF features retain the highly discriminative nature of LBP histograms.

2.2 Dataset

The images used in this paper were taken with an LG 4000 sensor. The LG 4000 uses near-infrared illumination and acquires a 480x640, 8-bit/pixel image. Example LG 4000 iris images appear in Figures 5. We used the UND iris database to train and test a gender classifier. The image dataset for this work consists of one left eye image and one right eye image for each of 750 males and 750 females, for a total of 3,000 images. This dataset is available to other researcher. Additional details and the release agreement are available at: http://www3.nd.edu/~cvrl/CVRL/Data_Sets.html.

For each subject, one left eye image was selected at random from their set of left eye images, and one right eye image was selected at random from their right eye images.

A training portion of the dataset was created by randomly selecting 80% of the males and 80% of the females, and the images for the remaining 20% of males and 20% of females was set aside as the test portion.

In this paper, experiments are conducted separately for the left eye and the right eye. This reflects the fact that historically many iris recognition applications use an image from only one eye rather than from both eyes. Because the left eye image and the right eye image for a given subject were generally not acquired in the same session, there may be differences in illumination, eyelid occlusion, or pose between the left and right eye images of a person. (For example, see Figure 5.)

2.3 Experiments

In this paper, we present different experiments for gender classification from the iris image. A significant limitation of the original LBP operator is its small spatial support area. Features calculated in a local 3x3 neighborhood cannot capture large-scale structures that may be the dominant features of some textures. A straightforward way of enlarging the spatial support area is to combine the information provided by N LBP operators with varying windows size. This way, each pixel in an image gets N different LBP codes. The most accurate information would be obtained by using the joint distribution of these codes.

The first approach that we explore is based on histogram of LBP features (LBPH) using uniform features $ULBP\ (8, 1)$, where we use 48 windows with size of 10x10 pixels. This represents two vertical regions each with 24 horizontal regions without overlap between regions and concatenated histograms. This approach results in the feature vector for an image having 2,582 values (2 vertical regions x 24 horizontal regions x 59 bins=2,582).

In the second approach, we use the same size of windows but using overlapping of 50%. This way more sub-windows over iris images could be obtained from each image (4 vertical regions x 48 horizontal region x 59 bins=11,328). Each pixel is labeled with the code of the texture primitive that best matches the local neighborhood. Thus each LBP code can be regarded as a micro-texton. Local primitives detected by the LBP include spots, flat areas, edges, edge ends, curves.

3 Results

Table 1 shows the gender classification rate (and standard deviation) obtained using different LBP implementations with the set of the left iris images. The first columns identifies the LBP implementation. The second columns lists the classification rate, which is also broken down by gender in columns 3 and 4.

The top row of Table 1 shows the gender classification accuracy obtained using the intensity values of the whole polar image (20x240), without any texture features extracted. The accuracy is 78.52% +/- 1.70.

The second row of Table 1 shows the classification accuracy using the traditional uniform LBP over the entire image, without overlapping and windows. The accuracy actually decreases substantially compared to using no feature extraction. Accuracy using LBP from this feature extraction method reaches about 71.33% +/- 0.80.

The third row of Table 1 shows the classification accuracy achieved using the Complete LBP using only the magnitude, over the entire image. The accuracy in this case decreases over that ULBP, reaching only 65.33% +/- 0.90.

The fourth row of Table 1 shows the classification accuracy achieved using the Complete LBP using only the sign, over the entire image. The accuracy in this case decreases over that ULBP and CLBP-Mag, reaching only 60.33% +/- 0.80.

The fifth and sixth rows of Table 1 show the classification accuracy achieved using the Complete uniform LBP using the magnitude and sign respectively, over the entire image. The accuracy in this case increases over that previous implementation, reaching 81.33% +/- 0.50 and 77.33% +/- 0.50 respectively.

The seventh row of Table 1 shows the classification accuracy achieved using the LBP-fourier (8,1), over the entire image. The accuracy in this case reach only 68.33% +/- 0.70.

The eighth row of Table 1 shows the classification accuracy achieved using the LBP-fourier (16,2), over the entire image. The accuracy in this case reach only 62.33% +/- 0.67.

The best results were obtained for the Uniform LBP (8,1) using windows of size 10x10 pixels without overlapping and Uniform LBP (8,1) with overlapping of the 50% reaching 90.33% +/- 0.35 and 91.33% +/- 0.40 respectively. These result are better than previously published.

It is important to notice that the highest gender classification rates were reached using the overlapping histograms. It may be that small windows contain more specific information for gender classification, or it may be that the information extracted from those windows is more exact due to segmentation accuracies and fusion of histograms.

For the best results in Table 1, using the ULBPh_ov(8,1) selection, the correct classification rate is substantially better for males than for females. For the left eye, the correct classification rate is 96.67% for males, versus 86% for females. This represents 145 correct male images out of 150, and 129 correct female images out of 150. For the second best method ULBPh(8,1), the correct classification

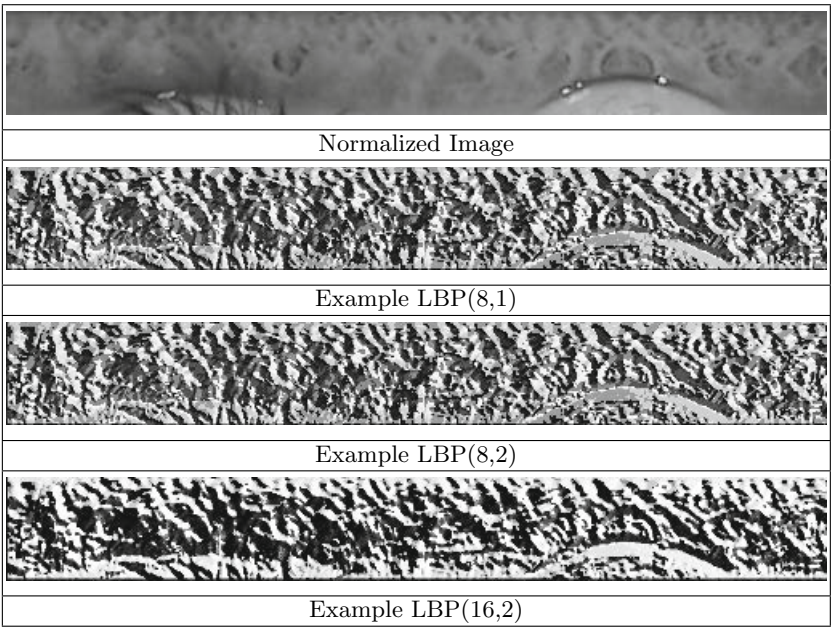


Fig. 4. Normalized iris image from segmentation stage and different LBP examples

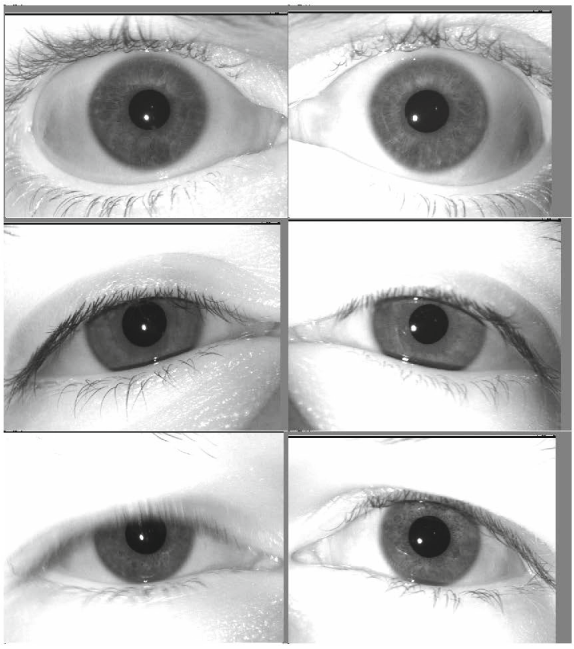


Fig. 5. Sample images showing right and left eye images. The image belong to the same person and shows the different illumination level for each eye.

Table 1. Gender Classification rate using different LBP implementation. In the first columns see the implementation methods and the second columns the classification rate for the left iris. Columns 3rd and 4 rd show the results by gender.

Implementation	Left eye (%)	Male (%)	Female (%)
Raw Image	78.52 +/- 1.70	77.50	79.53
LBP(8,1)	71.33 +/- 0.80	70.00	73.16
C-LBP-Mag(8,1)	65.33 +/- 0.90	68.25	62.35
C-LBP-Sign (8,1)	60.33 +/- 0.80	58.30	62.33
C-ULBP-Mag(8,1)	81.33 +/- 0.50	84.00	80.00
C-ULBP-Sign (8,1)	77.33 +/- 0.50	76.13	78.66
LBP-Fourier(8,1)	68.33 +/- 0.67	69.50	67.10
LBP-Fourier(16,2)	62.33 +/- 0.35	59.00	65,66
ULBPh(8,1)	90.33 +/- 0.35	92.67	88.00
ULBPh_ov(8,1)	**91.33** +/- 0.40	**96.67**	**86.00**

rate for males is 92.67% versus 88% for females. This represents 139 correct male images out of 150, and 132 correct female images out of 150.

4 Conclusions

This paper is the first to explore uniform LBP using fusion of histograms for predicting gender from the iris image using the polar representation.

The combination of the structural and statistical approaches stems from the fact that the distribution of micro-textons can be seen as statistical placement rules. The LBP distribution therefore has both of the properties of a structural analysis method: texture primitives and placement rules. On the other hand, the distribution is just a statistic of a non-linearly filtered image, clearly making the method a statistical one. For these reasons, the LBP distribution can be successfully used in gender classification using a wide variety of different textures, to which statistical and structural methods have normally been applied separately.

We found very large variations in accuracy based on using different implementations of LBP. The previous results motivate exploring more LBP implementation with different windows size and radii. Of the alternatives considered here, we found that using overlapping windows for histogram LBP(8,1) gave the best accuracy, obtaining 91.33%. This level of accuracy exceeds that of any other publication that we are aware of.

Several steps can be pursued to obtain even better accuracy in gender prediction from iris. We used the IrisBEE implementation in this work, and it is known to have as accurate of iris region segmentation as some other available implementations. Improving the accuracy of the iris region segmentation should naturally improve the accuracy of gender prediction. In this preliminary paper, we have presented results for only the left iris, we are still working on the results of the right iris and the fusion of the information from both irises. Older iris scanners (e.g., the LG 2200) and applications typically used just one iris, either

the left or right. But more modern sensors (e.g., the LG 4000) acquire both iris images, and so it makes sense to consider gender prediction based on the combination of left and right polar images.

Acknowledgments. Thanks to Vince Thomas, Mike Batanian, Steve Lagree and Yingjie Gu for the work they have previously done in this research topic.

This research was funded by FONDECYT 1120613 and by Department of Electrical Engineering, Universidad de Chile.

References

1. Bowyer, K.W., Hollingsworth, K., Flynn, P.J.: Image understanding for iris biometrics: A survey. Computer Vision and Image Understanding **110**(2), 281–307 (2008)
2. Lagree, S., Bowyer, K.: Predicting ethnicity and gender from iris texture. In: IEEE International Conference on Technologies for Homeland Security (HST), pp. 440–445, November 2011
3. Thomas, V., Chawla, N., Bowyer, K., Flynn, P.: Learning to predict gender from iris images. In: First IEEE International Conference on Biometrics: Theory, Applications, and Systems, BTAS 2007, pp. 1–5, September 2007
4. Tapia, J., Perez, C.: Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of LBP, Intensity, and Shape. IEEE Transactions on Information Forensics and Security **8**(3), 488–499 (2013)
5. He, Y., Feng, G., Hou, Y., Li, L., Micheli-Tzanakou, E.: Iris feature extraction method based on lbp and chunked encoding. In: Seventh International Conference on Natural Computation (ICNC) **3**, 1663–1667, July 2011
6. Yang, M.H., Moghaddam, B.: Gender classification using support vector machines. Proc. Int Image Processing Conf. **2**, 471–474 (2000)
7. Makinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(3), 541–547 (2008a)
8. Perez, C., Tapia, J., Estevez, P., Held, C.: Gender classification from face images using mutual information and feature fusion. International Journal of Optomechatronics **6**(1), 92–119 (2012)
9. Bansal, A., Agarwal, R., Sharma, R.K.: SVM based gender classification using iris images, pp. 425–429, November 2012
10. Lyle, J., Miller, P., Pundlik, S., Woodard, D.: Soft biometric classification using periocular region features. In: Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS) 2010, pp. 1–7, September 2010
11. Merkow, J., Jou, B., Savvides, M.: An exploration of gender identification using only the periocular region. In: Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS) 2010, pp. 1–5, September 2010
12. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1226–1238 (2005)
13. Liu, X., Bowyer, K., Flynn, P.: Experiments with an improved iris segmentation algorithm. In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies, pp. 118–123, October 2005

14. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 971–987 (2002)
15. Lei, Z., Ahonen, T., Pietikainen, M., Li, S.Z.: Local frequency descriptor for low-resolution face recognition. In: FG, pp. 161–166 (2011)
16. Guo, Z., Zhang, L., Zhang, D.: Rotation invariant texture classification using lbp variance (lbpv) with global matching. Pattern Recognition **43**(3), 706–719 (2010)
17. Guo, Z., Zhang, D., Zhang, D.: A completed modeling of local binary pattern operator for texture classification. IEEE Transactions on Image Processing **19**(6), 1657–1663 (2010)
18. Zhou, H., Wang, R., Wang, C.: A novel extended local-binary-pattern operator for texture analysis. Information Sciences **178**(22), 4314–4325 (2008)
19. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recognition Letters **33**(4), 431–437 (2012) Intelligent Multimedia Interactivity
20. Daugman, J.: How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology **14**(1), 21–30 (2004)
21. Bowyer, K.W., Hollingsworth, K.: The best bits in an iris code. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(6), 1–1 (2009)
22. Libor Masek, P.K.: Matlab source code for a biometric identification system based on iris patterns. The School of Computer Science and Software Engineering, The University of Western Australia (2003)
23. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 1–27 (2011)

Evaluation of Texture Descriptors for Automated Gender Estimation from Fingerprints

Ajita Rattani[1(✉)], Cunjian Chen[2], and Arun Ross[1]

[1] Michigan State University, East Lansing, USA
{ajita,rossarun}@cse.msu.edu
[2] West Virginia University, Morgantown, USA
cchen10@mix.wvu.edu

Abstract. Gender is an important demographic attribute. In the context of biometrics, gender information can be used to index databases or enhance the recognition accuracy of primary biometric traits. A number of studies have demonstrated that gender can be automatically deduced from face images. However, few studies have explored the possibility of automatically estimating gender information from fingerprint images. Consequently, there is a limited understanding in this topic. Fingerprint being a widely adopted biometrics, gender cues from the fingerprint image will significantly aid in commercial applications and forensic investigations. This study explores the use of classical texture descriptors - Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Binarized Statistical Image Features (BSIF) and Local Ternary Pattern (LTP) - to estimate gender from fingerprint images. The robustness of these descriptors to various types of image degradations is evaluated. Experiments conducted on the WVU fingerprint dataset suggest the efficacy of LBP descriptor in encoding gender information from good quality fingerprints. The BSIF descriptor is observed to be robust to partial fingerprints, while LPQ is observed to work well on blurred fingerprints. However, the gender estimation accuracy in the case of fingerprints is much lower than that of face, thereby suggesting that more work is necessary on this topic.

Keywords: Soft biometrics · Fingerprints · Gender estimation · LBP · LPQ · BSIF · LTP

1 Introduction

Gender[1] classification is a fundamental task for human beings, as many social interactions are gender-based [1]. The problem of gender classification has been investigated from both psychological [2] and computational perspectives [3].

[1] The more accurate term would be *sex* rather than *gender* in the context of this paper.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 764–777, 2015.
DOI: 10.1007/978-3-319-16181-5_58

It plays an important role in many applications such as human-computer interaction, surveillance, context-based indexing and searching, demographic studies and biometrics [1,4]. In the context of biometrics, gender can be viewed as a soft biometric trait that can be used to index databases or enhance the recognition accuracy of primary biometric traits.

The problem of automated gender estimation is typically treated as a two-class classification problem in which features extracted from a set of images corresponding to male and female subjects are used to train a two-class classifier. The output of the gender estimator is the classification of a test image as a male or female subject [1,4–6]. A number of studies suggest that gender can be robustly estimated from face images with relatively high accuracies [7,8].

However, only a limited number of studies have investigated the estimation of gender information from fingerprint images [9–12]. In most of these studies [9–11], gender estimation using fingerprints was based on the observation that females exhibit a higher ridge density due to finer epidermal ridge details compared to males. In [13], a method for gender classification based on Discrete Wavelet Transform (DWT) and Singular Value Decomposition (SVD) was proposed. Recently, in [14], quality-based features extracted from the frequency domain using Fourier Transform Analysis (FTA) and texture-based features captured by the Local Binary Pattern (LBP) and Local Phase Quantization (LPQ) descriptors, were used for gender estimation. These studies suggested that gender can be deduced from fingerprint images with an accuracy of about 82% [13,14]. However, these studies do not clearly indicate if the *subjects* in the training and test sets are non-overlapping - an important requirement for evaluating gender classifiers. The local texture information of a fingerprint should offer gender cues [13,14] because it can encode the ridge density structure that varies between males and females [9]. Deducing gender from fingerprints can be useful in forensic investigations and security applications where additional intelligence may be obtained from the fingerprint of a person. Further, gender information can also be used to enhance the recognition accuracy of a fingerprint matcher in commercial applications. However, there is a limited understanding of this topic which is partially due to the superficial nature of existing studies.

In this work, we investigate several aspects of gender estimation from fingerprint images. Firstly, we evaluate the ability of four commonly used texture descriptors to extract gender information from fingerprints. Secondly, we analyze if a gender estimator developed for one finger (e.g., left index) can be used to predict the gender of fingerprints originating from a different finger (e.g., right index). In previous studies, experiments were conducted by either training and testing the gender estimator on each finger individually [13] or by analyzing the differences in fingerprint ridge density between males and females over all the fingers and reporting aggregate statistics [9–11]. Thirdly, we evaluate the effect of degraded and partial fingerprint images on the performance of the gender estimator. To facilitate this analysis, we simulate noisy, blurred and partial fingerprint images. Finally, we investigate if the texture descriptors used for fingerprints can be used in the context of gender estimation from face images.

In summary, the contributions of this work are as follows:

- Exploring multiple textural descriptors to encode gender information from fingerprints.
- Evaluating the interoperability of the gender estimator across different fingers.
- Evaluating the performance of gender estimator on degraded fingerprint images.
- Utilizing the same set of texture descriptors for encoding gender information in both face and fingerprint images.

Experiments are conducted on the WVU multimodal face and fingerprint database [15].

This paper is organized as follows: Section 2 explains the textural descriptors used to encode gender information from fingerprint images. Section 3 presents the experimental investigations and results. Conclusions are drawn in section 4.

2 Texture Descriptors Used for Encoding Gender Information

The textural descriptors used to extract gender information from fingerprint images are summarized below.

1. **Local Binary Pattern (LBP).** It is a textural descriptor that assigns a label to every pixel of an image by thresholding the neighborhood of each pixel based on the center pixel value and converting the resultant binary number to a decimal value. Then histograms are computed from tessellated blocks and concatenated to form a descriptor [16]. The LBP operator can be extended with neighborhoods of different sizes. Using a circular neighborhood and bilinear interpolation at non-integer pixel coordinates allows for any radius and number of pixels in the neighborhood. The notation (P,R) will be used to denote a pixel neighborhood consisting of P points on a circular neighborhood of radius R. The LBP code of a pixel g_c is given by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, \tag{1}$$

$$s(x) = \begin{cases} 1 \text{ if } x \geq 0; \\ 0 \text{ otherwise.} \end{cases} \tag{2}$$

Here, g_c and g_p denote the center pixel and neighboring pixels, respectively. Further, $LBP^{u2}_{P,R}$ represents uniform rotation invariant LBP which can be used to reduce the number of codes, and hence the length of the feature vector[16].

In our work, each image is tessellated into non-overlapping blocks of size 18×25 and each block is represented with a feature vector which is a concatenation of histograms corresponding to $\text{LBP}^{u2}_{8,1}$, $\text{LBP}^{u2}_{16,2}$ and $\text{LBP}^{u2}_{24,3}$. The feature vectors of all block are concatenated to obtain the final feature descriptor.

2. **Local Phase Quantization (LPQ).** It is based on the quantization of Fourier transform phase in local neighborhoods [17]. Short Time Fourier Transform (STFT) is computed over a $M \times M$ neighborhood, N_x, at each pixel position x of the image $f(x)$ as follows:

$$F_{u,x} = \sum_{y \in N_x} f(x-y) e^{-j2\pi u^T y} = w_u^T f_x. \tag{3}$$

   Here, w_u is the basis vector of the 2-D DFT at frequency u, and f_x is a vector containing all M^2 image pixels from N_x. We used a window size of 5×5 to extract LPQ features. Then, a LPQ histogram was computed for each tessellated block of size 18×25 from an image.
3. **Binary Statistical Image Features (BSIF).** This method computes a binary code for each pixel by linearly projecting local image patches onto a subspace, whose basis vectors are learnt from natural images via independent component analysis, and by binarizing the coordinates in this basis via thresholding. The length of the binary code string is determined by the number of basis vectors. Image blocks are represented by histograms of binary codes. This method is different from other descriptors which produce binary codes, such as LBP and LPQ, in the sense that the proposed approach is based on statistics of natural images and this improves its modeling capacity [18]. We extracted BSIF features using a predefined filter of size 7×7 learnt from natural images and a 12-bit string.
4. **Local Ternary Pattern (LTP).** This is a texture descriptor that creates a ternary code for every pixel based on its neighborhood as follows [19].

$$s'(g_p, g_c, t) = \begin{cases} 1 & g_p \geq g_c + t \\ 0 & |g_p - g_c| < t \\ -1 & g_p \leq g_c - t \end{cases} \tag{4}$$

   Here, g_p is the neighborhood pixels, g_c is the center pixel and t is the threshold value. As stated in [19], LTP is less sensitive to noise since the threshold is not purely based on the center pixel, unlike LBP. A 59-bin histogram is extracted from each block of size 18×21.

To extract these textural features, a fingerprint image is first tessellated into non-overlapping blocks, and then textural histograms are computed from each block. The histograms of all the blocks are concatenated to obtain a final feature descriptor. The extracted feature vectors from a set of training images corresponding to male and female subjects are used to train a gender estimator based on two-class linear SVM [8]. Figure 1 illustrates the steps involved in gender estimation from a fingerprint image.

3 Gender Estimation from Fingerprints

In this section, we will describe the dataset used, the experiments conducted and the obtained results.

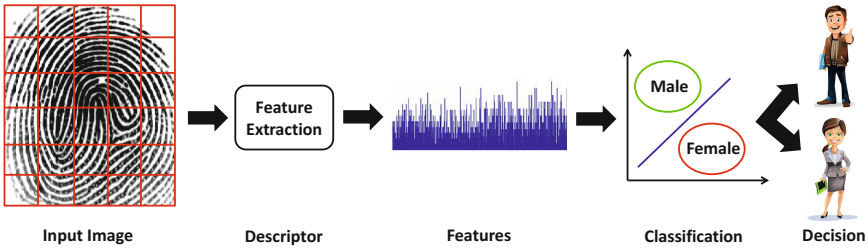


Fig. 1. The steps involved in gender estimation from a fingerprint image

3.1 WVU Multimodal Dataset

We utilized the WVU multimodal dataset consisting of face and fingerprint images of 166 male subjects and 71 female subjects. For every subject, five samples from each of four fingers (left index (L1), left middle (L2), right index (R1) and right middle (R2)) and five face samples were obtained. Sample images from this dataset are shown in Figure 2. Eye regions of the face have been masked to preserve the privacy of the subjects.

(a) Male Subjects (b) Female Subjects

Fig. 2. Sample images from the WVU multimodal dataset. Each subject has both face and fingerprint samples. The eye regions have been masked in order to preserve the privacy of users.

The images corresponding to 50 male and 50 female subjects were used to extract the histograms (LBP, LPQ, BSIF, LTP) and to train a two-class SVM based gender estimator. The remaining 111 male and 21 female subjects were used to evaluate the performance of the gender estimator. In order to perform cross-validation, this random partitioning into training and test sets was done 20 times. Each fingerprint image is 248×292 and the dimensionality of the obtained feature vectors are 7776, 36864, 4096, 14160 for LBP, LPQ, BSIF and

LTP, respectively. The performance of the gender estimator was evaluated using the correct overall classification rate (COCR), correct male classification rate (CMCR) and correct female classification rate (CFCR). Correct overall classification rate (COCR) is the percentage of test images whose gender was correctly estimated. Correct male (female) classification rate (CMCR and CFCR) is the percentage of images corresponding to males (females) correctly classified as males (females).

3.2 Evaluation of the Textural Descriptors to Encode Gender Information

First, we tested the performance of LBP, LPQ, BSIF and LTP based textural descriptors in extracting gender information from fingerprint images. Table 1 tabulates the COCR of the LBP, LPQ, BSIF and LTP based descriptors in estimating gender information from fingerprint images. These results are summarized over 20 test runs (as $\mu \pm \sigma^2$) and shown for the four fingers i.e., left index (indicated as L1), left middle (indicated as L2), right index (indicated as R1) and right middle (indicated as R2), individually. It can be seen that LBP performs marginally better than other textural descriptors in encoding gender information from fingerprint images (COCR is 71.7%). The second best performance is obtained by BSIF (COCR is 71.0%). The average COCR over all the four descriptors and fingers is 70.0%.

Table 1. Correct overall classification rate (COCR) of the LBP, LPQ, BSIF and LTP based textural descriptors in encoding gender information from fingerprint images

Methods	COCR [%]				
	L1	L2	R1	R2	**Average**
LBP+SVM	70.8 ± 2.1	72.4 ± 3.4	70.2 ± 3.6	73.4 ± 2.5	71.7 ± 2.9
LPQ+SVM	66.6 ± 3.3	66.2 ± 3.6	64.7 ± 3.2	65.7 ± 3.5	65.8 ± 3.4
BSIF+SVM	70.1 ± 2.7	72.2 ± 3.7	70.4 ± 2.9	70.5 ± 3.7	71.0 ± 3.3
LTP+SVM	70.9 ± 3.2	72.1 ± 2.9	69.1 ± 3.3	70.2 ± 2.5	70.0 ± 2.9

Further, Table 2 tabulates the correct male and female classification rates of the LBP, LPQ, BSIF and LTP based descriptors for the four fingers. The LBP based descriptor obtained the best correct male (71.2%) and female (74.6%) classification rates. The average correct male and female classification rates (CMCR and CFCR) over all four descriptors and four fingers are 69.5% and 72.9%, respectively.

