LNCS 6855

Pedro Real Daniel Diaz-Pernil
Helena Molina-Abril Ainhoa Berciano
Walter Kropatsch (Eds.)

Computer Analysis of Images and Patterns

14th International Conference, CAIP 2011
Seville, Spain, August 2011
Proceedings, Part II

2 Part II

Springer

Pedro Real Daniel Diaz-Pernil
Helena Molina-Abril Ainhoa Berciano
Walter Kropatsch (Eds.)

Computer Analysis of Images and Patterns

14th International Conference, CAIP 2011
Seville, Spain, August 29-31, 2011
Proceedings, Part II

Springer

Volume Editors

Ainhoa Berciano
Universidad del País Vasco
Euskal Herriko Unibertsitatea
Ramón y Cajal, 72, 48014 Bilbao, Spain
E-mail: ainhoa.berciano@ehu.es

Daniel Diaz-Pernil
Helena Molina-Abril
Pedro Real
University of Seville
Avenida Reina Mercedes s/n
41012 Seville, Spain
E-mail: {sbdani, habril, real}@us.es

Walter Kropatsch
Vienna University of Technology
Favoritenstraße 9/186-3
1040 Vienna, Austria
E-mail: krw@prip.tuwien.ac.at

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-23677-8 e-ISBN 978-3-642-23678-5
DOI 10.1007/978-3-642-23678-5
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011935054

CR Subject Classification (1998): I.5, I.4, I.3, I.2.10, I.2.6, H.2.8, I.2.7

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

This volume contains the papers presented at the 14th International Conference on Computer Analysis of Images and Patterns (CAIP 2011) held in Seville during August 29–31, 2011.

The first CAIP conference was in 1985 in Berlin. Since then CAIP has been organized biennially in different cities around Europe: Wismar, Leipzig, Dresden, Budapest, Prague, Kiel, Ljubljana, Warsaw, Groningen, Versailles, Vienna and Münster.

Following the spirit of the previous meetings, the 14th CAIP was conceived as a period of active interaction among the participants, with emphasis on exchanging ideas and on cooperation.

This year, 286 full scientific papers from 52 countries were submitted, of which 138 were accepted for presentation based on the positive scientific reviews. All the papers have been revised by, at least, two reviewers and, most of them by three.

The accepted papers were presented during the conference either as oral presentations or as posters in the single-track scientific program. Oral presentations allowed the authors to reach a large number of participants, while posters allowed for a more intense scientific interaction. We tried to continue the tradition of CAIP in providing a forum for scientific exchange at a high-quality level.

Two internationally recognized speakers accepted our invitation to present a stimulating research topic this year: Peter Sturm, INRIA Grenoble (France) and Facundo Memoli, Stanford University (USA).

Indeed, these proceedings are divided into two volumes, 6854 and 6855, where the index has been structured following the topics and program of the conference.

We are grateful for the great work realized by the Program Committee and additional reviewers. We especially thank the PRIP and CATAM members, who made a big effort to help.

We appreciate our sponsors for their direct and indirect financial support and Springer for giving us the opportunity to continue publishing CAIP proceedings in the LNCS series.

Finally, many thanks go to our local support team and, mainly, to María José Jiménez Rodríguez for her huge and careful work of supervision of almost all the tasks of the Organizing Committee.

August 2011

Ainhoa Berciano
Daniel Diaz-Pernil
Walter Kropatsch
Helena Molina-Abril
Pedro Real

CAIP 2011 Organization

Conference Chairs

Pedro Real	University of Seville, Spain
Walter Kropatsch	Vienna University of Technology, Austria

Steering Committee

André Gagalowicz (France)
Xiaoyi Jiang (Germany)
Reinhard Klette (New Zealand)
Walter Kropatsch (Austria)
Nicolai Petkov (The Netherlands)
Gerald Sommer (Germany)

Program Committee

Shigeo Abe
Ceyhun Burak Akgul
Mayer Aladjem
Sylvie Alayrangues
Madjid Allili
A. Antonacopoulos
Heider Araujo
Jonas August
Antonio Bandera
Elisa H. Barney Smith
Brian A. Barsky
Algirdas Bastys
E. Bayro Corrochano
Ardhendu Behera
Abdel Belaid
Olga Bellon
Ainhoa Berciano
Wolfgang Birkfellner
Dorothea Blostein
Gunilla Borgefors
Christian Breiteneder
Thomas Breuel
Luc Brun
Lorenzo Bruzzone
Martin Burger
Gustavo Carneiro
Kwok Ping Chan
Yung-Kuan Chan
Rama Chellappa
Sei-Wang Chen
Da-Chuan Cheng
Dmitry Chetverik
Jose Cortes Parejo
Bertrand Couasnon
Marco Cristani
Guillaume Damiand
Justin Dauwels
Mohammad Dawood
Gerard de Haan
Alberto Del Bimbo
Andreas Dengel
Joachim Denzler
Cecilia Di Ruberto
Daniel Diaz-Pernil
Philippe Dosch
Hazim Kemal Ekenel
Neamat El Gayar
Hakan Erdogan
Francisco Escolano
M. Taner Eskil
Chiung-Yao Fang
Miguel Ferrer
Massimo Ferri
Gernot Fink
Robert Fisher
Ana Fred
Patrizio Frosini
Laurent Fuchs
Xinbo Gao
Anarta Ghosh
Georgy Gimel'farb
Dmitry Goldgof
Rocio Gonzalez-Diaz
Cosmin Grigorescu
M.A. Gutierrez-Naranjo
Michal Haindl
Edwin Hancock
Changzheng He
Vaclav Hlavac
Zha Hongbin
Joachim Hornegger
Yo-Ping Huang
Yung-Fa Huang
Atsushi Imiya
Shuiwang Ji
Xiaoyi Jiang
Maria Jose Jimenez
Martin Kampel
Nahum Kiryati
Reinhard Klette
Andreas Koschan

Walter Kropatsch
James Kwok
Longin Jan Latecki
Xuelong Li
Pascal Lienhardt
Guo-Shiang Lin
Josep Llados
Jean-Luc Mari
Eckart Michaelse
Ioana Necula
Radu Nicolescu
Mads Nielsen
Darian Onchis-Moaca
Samuel Peltier
Petra Perner
Nicolai Petkov
Ioannis Pitas
Mario J. Perez Jimnez
Petia Radeva
Pedro Real
Jos Roerdink
Bodo Rosenhahn
Jose Ruiz-Shulcloper
Robert Sablatnig
Robert Sabourin
Hideo Saito
Albert Salah
Gabriella Sanniti Di Baja
Sudeep Sarkar
Oliver Schreer
Francesc Serratosa
Luciano Silva
Gerald Sommer
Mingli Song
K.G. Subramanian
Akihiro Sugimoto
Dacheng Tao
Klaus Toennies
Karl Tombre
Javier Toro
Andrea Torsello
Chwei-Shyong Tsai
Ernest Valveny
Mario Vento
Jose Antonio Vilches
Steffen Wachenfeld
Shengrui Wang
Michel Westenberg
Paul Whelan

Additional Reviewers

Nicole Artner
Facundo Bromberg
Christoph Brune
Javier Carnero
Andrea Cerri
Chao Chen
Wen-Chang Cheng
Michel Devy
Denis Enachescu
Yll Haxhimusa
Chih-Yu Hsu
Adrian Ion
Jiun-Jian Liaw
Helena Molina-Abril
Gennaro Percannella
Federico Schluter
Cheng-Ying Yang
Chih-Chia Yao

Local Organizing Committee

Ainhoa Berciano
Javier Carnero
Daniel Diaz-Pernil
Maria Jose Jimenez
Ioana Necula
Belen Medrano
Helena Molina-Abril
Ana Pacheco
Regina Poyatos
Angel Tenorio
Lidia de la Torre

Sponsoring Institutions

Vicerrectorado de Investigación, Universidad de Sevilla
Instituto de Matemáticas de la Universidad de Sevilla, A. de Castro Brzezicki
Fundación para la Investigación y el Desarrollo de las Tecnologías de la Información en Andalucía
Ministerio de Ciencia e Innovación (Spain)
Consejería de Economía, Ciencia e Innovación de la Junta de Andalucía
International Association for Pattern Recognition (IAPR)
Escuela Técnica superier de Ingeniería Informática, Universidad de Seville, Spain
Department of Applied Mathematics I, University of Seville, Spain

Table of Contents – Part II

Invited Lecture

Biometrics

Human and Face Detection and Recognition

Document Analysis

Applications

3D Vision

Image Restoration

Restoration

Natural Computation for Digital Imagery

Image and Video Processing

Calibration

Color and Texture

Tracking and Stereo Vision

Table of Contents – Part I

Invited Lecture

Motion Analysis

Image and Shape Models

Segmentation and Grouping

Shape Recovery

Graph-Based Methods and Representations

Curves, Surfaces and Objects beyond 2 Dimensions

Geo-topological Analysis of Images

Kernel Methods

Image and Video Indexing and Database Retrieval

Object Detection and Recognition

Medical Imaging

Structural Pattern Recognition

Metric Structures on Datasets: Stability and Classification of Algorithms

Facundo Mémoli

Department of Mathematics, Stanford University, California, USA, and
Department of Computer Science, The University of Adelaide, Australia
memoli@math.stanford.edu

Abstract. Several methods in data and shape analysis can be regarded as transformations between metric spaces. Examples are hierarchical clustering methods, the higher order constructions of computational persistent topology, and several computational techniques that operate within the context of data/shape matching under invariances.

Metric geometry, and in particular different variants of the Gromov-Hausdorff distance provide a point of view which is applicable in different scenarios. The underlying idea is to regard datasets as metric spaces, or metric measure spaces (a.k.a. mm-spaces, which are metric spaces enriched with probability measures), and then, crucially, at the same time regard the collection of all datasets as a metric space in itself. Variations of this point of view give rise to different taxonomies that include several methods for extracting information from datasets.

Imposing metric structures on the collection of all datasets could be regarded as a "soft" construction. The classification of algorithms, or the axiomatic characterization of them, could be achieved by imposing the more "rigid" category structures on the collection of all finite metric spaces and demanding functoriality of the algorithms. In this case, one would hope to single out all the algorithms that satisfy certain natural conditions, which would clarify the landscape of available methods. We describe how using this formalism leads to an axiomatic description of many clustering algorithms, both flat and hierarchical.

Keywords: metric geometry, categories and functors, metric spaces, Gromov-Hausdorff distance, Gromov-Wasserstein distance.

1 Introduction

Nowadays in the scientific community we are being asked to analyze and probe large volumes of data with the hope that we may learn something about the underlying phenomena producing these data. Questions such as "what is the shape of data" are routinely formulated and partial answers to these usually reveal interesting science.

An important goal of exploratory data analysis is to enable researchers to obtain insights about the organization of datasets. Several algorithms have been developed with the goal of discovering structure in data, and examples of the different tasks these algorithms tackle are:

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 1–33, 2011.

- Visualization, parametrization of high dimensional data
- Registration/matching of datasets: how different are two given datasets? what is a good correspondence between sub-parts of the datasets?
- What are the features present in the data? e.g. clustering, and number of holes in the data.
- How to agglomerate/merge (partial) datasets?

Some of the standard concerns about the results produced by algorithms that attempt to solve these tasks are: the dependence on a particular choice of coordinates, the invariance to certain uninteresting deformations, the stability/sensitivity to small perturbations, etc.

1.1 Visualization of Datasets

The *projection pursuit* method (see [42]) determines the linear projection on two or three dimensional space which optimizes a certain criterion. It is frequently very successful, and when it succeeds it produces a set in $\mathbb{R}^2$ or $\mathbb{R}^3$ which readily visualizable. Other methods (Isomap [85], locally linear embedding [74], multidimensional scaling [23]) attempt to find non-linear maps to Euclidean space which preserve the distance functions on the data set to as high a degree as possible. They also produce useful two and three dimensional versions of data sets when they succeed.

Other interesting methods are the *grand tour* of Asimov [2], the *parallel coordinates* of Inselberg [44], and the *principal curves* of Hastie and Stuetzle [38].

The Mapper algorithm [80] produces representations of data in a manner akin to the Reeb graph [71] and is based on the idea of *partial clustering* and can be considered as a hybrid method which combines the ability to parametrize and visualize data, with the the ability to extract features, see Figure 1. This algorithm has been used for shape matching tasks as well for studies of breast cancer [65] and RNA [6]. The mapper algorithm is also closely related to the *cluster tree* of Stuetzle [82].

1.2 Matching and Dissimilarity between Datasets

Measuring the *dissimilarity* between two objects is a task that is often performed in data and shape analysis, and summaries or features from each of the objects are typically compared to quantify this dissimilarity.

One important instance when computing the dissimilarity between is useful is the comparison of the three dimensional shape of proteins following the underlying scientific assumption that physically similar proteins have similar functional properties [52].

The notion of zero-dissimilarity between data-sets can be dependent on the application domain. For example, in object recognition, rigid motions specifically, and more generally isometries, are often uninteresting and not important. The

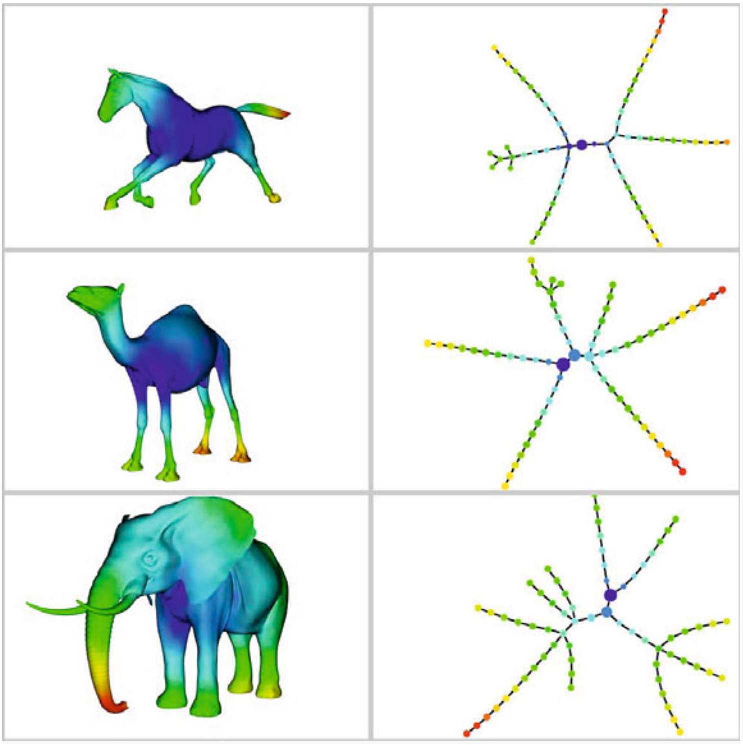


Fig. 1. A simplification of 3d models using the mapper algorithm [80]

same point applies to multidimensional data analysis, where particular choices of the coordinate system should not affect the result of algorithms. Therefore, the summaries/features extracted from the data must be insensitive to these unimportant changes.

There exists a plethora of practical methods for object comparison and matching, and most of them are based on comparing features. Given this rich and disparate collection of available methods, it seems that in order to obtain a deep understanding of the object matching problem and find possible avenues of improvement, it is of great importance to discover and establish relationships/connections between these methods. Theoretical understanding of these methods and their relationships will lead to expressing conditions of validity of each approach or family of approaches. This can no doubt help in

(a) guiding the choice of which method to use in a given practical application,
(b) deciding what parameters (if any) should be used for the particular method chosen, and
(c) clearly determining what are the *theoretical guarantees* of a particular method for the task at hand.

1.3 Features

Often, data-sets can be difficult to comprehend. One example of this is the case of high dimensional point clouds because our ability to visualize them is rather limited. To deal with this situation, one must attempt to extract *summaries* from the complicated data-set in order to capture robust global properties that signal important qualitative features present, but not apparent, in the data.