Next, we evaluated the performance when fusing the outputs of the gender estimators corresponding to the four fingers of a subject. The majority rule was used for fusion. Table 3 tabulates the correct male, female and overall classification rates of LBP, LPQ, BSIF and LTP. In case of ties, a label was randomly

Table 2. Correct male classification rate (CMCR) and correct female classification rate (CFCR) of the LBP, LPQ, BSIF and LTP based textural descriptors

Methods	CMCR [%]				CFCR [%]			
	L1	L2	R1	R2	L1	L2	R1	R2
LBP+SVM	69.7 ± 2.4	72.1 ± 4.6	70.1 ± 4.2	72.9 ± 3.2	76.8 ± 6.6	74.3 ± 2.4	70.7 ± 8.9	76.6 ± 6.2
LPQ+SVM	66.2 ± 3.8	65.9 ± 4.5	64.2 ± 4.1	65.3 ± 3.6	68.7 ± 5.3	68.1 ± 7.4	67.9 ± 6.5	68.7 ± 9.8
BSIF+SVM	69.9 ± 3.7	71.7 ± 4.6	69.7 ± 3.7	70.1 ± 4.3	71.0 ± 7.5	74.9 ± 5.8	73.6 ± 7.4	73.4 ± 7.5
LTP+SVM	70.5 ± 4.4	73.9 ± 4.5	69.4 ± 3.9	69.7 ± 3.2	72.9 ± 7.6	74.7 ± 6.1	68.4 ± 7.7	72.9 ± 6.8

assigned. It can be seen that fusion of gender cues from multiple fingerprints enhanced the accuracy of the gender estimator. For instance, COCR of the LBP based descriptor increased from 71.7% (see Table 1) to 80.4%. Similar observations can be made for other descriptors as well.

Table 3. Correct male classification rate (CMCR), correct female classification rate (CFCR) and correct overall classification rate (COCR) of the LBP, LPQ, BSIF and LTP based textural descriptors when outputs from the left index, left middle, right index and right middle fingerprints were fused using the majority rule

Methods	CMCR [%]	CFCR [%]	COCR [%]
LBP+SVM	82.2 ± 3.1	70.7 ± 8.4	80.4 ± 2.7
LPQ+SVM	80.3 ± 2.2	77.5 ± 1.8	77.5 ± 1.8
BSIF+SVM	83.7 ± 3.6	68.6 ± 7.3	81.4 ± 2.9
LTP+SVM	84.5 ± 2.7	67.3 ± 8.3	80.1 ± 2.5

3.3 Interoperability of the Gender Estimator Across Fingers

In this section, we evaluate the interoperability of the gender estimator across different fingers. The aim is to analyze if the gender can be estimated from the fingers different from those used for training the gender estimator.

Table 4 tabulates the COCR of the gender estimator trained using one finger (say left index) and tested on all others (say left middle, right index and right middle). It can be seen that performance of all the descriptors dropped across fingers. For instance, COCR of the LBP dropped from 71.7% (see Table 1) to 66.4%, and BSIF dropped from 71.0% (see Table 1) to 63.2%. However, LBP performed better than other descriptors in this case as well. Lowest average COCR was observed for LTP. Table 5 shows the CMCR and CFCR of these descriptors when evaluated across fingers.

Table 4. Correct overall classification rate (COCR) of the LBP, LPQ, BSIF and LTP based gender estimators across fingers

Training	Testing	LBP	LPQ	BSIF	LTP
L1	[L2 R1 R2]	72.9 ± 7.4	59.4 ± 10.9	66.2 ± 5.5	49.6 ± 6.8
L2	[L1 R1 R2]	63.8 ± 13.5	62.0 ± 9.5	64.2 ± 4.9	70.4 ± 5.1
R1	[L1 L2 R2]	78 ± 6.1	66.1 ± 10	69.4 ± 4.9	54.2 ± 6.0
R2	[L1 L2 R1]	50.8 ± 9.1	55.3 ± 10.1	52.8 ± 8.6	63.7 ± 4.9
Average		66.4 ± 9.1	60.7 ± 10.1	63.2 ± 5.9	59.4 ± 5.7

3.4 Performance of the Gender Estimator on Degraded and Partial Fingerprint Images

In this section, we evaluate the performance of the gender estimator when tested on degraded and partial fingerprint images. We simulated fingerprint degradations such as noise and blur. These type of fingerprint degradations are more likely to be encountered in some operational scenarios and forensic investigations. The process of lifting latent print by dusting the surface with fingerprint powder (black granular, aluminum flake, black magnetic, etc.) followed by photographing and lifting with clear adhesive tape also introduces noise and blur effect in the fingerprints. For this study, the gender estimator was always trained on the original (without degradations) fingerprints. Next, we evaluate the impact of degraded and partial prints on the gender estimator.

Table 5. Correct male classification rate (CMCR) and correct female classification rate (CFCR) of the LBP, LPQ, BSIF and LTP based gender estimators across fingers

Training	Testing	LBP		LPQ		BSIF		LTP	
		CMCR [%]	CFCR [%]	CMCR [%]	CFCR [%]	CMCR [%]	CFCR [%]	CMCR [%]	CFCR [%]
L1	[L2 R1 R2]	77.2 ± 9.2	48.2 ± 13.2	60.9 ± 16.1	50.6 ± 11.2	49.6 ± 6.8	55.8 ± 13.4	79.6 ± 6.8	44.2 ± 13.4
L2	[L1 R1 R2]	62.8 ± 18.6	69.3 ± 18.6	61.9 ± 13.5	62.1 ± 13.5	62.0 ± 9.5	69.8 ± 9.1	73.4 ± 4.1	53.6 ± 11.1
R1	[L1 L2 R2]	84.6 ± 8.9	39.3 ± 12.2	69.9 ± 15.1	44.7 ± 22.3	66.1 ± 10	55.9 ± 12.9	69.3 ± 7.4	51.5 ± 10.7
R2	[L1 L2 R1]	82.1 ± 9.1	43.2 ± 17.5	53.4 ± 14.5	54.2 ± 8.9	77.56 ± 5.7	55.3 ± 10.1	63.7 ± 4.9	60.8 ± 10.0

When Fingerprint Images are Noisy. We simulated noisy fingerprint images by applying a Gaussian noise with a mean value of 0.07. The variance varies from 0.04 to 0.07, with a step size of 0.01. An example of applying Gaussian noise to a fingerprint image is shown in Figure 3.

Table 6 shows the COCR of the LBP based gender estimator when test fingerprint images are noisy. The performance is evaluated for four noise levels (column 1). It can be seen that performance of the gender estimator drops significantly from 71.7% (see Table 1) to 33.1% (averaged over all the four noise levels and four fingers). The performance drops are obvious because LBP-based textural descriptors are not robust to noise [20].

Further, Table 7 shows the COCR of the BSIF based gender estimator when test fingerprint images are noisy. It can be seen that performance of the BSIF

Fig. 3. An illustration of applying Gaussian noise to a fingerprint image. From left to right, the noise level increases with the variance.

Table 6. Correct overall classification rate (COCR) of the **LBP** based gender estimator when test fingerprint images are noisy. Noisy fingerprint images are simulated by applying a Gaussian noise with a mean value of 0.07. The variance varies from 0.04 to 0.07, with a step size of 0.01 (shown in column 1).

Noise level	L1	L2	R1	R2	**Average**
$(\mu = 0.07, \sigma^2 = 0.04)$	32.9 ± 5.7	56.1 ± 3.8	28.6 ± 7.6	27.7 ± 8.8	36.3 ± 6.4
$(\mu = 0.07, \sigma^2 = 0.05)$	31.5 ± 6.2	42.7 ± 3.7	37.4 ± 5.6	26.2 ± 4.8	34.5 ± 5.2
$(\mu = 0.07, \sigma^2 = 0.06)$	31.4 ± 7.9	44.2 ± 4.5	20.9 ± 4.9	24.5 ± 5.3	30.3 ± 5.6
$(\mu = 0.07, \sigma^2 = 0.07)$	30.1 ± 5.6	45.7 ± 6.4	18.2 ± 7.4	31.6 ± 9.7	31.4 ± 7.2

based gender estimator also dropped from 71.0% (Table 1) to 52.6% (averaged over all four noise levels and four fingers). However, BSIF (COCR is 52.6%) performed better than LBP (COCR is 33.1%) on noisy fingerprint images. COCR of LPQ and LTP are 50.5% and 45.6%, respectively, over all the four fingers.

Table 7. Correct overall classification rate (COCR) of the **BSIF** based gender estimator when test fingerprint images are noisy. Noisy fingerprint images are simulated by applying a Gaussian noise with a mean value of 0.07. The variance varies from 0.04 to 0.07, with a step size of 0.01 (shown in column 1).

Noise level	L1	L2	R1	R2	**Average**
$(\mu = 0.07, \sigma^2 = 0.04)$	50.1 ± 9.7	44.5 ± 8.0	66.2 ± 5.8	52.9 ± 9.2	53.4 ± 8.2
$(\mu = 0.07, \sigma^2 = 0.05)$	50.5 ± 9.1	42.6 ± 9.9	62.3 ± 6.1	54.5 ± 9.7	52.5 ± 8.7
$(\mu = 0.07, \sigma^2 = 0.06)$	51.8 ± 9.5	42.1 ± 9.1	63.4 ± 6.1	51.4 ± 9.2	52.2 ± 8.4
$(\mu = 0.07, \sigma^2 = 0.07)$	50.8 ± 9.4	42.3 ± 9.5	64.4 ± 5.1	52.5 ± 9.7	52.5 ± 8.4

When Fingerprint Images are Blurred. We simulated blurred fingerprint images by applying a Gaussian low pass filter using a window size of 15×15. The variance varies from 3.0 to 24.0, with a step size of 6.0. An example of applying Gaussian blur to a fingerprint image is shown in Figure 4.

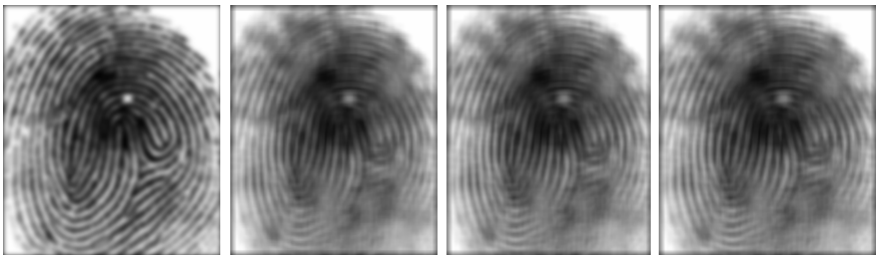


Fig. 4. An illustration of applying Gaussian low pass filter with a window size of 15×15 to a fingerprint image. From left to right, the blurring effect increases with the variance.

Table 8 shows the COCR of the LPQ-based gender estimator when test fingerprint images are blurred. The performance is evaluated at four different blur levels (column 1). It can be seen that the LPQ-based gender estimator is quite robust to blur (COCR is 68.4%). This is because the LPQ descriptor itself is resilient to blur [17].

Table 8. Correct overall classification rate (COCR) of the **LPQ** based gender estimator when test fingerprint images are blurred. Blurred fingerprint images are simulated by applying Gaussian low pass filter using a window size of 15×15. The variance varies from 3.0 to 24.0, with a step size of 6.0. (shown in column 1). The gender estimator was trained on original non-blurred fingerprint images.

Noise level	L1	L2	R1	R2	**Average**
$(\text{block} = 15, \sigma^2 = 3)$	66.5 ± 9.6	67.7 ± 9.5	74.7 ± 9.1	56.4 ± 9.5	66.3 ± 9.4
$(\text{block} = 15, \sigma^2 = 9)$	66.4 ± 9.2	74.5 ± 8.5	74.1 ± 9.3	66.2 ± 7.5	70.3 ± 8.6
$(\text{block} = 15, \sigma^2 = 15)$	69.1 ± 7.3	71.7 ± 8.2	68.5 ± 6.5	63.8 ± 7.7	68.3 ± 7.4
$(\text{block} = 15, \sigma^2 = 24)$	69.2 ± 9.5	71.8 ± 8.4	70.9 ± 9.7	63.8 ± 9.6	68.9 ± 9.3

Further, Table 9 shows the COCR of the LTP based gender estimator when test fingerprint images are blurred. It can be seen that performance of the LTP based gender estimator drops from 70.0% (see Table 1) to 61.6% (averaged over all four blur levels and four fingers). COCR of LBP and BSIF are 31.4% and 54.3%, respectively. LPQ (COCR is 68.4%) performs better than other descriptors on blurred fingerprint images.

When Fingerprint Images are Partial. Partial prints were generated by using half and one-fourth portion of the original fingerprint image as shown in Figure 5. The gender estimator was trained on full fingerprints.

It can be seen in Table 10 that COCR of LBP, BSIF, LPQ and LTP (averaged over all the four fingers) on partial prints generated using half of the original

Table 9. Correct overall classification rate (COCR) of the **LTP** based gender estimator when test fingerprint images are blurred. Blurred fingerprint images are simulated by applying Gaussian low pass filter with a window size of 15×15. The variance varies from 3.0 to 24.0, with a step size of 6.0. (shown in column 1). The gender estimator was trained on original non-blurred fingerprint images.

Noise level	L1	L2	R1	R2	**Average**
$(\text{block} = 15, \sigma^2 = 3)$	62.1 ± 9.3	66.6 ± 5.6	67.2 ± 7.5	56.3 ± 6.3	63.0 ± 7.2
$(\text{block} = 15, \sigma^2 = 9)$	67.1 ± 6.3	58.3 ± 4.3	62.3 ± 6.4	54.1 ± 7.1	60.4 ± 6.1
$(\text{block} = 15, \sigma^2 = 15)$	65.8 ± 6.1	63.4 ± 7.2	55.4 ± 5.6	59.8 ± 7.6	61.1 ± 6.6
$(\text{block} = 15, \sigma^2 = 24)$	65.6 ± 8.8	59.7 ± 9.7	65.9 ± 8.6	57.7 ± 7.5	62.2 ± 8.6

Fig. 5. An illustration of (a) Original print (b) One-half of original print and (c) One-fourth of original print (from left to right)

prints are 59.1%,71.7%, 65.4% and 64.2%, respectively. Further, COCR of these descriptors on partial prints generated using one-fourth of the original prints are 54.5%, 70.5%, 62.9% and 54.0%, respectively. BSIF is fairly robust to partial fingerprints compared to other descriptors. In fact, the COCR of the BSIF on partial prints is almost equal to those obtained on original fingerprints (see Table1). This clearly conveys the importance of the BSIF operator in estimating gender from fingerprint images.

These experimental results suggest that the performance of all four descriptors dropped when encountering degraded or partial fingerprints. However, BSIF performed better than LBP, LPQ and LTP on noisy and partial fingerprint images and LPQ performed better than LBP, BSIF and LTP on blurred fingerprint images.

3.5 Common Textural Descriptors to Encode Gender Information from Face and Fingerprints

Next we investigate if the same textural descriptors used for encoding gender information in fingerprint images can be used on face images. In this regard, we tested the performance of the LBP, LPQ, BSIF and LTP based textural descriptors on face images. The size of each cropped face image was 150×130

Table 10. Correct overall classification rate (COCR) of the LBP, BSIF, LPQ and LTP based gender estimator when tested on partial fingerprint images (generated using half and one-fourth portion of the original fingerprint). These gender estimators were trained on full fingerprint images.

Finger	LBP		BSIF		LPQ		LTP	
	Half	One-fourth	Half	One-fourth	Half	One-fourth	Half	One-fourth
L1	60.1 ± 7.7	60.6 ± 9.2	70.8 ± 2.7	71.5 ± 4.4	68.8 ± 3.8	67.7 ± 3.6	67.4 ± 5.8	45.7 ± 8.4
L2	65.2 ± 9.1	62.3 ± 9.8	70.7 ± 8.4	75.2 ± 9.1	62.6 ± 9.2	58.5 ± 10.6	62.6 ± 4.1	65.6 ± 9.4
R1	53.4 ± 6.7	52.4 ± 6.5	72.4 ± 6.2	69.1 ± 7.6	69.8 ± 9.8	70.4 ± 8.7	64.5 ± 3.9	55.6 ± 6.2
R2	57.7 ± 8.7	43.5 ± 4.7	73.3 ± 9.3	67.2 ± 8.4	60.5 ± 5.4	65.6 ± 4.2	62.1 ± 5.2	47.6 ± 8.3
Average	59.1 ± 8.1	54.5 ± 7.5	71.7 ± 6.6	70.5 ± 7.4	65.4 ± 7.1	62.9 ± 6.7	64.2 ± 4.7	54.0 ± 8

(block size was 18×21) and the dimensions of the obtained feature vectors were 2160, 12288, 4096 and 7434 for LBP, LPQ, BSIF and LTP, respectively. Further, to better understand the performance of these texture descriptors on face images, a state-of-the-art gender classifier named Intraface[2] was utilized for comparison.

Table 11 tabulates the CMCR, CFCR and COCR of the LBP, LPQ, BSIF and LTP based gender estimators from face images. It can be seen that LTP outperforms the other textural descriptors (LBP, LPQ and BSIF) in encoding gender information from face images. Further, the performance difference of LTP over Intraface is only 4.2%. The second best performance is obtained by LPQ. These results suggest that these textural descriptors can potentially be used for encoding gender cues from both face and fingerprint images.

Table 11. CMCR, CFCR and COCR rates of the LBP, LPQ, BSIF and LTP based textural descriptors for gender estimation from face images

Methods	CMCR [%]	CFCR [%]	COCR [%]
LBP+SVM	85.8 ± 3.27	85.3 ± 5.02	85.7 ± 2.65
LPQ+SVM	92.4 ± 2.36	93.6 ± 4.53	92.6 ± 1.91
BSIF+SVM	88.0 ± 3.11	91.4 ± 4.40	88.5 ± 2.39
LTP+SVM	92.5 ± 1.99	93.1 ± 4.04	92.6 ± 1.47
Intraface	98.4 ± 0.61	88.5 ± 5.32	96.8 ± 0.95

However, gender can be deduced from face images with relatively high accuracy than fingerprints. The COCR of the gender estimator based on face images is 89.8% (averaged over all the four descriptors), while the COCR of the gender estimator based on fingerprint is 71.7% (averaged over all the four descriptors (see Table 1)).

Further, the performance of individual texture descriptors varies across modalities. For instance, LBP outperformed LPQ, BSIF and LTP in encoding

[2] Intraface: http://www.humansensing.cs.cmu.edu/intraface/

gender information from fingerprint image. However, LTP outperformed LBP, BSIF and LTP in encoding gender information from face images. The possible reason is the radically different nature of the information used for sex determination from face and fingerprint images. Facial features play a dominant role in gender determination from face images [1,4], while studies suggest that ridge density reflects gender information in fingerprint images [9–12].

4 Conclusion and Discussion

This study evaluates the performance of four texture descriptors - LBP, LPQ, BSIF and LTP - for the task of gender estimation from fingerprint images. Further, the performance of the estimators is evaluated on degraded fingerprints that are noisy and blurred, as well as partial prints. Experimental results suggest that

- LBP descriptor is efficient in encoding gender information from high quality fingerprint images in comparison to LPQ, BSIF and LTP.
- The performance of all four gender estimators drops, when training is done using one set of fingers (e.g., left index) and testing is done on a different set of fingers (e.g., right index). LBP exhibited the least drop in performance.
- The performance of the gender estimator degrades when noisy and blurred fingerprint images are observed. BSIF performs much better than other descriptors on noisy and partial fingerprints. The reason could be that BSIF uses predefined filters learned from a set of natural images and this improves its modeling capacity. LPQ performs better than other descriptors on blurred images. This is because it is resilient to blur.
- Finally, the texture descriptors that were used to encode gender information in fingerprint images could also be used to encode face images. However, the performance of these descriptors varies depending on the modality used. This is because face and fingerprint contain different type of information used for gender determination.

As a part of future work, more robust features will be investigated for gender estimation from fingerprint images; experiments will be repeated on large scale multi-modal face and fingerprint datasets; and results will be compared against existing schemes for gender estimation from fingerprints. We will consider ways to fuse the outputs of the four descriptors in a systematic way. We will also investigate fusion of the gender estimators based on face and fingerprints at the feature, score and decision levels.

Acknowledgments. This work was funded by NSF Award 1066197 (CITeR).

References

1. Cao, D., Chen, C., Piccirilli, M., Adjeroh, D., Bourlai, T., Ross, A.: Can facial metrology predict gender?. In: Proc. of IJCB, pp. 1–8 (2011)
2. Burton, A.M., Bruce, V., Dench, N.: Perception. What's the difference between men and women? Evidence from facial measurement **22**, 153–176 (1993)
3. Castrillón-Santana, M., Vuong, Q.C.: An analysis of automatic gender classification. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 271–280. Springer, Heidelberg (2007)
4. Moghaddam, B., Yang, M.: Learning gender with support faces. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 707–711 (2002)
5. Makinen, E., Raisamo, R.: Evaluation of gender classification methods with automatically detected and aligned faces. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**, 541–547 (2008)
6. Moghaddam, B.: Gender classification with support vector machines. In: Proc. of AFGR, pp. 306–311 (2000)
7. Shan, C.: Learning local binary patterns for gender classification on real-world face images. Pattern Recognition Letters **33**(4), 431–437 (2012)
8. Chen, C., Ross, A.: Evaluation of gender classification methods on thermal and near-infrared face images. In: Proc. of IJCB, pp. 1–8 (2011)
9. Acree, M.A.: Is there a gender difference in fingerprint ridge density? Forensic Science International **102**(1), 35–44 (1999)
10. Gungadin, S.: Sex determination from fingerprint ridge density. Internet Journal of Medical Update **2**, 4–7 (2007)
11. Nigeria, Y.: Towards age and gender determination, ridge thickness to valley thickness ratio (RTVTR) and ridge count on gender detection. Analysis, Design and Implementation of Human Fingerprint Patterns System **1** (2012)
12. Gutierrez-Redomero, E., Alonso, C., Romero, E., Galera, V.: Variability of fingerprint ridge density in a sample of spanish caucasians and its application to sex determination. Forensic Science International **180**, 17–22 (2008)
13. Gnanasivam, P., Muttan, S.: Fingerprint gender classification using wavelet transform and singular value decomposition. International Journal of Biometrics and Bioinformatics (IJBB) **9** (2012)
14. Marasco, E., Lugini, L., Cukic, B.: Exploiting quality and texture features to estimate age and gender through fingerprint images. In: Proc. of SPIE 9075 (2014)
15. Crihalmeanu, S., Ross, A., Schuckers, S., Hornak, L.: A Protocol for Multibiometric Data Acquisition. Storage and Dissemination. Technical report, WVU (2007)
16. Ojala, T., Pietikinen, M., Menp, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**, 971–987 (2002)
17. Heikkila, J., Ojansivu, V.: Methods for local phase quantization in blur-insensitive image analysis. In: Proc. of LNLA (2009)
18. Kannala, J., Rahtu, E.: BSIF: binarized statistical image features. In: Proc. of ICPR (2012) 1363–1366
19. Tan, X., Triggs, B.: Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. In: Zhou, S.K., Zhao, W., Tang, X., Gong, S. (eds.) AMFG 2007. LNCS, vol. 4778, pp. 168–182. Springer, Heidelberg (2007)
20. Sintorn, I.M., Kylberg, G.: Evaluation of noise robustness for local binary pattern descriptors in texture classification. EURASIP Journal on Image and Video Processing **17** (2013)

Recognition of Facial Attributes Using Adaptive Sparse Representations of Random Patches

Domingo Mery[1(✉)] and Kevin Bowyer[2]

[1] Department of Computer Science,
Pontificia Universidad Católica de Chile, Santiago, Chile
dmery@ing.puc.cl
http://dmery.ing.puc.cl
[2] Department of Computer Science and Engineering,
University of Notre Dame, South Bend, USA
kwb@nd.edu
http://www.nd.edu/∼kwb

Abstract. It is well known that some facial attributes –like soft biometric traits– can increase the performance of traditional biometric systems and help recognition based on human descriptions. In addition, other facial attributes –like facial expressions– can be used in human–computer interfaces, image retrieval, talking heads and human emotion analysis. This paper addresses the problem of automated recognition of facial attributes by proposing a new general approach called Adaptive Sparse Representation of Random Patches (ASR+). In the learning stage, random patches are extracted from representative face images of each class (*e.g.*, in gender recognition –a two-class problem–, images of females/males) in order to construct representative dictionaries. In the testing stage, random test patches of the query image are extracted, and for each test patch a dictionary is built concatenating the 'best' representative dictionary of each class. Using this adapted dictionary, each test patch is classified following the Sparse Representation Classification (SRC) methodology. Finally, the query image is classified by patch voting. Thus, our approach is able to learn a model for each recognition task dealing with a larger degree of variability in ambient lighting, pose, expression, occlusion, face size and distance from the camera. Experiments were carried out on seven face databases in order to recognize facial expression, gender, race and disguise. Results show that ASR+ deals well with unconstrained conditions, outperforming various representative methods in the literature in many complex scenarios.

Keywords: Sparse representation · Soft biometrics · Gender recognition · Race recognition · Facial expression recognition

1 Introduction

Automated recognition of facial attributes has been a relevant area in computer vision, making many important contributions since the 1990s (see for example [19]). The relevance of this research field is twofold: First, the use of facial

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 778–792, 2015.
DOI: 10.1007/978-3-319-16181-5_59

attributes, like soft biometric traits (*e.g.*, gender [23], race [8], age [9], etc.), can increase the performance of traditional biometric systems [25] and help recognition based on human descriptions [27]. Second, other facial attributes, like facial expressions, can be used in human–computer interfaces, image retrieval, talking heads and human emotion analysis [35].