The term *feature* typically applies to the result of applying a certain simplification to a given dataset with the hope of retaining some useful information about the original data. The aim is that after this simplification it would become easier to quantify and/or visualize certain aspects of the dataset. Think for example of:

- computing the number of clusters in a given dataset, according to a given algorithm (e.g. linkage based methods, spectral clustering, k-means, etc);
- obtaining a dendrogram: the result of applying a hierarchical clustering algorithm to the data;
- computing the average distance to the barycenter of the dataset (assumed to be embedded in Euclidean space);
- computing the average distance between all pairs of points in the dataset;
- computing a histogram of all the interpoint distances between pairs of points in the dataset;
- computing persistent topology invariants of some filtration obtained from the dataset [33,17,81].

In the area of shape analysis a few examples are: the *size theory* of Frosini and collaborators [30,29,88,25,24,31]; the Reeb graph approach of Hilaga et al [39]; the *spin images* of Johnsson [49], the *shape distributions* of [68]; the *canonical forms* of [28]; the Hamza-Krim approach [36]; the spectral approaches of [72,76]; the *integral invariants* of [58,69,21]; the *shape contexts* of [3].

The theoretical question of proving that a given family of features is indeed able to signal proximity or similarity of objects in a reasonable way has hardly been addressed. In particular, the degree to which two objects with similar features are forced to be similar is in general does not seem to be well understood.

Conversely, one should ask the more basic question of whether the similarity between two objects forces their features to be similar.

Stability of features. Thus, a problem of interest is studying the extent to which a given feature is stable under perturbations of the dataset. In order to be able to say something precise in this respect we introduce some mathematical language.

To fix concepts we imagine that we have a collection $\mathcal{D}$ of all possible datasets, and a collection $\mathcal{F}$ of all possible features. A *feature map* will be any map $f : \mathcal{D} \to \mathcal{F}$. Assume further that $d_{\mathcal{D}}$ and $d_{\mathcal{F}}$ are metrics or distance functions on $\mathcal{F}$ and $\mathcal{D}$, respectively. One says that f is *quantitatively stable* whenever one can find a non-decreasing function $\Psi : [0, \infty) \to [0, \infty)$ with $\Psi(0) = 0$ such that for all $X, Y \in \mathcal{D}$ it holds that

$$d_{\mathcal{F}}(f(X), f(Y)) \leq \Psi\big(d_{\mathcal{D}}(X, Y)\big).$$

Note that this is stronger that the usual notion of *continuity* of maps, namely that $f(X_n) \to f(X)$ as $n \uparrow \infty$ whenever $(X_n)_n \subset \mathcal{D}$ is a sequence of datasets converging to X.

In subsequent sections of the paper we will describe instances of suitable metric spaces $(\mathcal{D}, d_{\mathcal{D}})$ and study the stability of different features.

2 Some Considerations

2.1 Importance of Stability and Classification of Algorithms

We claim that it would be desirable to elucidate the stability properties of the main methods used in data analysis. The underlying situation is that the output of data analysis algorithms are used in order to draw conclusions about the phenomenon producing the data, hence it is of extreme importance to make sure that these conclusions would not be grossly affected if the dataset were "noisy" or "slightly perturbed". In order to make sense of this question one needs to ascribe mathematical meaning to "data", "perturbations", "algorithms", etc.

In a similar vein, it would be clearly highly desirable to know what are the theoretical properties enjoyed by the main algorithms used in data analysis (such as clustering methods, for example). From a theoretical standpoint, it would be very nice to be able to derive algorithms from a list of desirable or required properties or axioms. In this respect, the works of Janowitz [47], Kleinberg [51], and von Luxburg [90] are very prominent.

2.2 Stability and Matching: A Duality

Assuming that datasets X and Y in $\mathcal{D}$ are given, a natural way of comparing them is to compute the $d_{\mathcal{D}}$ distance between them (whatever that distance is). Often times, however, features computed out of datasets constitute simpler structures than the datasets themselves, and as such, they are more readily amenable to direct comparisons.

So, for a family of indices A consider here the stable family $\{f_\alpha, \alpha \in A\}$ of feature maps $f_\alpha : \mathcal{D} \to \mathcal{F}$, where $\alpha \in A$ and $\mathcal{F}$ is some *feature space* which is metrized by the distance function $d_{\mathcal{F}}$. In line with the observation above, spaces of features tend to have simpler structure than the space of datasets, and in consequence the computation of $d_{\mathcal{F}}$ usually appears to be simpler. This suggests that in order to distinguish between two datasets X and Y one computes

$$\eta_A(X, Y) := \sup_{\alpha \in A} d_{\mathcal{F}}\big(f_\alpha(X), f_\alpha(Y)\big)$$

as a proxy for $d_{\mathcal{D}}(X,Y)$. This would be reasonable because since each of the features f_α, $\alpha \in A$ is stable, there exist functions Ψ_α such that

$$\eta_A(X,Y) \leq \sup_{\alpha\in A} \Psi_\alpha\big(d_{\mathcal{D}}(X,Y)\big).$$

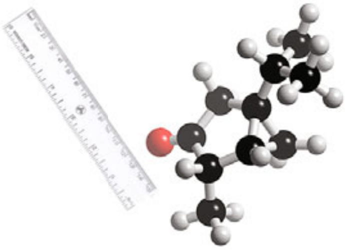
However, in order for this to be totally satisfactory it would be necessary to establish in the reverse direction! For a given subclass of datasets $\mathcal{O} \subset \mathcal{D}$, the main challenge is to find a stable family $\{f_\alpha, \alpha \in A\}$ that is rich enough so that it will discriminate all objects in $\mathcal{O}$: namely that if $X, Y \in \mathcal{O}$ and

$$f_\alpha(X) = f_\alpha(Y) \text{ for all } \alpha \in A \Longrightarrow X = Y.$$

In this respect the work of Olver [67], Boutin and Kemper [5] provide for example families of features that are able to discriminate certain datasets under rigid isometries. Other interesting and useful examples are ultrametric spaces, or in more generality trees.

3 Datasets as Metric Spaces or Metric Measure Spaces

In many applications datasets can be represented as metric spaces (see Figure 2), that is, as a pair (X, d_X) where $d_X : X \times X \to \mathbb{R}^+$ satisfies the three metric properties: (a) $d_X(x,x') = 0$ if and only if $x = x'$; (b) $d_X(x,x') = d_X(x',x)$ for all $x, x' \in X$; and (c) $d_{(}x,x') \leq d_X(x,x'') + d_X(x'',x')$ for all $x,x',x'' \in X$. Henceforth, $\mathcal{G}$ will denote the collection of all compact metric spaces.



$$\Longrightarrow \begin{pmatrix} 0 & d_{12} & d_{13} & d_{14} & \dots \\ d_{12} & 0 & d_{23} & d_{24} & \dots \\ d_{13} & d_{23} & 0 & d_{34} & \dots \\ d_{14} & d_{24} & d_{34} & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Fig. 2. Datasets as metric spaces: given the dataset, and a notion of "ruler", one induces a matrix containing the distance between all pairs of points; this distance is application dependent

We introduce some notation: for a finite metric space (X, d_X), its *separation* is the number $\mathbf{sep}\,(X) := \min_{x\neq x'} d_X(x,x')$. For any compact X, its diameter is $\mathbf{diam}\,(X) := \max_{x,x'} d_X(x,x')$.

For example in the case of Euclidean datasets, one has the following result:

Lemma 1 ([5]). *Let X and Y be finite subsets of $\mathbb{R}^k$ s.t. there exists $\phi : X \to Y$ a bijection with $\|x - x'\| = \|\phi(x) - \phi(x')\|$ for all $x, x' \in X$. Then, there exist a rigid isometry $\Phi : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $Y = \Phi(X)$.*

This lemma implies that representing a Euclidean dataset (e.g. a protein, a chemical compound, etc) as a metric space by endowing it with the ambient space distance, one retains the original information up to ambient space isometries (in this case, rotations, translations, and reflections). In particular, this is not restrictive in any way, because anyhow in most conceivable cases one would not want the output of an algorithm to depend on the coordinate system in which the data is represented.

In the context of protein structure comparison, some ideas regarding the direct comparison of distance matrices can be found for example in [40].