Usually, each single facial attribute has been recognized by a specific algorithm. Some examples are the following: a) Gender is identified using a SVM classifier with Gaussian RBF kernel [21], a Real AdaBoost classifier with texture features [34], an AdaBoost classifier with a low resolution image [3], and a SVM classifier of PCA representations [16]. b) Facial expressions are classified using a new feature called 'supervised locally linear embedding' [14], a decomposition into multiple two-class classification problems with 'salient feature vectors' [13], local binary patterns [28], a boosted deep belief network [15], active facial patches [36], and Gabor features [4]. c) Race is recognized using biologically inspired features [11], an ensemble framework with LDA [17], a probabilistic graphical model [22] and local binary patterns with wavelets features [26].

There are few approaches to estimate age, gender and race together (see for example [12]), however, to the best knowledge of the authors, there has been no reported approach, that can be used to recognize facial attributes in general. We believe that algorithms based on sparse representations can be used for this task because in many computer vision applications, under assumption that natural images can be represented using sparse decomposition, state-of-the-art results have been significantly improved [30]. Algorithms based on Sparse Representation Classification (SRC) [32] have been widely used in face recognition. In the sparse representation approach, a dictionary is built from the gallery images, and matching is done by reconstructing the query image using a sparse linear combination of the dictionary. The identity of the query image is assigned to the class with the minimal reconstruction error. Several variations of this approach were recently proposed. In [33], a sparse representation in two phases is proposed. In [7], sparse representations of patches distributed in a grid manner are used. These variations improve recognition performance as they are able to model various corruptions in face images, such as misalignment and occlusion.

Reflecting on the problems confronting recognition of facial attributes, we believe that there are some key ideas that should be present in new proposed solutions. First, it is clear that certain parts of the face are not providing any information about the class to be recognized. For this reason, such parts should be detected and should not be considered by the recognition algorithm. Second, in recognizing any class, there are parts of the face that are more relevant than other parts (for example the mouth when recognizing an expression like happiness). For this reason, relevant parts should be class-dependent, and could be found using unsupervised learning. Third, in the real-world environment, and given that face images are not perfectly aligned and the distance between camera and subject can vary from capture to capture, analysis of fixed sub-windows can lead to misclassification. For this reason, feature extraction should not be in fixed positions, and can be in several random positions, and use a selection

criterion that enables selection of the best regions. Fourth, the expression that is present in a query face image can be subdivided into 'sub-expressions', for different parts of the face (*e.g.*, eyebrows, nose, mouth). For this reason, when searching for images of the same class it would be helpful to search for image parts in all images of the gallery instead of similar gallery images.

Inspired by these key ideas, we propose a new general method for recognition of facial attributes.

Three main contributions of our approach are: 1) A new general algorithm that is able to recognize a wide range of facial attributes: it has been evaluated in the recognition of expressions, gender, race and disguise obtaining a performance at least comparable with that achieved by state-of-art techniques. 2) A new representation for the classes to be recognized: this is based on representative dictionaries learned for each class of the gallery images, which correspond to a rich collection of representations of selected relevant parts that are particular to a specific class. 3) A new representation for the query face image: this is based on *i*) a discriminative criterion that selects the ?best? test patches extracted randomly from the query image and *ii*) and an 'adaptive' sparse representation of the selected patches computed from the 'best' representative dictionary of each class. Using these new representations, the proposed method (ASR+) can achieve high recognition performance under many complex conditions, as shown in our extensive experiments.

The rest of the paper is organized as follows: in Section 2, the proposed method is explained in further detail. In Section 3, the experiments and results are presented. Finally, in Section 4, concluding remarks are given.

2 Proposed Method

According to the motivation of our work, we believe that facial attributes can be recognized using a patch-based approach. Thus, following a sparse representation methodology, in a learning stage a number of random patches can be extracted from each training image, and a dictionary can be built for each class by concatenating its patches (stacking in columns). In the testing stage, several patches can be extracted and each of them can be classified using its sparse representation. The final decision can be made by majority vote. This baseline approach, however, shows four important disadvantages: *i*) The location information of the patch is not considered, *i.e.*, a patch of one part of the face could be erroneously represented by a patch of a different part of the face. This first problem can be solved by considering the (x, y) location of the patch in its description. *ii)* The method requires a huge dictionary for reliable performance, *i.e.*, each sparse representation process would be very time consuming. This second problem can be remedied by using only a part of the dictionary *adapted* to each patch. Thus, the whole dictionary of a class can be subdivided into sub-dictionaries, and only the 'best' ones can be used to compute the sparse representation of a patch. *iii*) Not all query patches are relevant, *i.e.*, some patches of the face do not provide any discriminative information of the class (*e.g.*, sunglasses when

identifying gender). This third problem can be addressed by selecting the query patches according to a score value. *iv*) It is likely that many images of different classes has common patches, such as similar skin textures when identifying gender, which occur in most faces of all classes and are therefore not discriminating for a particular class. This fourth issue can be addressed using a text retrieval approach including a *visual vocabulary* and a *stop list* to reject those common words [29].

In this section we describe our approach taking into account the four mentioned improvements. As illustrated in Fig. 1, in the learning stage, for each class of the gallery, several random small patches are extracted and described from their images (using both intensity and location features). However, only those patches that are not filtered out by the stop list are considered to build representative dictionaries. In the testing stage, random test patches of the query image are extracted and described. A patch that belongs to the stop list is not considered. For each (considered) test patch a dictionary is built concatenating the 'best' representative dictionary of each class. Using this adapted dictionary, each test patch is classified in accordance with the Sparse Representation Classification (SRC) methodology [32]. Afterwards, the patches are selected according to a discriminative criterion. Finally, the query image is classified by voting for the selected patches. Both stages will be explained in this section in further detail.

2.1 Learning

In the training stage, a set of n face images of k classes is available, where $\mathbf{I}_j^i$ denotes image j of class i (for $i = 1 \dots k$ and $j = 1 \dots n$). In each image $\mathbf{I}_j^i$, m patches are randomly extracted. In this work, the description of a patch $\mathcal{P}$ is defined as vector:

$$\mathbf{y} = f(\mathcal{P}) = [\ \mathbf{z}\ ;\ \alpha x\ ;\ \alpha y\ ] \in \mathcal{R}^{d+2} \tag{1}$$

where $\mathbf{z} = g(\mathcal{P}) \in \mathcal{R}^d$ is a descriptor of patch $\mathcal{P}$; (x, y) are the image coordinates of the center of patch $\mathcal{P}$; and α is a weighting factor between description and location[1]. Using (1) all extracted patches are described as $\mathbf{y}_{jp}^i = f(\mathcal{P}_{jp}^i) = [\ \mathbf{z}_{jp}^i\ ;\ \alpha x_{jp}^i\ ;\ \alpha y_{jp}^i\ ]$, for $p = 1 \dots m$.

In order to eliminate non-discriminative patches, a *stop list* is computed from a *visual vocabulary*. The visual vocabulary is built using all descriptors $\mathbf{Z} = \{\mathbf{z}_{jp}^i\} \in \mathcal{R}^{d \times knm}$, for $i = 1 \dots k$, for $j = 1 \dots n$ and for $p = 1 \dots m$. Array $\mathbf{Z}$ is clustered using a k-means algorithm in N_v clusters. Thus, a visual vocabulary $\mathcal{V}$ containing N_v visual words is obtained. In order to construct the stop list, the *term frequency* 'tf' is computed: $\mathsf{tf}(d, v)$ is defined as the number of occurrences of word v in document d, for $d = 1 \dots K$, $v = 1 \dots N_v$. In our case, a document corresponds to a face image, and $K = kn$ is the number of faces in the gallery. Afterwards, the *document frequency* 'df' is computed: $\mathsf{df}(v) = \sum_d \{\mathsf{tf}(d, v) > 0\}$,

[1] In our experiments, the size of the patch is $w \times w$. The descriptor $\mathbf{z}$ corresponds to the intensity values of the patch subsampled by 2 in both directions, *i.e.*, $d = (w \times w)/4$ given by stacking its columns normalized to unit length in order to deal with different illumination conditions; (x, y) are normalized coordinates (values between 0 and 1).

i.e., the number of faces in the gallery that contain a word v, for $v = 1 \ldots N_v$. The stop list is built using words with highest and smallest df values: On one hand, visual words with highest df values are not discriminative because they occur in almost all images. On the other hand, visual words with smallest df are so unusual that they correspond in most of the cases to noise. Usually, the top 5% and bottom 10% are stopped [29]. Those patches of $\mathbf{Z}$ that belong to the stopped clusters are not considered in the following steps of our algorithm.

Now, for class i an array with the description of all (non stopped) patches $\mathbf{y}_{jp}^i$ is defined as $\mathbf{Y}^i$. The description $\mathbf{Y}^i$ of class i is clustered using a k-means algorithm in Q clusters that will be referred to as *parent* clusters:

$$\mathbf{c}_q^i = \text{kmeans}(\mathbf{Y}^i, Q) \tag{2}$$

for $q = 1 \ldots Q$, where $\mathbf{c}_q^i \in \mathcal{R}^{(d+2)}$ is the centroid of parent cluster q of class i. We define $\mathbf{Y}_q^i$ as the array with all samples $\mathbf{y}_{jp}^i$ that belong to the parent cluster

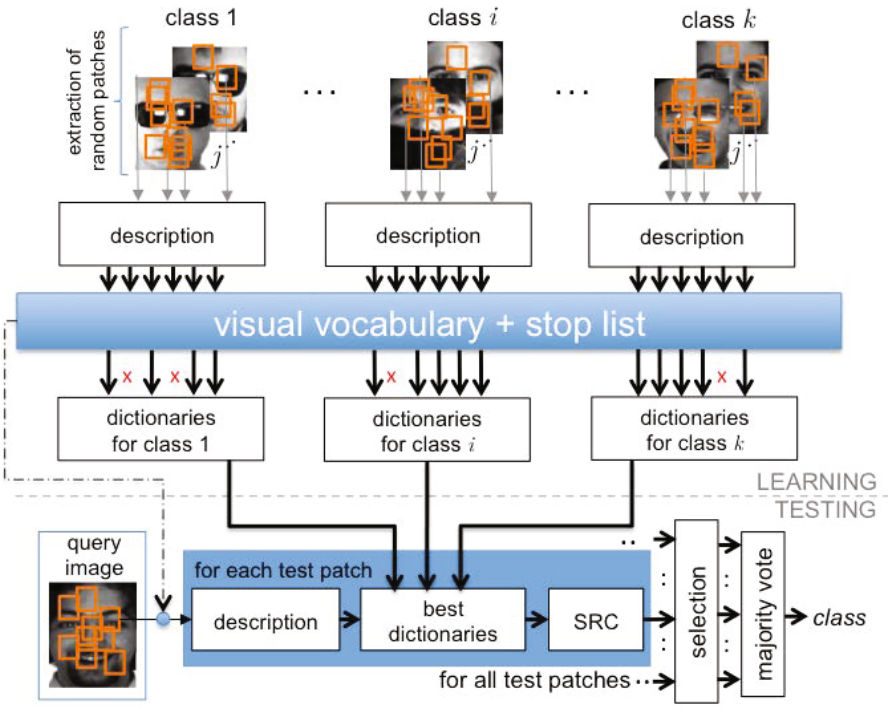


Fig. 1. Overview of the proposed method. The figure illustrates the recognition of disguise. The shown classes are three: sunglasses, scarf and no-disguise. The stop list is used to filter out patches that are not discriminating for these classes. The stopped patches are not considered in the dictionaries of each class and in the testing stage.

with centroid $\mathbf{c}_q^i$. In order to select a reduced number of samples, each parent cluster is clustered again in R *child* clusters[2]

$$\mathbf{c}_{qr}^i = \text{kmeans}(\mathbf{Y}_q^i, R) \tag{3}$$

for $r = 1 \dots R$, where $\mathbf{c}_{qr}^i \in \mathcal{R}^{(d+2)}$ is the centroid of child cluster r of parent cluster q of class i. All centroids of child clusters of class i are arranged in an array $\mathbf{D}^i$, and specifically for parent cluster q are arranged in a matrix:

$$\bar{\mathbf{A}}_q^i = [\mathbf{c}_{q1}^i \dots \ \mathbf{c}_{qr}^i \dots \ \mathbf{c}_{qR}^i]^\mathsf{T} \in \mathcal{R}^{(d+2)\times R} \tag{4}$$

Thus, this arrangement contains R representative samples of parent cluster q of class i as illustrated in Fig. 2. The set of all centroids of child clusters of class i ($\mathbf{D}^i$), represents Q representative dictionaries with R descriptions $\{\mathbf{c}_{qr}^i\}$ for $q = 1 \dots Q, r = 1 \dots R$.

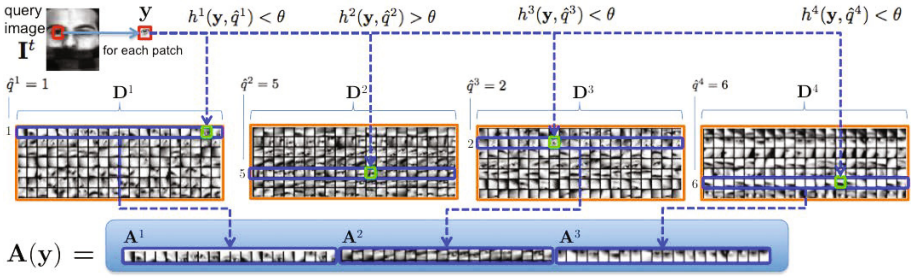


Fig. 2. Adaptive dictionary $\mathbf{A}$ of patch $\mathbf{y}$. In this example there are $k = 4$ classes in the gallery. For this patch only $k' = 3$ classes are selected. Dictionary $\mathbf{A}$ is built from those classes by selecting all child clusters (of a parent cluster -see blue rectangles-) which have a child with the smallest distance to the patch (see green squares). In this example, class 2 does not have child clusters that are similar enough to patch $\mathbf{y}$, *i.e.*, $h^2(\mathbf{y}, \hat{q}^2) > \theta$.

2.2 Testing

In the testing stage, the task is to determine the class of the query image $\mathbf{I}^t$ given the model learned in the previous section. From the test image, s selected test patches $\mathcal{P}_p^t$ of size $w \times w$ pixels are extracted and described using (1) as $\mathbf{y}_p^t = f(\mathcal{P}_p^t) = [\ \mathbf{z}_p^t;\ \alpha x_p^t;\ \alpha y_p^t\ ]$ (for $p = 1 \dots s$). The selection criterion of a test patch will be explained later in this section. For each selected test patch with description $\mathbf{y} = \mathbf{y}_p^t$, a distance to each parent cluster q of each class i of the gallery is measured:

$$h^i(\mathbf{y}, q) = \text{distance}(\mathbf{y}, \bar{\mathbf{A}}_q^i). \tag{5}$$

[2] If n_q^i, the number of samples of $\mathbf{Y}_q^i$, is less than R, $\mathbf{c}_{qr}^i$ is built by taking the R first samples of a replicated version of the samples $[\mathbf{Y}_q^i\ \mathbf{Y}_q^i \dots]$. This dictionary with R words is equivalent to have a dictionary of n_q^i words only.

We tested with several distance metrics. The best performance, however, was obtained by $h^i(\mathbf{y},q) = \min_r ||\mathbf{y}-\mathbf{c}^i_{qr}||$, which is the smallest distance to centroids of child clusters of parent cluster q as illustrated in Fig. 2. Normalizing $\mathbf{y}$ and $\mathbf{c}^i_{qr}$ to have unit ℓ_2 norm, (5) can be rewritten as:

$$h^i(\mathbf{y},q) = 1 - \max < \mathbf{y}, \mathbf{c}^i_{qr} > \text{ for } r = 1 \dots R \tag{6}$$

where the term $< \bullet >$ corresponds to scalar product that provides a similarity (cosine of angle) between vectors $\mathbf{y}$ and $\mathbf{c}^i_{qr}$. The parent cluster that has the minimal distance is searched:

$$\hat{q}^i = \underset{q}{\text{argmin}} \; h^i(\mathbf{y},q), \tag{7}$$

which minimal distance is $h^i(\mathbf{y},\hat{q}^i)$. For patch $\mathbf{y}$, we select those gallery classes that have a minimal distance less than a threshold θ in order to ensure a similarity between the test patch and representative class patches. If k' classes fulfill the condition $h^i(\mathbf{y},\hat{q}^i) < \theta$ for $i = 1 \dots k$, with $k' \leq k$, we can build a new index $v_{i'}$ that indicates the index of the i'-th selected class for $i' = 1 \dots k'$. For instance in a gallery with $k = 4$ classes, if $k' = 3$ classes are selected (*e.g.*, classes 1, 3 and 4), then the indices are $v_1 = 1$, $v_2 = 3$ and $v_3 = 4$ as illustrated in Fig. 2. The selected class i' for patch $\mathbf{y}$ has its dictionary $\mathbf{D}^{v_{i'}}$, and the corresponding parent cluster is $u_{i'} = \hat{q}^{v_{i'}}$, in which child clusters are stored in row $u_{i'}$ of $\mathbf{D}^{v_{i'}}$, *i.e.*, in $\mathbf{A}^{i'} := \bar{\mathbf{A}}^{v_{i'}}_{u_{i'}}$.

Therefore, a dictionary for patch $\mathbf{y}$ is built using the best representative patches as follows (see Fig. 2):

$$\mathbf{A}(\mathbf{y}) = [\ \mathbf{A}^1 \dots \mathbf{A}^{i'} \dots \mathbf{A}^{k'}\ ] \in \mathcal{R}^{(d+2)\times Rk'} \tag{8}$$

With this adaptive dictionary $\mathbf{A}$, built for patch $\mathbf{y}$, we can use SRC methodology [32]. That is, we look for a sparse representation of $\mathbf{y}$ using the ℓ_1-minimization approach:

$$\hat{\mathbf{x}} = \text{argmin}||\mathbf{x}||_1 \quad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{y} \tag{9}$$

The residuals are calculated for the reconstruction for the selected classes $i' = 1 \dots k'$:

$$r_{i'}(\mathbf{y}) = ||\mathbf{y} - \mathbf{A}\delta_{i'}(\hat{\mathbf{x}})|| \tag{10}$$

where $\delta_{i'}(\hat{\mathbf{x}})$ is a vector of the same size as $\hat{\mathbf{x}}$ whose only nonzero entries are the entries in $\hat{\mathbf{x}}$ corresponding to class $v(i') = v_{i'}$. Thus, the class of selected test patch $\mathbf{y}$ will be the class that has the minimal residual, that is it will be

$$\hat{i}(\mathbf{y}) = v(\hat{i'}) \tag{11}$$

where $\hat{i}' = \text{argmin}_{i'} r_{i'}(\mathbf{y})$. Finally, the identity of the query class will be the majority vote of the classes assigned to the s selected test patches $\mathbf{y}_p^t$, for $p = 1 \dots s$:

$$\text{identity}(\mathbf{I}^t) = \text{mode}(\hat{i}(\mathbf{y}_1^t), \dots \hat{i}(\mathbf{y}_p^t), \dots \hat{i}(\mathbf{y}_s^t)) \tag{12}$$

The selection of s patches of query image is as follows:
i) From query image $\mathbf{I}^t$, m^t patches are randomly extracted and described using (1): $\mathbf{y}_j^t$, for $j = 1 \dots m^t$, with $m^t \geq s$.
ii) Those patches $\mathbf{y}_j^t$ that belong to the stopped clusters of our visual vocabulary $\mathcal{V}$ are not considered.
iii) Each remaining patch $\mathbf{y}_j^t$ is represented by $\hat{\mathbf{x}}_j^t$ using (9).
iv) The *sparsity concentration index* (SCI) of each patch is computed in order to evaluate how spread are its sparse coefficients [32]. SCI is defined by

$$S_j := \text{SCI}(\mathbf{y}_j^t) = \frac{k \max(||\delta_{i'}(\hat{\mathbf{x}}_j^t)||_1)/||\hat{\mathbf{x}}_j^t||_1 - 1}{k - 1} \tag{13}$$

If a patch is discriminative enough it is expected that its SCI is large. Note that we use k instead of k' because the concentration of the coefficients related to k classes must be measured.
iv) Array $\{S\}_{j=1}^m$ is sorted into a descended order of SCI value. The first s patches in this sorted list in which SCI values are greater than a τ threshold are then selected. If only s' patches are selected, with $s' < s$, then the majority vote decision in (12) will be taken with the first s' patches.

3 Experimental Results

ASR+ was evaluated in the recognition of several facial attributes: facial expressions (Section 3.1), gender (Section 3.2), race (Section 3.3) and disguise (Section 3.4). Experiments were carried out on seven databases under varying conditions. We demonstrate the performance of our ASR+ approach with a combination of two types of experiments: 1) When it is possible, we compare performance of ASR+ against recent published performance results of a variety of algorithms using the database and similar experimental protocol used in the paper about each algorithm. 2) We compare performance of ASR+ to performance of five 'baseline methods'. They are re-implemented versions of five well-known general recognition algorithms that have been used in face recognition problems. In this case, the methods are the following: *i)* NBNN [6] using intensity features normalized to the unit length in 6×6 partitions, *ii)*, NBNN using LBP-based features [1] with 6×6 partitions, *iii)* SRC [32] where the images were sub-sampled to 22×18 pixels building features of dimension $d = 396$, *iv)* TPTSR based on a two-phase test sample sparse representation approach [33], and *v)* LAD [7] based on locally adaptive sparse representation of patches distributed in a grid. We coded these methods in Matlab according to the specifications given by the authors in their papers.

The used protocol –when evaluating our proposed approach and the baseline methods– is the following: In the databases, there were face images from k classes (*e.g.*, in gender recognition $k = 2$, for female and male) and more than n images per class. All face images were resized to 110 × 90 pixels and converted to a grayscale image if necessary. From each class, n images were randomly chosen for training and one for testing. In order to obtain a better confidence level in the accuracy, the test was repeated N times by randomly selecting $n + 1$ faces images per class each time. The reported accuracy η in all of our experiments is the average calculated over the N tests. In order to report the number of training images and runs of each experiment, we use the notation '$[n|N]$'.

In addition, we report other parameters of our method that depend on the alignment of the face images, the number of training images and the size of the local information of the face that is used in the recognition task. They are the number of parent and child clusters (Q and R), the number of patches extracted in each training image (m), the weighting factor for location coordinates (α), the size of patches (w) and the size of the visual vocabulary (N_v). We use the notation '$\{Q, R, m, \alpha, w, N_v\}$'.

3.1 Facial Expression

The performance of our method was evaluated on three databases: *i)* JAFFE database [20]: It contains 7 expressions ('neutral' and six basic emotions: 'anger', 'disgust', 'fear', 'happiness', 'sadness' and 'surprise') captured from 10 Japanese women. For each subject, there are 3–4 face images for the non-neutral and one for the neutral expressions, *i.e.*, the database consists of 213 images. Results are summarized in Tab. 1. In our case, we used $[n = 29|N = 50]$ and $\{Q = 100, R = 80, m = 250, \alpha = 3, w = 40, N_v = 400\}$. *ii)* CK+ database [18]: It consists of 8 expressions ('contempt' was added to the six basic emotions) captured from 100 subjects as sequences (starting with a neutral face and ending with the peak of a facial expression). In order to compare our method with other methods fairly, a common experimental protocol was followed: The first frame of the sequence (neutral face) and the three last frames (emotion faces) were used. Experiments were carried out to recognize the 6 basic emotions using a leave-one out strategy. Results are summarized in Tab. 1. In our case, we used $[n = 74|N = 50]$ and $\{Q = 100, R = 80, m = 120, \alpha = 0.25, w = 18, N_v = 400\}$. *iii)* SmileFlick (own database): In this experiment, the idea was to detect smiling faces. For this end, 52 face images with smile and 57 face images with neutral expression were collected manually from frontal portraits published in Flickr including subjects from different age, race, gender and illumination. The faces were detected automatically using Computer Vision Toolbox of Matlab[3]. In our experiments, we used $[n = 49|N = 60]$ and $\{Q = 80, R = 50, m = 300, \alpha = 3, w = 40, N_v = 400\}$. The results of our method compared with the baseline methods are summarized in Tab. 1.

[3] http://www.mathworks.com/products/computer-vision/

Table 1. Recognition of Expressions

Database	Method	Ref	η [%]
JAFFE	SLLE	[14]	86.8^{+}
	SFRCS	[13]	86.0^{+}
7	Ada+SVM(RBF)	[28]	81.0^{+}
classes	BDBN_{J}	[15]	91.8^{+}
	BDBN^{*}_{J+C}	[15]	93.0^{+}
	ASR+	(ours)	94.3
CK+	CSPL	[36]	89.9^{+}
	CPL	[36]	88.4^{+}
6	AdaGabor	[4]	93.3^{+}
classes	LBPSVM	[28]	95.1^{+}
	BDBN	[15]	96.7^{+}
	ASR+	(ours)	97.5
SmileFlick	NBNN	[6]	73.1
	LBP	[1]	87.5
2	SRC	[32]	96.8
classes	TPTSR	[33]	91.2
	LAD	[7]	97.5
	ASR+	(ours)	97.5

(*): It was improved using CK+ database.
(+): Result from cited paper.

Table 2. Recognition of Gender

Database	Method	Ref	η [%]
FERET	SVM-RBF	[21]	96.6^{+}
	Real AdaBoost	[34]	93.8^{+}
2	AdaBoost	[3]	94.4^{+}
classes	2DPCA-SVM	[16]	94.8^{+}
	ASR+	(ours)	95.0
GROUPS	NBNN	[6]	84.2
	LBP	[1]	83.3
2	SRC	[32]	86.9
classes	TPTSR	[33]	85.8
	LAD	[7]	87.5
	ASR+	(ours)	93.3

(+): Result from cited paper. Evaluation protocols are not exactly the same (see text).

3.2 Gender

The performance of our method was evaluated on two databases: *i)* FERET database [24]: It contains more than 3,500 face images from women and men (with different races such as African, Asian and Caucasian) involving different expressions and illumination conditions. We used a subset of 1,050 images (602 male and 448 female) where each subject has only one image. We used $[n = 440|N = 200]$ and $\{Q = 160, R = 80, m = 120, \alpha = 3, w = 36, N_v = 200\}$. Results are summarized in Table 2. In order to compare the performance of our approach, Table 2 shows the results obtained by other state-of-art methods, however, the evaluation protocols are not exactly the same. In [21], 1,044 males and 711 females were tested and the accuracy was estimated using a five-fold cross validation strategy. In [34], 3,529 images were used and the accuracy was estimated using a five-fold cross validation strategy. In [3], 2,409 images were used and 80% was used for training and 20% for testing ensuring that images of a particular individual appear only in the training set or test set. In [16], 400 males and 400 females were used and the accuracy was estimated using a five-fold cross validation strategy[4]. *ii)* GROUPS database [10]: It consists of 28,231 face images collected from Flickr images. It is a real–world database containing several facial expressions, face poses, illumination conditions and races. We used the labeled data contained in 'MATLAB DATA' file with 1978 face images (946 males and 1032 females). We used in this case $[n = 700|N = 100]$ and $\{Q =$

[4] There are other experiments on FERET database reported in the literature that are not included in Tab. 2 because the testing protocols are significantly different: In [3], there is an experiment where a subject may appear in both train and test set (in this case, the accuracy is 97.1%). Additionally, in [2], only 304 images (152 males and 152 females) were used for training and 107 images (60 males and 47 females) for testing (in this case, the reported accuracy is 99.1%).

Table 3. Recognition of Race

Database	Method	Ref	η [%]
WebRace	NBNN	[6]	61.3
	LBP	[1]	63.0
5	SRC	[32]	62.0
classes	TPTSR	[33]	65.3
	LAD	[7]	85.7
	ASR+	(ours)	87.1

Table 4. Recognition of Disguise

Database	Method	Ref	η [%]
AR	NBNN	[6]	97.8
	LBP	[1]	96.1
3	SRC	[32]	98.3
classes	TPTSR	[33]	97.8
	LAD	[7]	96.7
	ASR+	(ours)	97.8

$80, R = 50, m = 80, \alpha = 3, w = 16, N_v = 200\}$. Results are summarized in Tab. 2. Our method is compared with the basis methods[5].

3.3 Race

For human beings it is very difficult to distinguish a race, because it depends on how people self identify[6], however, in our paper, the term 'race' –as in [8]– refers to a person's physical appearance rather than sociological and cultural concepts like ethnicity. For this end, we manually built a database from frontal portraits from the web. The images were subjectively collected and categorized in five very different 'races'. The collected races and the number of images per class are the following: 'Asian' (80), 'Black' (89), 'Hispanic' (85), 'Indian' (84) and 'White' (90). We call this database WebRace. The faces were detected automatically using Computer Vision Toolbox of Matlab[3]. In this case, we used $[n = 79|N = 60]$ and $\{Q = 90, R = 90, m = 700, \alpha = 3, w = 48, N_v = 500\}$. The results of our method compared with the baseline methods are summarized in Tab. 3.

3.4 Disguise

In this experiment, the idea was to distinguish faces with certain kind of occlusion. For this purpose, the database AR [20] was used. The images of this database were taken from 100 subjects (50 women and 50 men) with different facial expressions, illumination conditions, and occlusions with sun glasses and scarf (we used the cropped version). The number of images per subject is 26. We divided the database into three groups: images with scarf (600), images with sunglasses (600) and the rest (1400). In this case, we used $[n = 19|N = 60]$ and $\{Q = 80, R = 50, m = 400, \alpha = 2, w = 16, N_v = 200\}$. The results of our method compared with the baseline methods are summarized in Tab. 4.

[5] There is another experiment on GROUPS database reported in [5], in which all 28,231 images were used (in this case, the reported accuracy is 76.0%). Since the evaluation protocol is very different, it is not included in Tab. 2.

[6] See for example the educational game 'Guess my race' which aims to show bias tendencies by presenting that race is the result of complex cultural and historical constructions (http://www.gamesforchange.org/play/guess-my-race/).

3.5 Smile Detection

3.6 Implementation Details

In the implementation of ASR+, we used open source libraries like VLFeat [31] for k-means and SPAMS for sparse representation[7]. Additional to the parameters $\{Q, R, m, \alpha, w, N_v\}$ given in each experiment, the other parameters were (for all experiments): Number of testing patches $m^t = 800$. Threshold for minimal distance between the test patch and child cluster: $\theta = 0.05$. Threshold for SCI $\tau = 0.1$. Number of selected patches $s = 300$. Additionally, the number of words ('atoms') selected from the dictionary in (9) is 20 k'/k, where k' is the number of selected classes for the adaptive sparse representation, and k is the number of classes in the gallery. The time computing depends on the number of classes and the size of the dictionary, however, in order to present a reference, the testing results for the recognition of race were obtained after 0.8s per subject on a Mac Mini Server OS X 10.9.3, processor 2.6 GHz Intel Core i7 with 4 cores and memory of 16GB RAM 1600 MHz DDR3. The remaining algorithms were implemented in MATLAB. The code of the MATLAB implementation is available on our webpage[8].

4 Conclusions

In this paper, we have presented ASR+, a new general algorithm that is able to recognize facial attributes automatically in cases with less constrained conditions, including some variability in ambient lighting, pose, expression, size of the face and distance from the camera. The main contribution of our paper is that the same algorithm can be used in all recognition tasks obtaining a performance at least comparable with that achieved by state-or-art techniques. The robustness of our algorithm is due to three reasons: *i)* the dictionaries learned for each class in the learning stage corresponded to a rich collection of representations of relevant parts which were selected and clustered; *ii)* the testing stage was based on 'adaptive' sparse representations of several patches using the dictionaries estimated in the previous stage which provided the best match with the patches, and *iii)* a visual vocabulary and a stop list used to reject non-discriminative patches in both learning and testing stage.

It is worth mentioning that our extensive empirical evaluation has been performed in two directions: *i)* Other representative methods from the literature have been re-implemented and compared against using our methodology; and *ii)* our algorithm has been evaluated using the methodology of other papers to get a result that can be compared to their published result(s) on the selected datasets. In both scenarios, ASR+ can deal with the unconstrained conditions extremely well, achieving a high recognition performance in many complex conditions and obtaining similar or better performance.

[7] SPArse Modeling Software available on http://spams-devel.gforge.inria.fr
[8] See http://dmery.ing.puc.cl/index.php/material/.

We believe that ASR+ can be used to solve other kinds of recognition problems (*e.g.*, recognition of faces with glasses, mustaches or beards and estimation of age). Preliminary results have shown that ASR+ can be used to recognize specific individuals as well. The proposed model is very flexible and obviously it can be used with other descriptors.

Acknowledgments. This work was supported in part by Fondecyt grant 1130934 from CONICYT–Chile and in part by Seed Grant Program of The College of Engineering at the Pontificia Universidad Catolica de Chile and the College of Engineering at the University of Notre Dame.

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns Application to face recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006)
2. Alexandre, L.A.: Gender recognition: A multiscale decision fusion approach. Pattern Recognition Letters **31**(11), 1422–1427 (2010)
3. Baluja, S., Rowley, H.A.: Boosting sex identification performance. International Journal of Computer Vision **71**(1), 111–119 (2007)
4. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005) (2005)
5. Bekios-Calfa, J., Buenaposada, J.M., Baumela, L.: Robust gender recognition by exploiting facial attributes dependencies. Pattern Recognition Letters **36**, 228–234 (2014)
6. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008) (2008)
7. Chen, Y., Do, T.T., Tran, T.D.: Robust face recognition using locally adaptive sparse representation. In: IEEE International Conference on Image Processing (ICIP 2010), pp. 1657–1660 (2010)
8. Fu, S., He, H., Hou, Z.: Learning race from face: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence **36**(12), 2483–2509 (2014)
9. Fu, Y., Guo, G., Huang, T.S.: Age synthesis and estimation via faces: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence **32**(11), 1955–1976 (2010)
10. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 256–263 (2009)
11. Guo, G., Mu, G.: A study of large-scale ethnicity estimation with gender and age variations. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010), pp. 79–86 (2010)
12. Guo, G., Mu, G.: Joint estimation of age, gender and ethnicity: CCA vs. PLS. In: IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013), pp. 1–6. IEEE (2013)
13. Kyperountas, M., Tefas, A., Pitas, I.: Salient feature and reliable classifier selection for facial expression classification. Pattern Recognition **43**(3), 972–986 (2010)

14. Liang, D., Yang, J., Zheng, Z., Chang, Y.: A facial expression recognition system based on supervised locally linear embedding. Pattern Recognition Letters **26**(15), 2374–2389 (2005)
15. Liu, P., Han, S., Men, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014) (2014)
16. Lu, L., Shi, P.: Fusion of multiple facial regions for expression-invariant gender classification. IEICE Electronics Express **6**(10), 587–593 (2009)
17. Lu, X., Jain, A.K.: Ethnicity identification from face images. In: Proceedings of SPIE Defense and Security Symposium, pp. 114–123 (2004)
18. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: IEEE Workshop on CVPR for Human Communicative Behavior Analysis (2010)
19. Lyons, M.J., Budynek, J., Akamatsu, S.: Automatic classification of single facial images. IEEE Trans. on Pattern Analysis and Machine Intelligence **21**(12), 1357–1362 (1999)
20. Martinez, A., Benavente, R.: The AR face database (June 1998). cVC Tech. Rep., No. 24
21. Moghaddam, B., Yang, M.H.: Learning gender with support faces. IEEE Trans. on Pattern Analysis and Machine Intelligence **24**(5), 707–711 (2002)
22. Moon, H., Sharma, R., Jung, N.: Method and system for robust human ethnicity recognition using image feature-based probabilistic graphical models (2013). US Patent 8,379,937
23. Ng, C.B., Tay, Y.H., Goi, B.-M.: Recognizing human gender in computer vision: A survey. In: Anthony, P., Ishizuka, M., Lukose, D. (eds.) PRICAI 2012. LNCS, vol. 7458, pp. 335–346. Springer, Heidelberg (2012)
24. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Trans. on Pattern Analysis and Machine Intelligence **22**(10), 1090–1104 (2000)
25. Reid, D.A., Samangooei, S., Chen, C., Nixon, M.S., Ross, A.: Soft biometrics for surveillance: An overview. In: Handbook of Statistics, vol. 31, pp. 1–27. Elsevier (2013)
26. Salah, S.H., Du, H., Al-Jawad, N.: Fusing local binary patterns with wavelet features for ethnicity identification. In: Proceedings of IEEE International Conference on Signal and Image Processing (ICSIP 2013), pp. 330–336 (2013)
27. Samangooei, S., Guo, B., Nixon, M.S.: The use of semantic human description as a soft biometric. In: IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2008), pp. 1–7 (2008)
28. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing **27**(6), 803–816 (2009)
29. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: International Conference on Computer Vision (ICCV 2003), pp. 1470–1477 (2003)
30. Tosic, I., Frossard, P.: Dictionary learning. IEEE Signal Processing Magazine **28**(2), 27–38 (2011)
31. Vedaldi, A., Fulkerson, B.: VLfeat: An open and portable library of computer vision algorithms. In: MM 2010: Proceedings of the international conference on Multimedia, pp. 1469–1472, New York (October 2010)

32. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. on Pattern Analysis and Machine Intelligence **31**(2), 210–227 (2009)
33. Xu, Y., Zhang, D., Yang, J., Yang, J.Y.: A Two-Phase Test Sample Sparse Representation Method for Use With Face Recognition. IEEE Trans. on Circuits and Systems for Video Technology **21**(9), 1255–1262 (2011)
34. Yang, Z., Li, M., Ai, H.: An experimental study on automatic face gender classification. In: 18th International Conference on Pattern Recognition (ICPR 2006), vol. 3, pp. 1099–1102 (2006)
35. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. on Pattern Analysis and Machine Intelligence **31**(1), 39–58 (2009)
36. Zhong, L., Liu, Q., Yang, P., Liu, B., Huang, J., Metaxas, D.N.: Learning active facial patches for expression analysis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012) (2012)

Person Identification in Natural Static Postures Using Kinect

Vempada Ramu Reddy(✉), Kingshuk Chakravarty, and S. Aniruddha

Innovation Lab, Tata Consultancy Services Ltd., Kolkata, India
{ramu.vempada,kingshuk.chakravarty,aniruddha.s}@tcs.com

Abstract. Automatic person identification using un-obtrusive methods are of immense importance in the area of computer vision. Anthropometric approaches are robust to external factors including environmental illumination and obstructions due to hair, spectacles, hats or any other wearable. Recently, there have been efforts made on people identification using walking pattern of the skeleton data obtained from Kinect. In this paper we investigate the possibility of identification using static postures namely sitting and standing. Existing gait based identifications, mostly rely on the dynamics of the joints of the skeleton data. In case of static postures the motion information is not available, hence the identification mainly relies on the static distance information between the joints. Moreover, the variation of pose in a particular posture makes the identification more challenging. The proposed methodology, initially sub-divides the body-parts into static, dynamic and noisy parts followed by a combinatorial element responsible for selectively extracting features for each of those parts. Finally a radial basis function support vector machine classifier is used to perform the training and testing for the identification. Results indicate an identification accuracy of more than 97 % in terms of F-score for 10 people using a dataset created with various poses of natural sitting and standing posture.

Keywords: Person identification · Natural static posture · Skeleton joints · Kinect

1 Introduction

Human brain can discriminate between people based on their unique physical as well as behavioural characteristics [1]. Everyday the importance of non-intrusive person identification has been increasing as the technology that can serve several critical applications like video surveillance, people counting, server-room or datacenter authentication, audience measurement etc. Several modalities of person identification (PI) in terms of biometrics already exist in the current literature on computer vision. A few of them include behavioural characteristics like lip movement, typing pattern etc. or physiological signatures like speech,

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 793–808, 2015.
DOI: 10.1007/978-3-319-16181-5_60

face, iris, fingerprint etc. Unfortunately, these modalities are intrusive in nature, thus require direct human interaction for the authentication. Moreover extraction of fingerprint, iris or audio related biometric information (at recognizable form) from a large distance is definitely a challenging job. However, when other cues are not robust enough in discriminating between people, soft-biometrics like global shape [2] can be used to do the person identification. Global shape based approaches mainly utilize physical build of a person like body dimensions, height, length of limbs etc. for identifying a person. This type of systems is comparatively advantageous because it is very difficult to hide and conceal. In addition, global shape traits can also be extracted without making any user interaction, so it is non-intrusive in nature. They can be obtained either by using RGB-D images or by analysing skeleton joint co-ordinates of a particular subject. Fortunately, the Microsoft motion sensing device named Kinect directly provides RGB-D information and 3D co-ordinates of 20 skeleton joints like head, shoulder-center etc. In this paper, instead of storing image/video, we analyse structural build characteristics of a subject using only skeleton data which is more robust to illumination conditions. Skeleton joints can be obtained even if the face of the person is obstructed by hair, if person wear spectacles, hats or any other wearable. However, the skeleton joints obtained from Kinect is somewhat noisy only if the person wears black clothes which is mainly due to infrared sensor.

Several works have already been done on skeleton information based person identification using Kinect. Preis et al. [3] and Sinha et al. [4] did the same from side walking pattern using static as well as dynamic nature of gait features like length of arms, legs, velocity etc. Naresh et al. [5] had proposed a PI system from arbitrary unconstrained walking pattern. Though they [5] obtained 90% identification accuracy for 20 subjects, but the paths of the subjects were predefined (a front walking pattern with Kinect as the reference point) during training phase. Sinha et al. [6] investigated an interesting pose and subpose based concept for modeling arbitrary gait pattern using only skeleton data. They [6] employed unsupervised learning algorithm i.e., K-Means clustering, for identifying 3 poses and 8 subposes. Their method was able to achieve 94% recognition accuracy for 20 subjects. But, all of these skeleton based approaches aimed at identifying an individual based on only movement-pattern rather than static posture. Chakravarty et al. [7] proposed a PI system in static posture. Though they [7] got 96% identification accuracy for 10 subjects, but their method is mainly focused on frontal standing posture, rather than unconstrained natural static ones. In addition, they had carried out performance evaluation using training and testing at a fixed predefined position and posture. However as the subject is not very robotic and can assume variety of poses, their method performs very poorly in real-life. Identifying the person using global shape information obtained from RGB-D is quite easy compared to skeleton but it is quite challenging using skeleton data. For example, if two people are of same height and assume limb lengths are of similar size, still they can be easily discriminated from the width of hands, legs or body from RGB-D as it gives these additional clues. However, skeleton joints are single points we cannot get these crucial information like the

3D structure of person which is very unique. Therefore, identifying the persons of similar structures is quite challenging using skeleton data. Keeping all those problems in mind, we have developed a robust person identification system in static postures mainly sitting and/or standing using only global shape based features. In this work, we have defined sitting, standing, bending etc. as posture where a posture may have many poses. A pose is described as the attitude (e.g. orientation with respect to a reference point) of the body, or the position of the limbs (arms and legs) in a particular posture. While developing the robust PI system using skeleton data, we have explored physical build characteristics of a person in two phases where in the first phase we have explored feature sets related to constrained sitting and standing postures (method 1) and then, based on the drawback-analysis of method 1 in real-life scenario, method 2 is proposed in phase 2. The method 2 does not require any user cooperation and performs well in constrained as well natural sitting and standing postures. The contribution of this paper is mainly 4 folds

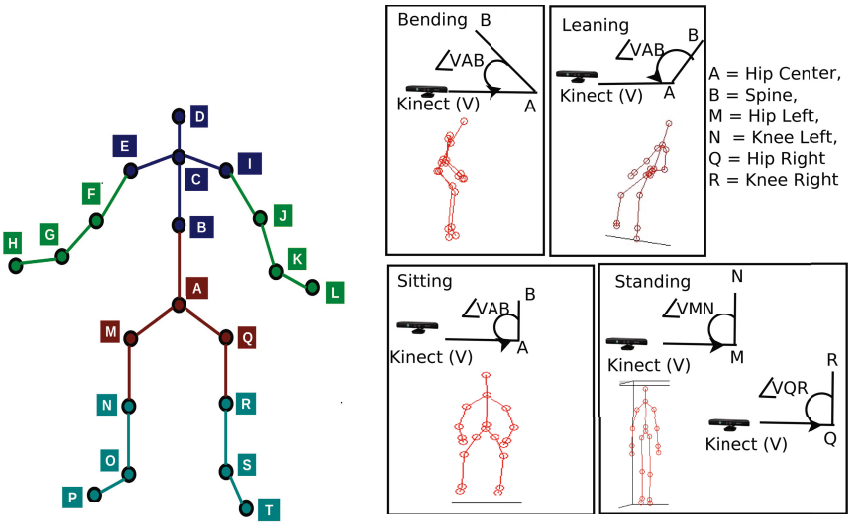
1. PI task is carried out in natural unconstrained static postures in real time using only skeleton data obtained from Kinect. The system is invariant to lighting condition and also ensures user's privacy.
2. Benefits and drawbacks of different feature sets are investigated for identifying an individual in natural and constrained static postures.
3. For robust PI, the pose invariant optimal feature vector is selected from different body-parts after examining combinatorial study on different feature sets.
4. Density based clustering approach is used for dividing entire skeleton structure based on static, dynamic and noisy nature of joints.

We have also evaluated performance of method 2 with respect to the state-of-the-art systems [6] [7] and it is shown that our method outperforms the existing systems in natural static postures.

Rest of the paper is organized as follows: Section 2 gives the brief explanation of posture and poses along with the details of database creation for static postures. Two phase implementation of our proposed PI system is presented in two sections Section 3 and Section 4 where Section 3 gives the performance analysis of different global shape based features on different datasets and Section 4 presents the proposed robust person identification system based on joint analysis of different body-parts. Conclusion of this paper is laid out in the final section.

2 Experimental Database

In this work, we have developed a person identification system using only skeleton information obtained from Kinect [8]. Here we are focusing on the PI task, only in static postures like sitting and standing. For this we have analyzed physical build characteristics in terms of skeleton data. Methods in [6] [4] [9] [3] [5] did the PI by analyzing the movement patterns in terms of spatio-temporal



(a) 20 skeleton joints with labels (b) Representation of postures using angles

Fig. 1. Representation of skeleton structure and postures

variation of skeleton joint co-ordinates. But, unfortunately no standard public database exists for person identification in static postures (specially sitting and standing) using skeleton data. Therefore, we have carefully designed our own database that suits to real-time scenario. In this study, we have used Kinect sensor which is placed at 6-10 ft distance from the subject to collect the skeleton data from sitting and standing posture. It mainly records $\{x, y, z\}$ co-ordinates (in meters) of different skeleton joints for a particular subject. The 20 skeleton joints namely Hip Center(A), Spine(B), Shoulder Center(C), Head(D), Shoulder Left(E), Elbow Left(F), Wrist Left(G), Hand Left(H), Shoulder Right (I), Elbow Right(J), Wrist Right(K), Hand Right(L), Hip Left(M), Knee Left(N), Ankle Left(O), Foot Left(P), Hip Right(Q), Knee Right(R), Ankle Right(S), Foot Right(T) obtained from the Kinect are shown in Fig. 1a. The data is collected from the sitting and standing postures in two modes - 1) *constrained static postures* - frontal sitting and standing pose, and 2) *unconstrained static postures* - natural sitting and standing pose. In both the modes, datasets are created from 10 people (3 female and 7 male). We have presented a brief discussion on posture and pose followed by the details on the corpus creation.

Overview of Posture and Pose

Before going into discussion about database creation on sitting and standing postures with different poses, we want to clarify the difference between posture and pose. Posture is viewed at macroscopic level whereas single posture can have multiple poses. The orientation of the posture with respect to some reference point is treated as pose. Therefore, pose can be viewed as containing microscopic level information. For example, the postures can be like sitting, standing,

sleeping, bending, leaning etc. Any particular posture is independent of person's orientation in the space. However, pose should be defined with respect to some reference point. In our case, if we consider Kinect as the reference, then the orientation of person with respect to Kinect is treated as pose. If the subject is straight towards camera i.e., perpendicular it is treated as straight pose or frontal pose. Else we consider pose (with some angle with respect to Kinect) as natural one. Not only that in natural poses, the subject may vary position of his/her limbs. The skeleton joints extracted from Kinect are represented by 3D world co-ordinates *(x,y,z)* where 'x' represents the left/right variation, 'y' represents up/down variation and 'z' represents to/from variation of subject with respect to Kinect. Scientifically, the angles formed by some joints with respect to Kinect in X-Y (coronal) and/or Z-Y (sagittal) plane can differentiate the posture. Once posture is fixed, the orientation of subject with respect to Kinect in Z-X (transverse) plane can differentiate poses within the posture. As shown in Fig. 1b, postures like leaning, sitting, bending are defined using the angle information which is obtained from the joints like A and B (Fig. 1a) with respect to Kinect(marked as V). However, the posture standing is discriminated from other postures using additional angle information made by the joints M and N or Q and R (Fig. 1a).

Dataset #1

This is created from the respective 10 people in the constrained static postures specifically frontal sitting and standing one. In this case, we have asked the subjects to view straight towards the Kinect. From each subject, we have collected 1 set of data for training and 3 sets of data for testing. Each set consists of 1 minute of data with approximately 30 frames per second. The training set is frontal one where legs are kept perpendicular to Kinect and hands are lied on both the legs at different locations which are varied from Knee to near Hip location. One set of test data is similar to the training set whereas for other two test sets we have requested the subjects to remain in the frontal standing or sitting pose but asked to produce small variations of dynamic joints like leg and hand positions (without crossing legs and folding hands).

Dataset #2

This is also created from the same 10 people of dataset #1 but in unconstrained static poses i.e., natural standing and sitting poses. From each subject, we have collected 2 sets of data, where one set is used for training and other set is for testing. In the training phase, we have asked the subject to sit and stand in some particular predefined poses (one example shown in Fig. 2) but in testing phase we have not restricted the subject in viewing Kinect. Instead the subject is encouraged to sit and stand with some angle to the Kinect. In fact, we have requested the subject to give arbitrary sitting and standing posture by making large variations of dynamic joints. While designing dataset #2 we have emphasized the fact that in real life, during testing, a subject may give totally different static pose that is not present in the training corpus.

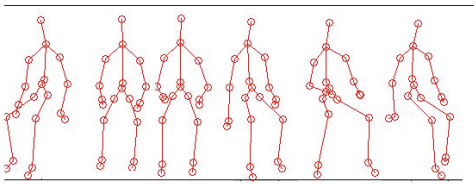


Fig. 2. Representation of different poses of sitting posture

3 Person Identification in Static Postures: Method 1

We have developed a person identification (PI) system from sitting and standing postures in two phases. In the first phase (method 1), the PI system is implemented in two steps (i) **feature extraction** (ii) **decision making and performance analysis**. The method 2 is proposed based on the performance analysis of method 1. In other words, we have critically analysed drawbacks of method 1 and proposed a robust PI system in phase 2 (method 2). It needs to be mentioned that the performance of method 1 & 2 is evaluated using both the datasets #1 & #2. The implementation details of method 1 are presented in the following subsections.