There are other types of datasets which are not Euclidean, but also fit in the metric framework. One example is given by phylogenetic trees. Indeed, it is well known [78] that trees are exactly those metric spaces (X, d_X) that satisfy the *four point condition*: for all $x, y, z, w \in X$

$$d_X(x,y) + d_X(z,w) \leq \max\big(d_X(x,z) + d_X(y,w), d_X(x,w) + d_X(z,y)\big).$$

Another rich class of examples where the metric representation of objects arises in problems in object recognition under invariance to bending transformations, see [55,28,63,64,12,41,70,11,9,10,8].

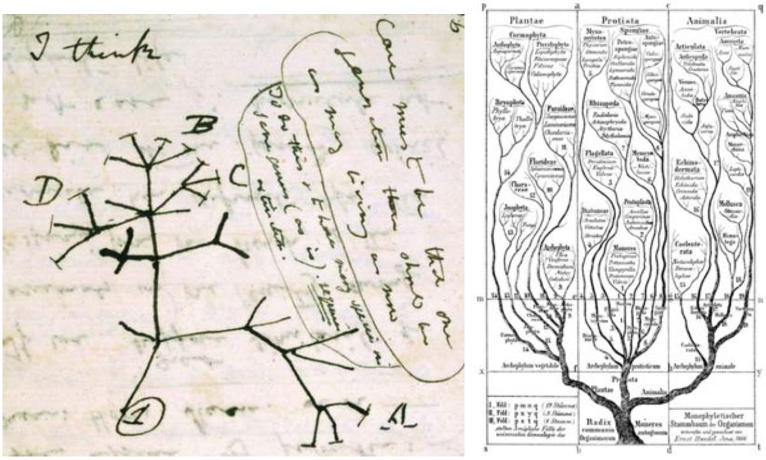


Fig. 3. Famous phylogenetic trees

mm-spaces. A *metric measure space* or *mm-space* for short, is a triple (X, d_X, μ_X) where (X, d_X) is a metric space and μ_X is a Borel probability measure on X. In the finite case, μ_X reduces to a collection of non-negative *weights*, one for each point $x \in X$, such that the sum of all weights equals 1. The interpretation is that $\mu_X(x)$ measures the "importance" of x: points with zero weight should not matter, points with lower values of the weight should be less prominent than points with larger values of the weight, etc. The representation of objects as mm-spaces can thus incorporate more information than the purely metric representation of data— when there is no application motivated choice of weights one can resort to the giving the points the *uniform distribution*, that is all points would have the same weight.

Henceforth, $\mathcal{G}_w$ will denote the collection of all compact mm-spaces.[1]

3.1 Equality of Datasets

What is the notion of equality between datasets? In the case when datasets are represented as metric spaces, we declare that $X, Y \in \mathcal{G}$ are equal whenever we cannot tell them apart by performing pairwise measurements of interpoint distances. In mathematical language, in order to check whether X and Y are equal we require that there be a surjective map $\phi : X \to Y$ which preserves distances and leaves no holes:

- $d_X(x,x') = d_Y(\phi(x),\phi(x'))$ for all $x, x' \in X$; and
- $\phi(X) = Y$.

Such maps (when X and Y are compact) are necessarily bijective, and are called *isometries*.

When datasets are represented as mm-spaces the notion of equality between them must take into account the preservation of not only the pair-wise distance information, but also that of the weights. One considers $X, Y \in \mathcal{G}_w$ to be equal, whenever there exists an isometry $\phi : X \to Y$ that *also preserves the weights*: namely that (assume that X and Y are finite for simplicity) $\mu_X(x) = \mu_Y(\phi(x))$, for all $x \in X$, see [60].

4 Metric Structures on Datasets

We now wish to produce a notion of distance between datasets that is not "too rigid" and allows substantiating a picture such as that emerging from §2.2. We will now describe the construction of distances in both $\mathcal{G}$ and $\mathcal{G}_w$.

4.1 The Case of $\mathcal{G}$

A suitable notion of distance between objects in $\mathcal{G}$ is the *Gromov-Hausdorff* distance, which can be defined as follows. We first introduce the case of finite objects and then explain the general construction.

Given objects $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ with metrics d_X and d_Y, respectively, let $R = (\!(r_{ij})\!) \in \{0,1\}^{n \times m}$ be such that

$$\sum_i r_{ij} \geq 1 \text{ for all } j \text{ and } \sum_j r_{ij} \geq 1 \text{ for all } i.$$

[1] The sub-index w is meant to suggest "weighted metric spaces".

The interpretation is that any such binary matrix R represents a notion of *friendship* between points in X and points in Y: namely, that x_i and y_j are friends if and only if $r_{ij} = 1$. Notice that the conditions above imply that every point in X has at least one friend in Y, and reciprocally, that every point in Y has at least one friend in X.

Denote by $\mathcal{R}(X, Y)$ the set of all such possible matrices, which we shall henceforth refer to as *correspondences* between X and Y.

Then, one defines the Gromov-Hausdorff distance between (X, d_X) and (Y, d_Y) as

$$d_{\mathcal{GH}}(X, Y) := \frac{1}{2} \min_{R} \max_{i,i',j,j'} \big|d_X(x_i, x_{i'}) - d_X(x_j, x_{j'})\big| r_{ij} r_{i'j'},$$

where the minimum is taken over $R \in \mathcal{R}(X, Y)$.

The definition above has the interpretation that one is trying to match points in X to points in Y in such a way that the metrics of X and Y are optimally aligned.

The general case. In the full case of any pair of datasets X and Y (not necessarily finite) in $\mathcal{G}$, one needs to generalize the definition above. Let $\mathcal{R}(X, Y)$ denote now the collection of all subsets R of the Cartesian product $X \times Y$ with the property that the canonical coordinate projections $\pi_1 : X \times Y \to X$ and $\pi_2 : X \times Y \to Y$ are *surjective*, when restricted to R.

Then the Gromov-Hausdorff distance between compact metric spaces X and Y is defined as

$$d_{\mathcal{GH}}(X, Y) := \frac{1}{2} \inf_{R \in \mathcal{R}(X,Y)} \sup_{(x,y),(x',y') \in R} \big|d_X(x, x') - d_Y(y, y')\big|. \tag{1}$$

This definition indeed respects the notion of equality of objects that we put forward in §3.1:

Theorem 1 ([35]). *$d_{\mathcal{GH}}$ is a metric on the isometry classes of $\mathcal{G}$.*

Another expression for the GH distance. Recall the definition of the Hausdorff distance between (closed) subsets A and B of a metric space (Z, d_Z):

$$d_{\mathcal{H}}^{Z}\big(A, B\big) := \max\big(\max_{a \in A} \min_{b \in B} d_Z(a, b), \max_{b \in B} \min_{a \in A} d_Z(a, b)\big).$$

Given compact metric spaces (X, d_X) and (Y, d_Y), consider all metrics d on the disjoint union $X \sqcup Y$ s.t.

- $d(x, x') = d_X(x, x')$, all $x, x' \in X$;
- $d(y, y') = d_Y(y, y')$, all $y, y' \in Y$.

Then, according to [13, Chapter 7]

$$d_{\mathcal{GH}}(X, Y) := \inf_{d} d_{\mathcal{H}}^{(X \sqcup Y, d)}\big(X, Y\big),$$

where the infimum is taken over all the metrics d that satisfy the conditions above.

Remark 1. According to this formulation, computing the GH distance between two finite metric spaces can be regarded as a **distance matrix completion problem**. The functional is $J(d) = \max\big(\max_x \min_y d(x,y), \max_y \min_x d(x,y)\big)$ [60]. The number of constraints is roughly of order n^3 for all the **triangle inequalities**, where $n = |X| \simeq |Y|$.

Example: Euclidean datasets. Endowing objects embedded in $\mathbb{R}^d$ with the (restricted) Euclidean metric makes the Gromov-Hausdorff distance invariant under ambient rigid isometries [59]. In order to argue that similarity in the Gromov-Hausdorff sense has a meaning which is compatible and comparable with other notions of similarity that we have already come to accept as natural, it is useful to look into the case of similarity of objects under rigid motions. One of the most commonplace notions of rigid similarity is given by the Hausdorff distance under rigid isometries [43] for which one has

Theorem 2 ([59]). *Let $X, Y \subset \mathbb{R}^d$ be compact. Then*

$$d_{\mathcal{GH}}((X, \|\cdot\|), (Y, \|\cdot\|)) \leq \inf_T d_{\mathcal{H}}^{\mathbb{R}^d}(X, T(Y)) \leq c_d \cdot M^{\frac{1}{2}} \cdot \big(d_{\mathcal{GH}}((X, \|\cdot\|), (Y, \|\cdot\|))\big)^{\frac{1}{2}},$$

where $M = \max(\mathbf{diam}\,(X)\,, \mathbf{diam}\,(Y))$ and c_d is a constant that depends only on d. The infimum over T above is taken amongst all Euclidean isometries.