3.1 Feature Extraction

The feature extraction module generates different sets of features for identifying the person in sitting and standing posture. Therefore, identifying appropriate salient features from the 3D world co-ordinates of 20 joints, which can discriminate the individual characteristics, is a very crucial step for any high performance system. The details of features for PI are as follows:

In static postures, meaningful information about identity or uniqueness of any individual can be obtained by extracting the features related to the structural or physical build of the subject (e.g. height, length of limbs etc.). So keeping this fact in mind, we have used differences of 3D world co-ordinates between every pair of joints (physically connected and unconnected) as a candidate feature vector ($\boldsymbol{F}_{cu}$) and $\boldsymbol{F}_{cu}$ is extracted at frame level. The feature set $\boldsymbol{F}_{cu}$ contains all the necessary and unique information about the physical build of a subject whereas, differences of co-ordinate 'x', 'y' and 'z' for every joint-pair capture the width, height and depth information, respectively. From Fig. 1a it is observed that there are 20 joints with 19 physically connected pairs, where the differences of co-ordinates 'y' inherently give the information about length of limbs. The features $\boldsymbol{F}_c$ and $\boldsymbol{F}_y$ represent the differences of 3D world co-ordinates and differences of 'y' co-ordinate between every '*connected*' pair of joints, respectively. In the first phase of our implementation, the candidate features such as $\boldsymbol{F}_c$ and $\boldsymbol{F}_y$ ($\boldsymbol{F}_y$, $\boldsymbol{F}_c \subset \boldsymbol{F}_{cu}$) are extracted from each frame for analysing how they affect PI in

different posing conditions. $\boldsymbol{F}_{cu}$, $\boldsymbol{F}_c$ and $\boldsymbol{F}_y$ are formulated using equations (1), (2) and (3) where J is the total number of joints in D dimensional co-ordinate system and CP represents number of physically connected joint-pairs.

$$\begin{aligned} &\boldsymbol{F}_{cu} = abs((x^j, y^j, z^j) - (x^k, y^k, z^k)) \ \forall \ j = [1, 20], k = [1, 20], j \neq k, \\ &\boldsymbol{F}_{cu} \in \mathbf{R}^{(\mathbf{D} \times \mathbf{J}\mathbf{C_2})}, \text{where } \mathbf{D}{=}3 \text{ and } \mathbf{J}{=}20 \end{aligned} \tag{1}$$

$$\begin{aligned} &\boldsymbol{F}_c = abs((x^j, y^j, z^j) - (x^k, y^k, z^k)) \ \forall \ j, k = \{1, \ldots, 20 | j, k \text{ connected}\}, \\ &\boldsymbol{F}_c \in \mathbf{R}^{(\mathbf{D} \times \mathbf{CP})} \text{ and } \boldsymbol{F}_c \subset \boldsymbol{F}_{cu}, \text{where } \mathbf{D}{=}3 \text{ and } \mathbf{CP}{=}19 \end{aligned} \tag{2}$$

$$\begin{aligned} &\boldsymbol{F}_y = abs((y^j) - (y^k)), \ \forall \ j, k = \{1, \ldots, 20 | j, k \text{ connected}\}, \\ &\boldsymbol{F}_y \in \mathbf{R}^{\mathbf{CP}} \text{ and } \boldsymbol{F}_y \subset \boldsymbol{F}_{cu}, \text{where } \mathbf{CP}{=}19 \end{aligned} \tag{3}$$

3.2 Decision Making and Performance Analysis

The decision making task is carried out using a supervised learning algorithm with feature sets $\boldsymbol{F}_{cu}$, $\boldsymbol{F}_c$ and $\boldsymbol{F}_y$ separately. A classification algorithm is used to map feature vectors to a particular object class representing a person. We have realized the classifier using multi-class support vector machine (SVM) with Radial Basis Function (RBF) as kernel [10] [11]. SVM classification is an example of supervised learning. SVMs are useful due to their wide applicability for classification tasks in many applications [12]- [18]. The main goal of SVM for classification problem is to produce a model which predicts target class label of data instances in the testing set, given only the attributes. The intuition to use RBF kernel function is due to its universal approximation properties. Also, it offers good generalization as well as good performance in solving practical problems [15] [16].

In this study, the statistical measure F-score [6], which is defined as the harmonic mean of precision and recall is used for performance evaluation. For N subjects F-Score is defined by the equation (4).

$$\text{F-score}_i = \frac{2 * precision_i * recall_i}{(precision_i + recall_i)} \quad \forall i, 1 \leq i \leq N \tag{4}$$

For method 1, various types of experiments are then carried out on the datasets explained in the section 2. These are described as follows:

(A) Trained and Tested at Frontal Static Posture.
As an initial step of our experimentation, we have used only dataset #1 for PI in frontal static posture. The identification accuracy in the form of confusion matrix for test set 1 using feature vector $\boldsymbol{F}_{cu}$ is given in Table 1. Table 2 represents the average F-scores of the PI system using feature vectors $\boldsymbol{F}_{cu}$, $\boldsymbol{F}_c$ and $\boldsymbol{F}_y$ separately on all the 3 test sets.

Analysis: The average performance of PI system shown in the diagonal of Table 1 indicates that almost all persons are well classified. But it is also observed from Table 2 that for all the features, performance of method 1 is better on test set

Table 1. Confusion matrix for 10 subjects trained and tested at frontal sitting posture using feature vector F_{cu}. Entries in table indicate F-scores in (%)

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P_2	1.85	**98.15**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P_3	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P_4	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00	0.00	0.00
P_5	0.00	0.00	0.00	0.00	**97.23**	2.01	0.76	0.00	0.00	0.00
P_6	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00	0.00
P_7	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00	0.00	0.00
P_8	0.00	0.65	0.00	0.00	0.00	0.00	0.00	**99.35**	0.00	0.00
P_9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**	0.00
P_{10}	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**100.00**

Table 2. Average F-scores(%) of PI system for frontal training and testing

Test on dataset #1	**Sitting**			**Standing**		
	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$
Set 1	99.47	98.61	97.79	100.00	99.42	98.18
Set 2	92.10	96.78	94.56	91.27	95.65	92.18
Set 3	87.45	90.19	89.20	90.13	92.90	92.14

Table 3. Average F-score(%) of PI system for frontal training and natural testing

Test on dataset #2	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$
Natural Sit	54.60	62.17	58.19
Natural Stand	63.22	68.98	65.37

1 compared to test sets 2 and 3. This is mainly because the test set 1 and the training set have similar poses for the postures. However, if the subject even slightly varies his/her frontal sitting or standing pose (dataset #1–> test sets 2 & 3) like keeping the arm and leg positions different from that of training model, the performance of this implementation degrades (F-scores for set 2 and set 3 in Table 2). Moreover, as $\boldsymbol{F}_{cu}$ includes differences of 3D co-ordinates for both connected and unconnected pairs, it is obvious that $\boldsymbol{F}_{c}$ and $\boldsymbol{F}_{y}$ perform relatively better on test sets 2 & 3 than $\boldsymbol{F}_{cu}$. Therefore, slight variation in legs and arm positions in testing phase largely affects feature vectors related to the unconnected joint-pairs which is present $\boldsymbol{F}_{cu}$.

(B) Trained at Frontal and Tested Using Unconstrained Static Posture. To make our PI system more realistic, we have used frontal sitting and standing data from dataset #1 for training and unconstrained (natural) pose data from dataset #2 for testing. The average F-scores for all feature vectors are compared in Table 3.

Analysis: Table 3 clearly tells us that the results are more worse compared to Table 2. Our analysis suggests that the system performs poorly because of lack of pose variation information in the training data.

(C) Trained and tested at natural static posture
Next, both the training and testing data are taken from dataset #2. The diagonal entries in Table 4 show the average PI performance for 10 subjects using feature vector $\boldsymbol{F}_{cu}$. We have also compared the performance using feature vectors $\boldsymbol{F}_{cu}$, $\boldsymbol{F}_{c}$ and $\boldsymbol{F}_{y}$ in Table 5.

Table 4. Confusion matrix for 10 subjects trained and tested at natural sitting posture using the feature vector F_{cu}. Entries in table indicate F-scores in (%)

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	**62.07**	13.12	0.00	0.00	0.00	14.81	0.00	4.48	0.00	5.52
P_2	6.35	**91.29**	0.00	0.05	2.31	0.00	0.00	0.00	0.00	0.00
P_3	21.48	0.00	**52.10**	6.36	0.00	0.00	0.00	5.62	13.18	1.26
P_4	1.83	0.00	6.67	**89.58**	0.00	0.00	1.92	0.00	0.00	0.00
P_5	20.40	0.00	0.00	0.00	**61.16**	0.00	3.60	0.00	14.84	0.00
P_6	0.00	2.62	20.10	0.00	0.00	**75.23**	0.00	0.00	2.05	0.00
P_7	0.00	0.00	45.17	0.00	0.00	0.00	**54.83**	0.00	0.00	0.00
P_8	0.00	9.22	2.67	16.46	0.00	0.00	0.00	**71.65**	0.00	0.00
P_9	0.00	0.06	28.70	0.00	0.00	2.80	0.85	0.00	**67.59**	0.00
P_{10}	0.00	0.00	2.63	0.98	15.28	0.06	0.00	0.08	0.00	**80.97**

Table 5. Average F-score(%) of PI system for natural training and testing

Test on dataset #2	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$
Natural Sit	70.65	69.80	65.57
Natural Stand	79.12	73.29	69.65

Analysis: From Tables 4 and 5, it is observed that the average performance of method 1 is slightly improved compared to the previous approach. However, the performance of PI is still not satisfactory and we have got maximum 70.65% and 79.12% PI accuracies in natural sitting and standing postures, respectively. Our analysis suggests that even the subject maintains different pose but may not have good control on hands and leg positions due to flexibility of more dynamic nature of joints in natural scenario. It is also seen that some of the joints exhibit noise in some viewing angles due to occlusion and thus make the PI system more erroneous. From the above analysis, we conclude that different features perform better in different conditions for method 1. Hence, if we can carefully select the features based on the orientation of joints, it will definitely improve the system performance. This gives us the motivation to develop more robust PI system by modifying method 1. The following section 4 describes our modified approach.

4 Person Identification in Static Postures: Method 2

We always keep in mind that we have to design a PI system in natural static posture so that it perfectly matches any real-life scenario. Therefore we have developed method 2 by modifying method 1 to overcome the above limitations (described in the section 3). In method 2, we have analyzed joints belong to different body-parts, extracted relevant features and then finally evaluated the performance. The frame-work of method 2 contains 5 modules (i) **Data Acquisition (DA)**, (ii) **Skeleton Divider (SD)**, (iii) **Feature Generator (FG)**, (iv) **Combinatorics Engine (CE)** and (v) **Model Generator (MG)** in its functional architecture which is shown in Fig. 3.

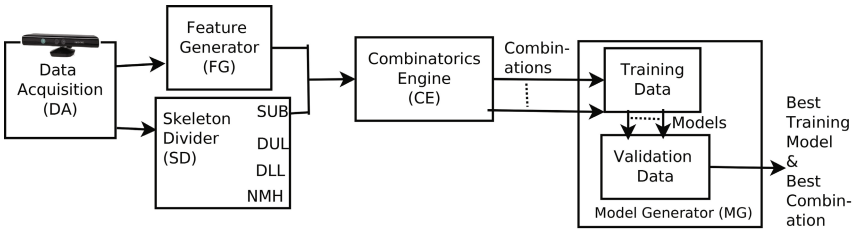


Fig. 3. Functional architecture of method 2

The DA module captures co-ordinates of 20 skeleton-joints using Kinect and forwards these 20 joints to SD module. In SD module, entire skeleton structure is divided into different body-parts based on the static upper, dynamic lower, dynamic upper and noisy middle nature of joints (labeled with different colors in Fig. 1a). FG module extracts the candidate features $\boldsymbol{F}_{cu}$, $\boldsymbol{F}_{c}$ and $\boldsymbol{F}_{y}$ (explained in the section 3) from the skeleton joints. Once the feature generation is done, CE module explores all possible combination of features extracted from different body-parts and finally, all these combinations are feeded to MG module to generate the training models. In method 2, the training and testing are done only on dataset #2 where predefined poses are used for training but testing is carried out with unconstrained natural static postures. The key contribution of this approach is mainly dividing 20 skeleton joints into different body-parts and automatic selection of optimal features from the respective body-parts. In addition to this we have also analyzed the influence of certain angles in capturing the pose related information. The details of the proposed methodology and influence of the angles on the proposed system are presented in the following subsections.

4.1 Methodology

In the static posture like sitting or standing, a person can be oriented in any direction with respect to Kinect exhibiting natural pose. However, for a given posture a person can not move some of the joints flexibly irrespective of poses.

For example, upper body joints like Spine, Hip Center etc. are fixed for any pose in a particular posture. We define those joints as static one. On the contrary, in a single pose, a subject can move his joints like Knee Left, Wrist Left, Foot Left etc. very flexibly. Therefore, we name them as dynamic joints. It is also noticed that some of the joints are more prone to noise due to occlusion effect. For example, in most of the poses of sitting posture, Hip Left and Hip Right are occluded with Knee Left and Knee Right, respectively. These types of joints are considered as noisy joints. This is also verified by grouping the co-ordinates of different joints from upper and lower body-parts using density based clustering algorithm DBSCAN [19]. The results of DBSCAN for some joints (Left portion of the body) for both the postures are illustrated in Table 6. Right portion of the body joints also exhibited the similar trend. Table 6 indicates that for sitting posture(s) DBSCAN identifies 6 clusters whereas for standing posture(s) it forms only 2 clusters. It can also be noticed from the results that in both postures, certain joints of upper body like Shoulder Center, Shoulder Left, Spine etc. form one cluster (static cluster) and Elbow Left, Wrist Left, Knee Left, Ankle Left form another cluster (dynamic cluster). The joints which are varying over frames mainly belong to dynamic cluster and the joints which are static over frames form the static cluster. However, Table 6 also captures an interesting fact that for Hip portion joints like Hip Left, Hip Center, the frames are not clearly separable as pure static or dynamic ones because Hip portion joints are occluded by Knee portion joints while sitting, this causes the noisy nature of Hip joints. 35.02% and 16.09% of HipLeft frames (Table 6) are moved to dynamic cluster in sitting and standing posture. This is mainly because in sitting, the occlusion is more compared to standing. The dynamic joints are further divided into two portions namely dynamic upper and dynamic lower. So, based on the above observation, entire skeleton structure is divided into four parts:

1. *Static Upper Body (SUB)*: The joints B, C, D, E and I representing the main body portion (color coded in dark blue in Fig. 1a) are more static in nature during any pose for a particular posture.
2. *Dynamic Upper Limbs (DUL)*: Based on the subject's flexibility of changing the arm positions in natural static postures, the joints F, G, H, J, K and L are considered as dynamic upper limbs (color coded in green in Fig. 1a).
3. *Dynamic Lower Limbs (DLL)*: Based on the subject's flexibility of changing leg positions in natural static postures, we have considered the joints N, O, P, R, S and T as dynamic lower limbs (color coded in sky blue in Fig. 1a).
4. *Noisy Middle Hip (NMH)*: It is also noticed that if the person varies his/her pose in a particular posture, some joints are reliable and some are noisy. It is mainly due to occlusion of some body portions. This effect is very much vivid in middle hip portion. Therefore, we name the joints A, M and Q as noisy middle hip joints (color coded in deep red in Fig. 1a).

When the body-part segmentation is done, we have explored all the possible combination of features ($\boldsymbol{F}_{cu}$, $\boldsymbol{F}_{c}$ and $\boldsymbol{F}_{y}$) extracted from those body-parts

Table 6. Division of body parts using clustering of joints where Cl.=Cluster

Sitting						Joint	Standing	
Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 6		Cl. 1	Cl. 2
29.96	0.67	*64.66*	0.67	0.00	4.04	HipCenter	*18.15*	*77.85*
4.71	0.54	**93.67**	0.40	0.00	0.67	Spine	13.00	**87.00**
3.10	0.27	**95.69**	0.40	0.13	0.40	ShoulderCenter	13.34	**86.66**
12.11	0.00	**87.08**	0.40	0.13	0.27	ShoulderLeft	4.26	**95.74**
88.83	2.56	3.90	0.67	0.67	3.36	ElbowLeft	**91.20**	8.80
97.98	0.40	0.40	0.00	0.27	0.94	WristLeft	**91.20**	8.80
94.75	0.54	2.83	1.21	0.13	0.54	KneeLeft	**89.96**	10.04
99.33	0.13	0.13	0.00	0.27	0.13	AnkleLeft	**88.72**	11.28
35.02	1.08	*60.94*	0.27	0.54	2.15	HipLeft	*16.09*	*83.91*

(SUB, DUL, DLL and NMH). This is carried out by CE module and it generates total number of combinations $= \sum_{k=1}^{TP} p^k \times \binom{TP}{k}$, where TP= total number of body-parts and p = total number of features. With 3 type of features and 4 body parts, different features extracted from single body part result to 12 combinations $(3^1 \times \binom{4}{1})$. For example, if 3 feature vectors is extracted from single body part at a time and no features are extracted from other body parts, it can be done in three ways. In the same way, three feature vectors extracted from rest of the body parts can be done in 9 ways. Therefore, features extracted from single body part scheme results to total 12 combinations. Similarly, feature vectors extracted from two body parts at a time while maintaining other two body-parts features none can result 54 combinations $(3^2 \times \binom{4}{2})$. Three feature vector combinations for 3 different body parts and no features from left body part will result 108 combinations $(3^3 \times \binom{4}{3})$. Finally, different feature vector combinations including all the body parts result 81 $(3^4 \times \binom{4}{4})$. Thus the system has result to 255 combinations in total. Then all these combinations are feeded to multi-class SVM to generate different models. Now to do the evaluation, we have done 5 fold cross-validation using the training corpus from dataset #2. The average of top 10 PI F-scores are listed in Table 7 for both sitting and standing postures. In Table 7, 'NOT' indicates that none of the feature vectors are employed for that particular body-part. It is found that among all the 255 models, the combination (F_{sit}^{best})– $\boldsymbol{F}_{cu}$ for SUB, $\boldsymbol{F}_c$ for DUL, $\boldsymbol{F}_y$ for DLL and 'NOT' for NMH produces the best F-Score in sitting posture. Similarly for standing posture, we compute the same i.e. F_{stand}^{best}. To test the robustness of method 2, these combinations are applied on the test data of dataset #2 and we able to achieve average 93.00% & 95.33% identification accuracy in sitting and standing, respectively. Table 8 shows the confusion matrix in natural sitting posture for the combination F_{sit}^{best}.

Table 7. Top 10 F-scores(%) of method 2 using combination of features and bodyparts (Cross-validation performance)

Sl. No.	Sitting					Standing				
	SUB	DUL	DLL	NMH	F-score (%)	SUB	DUL	DLL	NMH	F-score(%)
1	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	NOT	95.51	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{cu}$	96.02
2	$\boldsymbol{F}_{cu}$	NOT	$\boldsymbol{F}_{y}$	NOT	94.98	$\boldsymbol{F}_{cu}$	NOT	$\boldsymbol{F}_{c}$	NOT	95.73
3	$\boldsymbol{F}_{cu}$	NOT	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{cu}$	93.01	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	NOT	94.56
4	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{y}$	NOT	91.71	$\boldsymbol{F}_{cu}$	NOT	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{cu}$	94.05
5	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{cu}$	91.27	$\boldsymbol{F}_{cu}$	NOT	$\boldsymbol{F}_{y}$	NOT	93.72
6	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	90.86	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{y}$	91.00
7	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	NOT	$\boldsymbol{F}_{cu}$	89.98	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{cu}$	90.72
8	$\boldsymbol{F}_{c}$	NOT	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	89.45	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{y}$	90.57
9	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	NOT	NOT	89.38	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{c}$	NOT	90.57
10	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{c}$	$\boldsymbol{F}_{cu}$	89.29	$\boldsymbol{F}_{cu}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{y}$	$\boldsymbol{F}_{cu}$	90.36

Table 8. Confusion matrix for 10 subjects using F_{sit}^{best} on test-set of dataset #2

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
P_1	**94.21**	0.00	1.21	2.43	0.00	2.15	0.00	0.00	0.00	0.00
P_2	0.00	**89.79**	0.00	6.14	2.00	0.00	0.00	2.07	0.00	0.00
P_3	3.33	0.00	**90.18**	0.00	3.16	0.00	0.00	0.00	0.00	3.33
P_4	0.00	0.00	0.00	**96.67**	0.00	0.00	0.00	3.33	0.00	0.00
P_5	0.00	6.66	3.33	0.00	**88.02**	0.00	0.00	0.00	1.99	0.00
P_6	2.00	0.00	0.00	0.00	0.00	**98.00**	0.00	0.00	0.00	0.00
P_7	0.00	0.00	0.00	0.00	5.02	0.00	**92.50**	0.00	0.38	2.10
P_8	0.00	0.00	0.00	5.33	0.00	0.00	0.00	**94.67**	0.00	0.00
P_9	1.50	0.00	0.00	6.67	0.00	4.02	0.00	0.00	**87.41**	0.40
P_{10}	0.00	0.00	1.50	0.00	0.00	0.00	0.00	0.00	0.00	**98.50**

Analysis: The top 10 results indicate that in many cases if the features extracted from the body-part NMH are not considered then the performance is better. Even best PI accuracy in sitting posture is obtained without using NMH joints (Row 1 Table 7). It is observed in Table 7 that 'NOT' for NMH joints appeared four times. However, it is observed from the results of 255 combinations this effect is less in standing posture due to less occlusion of NMH joints. Due to space constraint we have not given all 255 combinations. In some cases, it is also seen that if we do not use features from DUL and DLL, method 2 provides good results. This analysis helps us to conclude that the joints belong to NMH are more noisy than the others. Moreover, some joints of DUL and DLL produce noise when the person sits or stands with some orientation other than frontal pose. It is mainly because of occlusion of joints by other body-parts. Therefore, some frames get misclassified which results in slightly reduced performance. Table 7 also emphasizes that in all top 10 results most of the times, DLL and DUL use features $\boldsymbol{F}_{y}$

and $\boldsymbol{F}_c$ but not $\boldsymbol{F}_{cu}$. It is because, the features related to connected joint pairs are sufficient enough to capture the dynamic nature of joints specially arms and legs. Similarly, $\boldsymbol{F}_{cu}$ captures all the information related to static nature of upper main body portion. As the joints belonging to the SUB part are more static in nature, the body segment is proved to be most stable one across all poses in any postures. Not only that, we also explore different angles made by SUB-joints to capture the variation of poses in any natural static posture.

4.2 Influence of Angles

If the person sits or stands in natural pose, the orientation of main body is very crucial for defining a pose. It can be easily captured by computing angles formed by the joints C, E and I from shoulder portion and A, M and Q from hip portion (with respect to Kinect (V)) in Z-X plane. Table 9 shows the effect of these four angles namely ∠VCE, ∠VCI, ∠VAM and ∠VAQ on PI system.

Table 9. F-scores(%) of method 2 without and with angles and F-scores(%) with the methods proposed in [6] & [7]. In sitting $F^{best} = F^{best}_{sit}$ & in standing $F^{best} = F^{best}_{stand}$, and V is Kinect position

Posture	F^{best}	F^{best}, ∠VCE, ∠VCI, ∠VAM and ∠VAQ	F^{best}, ∠VCE and ∠VCI	[6]	[7]
Sitting	93.00	*89.63*	**96.81**	11.16	15.29
Standing	95.33	*93.61*	**97.65**	31.28	20.73

Analysis: Table 9 clearly shows that in both static postures, the performance of method 2 is degraded with the inclusion of these four angles along with the optimal combination F^{best}_{sit} & F^{best}_{stand} (shown in italics column 3 in Table 9). This is mainly due to the inclusion of angles formed by more noisy joints like A, M and Q. However the degradation in performance is less in standing posture compared to sitting one as the occlusion of hip portion is less in natural standing. After removal of these angles (∠VAM and ∠VAQ), it is observed that the performance of method 2 is enhanced further compared to previous one (shown in column 2 and 4 in Table 9). From this we infer that angles formed by the shoulder joints are the key contributors in capturing the variation of pose information in the natural unconstrained static postures. In this study, we have done the step-by-step analysis for making the PI system in static postures more robust and realistic. It needs to be mentioned that using method 2, the feature set F^{best}, ∠VCE and ∠VCI is able to achieve average 96.81% and 97.65% identification accuracy, in sitting and standing postures, respectively.

For the sake of completion of the analysis, we compare with the features proposed earlier for walking pattern in [6]. Sinha et al. had done the pose and subpose based modeling using static and dynamic gait features in [6]. We have also tested their approach on our dataset #2. But their performance on our dataset is not

very satisfactory. It is mainly because their proposed features related to poses and subposes [6] fail to model pose variations in static postures. In addition, we have explored the method mentioned by Chakravarty et al. [7] on dataset #2. As the feature vector used in [7], is strictly focused on constrained frontal standing pose, their system fails to identify most of the subjects in natural sitting and standing poses. The performance comparison of method 2 with the state-of-the-art systems [6] & [7] is presented in the last 2 columns of Table 9. As expected the features for walking or constrained standing posture are not good for the unconstrained natural static scenario.

5 Conclusions

In this work, we have proposed a PI system in 2 phases. In the first phase, different sets of global shape based features which represent the identity of the person are explored. These features are then extracted from constrained and unconstrained datasets of sitting and standing postures. Based on the analysis and drawbacks of certain features for different body-parts in different poses, robust PI system is proposed in phase 2. In phase 2, clustering algorithm is used to identify static, dynamic and noisy joints. From that analysis, entire skeleton body is divided into four segments and we have explored all possible combinations of features from these segments. It greatly improves PI accuracy from 70.65% to 93% in sitting and 79.12% to 95.33% in standing posture. The effect of angle information from shoulder and hip portions is also analysed and it is found that inclusion of angles from hip portion degrades the system performance whereas angles extracted from shoulder portion enhances PI accuracy to 96.81% and 97.65% for both sitting and standing postures, respectively. Performance evaluation matrices also portray the significant improvement of identification accuracy in static postures over the contemporary systems. In future, we like to incorporate more static postures in our proposed system. We have also like to improve the system performance accuracy use angle information obtained from different joints i.e, transforming all poses to frontal pose using angle information and then extracted the features. Moreover we have a plan to combine our approach with other soft-biometric traits like gait, skin color etc. to build a multimodal PI system.