Note that this theorem is a natural relaxation of the statement of Lemma 1.

4.2 The Case of $\mathcal{G}_w$

Using ideas from mass transport it is possible to define a version of the Gromov-Hausdorff distance that applies to datasets in $\mathcal{G}_w$.

Fix a metric space (Z, d_Z) and let $\mathcal{P}(Z)$ denote the collection of all the Borel probability measures. For $\alpha, \beta \in \mathcal{P}(Z)$, the **Wasserstein distance** (or order $p \geq 1$) on $\mathcal{P}(Z)$ is given by:

$$d_{\mathcal{W},p}^{(Z,d_Z)}(\alpha, \beta) := \left(\iint_{Z\times Z} \big(d_Z(z,z')\big)^p \mu(dz \times dz')\right)^{1/p},$$

where $\mu \in \mathcal{P}(Z \times Z)$ is a probability measure with marginals α and β. An excellent reference for these concepts is the book of Villani [89].

An interpretation of this definition comes from thinking that one has a pile of sand/dirt that must be moved from one location to another, where in the destination one wants build something with this material, see Figure 4. In the finite case (i.e. when all the probability measures are linear combinations of deltas), $\mu_{i,j}$ encodes information about how much of the mass initially at x_i must be moved to x_j, see Figure 4.

The **Gromov-Wasserstein distance** between mm-spaces X and Y is defined as an *optimal mass transportation* problem on $X \sqcup Y$: for $p \geq 1$

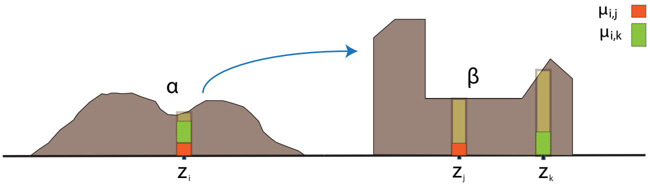


Fig. 4. An optimal mass transportation problem (in the Kantorovich formulation): the pile of sand/dirt on the left must be moved to another location on the right with the purpose of assembling a building or structure

$$d_{\mathcal{GW},p}(X,Y) := \inf_d d_{\mathcal{W},p}^{(X\sqcup Y,d)}(\mu_X,\mu_Y),$$

where as before d is a metric on $X \sqcup Y$ gluing X and Y.

The definition above is due to Sturm [83]. Notice that the underlying optimization problems that one needs to solve now are of continuous nature as opposed to the combinatorial optimization problems yielded by the GH distance. Another non-equivalent definition of the Gromov-Wasserstein distance is proposed in [60] whose discretization is more tractable.

As we will see ahead, several features become stable in the GW sense.

4.3 Stability of Hierarchical Clustering Methods

Denote by $\mathbf{P}(X)$ the set of all partitions of the finite set X.

A *dendrogram* over a finite set X is a function $\theta_X : [0,\infty) \to \mathbf{P}(X)$ with the following properties:

1. $\theta_X(0) = \{\{x_1\},\ldots,\{x_n\}\}$.
2. There exists t_0 s.t. $\theta_X(t)$ is the *single block partition* for all $t \geq t_0$.
3. If $r \leq s$ then $\theta_X(r)$ *refines* $\theta_X(s)$.
4. For all r there exists $\varepsilon > 0$ s.t. $\theta_X(r) = \theta_X(t)$ for $t \in [r, r+\varepsilon]$.

Let $\mathbf{D}(X)$ denote the collection of all possible dendrograms over a given finite set X.

Hierarchical clustering methods are maps $\mathfrak{H}$ from the collection of all finite metric spaces into the collection of all dendrograms, such that (X, d_X) is mapped into an element of $\mathbf{D}(X)$.

Standard examples of clustering methods are *single, complete and average linkage methods* [46].

A question of great interest is whether any of these clustering methods is stable to perturbations in the input metric spaces.

Linkage based agglomerative HC methods. Here we review the basic procedure of linkage based hierarchical clustering methods:

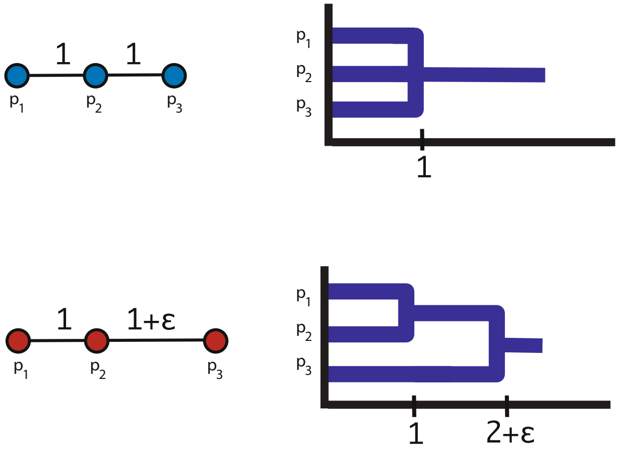


Fig. 5. Complete Linkage is not stable to small perturbations in the metric. On the left we show two metric spaces that are metrically very similar. To the right of each of them we show their CL dendrogram outputs. Regardless of $\varepsilon > 0$, the two outputs are always very dissimilar.

Assume (X, d_X) is a given finite metric space. In this example, we use the formulas for CL but the structure of the iterative procedure in this example is common to all HC methods [46, Chapter 3]. Let θ be the dendrogram to be constructed in this example.

1. Set $X_0 = X$ and $D_0 = d_X$ and set $\theta(0)$ to be the partition of X into singletons.
2. Search the matrix D_0 for the smallest non-zero value, i.e. find $\delta_0 = \mathbf{sep}\,(X_0)$, and find all pairs of points $\{(x_{i_1}, x_{j_1}), (x_{i_2}, x_{j_2}) \ldots, (x_{i_k}, x_{j_k})\}$ at distance δ_0 from eachother, i.e. $d(x_{i_\alpha}, x_{j_\alpha}) = \delta_0$ for all $\alpha = 1, 2, \ldots, k$, where one orders the indices s.t. $i_1 < i_2 < \ldots < i_k$.
3. Merge the first pair of elements in that list, (x_{i_1}, x_{j_1}), into a single group. The procedure now removes (x_{i_1}, x_{j_1}) from the initial set of points and adds a point c to represent the cluster formed by both: define $X_1 = \big(X_0 \backslash \{x_{i_1}, x_{j_1}\}\big) \cup \{c\}$. Define the dissimilarity matrix D_1 on $X_1 \times X_1$ by $D_1(a,b) = D_0(a,b)$ for all $a, b \neq c$ and $D_1(a,c) = D_1(c,a) = \max\big(D_0(x_{i_1}, a), D_0(x_{j_1}, a)\big)$ (this step is the only one that depends on the choice corresponding to CL). Finally, set
$$\theta(\delta) = \{x_{i_1}, x_{j_1}\} \cup \bigcup_{i \neq i_1, j_1} \{x_i\}.$$
4. The construction of the dendrogram θ is completed by repeating the previous steps until all points have been merged into a single cluster.

The *tie breaking* strategy used in step 3 results in the algorithm producing different non-isomorphic outputs depending on the labeling of the points. This is undesirable, but can be remedied by defining certain versions of all the linkage based HC methods that behave well under permutations [18] .

Unfortunately, even these "patched" versions of AL and CL fail to exhibit stability, see Figure 5.

It turns out, however, that single linkage does enjoy stability. Before we phrase the precise result we need to introduce the ultrametric representation of dendrograms. Furthermore, as we will see in 5, there's a sense in which SLHC is *the only HC method that can be stable.*

Dendrograms as ultrametric spaces. The representation of dendrograms as ultrametrics is well known [48,37,46].