References

1. Anastassiou, G.A., Duman, O.: Introduction. In: Anastassiou, G.A., Duman, O. (eds.) Towards Intelligent Modeling: Statistical Approximation Theory. ISRL, vol. 14, pp. 1–8. Springer, Heidelberg (2011)
2. Jain, A.K., Dass, S.C., Nandakumar, K.: Can soft biometric traits assist user recognition? In: Defense and Security, International Society for Optics and Photonics, pp. 561–572 (2004)
3. Preis, J., Kessel, M., Werner, M., Linnhoff-Popien, C.: Gait recognition with kinect. In: 1st International Workshop on Kinect in Pervasive Computing (2012)

4. Sinha, A., Chakravarty, K., Bhowmick, B.: Person identification using skeleton information from kinect. In: ACHI 2013, The Sixth International Conference on Advances in Computer-Human Interactions, pp. 101–108 (2013)
5. Kumar, M., Babu, R.V.: Human gait recognition using depth camera: a covariance based approach. In: Proc. of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP), vol. 20. ACM (2012)
6. Sinha, A., Chakravarty, K.: Pose based person identification using kinect. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC) 2013, pp. 497–503 (2013)
7. Chakravarty, K., Chattopadhyay, T.: Frontal-standing pose based person identification using kinect. In: Kurosu, M. (ed.) HCI 2014, Part II. LNCS, vol. 8511, pp. 215–223. Springer, Heidelberg (2014)
8. Microsoft: Kinect sdk (2012). http://www.microsoft.com/en-us/kinectforwindows/develop/developer-downloads.aspx. Accessed 29 June 2014
9. Ball, A., Rye, D., Ramos, F., Velonaki, M.: Unsupervised clustering of people from skeletondata. In: 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 225–226. IEEE (2012)
10. Cortes, C., Vapnik, V.: Support-vector networks, vol. 20, pp. 273–297. Springer (1995)
11. Dietterich, T.G., Bakiri, G.: Solving multiclass learning problems via error-correcting output codes (1995)
12. Koolagudi, S.G., Reddy, R., Rao, K.S.: Emotion recognition from speech signal using epoch parameters. In: 2010 International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2010)
13. Rao, K.S., Reddy, R., Maity, S., Koolagudi, S.G.: Characterization of emotions using the dynamics of prosodic. Proc. speech prosody, vol. 4 (2010)
14. Rao, K.S., Koolagudi, S.G., Vempada, R.R.: Emotion recognition from speech using global and local prosodic features. International Journal of Speech Technology **16**(2), 143–160 (2013)
15. Reddy, V.R., Sinha, A., Seshadri, G.: Fusion of spectral and time domain features for crowd noise classification system. In: 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 1–6. IEEE (2013)
16. Reddy, V.R., Chattopadhyay, T.: Human activity recognition from kinect captured data using stick model. In: Kurosu, M. (ed.) HCI 2014, Part II. LNCS, vol. 8511, pp. 305–315. Springer, Heidelberg (2014)
17. Vempada, R., Kumar, B., Rao, K.: Characterization of infant cries using spectral and prosodic features. In: 2012 National Conference on Communications (NCC), pp. 1–5. IEEE (2012)
18. Chattopadhyay, T., Reddy, V.R., Garain, U.: Automatic selection of binarization method for robust ocr. In: 2013 12th International Conference on Document Analysis and Recognition (ICDAR), pp. 1170–1174. IEEE (2013)
19. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. **96**, 226–231 (1996)

Activity-Based Person Identification Using Discriminative Sparse Projections and Orthogonal Ensemble Metric Learning

Haibin Yan[1], Jiwen Lu[2], and Xiuzhuang Zhou[3(✉)]

[1] National University of Singapore, Singapore, Singapore
[2] Advanced Digital Sciences Center, Singapore, Singapore
[3] Captial Normal University, Beijing, China
zxz@xeehoo.com

Abstract. In this paper, we propose an activity-based human identification approach using discriminative sparse projections (DSP) and orthogonal ensemble metric learning (OEML). Unlike gait recognition which recognizes person only from his/her walking activity, this study aims to identify people from more general types of human activities such as eating, drinking, running, and so on. That is because people may not always walk in the scene and gait recognition fails to work in this scenario. Given an activity video, human body mask in each frame is first extracted by background substraction. Then, we propose a DSP method to map these body masks into a low-dimensional subspace and cluster them into a number of clusters to form a dictionary, simultaneously. Subsequently, each video clip is pooled as a histogram feature for activity representation. Lastly, we propose an OEML method to learn a similarity distance metric to exploit discriminative information for recognition. Experimental results show the effectiveness of our proposed approach and better recognition rate is achieved than state-of-the-art methods.

Keywords: Human identification · Activity analysis · Subspace learning · Sparse coding · Metric learning

1 Introduction

Over the past two decades, gait recognition has attracted much attention in computer vision [11], [36], [9], [28], [1], [18], [31], [12], [17], [16], [13], [35], [20], [23] because human gait provides a noninvasive way to human identification at a distance. One key shortcoming of gait recognition is that only the walking activity is exploited for human identification and these gait recognition systems are likely to fail to work when people perform other activities such as eating, drinking, and running rather than walking. In many real-world applications, people may not always walk in the scene and it is very likely that they are performing other activities besides walking in the scene. Since gait can provide enough discriminative information for human identification, a natural question

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 809–824, 2015.
DOI: 10.1007/978-3-319-16181-5_61

arises: is it possible to identify people from different types of activities rather than gait since gait can be considered as a special case of general human activities? If so, how to effectively explore discriminative features of these activities to achieve this goal? In this paper, we provide a positive answer to these two questions.

Intuitively, the manner with which humans perform different activities can provide some distinctive information for human identification because human body information is generally distinct for different persons. Moreover, different dynamic information observed in other activities are also discriminative. Similar to gait recognition, people may perform the same activity in different manners. While gait recognition [11], [36], [9], [28], [1], [31], [35] has been extensively studied over the past decade, there has been extremely few attempts on using other activities rather than gait for human identification. In this paper, we present a new approach to activity-based human identification. For each activity video, human body mask in each frame is extracted by background substraction. Then, we project these body masks into a low-dimensional subspace and cluster them into a number of clusters, simultaneously. Subsequently, each video clip is pooled as a histogram feature for activity representation. Finally, we propose an OEML method to learn a discriminative distance metric for discriminative feature extraction. Experimental results show the effectiveness of our proposed approach.

2 Related Work

Human Activity Analysis: In computer vision, a large number of activity recognition methods have been proposed in recent years [38], [24], [34], [27], [33], [25], [8], [30]. Unlike activity recognition which aims to recognize the type of human activity from videos, activity-based human identification is a relatively new research topic, and there has been only a few seminal studies in recent years [4], [7], [14]. To our best knowledge, Gkalelis *et al.* [4] was the first attempt to formally address the problem of activity-based human identification by using fuzzy c-means (FCM) and linear discriminant analysis (LDA). Their method was further evaluated on more activity datasets and encouraging results were achieved to show the feasibility of human identification using activities [7]. More recently, Lu *et al.* [14] presented a sparse coding method for activity-based human identification. Since the the quantization error is reduced, their method achieved better performance than [4]. However, both FCM and sparse coding are not discriminative enough since they are generative methods. Moreover, these methods performed feature quantization in the original feature space, which may not be effective enough because some irrelevant and redundancy information are contained in this space. To address these shortcomings, we propose a discriminative sparse projections (DSP) method to learn a low-dimensional subspace for feature quantization, so that the irrelevant information of human body masks is discarded in the learned subspace and a discriminative codebook can be obtained for feature encoding.

Metric Learning: Metric learning has been proven to be an effective tool for visual analysis and many such algorithms have been presented over the past decade [5], [37], [2], [26], [19], [15], [22], [21], [39]. While these methods have achieved reasonably good performance in many computer vision applications, these methods usually suffer from high-dimensional feature representations. To address this, PCA is usually applied to reduce the feature dimensionality before metric learning. However, such a preprocessing may lose some discriminative information. In this paper, we propose a new OEML method to learn multiple projections from randomly sampled subsets of training samples, and orthogonalize these projections and combine them into a distance metric. Hence, no PCA preprocessing is required in our method. Moreover, the basic vectors of our learned distance metric are orthogonal to each other such that they are more compact than those of most existing metric learning methods [5], [37], [2], [26], [15].

3 Proposed Approach

Our key objective of this work is to learn discriminative identity information from activities for person recognition. Such information can be exploited at two levels: the single frame level and the whole video level. To extract discriminative information at the single frame level, we propose simultaneously learning a low-dimensional subspace and a discriminative dictionary, so that the irrelevant and redundancy information of body masks are discarded in the learned subspace and discriminative information can be exploited in the learned dictionary. To extract discriminative information at the whole video level, we propose OEML to learn a discriminative distance metric to enhance their separability. We will detail the proposed approach in the following subsections.

3.1 Body Mask Extraction

For each activity video, we first extract human body silhouette in each frame by background subtraction by using the method in [28]. Then, we align each body mask into 64×48 in each frame to make all body masks in different frames are of the same size. Fig. 1 shows several extracted body masks from different types of activities.

3.2 Discriminative Sparse Projections

Let $Y = [y_1, y_2, \cdots, y_N] \in R^{d \times N}$ be a training set of binary masks, where $y_i \in R^d$ is the ith sample, d is the feature dimension of each y_i, and N is the number of training samples. The aim of DSP is to learn a low-dimensional subspace $P \in R^{l \times d}$ and a codebook $U \in R^{l \times K}$, under which each sample y_i is encoded as $v_i \in R^K$ so that

1. Each sample y_i is sparsely reconstructed by v_i over U;
2. The intraclass and interclass variations of each y_i are minimized and maximized, simultaneously.

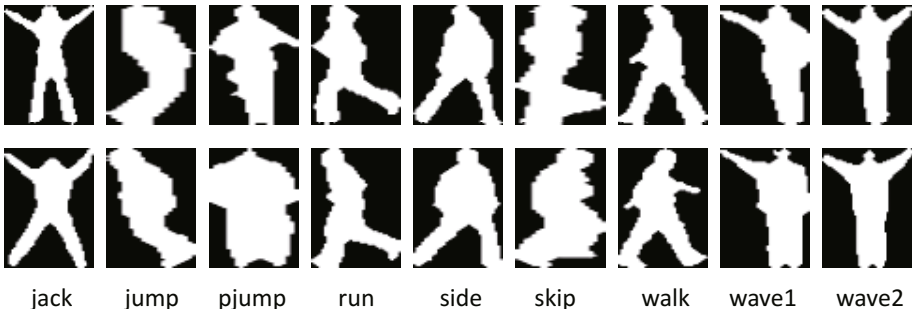


Fig. 1. Extracted and aligned body masks from different activities in the Weizmann dataset

We present the following optimization objective function to achieve the above goals:

$$\begin{aligned} \min_{P,U,V} \ & \|PY - UV\|_F^2 + \alpha\|Y - P^T PY\|_F^2 \\ & +\beta\Big(\sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^c - \sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^p\Big) \\ \text{subject to } \ & PP^T = I, \|v_i\|_0 \le T_0, \text{and} \|u_i\|_F^2 \le 1, \forall i. \end{aligned} \tag{1}$$

where $I \in R^{l\times l}$ is the identity matrix, α and β are non-negative constants and they were empirically set as 1.0 and 1.0 in our experiments, P is the learned low-dimensional subspace, and rows of P are enforced to be orthogonal and normalized to unit norm. U is the dictionary learned in the low-dimensional subspace, $\|u_i\|_F^2 \le 1$ is to constrain the scale of u_i, V is the sparse representation of Y over U, and T_0 is the sparsity level, W^c and W^p are two affinity matrices to characterize the geometrical structure of the samples in the training set, which are defined as [40]:

$$W_{ij}^c = \begin{cases} 1 & \text{if } x_i \in N_{k_1}^+(x_j) \text{ or } x_j \in N_{k_1}^+(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

and

$$W_{ij}^p = \begin{cases} 1 & \text{if } x_i \in N_{k_2}^-(x_j) \text{ or } x_j \in N_{k_2}^-(x_i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $N_{k_1}^+(x)$ and $N_{k_1}^-(x)$ denote the k_1-intra-class and k_2-inter-class nearest neighbors of x, respectively, and k_1 and k_2 are two empirically pre-specified

parameters to define the sizes of the local neighborhoods. With some algebraic deduction, the third term of Eq. (1) can be simplified as

$$\begin{aligned} &\sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^c - \sum_{ij}^{N} \|v_i - v_j\|^2 W_{ij}^p \\ =\ & tr(V^T L^c V) - tr(V^T L^P V) \end{aligned} \tag{4}$$

where $L^c = D^c - W^c$ and $L^p = D^p - W^p$ are two Laplacian matrices, $D_{ii}^c = \sum_j W_{ij}^c$ and $D_{ii}^p = \sum_j W_{ij}^p$ are two diagonal matrices to reflect the degree of the ith sample, respectively.

In Eq. (1), the first term aims to seek sparse signals in the low-dimensional subspace, the second term preserves the energy of the samples in the learned subspace as much as possible, the third term aims to maximize the between-class margin in a local neighborhood.

While the objective function in Eq. (1) is not convex over P, U and V, it is convex to one of them when the other two are fixed. Following the work [10], we iteratively optimize P, U and V using the following three-stage method:

Step 1: Solve P with fixed U and V: when U and V are fixed, Eq. (1) can be rewritten as

$$\begin{aligned} &\min_P \ \|PY - UV\|_F^2 + \alpha \|Y - P^T PY\|_F^2 \\ &\text{subject to } \ PP^T = I. \end{aligned} \tag{5}$$

Let $Q = UVY^{-1}$. Eq. (5) can be formulated as

$$\begin{aligned} &\min_P \ \|P - Q\|_F^2 + \alpha \|I - P^T P\|_F^2 \\ &\text{subject to } \ PP^T = I. \end{aligned} \tag{6}$$

We construct a Lagrange function as follows

$$\mathcal{L}(P, \mu) = \|P - Q\|_F^2 + \alpha \|I - P^T P\|_F^2 - \mu(PP^T - I) \tag{7}$$

Let $\frac{\partial \mathcal{L}(P,\mu)}{\partial P} = 0$ and $\frac{\partial \mathcal{L}(P,\mu)}{\partial \mu} = 0$, we have

$$\frac{\partial \mathcal{L}(P, \mu)}{\partial P} = (1 - \alpha - \mu)P - 2Q = 0 \tag{8}$$

$$\frac{\partial \mathcal{L}(P, \mu)}{\partial \mu} = PP^T - I = 0 \tag{9}$$

According to Eqs. (8) and (9), P can be obtained as

$$P = \frac{UVY^{-1}}{2\|UVY^{-1}\|_F^2} \tag{10}$$

Input: Training set $Y = [y_1, y_2, \cdots, y_N] \in R^{d \times N}$, affinity matrices W^c and W^p, parameters α, β, T_0, iteration number R, convergence error ϵ.
Output: Projection matrix P, dictionary U, and sparse coefficient matrix V.
Step 1 (Initialization):
Compute the initiations: P^0, U^0 and V^0.
Step 2 (Local optimization):
For $r = 1, 2, \cdots, R$, repeat
2.1. Solve P with fixed U and V via Eq. (10).
2.2. Solve U with fixed P and V via Eq. (11).
2.3. Solve V with fixed P and U via Eq. (13).
2.3. If $r > 2$ and $|U^r - U^{r-1}| < \epsilon$, go to Step 3.
Step 3 (Output):
Output P^r, U^r, and V^r.

Algorithm 1. DSP

Step 2: Solve U with fixed P and V: when P and V are fixed, Eq. (1) can be rewritten as

$$\begin{aligned} \min_{U} \ & \|PY - UV\|_F^2 \\ \text{subject to } & \|u_i\|_F^2 \leq 1, \forall i. \end{aligned} \tag{11}$$

Eq. (11) is a least square problem with quadratic constraints. There are many possible methods to solve this problem. Following [10], we use the conjugate gradient decent method to learn the dictionary U.

Step 3: Solve V with fixed P and U: when P and U are fixed, Eq. (1) can be rewritten as

$$\begin{aligned} \min_{V} \ & \|PY - UV\|_F^2 + \beta(tr(V^T L^c V) - tr(V^T L^P V)) \\ \text{subject to } & \|v_i\|_0 \leq T_0, \forall i. \end{aligned} \tag{12}$$

Following the work in [10], we optimize each v_i individually by fixing other coefficients v_j $(j \neq i)$. We rewrite Eq. (12) as

$$\begin{aligned} \min_{v_i} \ & \|PY - Uv_i\|_F^2 + \beta G(v_i) \\ \text{subject to } & \|v_i\|_0 \leq T_0, \forall i. \end{aligned} \tag{13}$$

where

$$G(v_i) = (v_i V L_i^c + (V L_i^c)^T v_i - v_i L_{ii}^c v_i) - (v_i V L_i^p + (V L_i^p)^T v_i - v_i L_{ii}^p v_i) \tag{14}$$

We apply the feature sign search algorithm [10] to solve each v_i.

Now, we discuss how to set the initiations of our proposed DSP method. According to the second term of Eq. (1), the objective of P is to preserve the energy of the samples in the learned subspace as much as possible. Hence, we first learn a PCA subspace on Y as the initiation of P^0. Then, we apply P_0 to map Y into a low-dimensional subspace Y_1. Lastly, we employ the conventional

sparse coding method [41] on Y_1 to learn U^0 and V^0 as the initiations of U and V. The proposed DSP method is summarized in **Algorithm 1**.

Having obtained $V = [v_1, v_2, \cdots, v_M]$ for a set of human body masks extracted from one activity video clip, we represent it as $S = [s_1, s_2, \cdots, s_K]$ by a predefined pooling function:

$$s = \mathcal{F}(V) \tag{15}$$

where

$$s_j = \max\{|v_{1j}|, |v_{2j}|, \cdots, |v_{Mj}|\} \tag{16}$$

s_j is the jth element of s, $j = 1, \cdots, K$, K is the size of the codebook U, which is empirically set as 200 in our implementations.

3.3 Orthogonal Ensemble Metric Learning

Let $S = [s_1, s_2, \cdots, s_n]$ be the training set of C different persons, where $s_i \in R^K$ is the feature of the ith sample and n is the number of activity video clips, $L = [l_1, l_2, \cdots, l_n]$ be the labels of the training samples, where $l_i \in [1, 2, \cdots, C]$. OEML aims to seek a distance metric M which pushes s_i and s_j $(l_i = l_j)$ as close as possible, and pull s_i and s_j $(l_i \neq l_j)$ as far as possible, simultaneously, where

$$d_M(s_i, s_j) = \sqrt{(s_i - s_j)^T M (s_i - s_j)} \tag{17}$$

where M is a $K \times K$ square matrix, and $1 \leq i, j \leq n$. Since M is a distance metric, it should be symmetric and positive semi-definite. hence, we can seek a non-square matrix Q of size $K \times K'$, where $K' \leq K$, such that

$$M = QQ^T \tag{18}$$

Then, Eq. (17) can be rewritten as

$$\begin{aligned} d_M(s_i, s_j) &= \sqrt{(s_i - s_j)^T M (s_i - s_j)} \\ &= \sqrt{(s_i - s_j)^T QQ^T (s_i - s_j)} \\ &= \sqrt{(t_i - t_j)^T (t_i - t_j)} \end{aligned} \tag{19}$$

where $t_i = Q^T s_i$ and $t_j = Q^T s_j$.

Different from most existing distance metric learning methods [5], [37], [2], [26] which learn the distance metric over the whole training samples, we randomly sample two groups of samples from the training set and consider them as positive and negative samples for SVM learning. Assume there are C persons in the training set, we generate one group by randomly sampling F $(F \leq \frac{C}{2})$ classes from the whole training samples as positive samples. Then, we generate another group by randomly sampling F classes from the remaining training samples as

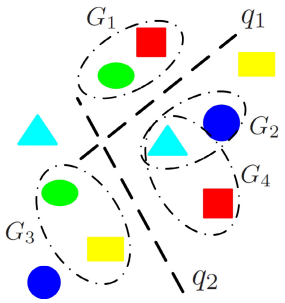


Fig. 2. Learning different projection vectors by SVM from different subsets of the training samples, where q_1 are learned from G_1 and G_2, and q_2 are learned from G_3 and G_4, respectively

Input: Training set: $S = [s_1, s_2, \cdots, s_n]$, label vector $L = [l_1, l_2, \cdots, l_n]$, parameter K'.
Output: Projection matrix Q.
Step 1 (Learning projection vectors with SVM):
For $k = 1, 2, \cdots, K'$, repeat
1.1. Sampling two groups of samples from S.
1.2. Obtain z_k with SVM.
Step 2 (Orthogonazation):
Orthogonalize Z to obtain Q.
Step 3 (Output projection matrix):
Output projection matrix Q.

Algorithm 2. OEML

negative samples. Hence, these two groups don't share any the same-class sample because we need to learn a projection vector to distinguish them.

Then, we learn a linear SVM on these two groups of samples and seek a projection vector $p_i = (w_i - b_i)^T$ to maximize the margin of these two groups of samples, where w_i and b_i are the normal vector and bias of the SVM model. We randomly iterate this procedure K' times and have multiple projection vectors $Z = [z_1, z_2, \cdots, z_{K'}]$. Fig 2 illustrates the basic idea of the learning procedure. In our experiments, we empirically set K' as 200.

Since the projection vectors are learned from the randomly sampled samples, they are not orthogonal. To reduce the redundancy of these projection vectors, we orthogonalize them to make more succinct feature extraction as follows.

Assume $Q = [q_1, q_2, \cdots, q_{K'}]$ be the orthogonal basis vectors of Z. Let $q_1 = p_1$. The ith projection vector q_i can be computed as follows:

$$q_i = z_i - \sum_{j=1}^{i-1} \frac{(q_j)^T z_i}{(q_j)^T q_j} q_j \tag{20}$$

Algorithm 2 summarizes the proposed OEML method.

4 Experimental Results

In this section, we conduct experiments on five different activity databases including the Weizmann [6], AIIA-MOBISERV [7], KTH [29], MSR [35] and TUM [32] databases to evaluate the performance of our proposed approach.

4.1 Datasets and Settings

The Weizmann dataset [6] contains 9 persons and each person performed 10 different activities including bending, jumping-jack, jumping-forward-on-two-legs, jumping in place-on-two-legs, running, galloping-sideways, skipping, walking, waving-one-hand, and waving-two-hands, respectively. There are 93 video clips in this database. Since some videos contain two or more cycles of a specific action performed by some subjects, we break up these videos into several single period activity videos. Hence, we obtain a database of 216 videos in total. For each person, we randomly selected 5 activities for training and the remaining 5 activities were used for testing.

The AIIA-MOBISERV dataset [7] was specifically designed for the activity-based human identification task. It contains 12 persons and each person performed eating and drinking activities with two different clothing in four different days. There are totally 96 videos in this database. Since some videos contain two or more cycles of a specific activity performed by some subjects, these sequences were segmented into several single-period activities. Following the settings in [7], we consider drinking with a cup and eating with a fork for human identification, where 776 video clips in total were selected. We use the eating activity for training and the drinking activity for testing.

The KTH dataset [29] contains 25 persons, and each person performed 6 different activities, including boxing, handclapping, handwaving, jogging, running, and walking, respectively. For each activity, it is captured at 4 different scenarios such as outdoor, indoor, outdoor with a scale variation, and outdoor with different clothes, respectively. In our experiments, we randomly chose 3 activities as training examples for each scenario and the remaining 3 activities as testing examples.

The MSR dataset [35] was captured by a Kinect device. There are 10 subjects in this dataset. For each subject, there are 16 activities: drinking, eating, reading a book, calling a cellphone, writing on a paper, using a laptop, using a vacuum cleaner, cheering up, sitting still, tossing paper, playing game, lying down on sofa, walking, playing guitar, standing up, and sitting, respectively. Each subject performed each activity twice: one in standing position and the other one in sitting position. For each person, both color and depth videos are captured. Hence, there are 320 videos in total. In our experiments, we only use the color videos to evaluate the performance of our approach. We randomly selected 8 activities for each person as training examples and the remaining 8 activities as testing examples.

The TUM dataset [32] is a collection of activity sequences recorded in a kitchen environment equipped with multiple complementary sensors. The recorded data consists of 4 subjects who naturally performed manipulation tasks in a kitchen environment with different manners. Different from previous activity datasets, this dataset offers more natural activities for evaluating activity recognition and motion tracking. There are multiple sensors used to capture human activities such as web camera, RFID and Magnetic (reed) sensors. In our experiments, we only use the video data for human identification. For each person, we selected 4 video sequences captured from 4 synchronized cameras which were installed at 4 different viewpoints. We randomly selected videos from two viewpoints as training examples and the remaining two viewpoints as testing examples.

We also construct a hybrid dataset which combines the Weizmann, AIIA-MOBISERV, KTH, MSR, and TUM databases into a larger dataset to evaluate the performance of our approach. Intuitively, this hybrid dataset is more challenging because there are 50 persons and different persons may perform different types of activities in the hybrid dataset. We followed the above experimental protocol for different datasets to construct the training and test datasets. Specifically, all training sets from each dataset which used in the above experiments were used for training and the remaining videos were used for testing.

We conducted experiments 10 times with different randomly selected training and testing samples, and the final result was shown as the mean of the correct identification rate[1]. In our experiments, the nearest neighbor classifier is used for classification. Since the advantage of our proposed approach results from two different stages: DSP feature encoding and OEML metric learning, we evaluate the performance where only one is applied to reveal their respective effects, respectively.

4.2 Results and Analysis

Comparison with Existing Feature Encoding Methods: We compare our proposed DSP method with different feature encoding methods including the K-means (KM), FCM, sparse coding (SC) [41], Laplacian sparse coding (LSC) [3] on the activity-based human identification task. For the SC and LSC methods, the maximal pooling was also used. The codebook size was set as 300 and the nearest neighbor (NN) classifier with the Euclidian distance was used for identification. Table 1 shows the rank-one identification rate of different feature encoding methods. We can see that our DSP performs better than the other four compared methods. This is because the other compared feature encoding methods are unsupervised and our DSP method is supervised, such that more discriminative information can be exploited in our method. Moreover, our DSP method performs feature encoding in the low-dimensional subspace, which can remove the noisy and irrelevant information in the learned codebook.

[1] The AIIA-MOBISERV dataset was not repeated 10 times because the training and testing sets are fixed in this dataset.