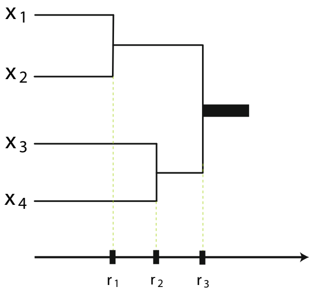
Theorem 3 ([18]). *Given a finite set X, there is a bijection $\Psi : \mathbf{D}(X) \to \mathbf{U}(X)$ between the collection $\mathbf{D}(X)$ of all dendrograms over X and the collection $\mathbf{U}(X)$ of all ultrametrics over X such that for any dendrogram $\theta \in \mathbf{D}(X)$ the ultrametric $\Psi(\theta)$ over X generates the same hierarchical decomposition as θ, i.e.*

$$(*) \quad \textit{for each } r \geq 0,\ x, x' \in B \in \theta(r) \iff \Psi(\theta)(x, x') \leq r.$$

Furthermore, this bijection is given by

$$\Psi(\theta)(x, x') = \min\{r \geq 0 |\, x, x' \textit{ belong to the same block of } \theta(r)\}. \qquad (2)$$

See Figure 6.



$$((u_\theta)) = \begin{array}{c|cccc} & x_1 & x_2 & x_3 & x_4 \\ \hline x_1 & 0 & r_1 & r_3 & r_3 \\ x_2 & r_1 & 0 & r_3 & r_3 \\ x_3 & r_3 & r_3 & 0 & r_2 \\ x_4 & r_3 & r_3 & r_2 & 0 \end{array}$$

Fig. 6. A graphical representation of a dendrogram θ over $X = \{x_1, x_2, x_3, x_3\}$ and the corresponding ultrametric $u_\theta := \Psi(\theta)$. Notice for example, that according to (2), $u_\theta(x_1, x_2) = r_1$ since r_1 is the first value of the (scale) parameter for which x_1 and x_2 are merged into the same cluster. Similarly, since x_1 and x_3 are merged into the same cluster for the first time when the parameter equals r_3, then $u_\theta(x_1, x_3) = r_3$.

Let $\mathcal{U} \subset \mathcal{G}$ denote the collection of all (compact) ultrametric spaces. It follows from Theorem 3 that one can regard HC methods as maps $\mathfrak{H} : \mathcal{G} \to \mathcal{U}$. In particular [18], SLHC can be regarded as the map $\mathfrak{H}^{\mathrm{SL}}$ that assigns (X, d_X)

with (X, u_X), where u_X is the *maximal subdominant ultrametric* relative to d_X. This is given as

$$u_X(x, x') := \min \left\{ \max_{i=0,\ldots,k-1} d_X(x_i, x_{i+1}), \text{ s.t. } x = x_0, \ldots, x_k = x' \right\}. \qquad (3)$$

Stability and convergence of SLHC. In contrast with the situation for complete and average linkage HCMs, we have the following statement concerning the quantitative stability of SLHC:

Theorem 4 ([18]). *Let (X, d_X) and (Y, d_Y) be two finite metric spaces. Then,*

$$d_{\mathcal{GH}}(\mathfrak{H}^{\mathrm{SL}}(X, d_X), \mathfrak{H}^{\mathrm{SL}}(Y, d_Y)) \leq d_{\mathcal{GH}}((X, d_X), (Y, d_Y)).$$

Invoking the ultrametric representation of dendrograms and using Theorem 4, [18] proves the following convergence result, see Figure 7.

Theorem 5. *Let (Z, d_Z, μ_Z) be an mm-space and write supp $[\mu_Z] = \bigcup_{\alpha \in A} Z^{(\alpha)}$ for a finite index set A and $\{Z^{(\alpha)}\}_{\alpha \in A}$ a collection of disjoint, compact, path-connected subsets of Z. Let (A, u_A) be the ultrametric space where u_A is the maximal subdominant ultrametric with respect to $W_A(\alpha, \alpha') := \min_{z \in Z^{(\alpha)}, z' \in Z^{(\alpha')}} d_Z(z, z')$, for $\alpha, \alpha' \in A$.*

For each $n \in \mathbb{N}$, let $X_n = \{z_1, z_2, \ldots, z_n\}$ be a collection of n independent random variables (defined on some probability space Ω with values in Z) with distribution μ_Z, and let d_{X_n} be the restriction of d_Z to $X_n \times X_n$. Then, $\mathfrak{H}^{\mathrm{SL}}(X_n, d_{X_n}) \xrightarrow{n} (A, u_A)$ in the Gromov-Hausdorff sense μ_Z-almost surely.

4.4 Stability of Vietoris-Rips Barcodes

Much in the same way as standard flat clustering can be understood as the zero-dimensional version of the notion of homology, hierarchical clustering can be regarded as the zero-dimensional version of persistent homology [27].

The notion of Vietoris-Rips persistent barcodes provides a precise sense in which the above statement is true. For a given finite metric space (X, d_X) and $r \geq 0$, let $R_r(X)$ denote the simplicial complex with vertex set X where $\sigma = [x_0, x_1, \ldots, x_k] \in R_r(X, d_X)$ if and only if $\max_{i,j} d_X(x_i, x_j) \leq r$. This is called the Vietoris-Rips simplicial complex (with parameter r). Then, the family

$$\mathcal{R}(X, d_X) := \{R_r(X, d_X),\, r \geq 0\}$$

constitutes a *filtration*, in the sense that

$$R_r(X, d_X) \subseteq R_s(X, d_X), \text{ whenever } s \geq r.$$

In the sequel we may abbreviate $R_r(X)$ for $R_r(X, d_X)$, and similarly for $\mathcal{R}(X)$. Now, passing to homology with field coefficients, this inclusion gives rise to a pair of vector spaces and a linear map between them:

$$\phi_r^s : H_*(R_r(X) \longrightarrow H_*(R_s(X)).$$

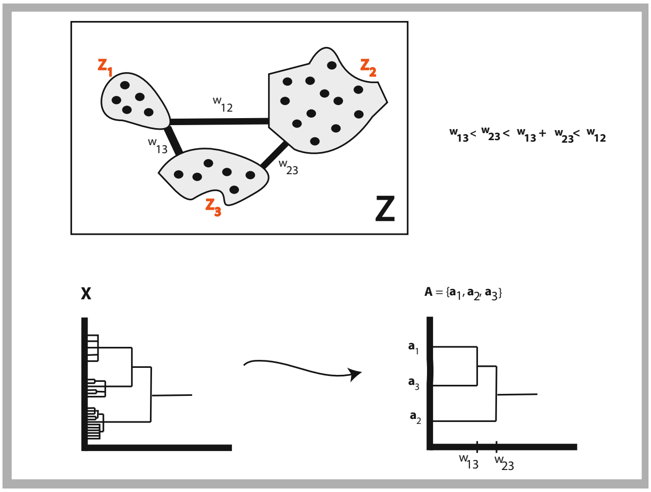


Fig. 7. Illustration of Theorem 5. *Top*: A space Z composed of 3 disjoint path connected parts, $Z^{(1)}, Z^{(2)}$ and $Z^{(3)}$. The black dots are the points in the finite sample X_n. In the figure, $w_{ij} = W_A(a_i, a_j)$, $1 \leq i \neq j \leq 3$. *Bottom Left*: The dendrogram representation of $(X_n, u_{X_n}) := \mathfrak{H}^{\mathrm{SL}}(X_n)$. *Bottom Right*: The dendrogram representation of (A, u_A). Note that $u_A(a_1, a_2) = w_{23}$, $u_A(a_1, a_3) = w_{13}$ and $u_A(a_2, a_3) = w_{23}$. As $n \to \infty$, $(X_n, u_{X_n}) \to (A, u_A)$ a.s. in the Gromov-Hausdorff sense, see text for details.

In more detail, if $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = \mathbf{diam}\,(X)$ are the distinct values assumed by d_X, then one obtains the *persistent vector space*:

$$H_*(R_{\alpha_0}(X)) \xrightarrow{\phi_0^1} H_*(R_{\alpha_1}(X)) \xrightarrow{\phi_1^2} H_*(R_{\alpha_2}(X)) \xrightarrow{\phi_2^3} \cdots$$

$$\cdots \xrightarrow{\phi_{m-2}^{m-1}} H_*(R_{\alpha_{m-1}}(X)) \xrightarrow{\phi_{m-1}^{m}} H_*(R_{\alpha_m}(X)).$$

It is well known [91] that there is a classification of such objects in terms of a finite multisets of points in the extended plane $\overline{\mathbb{R}}^2$, called the *persistence diagram* of $\mathcal{R}(X)$, and denoted $\mathrm{D}_*\mathcal{R}(X)$ which is contained in the union of the extended diagonal $\Delta = \{(x,x) : x \in \overline{\mathbb{R}}\}$ and of the grid $\{\alpha_0, \cdots, \alpha_m\} \times \{\alpha_0, \cdots, \alpha_m, \alpha_\infty = +\infty\}$. The multiplicity of the points of Δ is set to $+\infty$, while the multiplicities of the (α_i, α_j), $0 \leq i < j \leq +\infty$, are defined in terms of the ranks of the linear transformations $\phi_i^j = \phi_{j-1}^j \circ \cdots \circ \phi_i^{i+1}$ [20].

The *bottleneck distance* $d_{\mathrm{B}}^\infty(A,B)$ between two multisets in $(\overline{\mathbb{R}}^2, l^\infty)$ is the quantity $\min_\gamma \max_{p \in A} \|p - \gamma(p)\|_\infty$, where γ ranges over all bijections from A to B. Then, one obtains the following generalization of Theorem 4.

Theorem 6 ([20]). *Let* (X, d_X) *and* (Y, d_Y) *be any two finite metric spaces. Then for all* $k \geq 0$,

$$\frac{1}{2} d_{\mathrm{B}}^\infty\left(\mathrm{D}_k\mathcal{R}(X), \mathrm{D}_k\mathcal{R}(Y)\right) \leq d_{\mathcal{GH}}(X,Y).$$

This type of results are of great importance for applications of the Vietoris-Rips barcodes to data analysis.

4.5 Object Matching: More Details

Some features of mm-spaces. We define a few simple isomorphism invariants, or features, of mm-spaces, many of which will be used in §4.5 to establish lower bounds for the metrics we will impose on $\mathcal{G}_w$. All the features we discuss below have are routinely used in the data analysis and object matching communities.

Definition 1 (p-diameters). *Given a mm-space (X, d_X, μ_X) and $p \in [1, \infty]$ we define its p-**diameter** as*

$$\mathbf{diam}_p(X) := \left(\int_X \int_X \big(d_X(x,x')\big)^p \mu_X(dx) \mu_X(dx') \right)^{1/p}$$

for $1 \leq p < \infty$.

Definition 2. *Given $p \in [1, \infty]$ and an mm-space (X, d_X, μ_X) we define the p-**eccentricity function** of X as*

$$s_{X,p} : X \to \mathbb{R}^+ \quad \textit{given by} \quad x \mapsto \left(\int_X d_X(x,x')^p \mu(dx') \right)^{1/p}$$

for $1 \leq p < \infty$.

Hamza and Krim proposed using eccentricity functions (with $p = 2$) for describing objects in [36]. Ideas similar to those proposed in [36] have been revisited recently in [45]. See also Hilaga et al. [39]. Eccentricities are also routinely used as part of topological data analysis algorithms such as mapper [80].

Definition 3 (Distribution of distances). *To an mm-space (X, d_X, μ_X) we associate its **distribution of distances**:*

$$f_X : [0, \mathbf{diam}(X)] \to [0,1] \quad \textit{given by} \quad t \mapsto \mu_X \otimes \mu_X \big(\{(x,x') | d_X(x,x') \leq t\} \big).$$

See Figure 8 and [5,68].

Definition 4 (Local distribution of distances). *To a mm-space (X, d_X, μ_X) we associate its* local distribution of distances *defined by:*

$$h_X : X \times [0, \mathbf{diam}(X)] \to [0,1] \quad \textit{given by} \quad (x,t) \mapsto \mu_X \left(\overline{B_X(x,t)} \right).$$

See Figure 9. The earliest use of an invariant of this type known to the author is in the work of German researchers [4,50,1]. The so called **shape context** [3,79,75,14] invariant is closely related to h_X.

More similar to h_X is the invariant proposed by Manay et al. in [58] in the context of planar objects. This type of invariant has also been used for three dimensional objects [21,32]. More recently, in the context of planar curves, similar constructions have been analyzed in [7]. See also, [34].

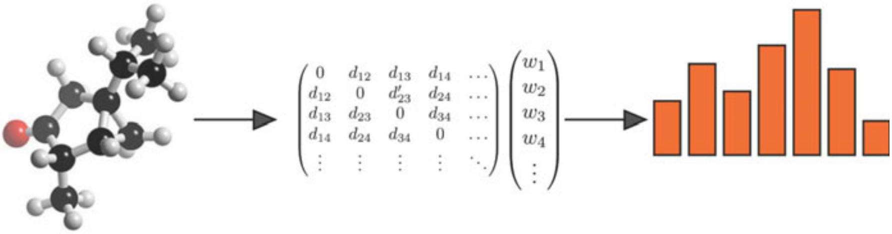


Fig. 8. Distribution of distances: from a dataset to the mm-space representation and from it to the distribution of distances

Remark 2 (**Local distribution of distances as a proxy for scalar curvature).** There is an interesting observation that in the class Riem $\subset \mathcal{G}_w$ of closed Riemannian manifolds local distributions of distance are intimately related to curvatures. Let M be an n-dimensional closed Riemannian manifold which we regard as an mm-space by endowing it with the geodesic metric and with probability measure given by the normalized volume measure. Using the well known expansion [77] of the Riemannian volume of a ball of radius t centered at $x \in M$ one finds:

$$h_M(x,t) = \frac{\omega_n(t)}{\mathbf{Vol}\,(M)}\left(1 - \frac{S_M(x)}{6(n+2)}t^2 + O(t^4)\right),$$

where $S_M(x)$ is the *scalar curvature* of M at x, $\omega_n(t)$ is the volume of a ball of radius t in $\mathbb{R}^n$ and $O(t^4)$ is a term whose decay to 0 as $t \downarrow 0$ is faster than t^4.

One may then argue that local shape distributions play a role of generalized notions of curvature.

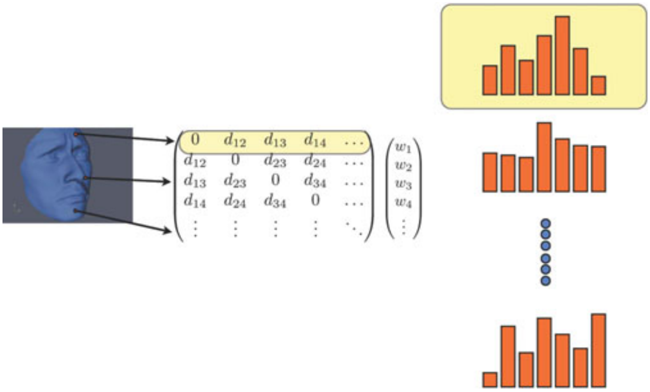


Fig. 9. Local distribution of distances: from a dataset to the mm-space representation and from it the local distribution of distances. To each point on the object one assigns the distribution of distance from this point to all other points on the object.

Precise bounds

Definition 5. *For $X, Y \in \mathcal{G}_w$ define*

$$\mathbf{FLB}(X,Y) := \frac{1}{2} \inf_{\mu \in \mathcal{M}(\mu_X,\mu_Y)} \left(\int_{X \times Y} |s_{X,1}(x) - s_{Y,1}(y)| \; \mu(dx \times dy) \right);$$

$$\mathbf{SLB}(X,Y) := \int_0^\infty |f_X(t) - f_Y(t)| \, dt;$$

$$\mathbf{TLB}(X,Y) := \frac{1}{2} \min_{\mu \in \mathcal{M}(\mu_X,\mu_Y)} \int_{X \times Y} \left(\int_0^\infty \big| h_X(x,t) - h_Y(y,t) \big| \, dt \right) \mu(dx \times dy).$$

For finite X and Y, computing the (exact) value of each of the quantities in the definition reduces to solving linear programming problems [60].