Table 1. Rank-1 identification rate (%) of different feature encoding methods on different datasets

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
KM	64.5	55.4	20.5	24.7	41.7	40.0
FCM	68.3	57.6	24.5	28.6	50.0	43.0
SC	72.1	59.3	27.5	30.6	50.0	45.5
LSC	73.4	61.3	30.4	32.5	58.3	48.8
DSP	**78.5**	**64.5**	**32.7**	**35.6**	**66.7**	**51.3**

Table 2. Rank-1 identification rate (%) of different metric learning methods on different datasets

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
LMNN	75.5	59.5	22.5	27.8	58.3	45.0
NCA	74.3	58.3	21.8	26.9	50.0	44.5
ITML	74.6	58.0	21.6	27.3	50.0	44.0
CSML	76.3	60.5	25.7	30.4	66.7	47.5
NRML	77.5	61.7	28.6	33.5	66.7	49.0
OEML	**80.2**	**65.1**	**32.5**	**36.2**	**75.0**	**52.5**

Comparison with Existing Metric Learning Methods: To investigate the effectiveness of the proposed OEML method in the activity-based human identification task, we compare it with five state-of-the-art metric learning methods including large margin nearest neighbor (LMNN) [37], neighborhood component analysis (NCA) [5], information theoretic metric learning [2], cosine similarity metric learning (CSML) [26], and neighborhood repulsed metric learning (NRML) [15]. For the first four compared methods, we empirically set the number of the nearest neighbors as 5. For the NRML method, two neighborhood sizes were set as 5 and 20, respectively. We also applied principal component analysis (PCA) to reduce each encoded histogram feature learned into 100 dimensions for these five metric learning methods. For the proposed OEML method, we learned the distance metric directly from the original feature space. The FCM method was used for feature encoding. Table 2 compares the rank-1 identification rate of different metric learning methods. We can clearly see from this table that our OEML performs better than the other five compared metric learning. The reason is that the other compared metric learning methods learn the distance metric in the PCA reduced subspace and some discriminative information may be removed in the subspace because the objectives of PCA and these metric learning methods are usually not consistent. However, our OEML method learns the distance metric in the original high-dimensional feature space, which can exploit more discriminative information in the high-dimensional feature space directly.

Comparison with State-of-the-Art Activity-Based human identification Methods: we compare our approach with the state-of-the-art activity-based human identification methods in [4], [7] and [14]. We implemented the three compared methods [4], [7], [14] ourselves. For a fair comparison, the num-

Table 3. Rank-1 identification rate (%) of different activity-based human identification methods on different datasets

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
Method in [4]	70.4	58.6	25.8	29.4	50.0	44.9
Method in [7]	74.3	60.3	27.6	32.3	50.0	48.5
Method in [14]	75.4	62.5	31.4	35.7	66.7	50.2
Ours	**83.3**	**67.5**	**35.8**	**40.3**	**83.3**	**54.9**

Table 4. Rank-1 identification rate (%) of different combinations of feature encoding and metric learning methods on different datasets

Method	Weizmann	AIIA-MOBISERV	KTH	MSR	TUM	Hybird
Baseline	76.1	61.3	24.9	31.2	66.7	50.6
Baseline+DSP	78.5	64.5	32.7	35.6	66.7	52.2
Baseline+OEML	81.3	65.8	33.9	37.5	75.0	52.6
DSP+OEML	**83.3**	**67.5**	**35.8**	**40.3**	**83.3**	**54.9**

ber of clusters is set as 300 in our implementations for all methods. Table 3 compares the rank-1 identification rate of different methods. As can be seen from this table, our approach significantly outperforms the compared activity-based human identification methods because our approach adopts supervised feature encoding and high-dimensional metric learning, such that more discriminative information can be extracted for recognition.

Performance Analysis of Different Stages in Our Approach: We conduct experiments to analyze our approach when different modules are used. We create the baseline method which performs dictionary learning in the original feature space and uses NN for recognition without metric learning. Then, we include different modules in our approach. Table 4 compares the rank-1 identification rate

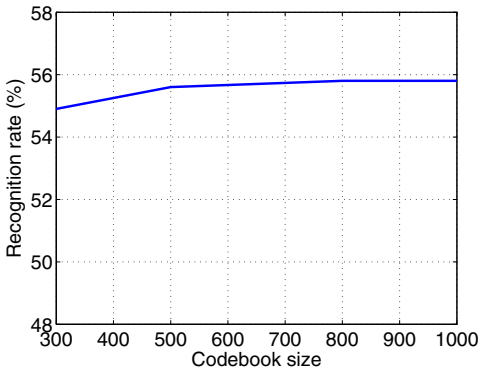


Fig. 3. Rank-1 identification rate (%) of our approach versus different codebook sizes on the hybrid dataset

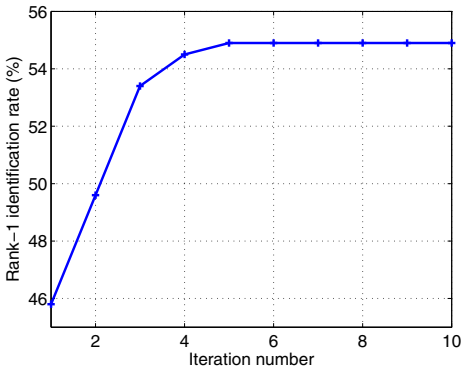


Fig. 4. Rank-1 identification rate versus different number of iterations of DSP on the hybrid dataset

of different combinations of feature encoding and metric learning methods. We see that all modules including low-dimensional subspace, discriminative dictionary learning, and discriminative metric learning contribute the final recognition rate of our approach.

Parameter Analysis: We first evaluate the effect of the codebook sizes of our approach on the hybrid dataset. Fig. 3 shows the rank-1 identification rate of our approach versus different codebook sizes on the hybrid dataset. We see that the performance of our approach continues to increase as the increasing of the codebook size. However, the improvement is marginal, which indicates that the performance of our approach is not sensitive to the codebook size.

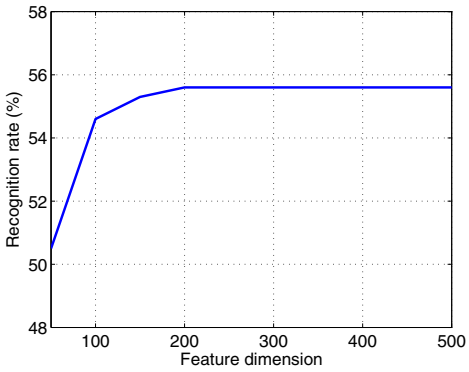


Fig. 5. Rank-1 identification rate (%) of our approach versus different number of feature dimensions on the hybrid dataset

Fig. 4 shows the rank-1 identification rate versus different number of iterations on the hybrid database. We see that the recognition performance of our proposed DSP method can converge to a local optimal peak in a few iterations.

Lastly, we investigated the effect of the parameter K' in OEML. Fig. 5 shows the rank-1 identification rate versus different number of feature dimensions on the hybrid database. We see that our OEML can reach stable performance when the number of K' is above 100.

5 Conclusion

This paper presented a new activity-based human identification approach by using discriminative sparse projections and orthogonal ensemble metric learning (OEML). Experimental results demonstrate the effectiveness of the proposed approach. How to apply our proposed approach to other visual recognition applications such as face identification, object recognition, and visual tracking to further demonstrate its effectiveness seems an interesting future work.

Acknowledgements.. This study is partially supported by the research grant for the Human Cyber Security Systems (HCSS) Program at the Advanced Digital Sciences Center (ADSC) from the Agency for Science, Technology and Research (A*STAR) of Singapore, and the research grant from the National Natural Science Foundation of China under Grant 61373090.

References

1. Boulgouris, N., Hatzinakos, D., Plataniotis, K.: Gait recognition: a challenging signal processing technology for biometric identification. IEEE Signal Processing Magazine **22**(6), 78–90 (2005)
2. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-theoretic metric learning. In: International Conference on Machine Learning, pp. 209–216 (2007)
3. Gao, S., Tsang, I.W., Chia, L.T., Zhao, P.: Local features are not lonely-laplacian sparse coding for image classification. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 3555–3561 (2010)
4. Gkalelis, N., Tefas, A., Pitas, I.: Human identification from human movements. In: IEEE International Conference on Image Processing, pp. 2585–2588 (2009)
5. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighborhood component analysis. In: Advances in Neural Information Processing Systems, pp. 2539–2544 (2004)
6. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(12), 2247–2253 (2007)
7. Iosifidis, A., Tefas, A., Pitas, I.: Activity-based person identification using fuzzy representation and discriminant learning. IEEE Transactions on Information Forensics and Security **7**(2), 530–542 (2012)
8. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(1), 172–185 (2011)

9. Kale, A., Sundaresan, A., Rajagopalan, A., Cuntoor, N., Roy-Chowdhury, A., Kruger, V., Chellappa, R.: Identification of humans using gait. IEEE Transactions on Image Processing **13**(9), 1163–1173 (2004)
10. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, vol. 19, p. 801 (2006)
11. Lee, L., Grimson, W.: Gait analysis for recognition and classification. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 148–155 (2002)
12. Li, X., Maybank, S., Yan, S., Tao, D., Xu, D.: Gait components and their application to gender recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **38**(2), 145–155 (2008)
13. Liu, N., Lu, J., Tan, Y.P.: Joint subspace learning for view-invariant gait recognition. IEEE Signal Processing Letters **18**(7), 431–434 (2011)
14. Lu, J., Hu, J., Zhou, X., Shang, Y.: Activity-based person identification using sparse coding and discriminative metric learning. In: ACM International Conference on Multimedia, pp. 1061–1064 (2012)
15. Lu, J., Hu, J., Zhou, X., Shang, Y., Tan, Y.P., Wang, G.: Neighborhood repulsed metric learning for kinship verification. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2594–2601 (2012)
16. Lu, J., Tan, Y.P.: Gait-based human age estimation. IEEE Transactions on Information Forensics and Security **5**(4), 761–770 (2010)
17. Lu, J., Tan, Y.: Uncorrelated discriminant simplex analysis for view-invariant gait signal computing. Pattern Recognition Letters **31**(5), 382–393 (2010)
18. Lu, J., Zhang, E.: Gait recognition for human identification based on ica and fuzzy svm through multiple views fusion. Pattern Recognition Letters **28**(16), 2401–2411 (2007)
19. Lu, J., Tan, Y.P.: Regularized locality preserving projections and its extensions for face recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **40**(3), 958–963 (2010)
20. Lu, J., Tan, Y.P.: Ordinary preserving manifold analysis for human age and head pose estimation. IEEE Transactions on Human-Machine Systems **43**(2), 249–258 (2013)
21. Lu, J., Tan, Y.P., Wang, G.: Discriminative multimanifold analysis for face recognition from a single training sample per person. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 39–51 (2013)
22. Lu, J., Wang, G., Moulin, P.: Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning. In: IEEE International Conference on Computer Vision, pp. 329–336 (2013)
23. Lu, J., Wang, G., Moulin, P.: Human identity and gender recognition from gait sequences with arbitrary walking directions. IEEE Transactions on Information Forensics and Security **9**(1), 51–61 (2014)
24. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2929–2936 (2009)
25. Mikolajczyk, K., Uemura, H.: Action recognition with motion-appearance vocabulary forest. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
26. Nguyen, H., Bai, L.: Cosine similarity metric learning for face verification. Asian Conference on Computer Vision, pp. 709–720 (2011)
27. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6), 976–990 (2010)

28. Sarkar, S., Phillips, P., Liu, Z., Vega, I., Grother, P., Bowyer, K.: The humanid gait challenge problem: Data sets, performance, and analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(2), 162–177 (2005)
29. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: IEEE International Conference on Pattern Recognition, vol. 3, pp. 32–36 (2004)
30. Seo, H., Milanfar, P.: Action recognition from one example. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 867–882 (2011)
31. Tao, D., Li, X., Wu, X., Maybank, S.: General tensor discriminant analysis and gabor features for gait recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(10), 1700–1715 (2007)
32. Tenorth, M., Bandouch, J., Beetz, M.: The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In: IEEE International Conference on Computer Vision Workshops, pp. 1089–1096 (2009)
33. Tran, D., Sorokin, A.: Human Activity Recognition with Metric Learning. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 548–561. Springer, Heidelberg (2008)
34. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Circuits and Systems for Video Technology **18**(11), 1473–1488 (2008)
35. Wang, C., Zhang, J., Wang, L., Pu, J., Yuan, X.: Human identification using temporal information preserving gait template. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2164–2176 (2012)
36. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(12), 1505–1518 (2003)
37. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: Advances in Neural Information Processing Systems (2005)
38. Wu, J., Osuntogun, A., Choudhury, T., Philipose, M., Rehg, J.: A scalable approach to activity recognition based on object use. In: IEEE Interntaional Conference on Computer Vision, pp. 1–8 (2007)
39. Yan, H., Lu, J., Deng, W., Zhou, X.: Discriminative multimetric learning for kinship verification. IEEE Transactions on Information Forensics and Security **9**(7), 1169–1178 (2014)
40. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(1), 40–51 (2007)
41. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1794–1801 (2009)

Facial Ethnic Appearance Synthesis

Felix Juefei-Xu(✉) and Marios Savvides

Carnegie Mellon University, Pittsburgh, USA
felixu@cmu.edu, msavvid@ri.cmu.edu

Abstract. In this work, we have explored several subspace reconstruction methods for facial ethnic appearance synthesis (FEAS). In our experiments, our proposed dual subspace modeling using the Fukunaga Koontz transform (FKT) yields much better facial ethnic synthesis results than the ℓ_1 minimization, the ℓ_2 minimization and the principal component analysis (PCA) reconstruction method. With that, we are able to automatically and efficiently synthesize different facial ethnic appearance and alter the facial ethnic appearance of the query image to any other ethnic appearance as desired. Our technique well preserves the facial structure of the query image and simultaneously synthesize the skin tone and ethnic features that best matches target ethnicity group. Facial ethnic appearance synthesis can be applied to synthesizing facial images of a particular ethnicity group for unbalanced database, and can be used to train ethnicity invariant classifiers by generating multiple ethnic appearances of the same subject in the training stage.

Keywords: Soft biometrics · Ethnicity · Face synthesis · Fukunaga Koontz transform

1 Introduction

Within this decade, soft biometrics identification has gained more and more attention as an aid for the traditional face recognition in the biometrics world. Different from traditional hard biometrics such as iris, fingerprints, palmprints, and face [10–19,31] that are difficult to change with the time and living behaviors and have high confidence in identifying subjects, the soft biometrics [8,20,25,28–30,34], on the other hand, focuses more on the physical and behavioral traits that are more prone to change with time and life style, and is less confident in subject identification if used alone. For example, the shape of the eyebrows, the presence of the beard and moustache, skin color, skin texture, color of the pupil, facial marks, gait patterns [9], and so forth, can all be considered as traits of soft biometrics. With the correct identification of these soft biometrics traits, we can infer the age, gender and the ethnicity of the subject. By doing so, we can dramatically narrow down the search space in the scenario of identifying or verifying the subject against a huge gallery database.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part II, LNCS 8926, pp. 825–840, 2015.
DOI: 10.1007/978-3-319-16181-5_62

For ethnicity classification [1,5,7,26,33,35], it is crucial that the researchers obtain a balanced database with subjects from all ethnic groups[1] equally presented for both genders. This is one of the priorities before any learning algorithms are applied for ethnicity classification. Moreover, subjects in the database should be uniquely presented. In this way, the learning machine learns an ethnicity classifier instead of subject-dependent classifier. But unfortunately, database collection with high quality images covering all the ethnic group for both genders is pretty hard to accomplish[2]. That is why in this paper, we will be focusing on solving one of the biggest problems concerning the database creation for ethnicity classification by the synthesis of facial ethnic appearance. In this way, we can automatically and efficiently achieve the balance in the database while also keeping the subject uniqueness in the synthesized database.

The rest of this paper is organized as follows: in Section 2, we will describe our database with which the ethnicity-specific subspaces are built. Section 3 details the facial ethnic appearance synthesis using single subspace modeling methods such as the ℓ_2 minimization, principal components analysis reconstruction and the ℓ_1 minimization. Section 4 details the facial ethnic appearance synthesis using dual subspace modeling with Fukunaga Koontz transform. Experimental setup and results are discussed and analyzed in Section 5. Finally we present some conclusions of our work in Section 6.

2 Database

2.1 Database Collection

We have collected a database with a total of 6849 frontal mugshot-like images from 4 different ethnic groups: east asian, south asian, white and black. The statistics of our database is shown in Table 1.

Table 1. Statistics of our FEAS database

	Female	**Male**	**Total**
East Asian	559	477	1036
South Asian	86	138	224
White	2284	2256	4540
Black	482	567	1049
Total	3411	3438	6849

As can be seen from Table 1, the database is not balanced, the majority of the subjects are white people and there are very few south asians. Because of

[1] As is commonly adopted in the literature, the classification of ethnicity boils down to the 3-class case (asian, black and white), or the 4-class case (east asian, south asian, black and white). In this work, we consider the 4-class case, where we specifically separate south asians from east asians.

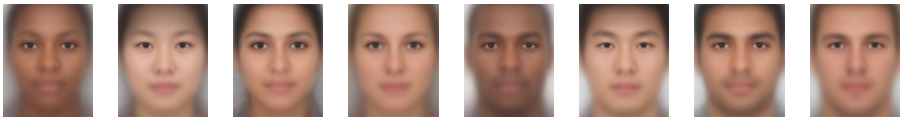
[2] White people dominates most of the ethnicity database publicly available, followed by black people. There are fewer east asian people and south asians are the rarest.

this limitation, the reconstruction performs worse if the target ethnic group is set to be south asian since there are not sufficient images to learn from. Our database also has a bias in age distribution. The majority of the subjects in the database are young adults from 20 to 30 years old. This bias tends to jeopardize the synthesis of the facial ethnic appearance when the query image is an aged subject. Figure 1 shows the mean faces from each of the ethnic group in our database for both genders.

2.2 Preprocessing

We localize and center the eye of each facial image using the modified active shape model (MASM) [32] and crop the rectified full image to be size of 84×68. The original size of the face in the image varies, and the reason we crop the face using this dimension is two-fold. First, this is a reasonable size to compute with using the reconstruction methods to be discussed. Second, some images in the database have low resolution and our cropping dimension of choice suppresses the artifacts and errors caused by up-sampling.

If high resolution synthesis is indeed desired in some applications, we can easily port the algorithms to GPUs using CUDA.



(a) (b) (c) (d) (e) (f) (g)

Fig. 1. Mean faces from our database. (a) Female black, (b) female east asian, (c) female south asian, (d) female white, (e) male black, (f) male east asian, (g) male south asian, and (h) male white.

3 Single Subspace Modeling for Facial Ethnic Appearance Synthesis

In this section, we show the use of single subspace for the synthesis of facial ethnic appearance. We construct a subspace using images from the target ethnic group. This ensures that any reconstruction obtained using components of this subspace has rich ethnic features of this particular ethnic group. We then reconstruct a given query face from the source ethnic group using this target subspace in order to synthesize the ethnic appearance.

Following this procedure, the synthesized facial image is supposed to preserve the subject identity as well as appear closer to the target ethnic group.

We first detail 3 well-established subspace methods using single subspace, namely the ℓ_2 minimization, principal component analysis reconstruction, and the ℓ_1 minimization and then show some results obtained by each of the methods, followed by analysis and discussion.

3.1 ℓ_2 Minimization

Here, we discuss the reconstruction of the given query face in the target subspace using the ℓ_2 minimization. We first build a subspace of faces with target ethnicity and find the linear combination of basis vectors from this subspace that matches the query image the closest. The weight vector $\mathbf{w}$ for the images spanning the subspace is found by minimizing the ℓ_2-norm.

Let $\mathbf{R}$ be a matrix of dimensions $d \times n$ where d is the number of pixels in each face image and n is the number of images spanning the subspace. In other words, each column of $\mathbf{R}$ is a vectorized image of a face from the target ethnic group. Let $\mathbf{x}$ be an incoming query image which is resized to the same size as the face images in the subspace and vectorized. Let $\mathbf{x}^*$ be the reconstructed image using the subspace $\mathbf{R}$. Let $\mathbf{w}^*$ be the $n \times 1$ array of optimal weights for each image in the subspace. The equations used in reconstruction are shown below:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{x} - \mathbf{R}\mathbf{w}\|_2^2 = (\mathbf{R}^\top \mathbf{R})^{-1}\mathbf{R}^\top \mathbf{x} \tag{1}$$

$$\mathbf{x}^* = \mathbf{R}\mathbf{w}^* \tag{2}$$

3.2 Principal Component Analysis

In this part, we outline a PCA based synthesis of a facial image from the target ethnic group. The subspace is built using the same samples as in the ℓ_2 minimization method. PCA is applied to the data and the eigenvectors obtained from the subspace are used for projection and reconstruction. The matrix $\mathbf{R}$, with which the PCA subspace is built is of dimension $d \times n$ where d is the number of pixels in each image and n is the number of images used to construct the subspace. As before, each column of $\mathbf{R}$ is a vectorized image of the images from target set. Let $\mathbf{V}$ be the matrix of eigenvectors generated after performing PCA on $\mathbf{R}$ and $\boldsymbol{\mu}$ be the mean of the images in $\mathbf{R}$. The following equations show the projection and reconstruction of a query facial image $\mathbf{x}$ using this PCA subspace:

$$\mathbf{w} = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{V} \tag{3}$$

$$\mathbf{x}^* = \mathbf{V}\mathbf{w}^\top + \boldsymbol{\mu} \tag{4}$$

3.3 ℓ_1 Minimization

In this subsection, we discuss the application of the ℓ_1 minimization to the problem posed above to obtain a better reconstruction. It is done by using the basis pursuit or the basis pursuit de-noising (BPDN) [2,4,6,22–24,27] and provides a sparse set of weights. This indicates that we use a relatively sparse set of images from the given training set while trying to reconstruct the image with target ethnicity that is closest in an ℓ_2 sense to the query image. In other words, we use the same optimization function as for the ℓ_2 minimization case but add a regularization term which regularizes the ℓ_1-norm of the weights. Let $\mathbf{w}$ be the

optimal set of sparse weights for images in the subspace and let $\mathbf{R}$ be the matrix containing images in the subspace. The modified optimization function is shown below:

$$\text{minimize} \quad \|\mathbf{w}\|_1 \quad \text{subject to} \quad \|\mathbf{x} - \mathbf{R}\mathbf{w}\|_2^2 \leq \epsilon \tag{5}$$

The above optimization can be rewritten using Lagrange multiplier λ as shown below:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \|\mathbf{x} - \mathbf{R}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \tag{6}$$

For this approach to perform well, we need a large number of examples spanning our subspace since sparse reconstruction typically generalizes well when the dictionary is over-complete. In both the previous approaches, the lack of a constraint on $\mathbf{x}$ distributes the energy of $\mathbf{w}$ across a large number of its elements which results in a lot of artifacts left behind primarily from the boundaries of the face where there is a sharp change in intensity. The reconstruction is also more blurred. The ℓ_1 minimization is expected to perform better than the ℓ_2 minimization and the PCA reconstruction since it avoids these artifacts left behind and also produces a smoother reconstruction than the ℓ_2 minimization and PCA reconstruction.

4 Dual Subspace Modeling for Facial Ethnic Appearance Synthesis Using Fukunaga Koontz Transform

Ever since Fukunaga Koontz transform (FKT) came along in 1970 [3], it has been widely used for feature selection especially for general pattern recognition and image processing problems. Unlike traditional principal component analysis (Karhunen-Loève transform), the FKT incorporates data form both positive and negative classes and using eigen decomposition on the joint covariance matrix, in order to find the optimal basis vectors that very well represent one class while have least representation power on the other class. The intrinsic nature of the FKT formulation makes it a very good feature selection tool for two-class problems. More recently, Li *et al.* [21] managed to generalize the FKT to be applied to multi-class problems. We start with the basics of the FKT and We will take a further step in the FKT analysis to explore some very nice properties of the dual subspace modeling which may not be found in other literatures.

Let $\mathbf{X} \in \mathbb{R}^{d \times m}$ be the data set containing the source ethnic facial images, with each column a vectorized image with dimension d. Let $\mathbf{Y} \in \mathbb{R}^{d \times n}$ be the data set containing all the target ethnic facial images. Both $\mathbf{X}$ and $\mathbf{Y}$ are mean removed. The covariance $\mathbf{\Sigma}$ of both the source and target images are the summation of the covariance for each set $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_Y}$. The total covariance matrix $\mathbf{\Sigma}$ is symmetric and can be diagonalized using eigen-decomposition as:

$$\mathbf{\Sigma} = \mathbf{\Sigma_X} + \mathbf{\Sigma_Y} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^\top \tag{7}$$

where $\mathbf{\Phi}$ contains the entire span of eigenvectors of $\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ houses the corresponding eigenvalues of $\mathbf{\Sigma}$ on its diagonal.

Next, a pre-whitening step is applied in the FKT. Both the source and target data are transformed by a pre-whitening matrix $\mathbf{P} = \mathbf{\Phi}\mathbf{\Lambda}^{-\frac{1}{2}}$. So the transformed data $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ becomes:

$$\widehat{\mathbf{X}} = \mathbf{P}^\top \mathbf{X} \quad \text{and} \quad \widehat{\mathbf{Y}} = \mathbf{P}^\top \mathbf{Y} \tag{8}$$

Therefore, the covariance matrices of the transformed source data $\widehat{\mathbf{X}}$ and target data $\widehat{\mathbf{Y}}$ become:

$$\mathbf{\Sigma}_{\widehat{\mathbf{X}}} = \widehat{\mathbf{X}}\widehat{\mathbf{X}}^\top = \mathbf{P}^\top \mathbf{X}\mathbf{X}^\top \mathbf{P} = \mathbf{P}^\top \mathbf{\Sigma}_{\mathbf{X}} \mathbf{P} \tag{9}$$

$$\mathbf{\Sigma}_{\widehat{\mathbf{Y}}} = \widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^\top = \mathbf{P}^\top \mathbf{Y}\mathbf{Y}^\top \mathbf{P} = \mathbf{P}^\top \mathbf{\Sigma}_{\mathbf{Y}} \mathbf{P} \tag{10}$$

The transformed covariance matrix for both source and target data becomes:

$$\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma}_{\widehat{\mathbf{X}}} + \mathbf{\Sigma}_{\widehat{\mathbf{Y}}} = \mathbf{P}^\top \mathbf{\Sigma}_{\mathbf{X}} \mathbf{P} + \mathbf{P}^\top \mathbf{\Sigma}_{\mathbf{Y}} \mathbf{P} \tag{11}$$

$$= \mathbf{P}^\top (\mathbf{\Sigma}_{\mathbf{X}} + \mathbf{\Sigma}_{\mathbf{Y}}) \mathbf{P} = \mathbf{P}^\top \mathbf{\Sigma} \mathbf{P} = \mathbf{I} \tag{12}$$

So, the new covariance matrix is actually an identify matrix. This is because we have performed a global pre-whitening transformation instead of a class-specific pre-whitening transformation to de-correlate the data.

Here, we again perform an eigen-decomposition on the source covariance $\mathbf{\Sigma}_{\widehat{\mathbf{X}}}$, which yields:

$$\mathbf{\Sigma}_{\widehat{\mathbf{X}}} \mathbf{w} = \lambda \mathbf{w} \tag{13}$$

From Equation 12 we can obtain the following by multiplying $\mathbf{w}$ on both sides:

$$\mathbf{\Sigma}_{\widehat{\mathbf{X}}} \mathbf{w} + \mathbf{\Sigma}_{\widehat{\mathbf{Y}}} \mathbf{w} = \mathbf{w} \tag{14}$$

With Equation 12 and 14, we have:

$$\mathbf{\Sigma}_{\widehat{\mathbf{Y}}} \mathbf{w} = \mathbf{w} - \mathbf{\Sigma}_{\widehat{\mathbf{X}}} \mathbf{w} = \mathbf{w} - \lambda \mathbf{w} = (1 - \lambda)\mathbf{w} \tag{15}$$

This effectively means that the covariance matrix from two classes share the same eigenvectors $\mathbf{w}$ and the eigenvalue of one class is exactly the complement of the eigenvalue of the other class. Because of the complementary property of the eigenvalues, the eigenvectors that are the most dominant in one class, is the least dominant in the other class. So in the traditional FKT method as applied to any two-class problem, a discriminative subspace is created by selecting a few of the most dominant eigenvectors for one class and the least dominant ones for the other class. By ignoring the eigenvectors in the middle range, the subspace we obtain contains basis that are very discriminative and will yield discriminative feature selection after projection as shown in Figure 2.

4.1 Dual Subspace Modeling

The aforementioned case is only true for the ideal scenario where the covariance matrices $\mathbf{\Sigma}_{\widehat{\mathbf{X}}}$ and $\mathbf{\Sigma}_{\widehat{\mathbf{Y}}}$ are full rank and have non-zero eigenvalues. But in the

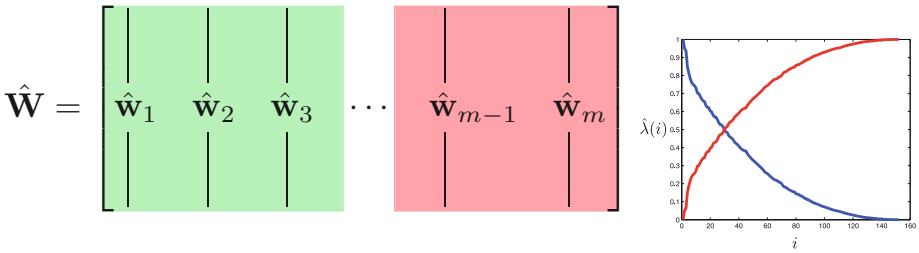


Fig. 2. (left) The most and least dominant vectors for one class are used for better classifying two classes; (right) complementary property of the eigenvalues, where they sum up to 1

real application, this is seldom the case. When the covariance matrix is not full rank, there will be k least dominant eigenvectors in class 1 that all have 0 eigenvalues. According to the FKT, there will be k most dominant eigenvectors with eigenvalues being 1. That means, in the eigenvector $\mathbf{w}$, there are $2k$ eigenvectors that are not properly ranked, and thus the complementary paring of eigenvalues and the sharing of eigenvectors is no longer valid.

In this case, instead of decomposing the covariance of only one class, we propose to decompose both the covariance matrices $\mathbf{\Sigma}_{\widehat{\mathbf{X}}}$ and $\mathbf{\Sigma}_{\widehat{\mathbf{Y}}}$, and instead of modeling both covariance matrices using the same eigenvector $\mathbf{w}$, we propose a dual subspace model using class-specific eigenvectors $\mathbf{w}_x$ and $\mathbf{w}_y$ for decomposition:

$$\mathbf{\Sigma}_{\widehat{\mathbf{X}}}\mathbf{w}_x = \lambda_x \mathbf{w}_x \tag{16}$$

$$\mathbf{\Sigma}_{\widehat{\mathbf{Y}}}\mathbf{w}_y = \lambda_y \mathbf{w}_y \tag{17}$$

The complementary relationship now becomes:

$$\mathbf{\Sigma}_{\widehat{\mathbf{Y}}} = \mathbf{w}_x - \mathbf{\Sigma}_{\widehat{\mathbf{X}}} = \mathbf{w}_x - \lambda_x \mathbf{w}_x = (1 - \lambda_x)\mathbf{w}_x \tag{18}$$

$$\mathbf{\Sigma}_{\widehat{\mathbf{X}}} = \mathbf{w}_y - \mathbf{\Sigma}_{\widehat{\mathbf{Y}}} = \mathbf{w}_y - \lambda_y \mathbf{w}_y = (1 - \lambda_y)\mathbf{w}_y \tag{19}$$

Here, the most dominant eigenvector in $\mathbf{w}_x$ and the least dominant eigenvectors in $\mathbf{w}_y$ are not necessarily the same. Instead of keeping the first and the last tier of eigenvectors from $\mathbf{w}$ for subspace modeling, we now take the most dominant eigenvectors from $\mathbf{w}_x$ as well as from $\mathbf{w}_y$ to create the subspace. In this way, we essentially remove the eigenvectors corresponding to 0 eigenvalues, while still keeping high discriminative power.

5 Experiments

In this section, we first describe the experimental setup for the aforementioned facial ethnic appearance synthesis techniques: (1) the ℓ_2 minimization, (2) PCA reconstruction, (3) the ℓ_1 minimization, and (4) the dual subspace modeling using FKT. Second, we show and analyze the experimental results using all four synthesis techniques.

5.1 Experimental Setup

For the single subspace reconstruction methods (the ℓ_2 minimization, PCA reconstruction, and the ℓ_1 minimization), only the subspace obtained from the target ethnic group is needed in the reconstruction process. The query image from the source ethnic group is reconstructed and synthesized to best match the target ethnic group.

In the dual subspace reconstruction method using FKT, two subspaces from both the source and target ethnic group are acquired. By linearly combining the two subspaces using α blending, the query image from the source ethnic group is gradually transformed to the target ethnic group.

All the color images are in the **RGB** format, so the facial ethnic appearance synthesis is done by reconstructing each color channel individually and finally combined together to display the color synthesis images.

One important characteristic of our proposed facial ethnic appearance synthesis is that the subject identity is very well preserved during the synthesis, and at the same time, the subject's facial ethnic features such as skin tone and eye contours are altered to best match the target ethnic group. In this way, we can transform people from one ethnic group to another, by keeping their own uniqueness. This is very good in the application of synthesizing new subjects from particular ethnic groups that are unique.

5.2 Experimental Results

We have trained our optimal projection coefficient $\mathbf{w}^*$ for each of the single subspace methods using the images only from the target ethnic group, and synthesize the facial appearance of the query image that best matches the target group.

Figure 3 and 4 show the synthesis using the ℓ_2 minimization. In this experiment, we pick query images of celebrities from the Internet and synthesize them to another ethnic group in terms of: (a) black to east asian, (b) black to south asian, (c) white to east asian, and (d) white to south asian, for both genders.

As can be seen, the reconstruction is of bad quality with many artifacts. This is partially due to the fact that the number of images in the target set is limited and when an unseen query image looks quite different from the images in the target set, the reconstruction would be jeopardized.

Figure 5 and 6 show the synthesis using PCA reconstruction. The same query images are selected as in the ℓ_2 minimization case to show the comparisons. The reconstruction, still not satisfactory. The results using PCA look similar to the ones using the ℓ_2 minimization because the way PCA finds the optimal projecting directions (principal components) is actually minimizing the variance: $\mathrm{Var}(\mathbf{w}^\top \mathbf{x})$, and both PCA and ℓ_2 minimization is dealing with second-order statistics of the data. Thus their optimal results are similar.

Figure 7 and 8 show the synthesis using the ℓ_1 minimization. The same query images and target ethnic groups are used. We can see that the reconstruction still is not as good as we have expected. Many of the synthesized faces are bluish.

(a) (b) (c) (d)

Fig. 3. Facial ethnic appearance synthesis using ℓ_2 minimization on female subjects. (a) Black to east asian, (b) black to south asian, (c) white to east asian, and (d) white to south asian. In each subfigure, the input is on the **left** and the reconstructed image is on the **right**.

(a) (b) (c) (d)

Fig. 4. Facial ethnic appearance synthesis using ℓ_2 minimization on male subjects. (a) Black to east asian, (b) black to south asian, (c) white to east asian, and (d) white to south asian. In each subfigure, the input is on the **left** and the reconstructed image is on the **right**.

This is because the ℓ_1 minimization gives a sparse solution, and since we are dealing each color channel individually, the **R** and **G** channel are overwhelmed by the **B** channel. Moreover, some target ethnic set does not have enough images to create an over-complete dictionary so the reconstruction quality using the ℓ_1 minimization is still questionable.

By applying our proposed dual subspace modeling using FKT, the facial ethnic appearance synthesis results are much better than the previously discussed single subspace methods. Figure 9 and 10 show the synthesis. The query images are on the far left and the following five images are reconstructed using the dual subspaces (one subspace is built using target ethnic group, and the other sub-

(a) (b) (c) (d)

Fig. 5. Facial ethnic appearance synthesis using PCA reconstruction on female subjects. (a) Black to east asian, (b) black to south asian, (c) white to east asian, and (d) white to south asian. In each subfigure, the input is on the **left** and the reconstructed image is on the **right**.

(a) (b) (c) (d)

Fig. 6. Facial ethnic appearance synthesis using PCA reconstruction on male subjects. (a) Black to east asian, (b) black to south asian, (c) white to east asian, and (d) white to south asian. In each subfigure, the input is on the **left** and the reconstructed image is on the **right**.

(a) (b) (c) (d)

Fig. 7. Facial ethnic appearance synthesis using ℓ_1 minimization on female subjects. (a) Black to east asian, (b) black to south asian, (c) white to east asian, and (d) white to south asian. In each subfigure, the input is on the **left** and the reconstructed image is on the **right**.

(a) (b) (c) (d)

Fig. 8. Facial ethnic appearance synthesis using ℓ_1 minimization on male subjects. (a) Black to east asian, (b) black to south asian, (c) white to east asian, and (d) white to south asian. In each subfigure, the input is on the **left** and the reconstructed image is on the **right**.

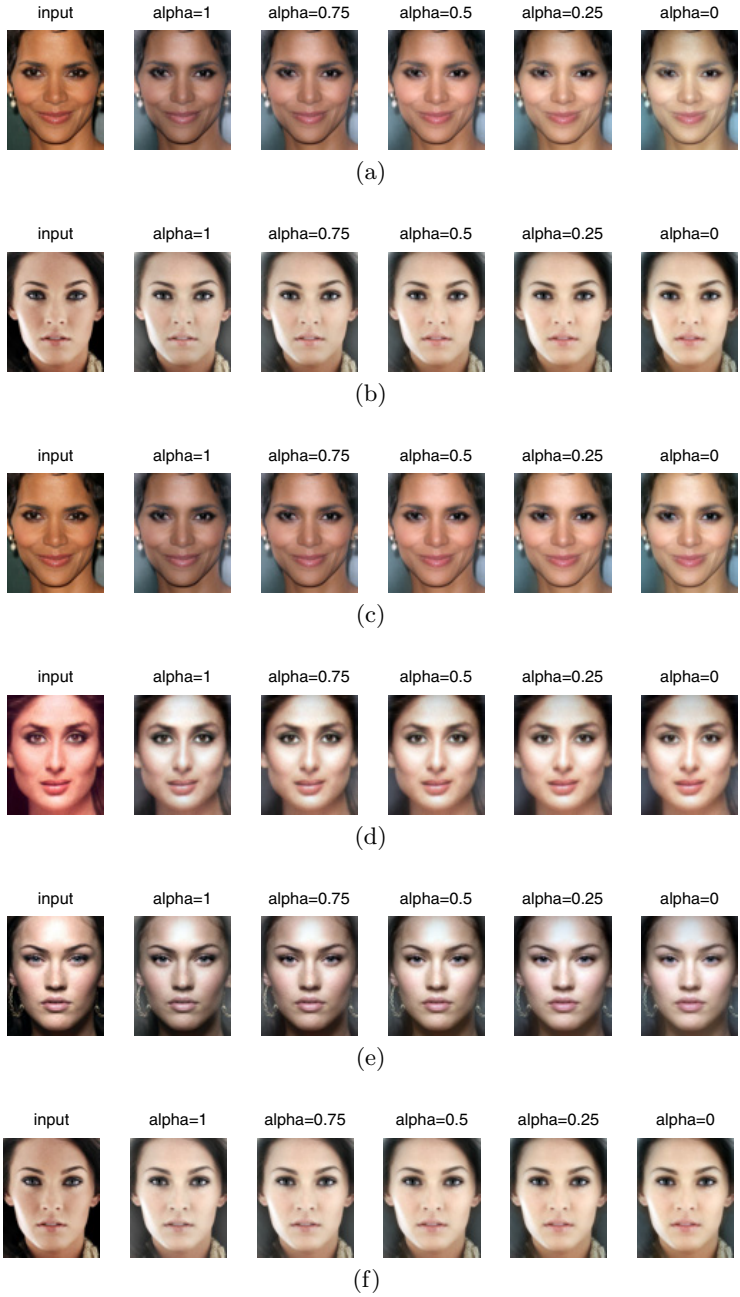


Fig. 9. Facial ethnic appearance synthesis using FKT with dual subspace modeling on female subjects. (a) Black to east asian, (b) black to south asian, (c) black to white, (d) south asian to east asian, (e) white to east asian, and (f) white to south asian. In each subfigure, the input is on the far left and the following five images are the synthesis with α blending of the dual subspaces.

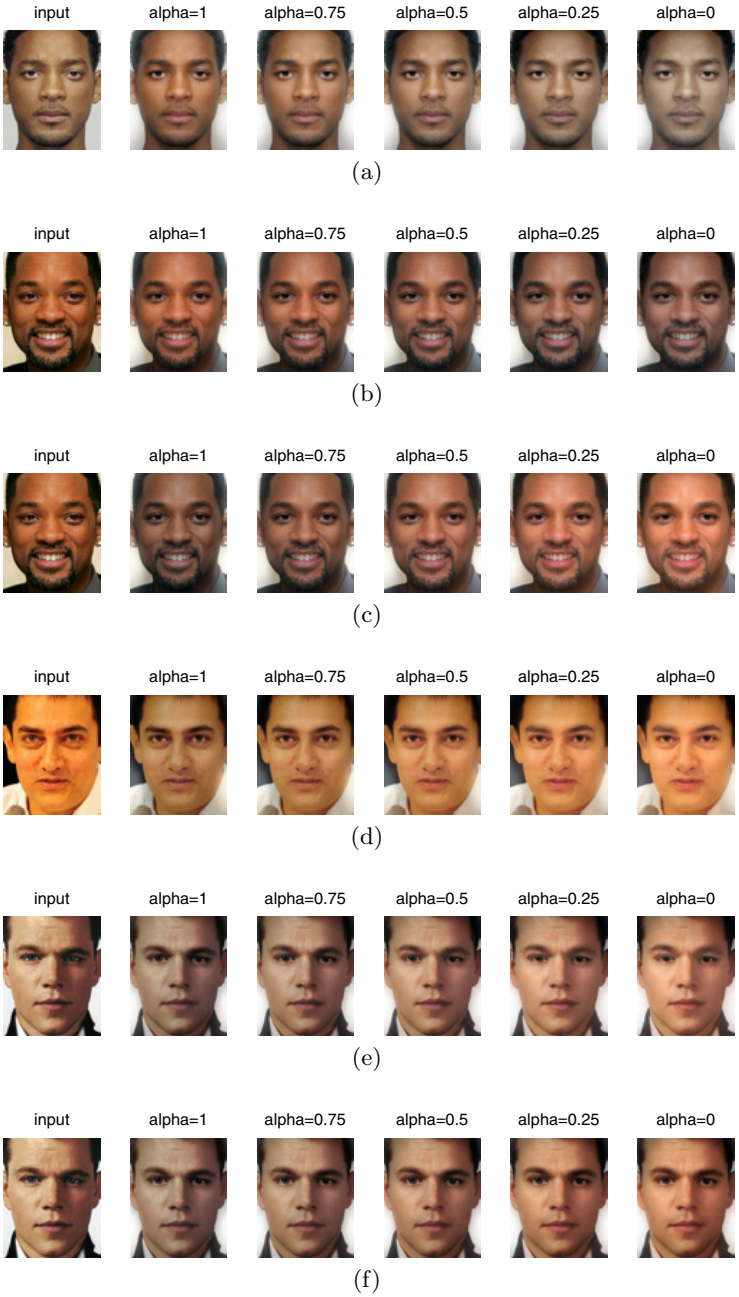


Fig. 10. Facial ethnic appearance synthesis using FKT with dual subspace modeling on male subjects. (a) Black to east asian, (b) black to south asian, (c) black to white, (d) south asian to east asian, (e) white to east asian, and (f) white to south asian. In each subfigure, the input is on the far left and the following five images are the synthesis with α blending of the dual subspaces.

space is built using the source ethnic group, the same as the query image to be ethnically altered). The α blending coefficient is shown in the 2 figures. When $\alpha = 1$, pure source group subspace is utilized and when $\alpha = 0$, pure target group subspace is used. By changing α from 1 to 0, a gradual transformation from the source ethnic group to the target ethnic group can be shown. The synthesis quality of FKT dual subspace modeling is much better than the single subspace method with no identifiable artifacts at all.

The query image is transformed to the target ethnic group while keeping his or her identity to the largest extent, meaning the distinctive identity features are well preserved. As are shown in Figure 9 and 10, the photometric features of the query images are also well kelp. For example, the highlights on the forehead and the shadow on the cheek and so forth are still well preserved in the ethnic appearance synthesized image.

From the mean faces of our database as shown earlier in Figure 1, the eye region is the best registered region on the faces. Compared with eye region, the mouth region is not as well aligned. This is due to the fact that images in our database may have different expression and the mouth region are not perfectly aligned and registered. Therefore, we should expect a better synthesis quality around the eye region and less around the mouth region. Even with that, we are still able to achieve a much better ethnic appearance synthesis than the single subspace techniques such as the ℓ_2 minimization, PCA and the ℓ_1 minimization.

In our experiments with the FKT dual subspace modeling, we apply our facial ethnic appearance synthesis to unseen facial images from the web. These celebrity images are actually very different from the images in our source and target database. The skin is usually highly polished due to makeups and photo re-touching. So, the synthesis results is not as genuine as the query images that are actually from the source data set. We cannot disclose the query images from our database in this paper, but the supplementary material available to the reviewers actually show more genuine results.

6 Conclusions

In this work, we have explored several subspace reconstruction methods for facial ethnic appearance synthesis (FEAS). In our experiments, our proposed dual subspace modeling using the Fukunaga Koontz transform (FKT) yields much better facial ethnic synthesis results than the ℓ_1 minimization, the ℓ_2 minimization and the principal component analysis (PCA) reconstruction method. With that, we are able to automatically and efficiently synthesize different facial ethnic appearance and alter the facial ethnic appearance of the query image to any other ethnic appearance as desired. Our technique well preserves the facial structure of the query image and simultaneously synthesize the skin tone and ethnic features that best matches target ethnicity group. Facial ethnic appearance synthesis can be applied to synthesizing facial images of a particular ethnicity group for unbalanced database, and can be used to train ethnicity invariant classifiers by generating multiple ethnic appearances of the same subject in the training stage.

References

1. Dhamecha, T.I., Sankaran, A., Singh, R., Vatsa, M.: Is gender classification across ethnicity feasible using discriminant functions? In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7 (September 2011)
2. Ekanadham, C., Tranchina, D., Simoncelli, E.P.: Recovery of sparse translation-invariant signals with continuous basis pursuit. IEEE Transactions on Signal Processing **59**(10), 4735–4744 (2011)
3. Fukunaga, K., Koontz, W.L.G.: Application of the karhunen-loève expansion to feature selection and ordering. IEEE Transactions on Computers **C-19**(4), 311–318 (1970)
4. Gill, P.R., Wang, A., Molnar, A.: The in-crowd algorithm for fast basis pursuit denoising. IEEE Transactions on Signal Processing **59**(10), 4595–4605 (2011)
5. Guo, G., Mu, G.: A study of large-scale ethnicity estimation with gender and age variations. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), pp. 79–86 (June 2010)
6. Hoang, T.V., Smith, E.H.B., Tabbone, S.: Edge noise removal in bilevel graphical document images using sparse representation. In: 2011 18th IEEE International Conference on Image Processing (ICIP), pp. 3549–3552 (September 2011)
7. Hosoi, S., Takikawa, E., Kawade, M.: Ethnicity estimation with facial images. In: Sixth IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 195–200 (May 2004)
8. Jain, A.K., Park, U.: Facial marks: Soft biometric for face recognition. In: 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 37–40 (November 2009)
9. Juefei-Xu, F., Bhagavatula, C., Jaech, A., Prasad, U., Savvides, M.: Gait-id on the move: pace independent human identification using cell phone accelerometer dynamics. In: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 8–15. IEEE (2012)
10. Juefei-Xu, F., Cha, M., Heyman, J.L., Venugopalan, S., Abiantun, R., Savvides, M.: Robust local binary pattern feature sets for periocular biometric identification. In: 2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), pp. 1–8. IEEE (2010)
11. Juefei-Xu, F., Cha, M., Savvides, M., Bedros, S., Trojanova, J.: Robust periocular biometric recognition using multi-level fusion of various local feature extraction techniques. In: IEEE 17th International Conference on Digital Signal Processing (DSP) (2011)
12. Juefei-Xu, F., Luu, K., Savvides, M., Bui, T.D., Suen, C.Y.: Investigating age invariant face recognition based on periocular biometrics. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7. IEEE (2011)
13. Juefei-Xu, F., Pal, D.K., Savvides, M.: Hallucinating the full face from the periocular region via dimensionally weighted K-SVD. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW). IEEE (2014)
14. Juefei-Xu, F., Savvides, M.: Can your eyebrows tell me who you are? In: 2011 5th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1–8. IEEE (2011)
15. Juefei-Xu, F., Savvides, M.: Unconstrained periocular biometric acquisition and recognition using COTS PTZ camera for uncooperative and non-cooperative subjects. In: 2012 IEEE Workshop on Applications of Computer Vision (WACV), pp. 201–208. IEEE (2012)

16. Juefei-Xu, F., Savvides, M.: An augmented linear discriminant analysis approach for identifying identical twins with the aid of facial asymmetry features. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) (2013)
17. Juefei-Xu, F., Savvides, M.: An image statistics approach towards efficient and robust refinement for landmarks on facial boundary. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2013)
18. Juefei-Xu, F., Savvides, M.: Subspace based discrete transform encoded local binary patterns representations for robust periocular matching on NIST's face recognition grand challenge. IEEE Transactions on Image Processing **23**(8), 3490–3505 (2014)
19. Juefei-Xu, F., Savvides, M.: Weight-optimal local binary patterns. In: European Conference on Computer Vision (ECCV) Workshops. Springer (2014)
20. Kashyap, A.L., Tulyakov, S., Govindaraju, V.: Facial behavior as a soft biometric. In: 2012 5th IAPR International Conference on Biometrics (ICB), pp. 147–151 (April 2012)
21. Li, Y., Savvides, M.: Kernel fukunaga-koontz transform subspaces for enhanced face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8 (June 2007)
22. Liu, C., Zakharov, Y.V., Chen, T.: Broadband underwater localization of multiple sources using basis pursuit de-noising. IEEE Transactions on Signal Processing **60**(4), 1708–1717 (2012)
23. Lu, W., Vaswani, N.: Modified basis pursuit denoising (modified-BPDN) for noisy compressive sensing with partially known support. In: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 3926–3929 (March 2010)
24. Lu, W., Vaswani, N.: Exact reconstruction conditions for regularized modified basis pursuit. IEEE Transactions on Signal Processing **60**(5), 2634–2640 (2012)
25. Lyle, J.R., Miller, P.E., Pundlik, S.J., Woodard, D.L.: Soft biometric classification using periocular region features. In: 2010 Fourth IEEE International Conference on Biometrics: Theory Application and Systems (BTAS), pp. 1–7 (September 2010)
26. Manesh, F.S., Ghahramani, M., Tan, Y.P.: Facial part displacement effect on template-based gender and ethnicity classification. In: 2010 11th International Conference on Control Automation Robotics and Vision (ICARCV), pp. 1644–1649 (December 2010)
27. Mota, J.F.C., Xavier, J.M.F., Aguiar, P.M.Q., Puschel, M.: Distributed basis pursuit. IEEE Transactions on Signal Processing **60**(4), 1942–1956 (2012)
28. Niinuma, K., Park, U., Jain, A.K.: Soft biometric traits for continuous user authentication. IEEE Transactions on Information Forensics and Security **5**(4), 771–780 (2010)
29. Park, U., Jain, A.K.: Face matching and retrieval using soft biometrics. IEEE Transactions on Information Forensics and Security **5**(3), 406–415 (2010)
30. Reid, D.A., Nixon, M.S.: Using comparative human descriptions for soft biometrics. In: 2011 International Joint Conference on Biometrics (IJCB), pp. 1–6 (October 2011)
31. Savvides, M., Juefei-Xu, F.: Image matching using subspace-based discrete transform encoded local binary patterns (September 2013). http://www.google.com/patents/US20140212044

32. Seshadri, K., Savvides, M.: An analysis of the sensitivity of active shape models to initialization when applied to automatic facial landmarking. IEEE Transactions on Information Forensics and Security **7**(4), 1255–1269 (2012)
33. Wu, B., Ai, H., Huang, C.: Facial image retrieval based on demographic classification. In: 17th International Conference on Pattern Recognition (ICPR), vol. 3, pp. 914–917 (August 2004)
34. Yan, S., Wang, H., Tang, X., Liu, J., Huang, T.S.: Regression from uncertain labels and its applications to soft biometrics. IEEE Trans. on Information Forensics and Security **3**(4), 698–708 (2008)
35. Zhang, G., Wang, Y.: Multimodal 2D and 3D facial ethnicity classification. In: 2009 Fifth International Conference on Image and Graphics (ICIG), pp. 928–932 (September 2009)

Author Index