We now can state the following theorem asserting the stability of the features discussed in this section:

Theorem 7 ([60]). *For all $X, Y \in \mathcal{G}_w$, and all $p \geq 1$*

$$d_{\mathcal{GW},p}(X,Y) \geq \begin{cases} \mathbf{TLB}(X,Y) \geq \mathbf{FLB}(X,Y) \geq \frac{1}{2} |\mathbf{diam}_1(X) - \mathbf{diam}_1(Y)|. \\ \mathbf{SLB}(X,Y). \end{cases}$$

Bounds of this type, besides establishing the quantitative stability of the different intervening features, have the added usefulness that in practice they may be utilized in *layered* comparison of objects: those bounds involving simpler invariants are frequently easier to compute, whereas those involving more powerful features most often require more effort. Furthermore, hierarchical bounds of this nature that interconnect different approaches proposed in the literature allow for a better understanding of the landscape of different existing techniques, see Figure 10.

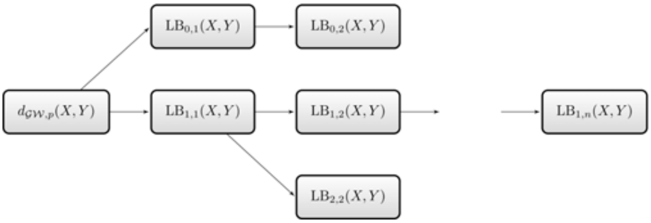


Fig. 10. Having a hierarchy (arrows should be read as $\geq$ symbols) of lower bounds such the one suggested in the figure can help in matching tasks: the strategy that suggests itself is to start the comparison using the weaker bounds and gradually increase the complexity

Also, different families of lower bounds for the GH distance have recently been found [61]; these incorporate features similar to those of [56,86].

Spectral versions of the GH and GW distances. It is possible to obtain a hierarchy of lower bounds similar to the ones above but in the context of *spectral methods* [62], see Figure 12. The motivation comes from the so called Varadhan's Lemma: if X is a compact Riemannian manifold without boundary, and k_X denotes the *heat kernel* of X, then one has

Lemma 2 ([66]). *For any compact Riemannian manifold without boundary X,*

$$\lim_{t\downarrow 0}\big(-4t\ln k_X(t,x,x')\big) = d_X^2(x,x'),$$

for all $x,x'\in X$. Here $d_X(x,x')$ is the geodesic distance between x and x' on X.

The spectral representation of objects (see Figure 12), and in particular shapes is interesting because it readily encodes a notion of *scale*. This scale parameter (the t parameter in the heat kernel) permits reasoning about similarity of shapes at different levels of "blurring" or "smoothing", see Figure 11. A (still not thoroughly satisfactory) interpretation of t as a scale parameter arises from the following observations:

- For $t\downarrow 0^+$, $k_X(t,x,x)\simeq(4\pi t)^{-d/2}\big(1+\frac{1}{6}S_X(x)+\dots\big)$, where d is the dimension of X. Recall that S_X is the scalar curvature— therefore for small enough t, one sees local information about X.
- For $t\to\infty$, $k_X(t,x,x')\to\frac{1}{\mathbf{Vol}(X)}$. Hence, for large t all points "look the same".
- Pick $n\in\mathbb{N}$ and $\varepsilon>0$ and let $L_g=(\mathbb{R},g,\lambda)$ for $g(x)=1+\varepsilon\cos(2\pi xn)$, then the *homogenized metric* is $\overline{g}=1$. Then, by results due to Tsuchida and Davies [87,26] one has that

$$\sup_{x,x'\in\mathbb{R}}\big|k_g(t,x,x')-k_{\overline{g}}(t,x,x')\big|\le\frac{C}{t}\text{ as }t\uparrow\infty.$$

Since for Riemannian manifolds X and Y, by Varadhan's lemma, the heat kernels k_X and k_Y determine the geodesic metrics d_X and d_Y, respectively, this suggests defining **spectral versions** of the GH and GW distances. For each $p\ge 1$, one defines [62]

$$d^{\text{spec}}_{\mathcal{GW},p}(X,Y):=\frac{1}{2}\inf_{\mu}\sup_{t>0}\mathbf{F}_p\big(k_X(t,\cdot,\cdot),k_Y(t,\cdot,\cdot),\mu\big),$$

where $\mathbf{F}_p$ is a certain functional that depends on both heat kernels and the measure coupling μ (see [62]).[2] The interpretation is that one takes the supremum over all t as way of choosing *the most discriminative scale.*

One has:

Theorem 8 ([62]). $d^{\text{spec}}_{\mathcal{GW},p}$ *defines a metric on the collection of (isometry classes of) Riemannian manifolds.*

[2] Here μ is a measure coupling between the *normalized* volume measures of X and Y.

A large number of spectral features are **quantitatively stable** under $d_{\mathcal{GH}}^{\mathrm{spec}}$ [62]. Examples are the spectrum of the Laplace-Beltrami operator [73], features computed from the diffusion distance, [53,22], and the heat kernel signature [84].

A more precise framework for the geometric scales of subsets of $\mathbb{R}^d$ is worked out in [54].

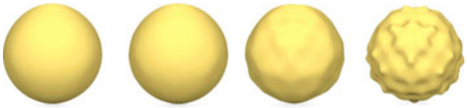


Fig. 11. A bumpy sphere at different levels of smoothing

5 Classification of Algorithms

In the next section, we will give a brief description of the theory of categories and functors, an excellent reference for these ideas is [57].

5.1 Brief Overview of Categories and Functors

Categories are mathematical constructs that encode the nature of certain objects of interest *together with a set of admissible maps between them.*

Definition 6. *A* ***category*** $\underline{C}$ *consists of:*

- *A collection of* ***objects*** $\mathrm{ob}(\underline{C})$ *(e.g. sets, groups, vector spaces, etc.)*
- *For each pair of objects* $X, Y \in \mathrm{ob}(\underline{C})$*, a set* $\mathrm{Mor}_{\underline{C}}(X,Y)$*, the* ***morphisms*** *from* X *to* Y *(e.g. maps of sets from* X *to* Y*, homomorphisms of groups from* X *to* Y*, linear transformations from* X *to* Y*, etc. respectively)*
- *Composition operations:* $\circ : \mathrm{Mor}_{\underline{C}}(X,Y) \times \mathrm{Mor}_{\underline{C}}(Y,Z) \to \mathrm{Mor}_{\underline{C}}(X,Z)$*, corresponding to* ***composition*** *of set maps, group homomorphisms, linear transformations, etc.*
- *For each object* $X \in \underline{C}$*, a distinguished element* $id_X \in \mathrm{Mor}_{\underline{C}}(X,X)$*, called the* ***identity*** *morphism.*

The composition is assumed to be associative in the obvious sense, and for any $f \in \mathrm{Mor}_{\underline{C}}(X,Y)$*, it is assumed that* $id_Y \circ f = f$ *and* $f \circ id_X = f$*.*

Definition 7 ($\underline{\mathcal{C}}$, a category of outputs of standard clustering schemes). *Let* Y *be a finite set,* $P_Y \in \mathcal{P}(Y)$*, and* $f : X \to Y$ *be a set map. We define* $f^*(P_Y)$ *to be the partition of* X *whose blocks are the sets* $f^{-1}(B)$ *where* B *ranges over the blocks of* P_Y*. We construct the category* $\underline{\mathcal{C}}$ *of outputs of standard clustering algorithms with* $\mathrm{ob}(\underline{\mathcal{C}})$ *equal to all possible pairs* (X, P_X) *where* X *is a finite set and* P_X *is a partition of* X*:* $P_X \in \mathcal{P}(X)$*. For objects* (X, P_X) *and* (Y, P_Y) *one sets* $\mathrm{Mor}_{\underline{\mathcal{C}}}\big((X, P_X), (Y, P_Y)\big)$ *to be the set of all maps* $f : X \to Y$ *with the property that* P_X *is a refinement of* $f^*(P_Y)$*.*

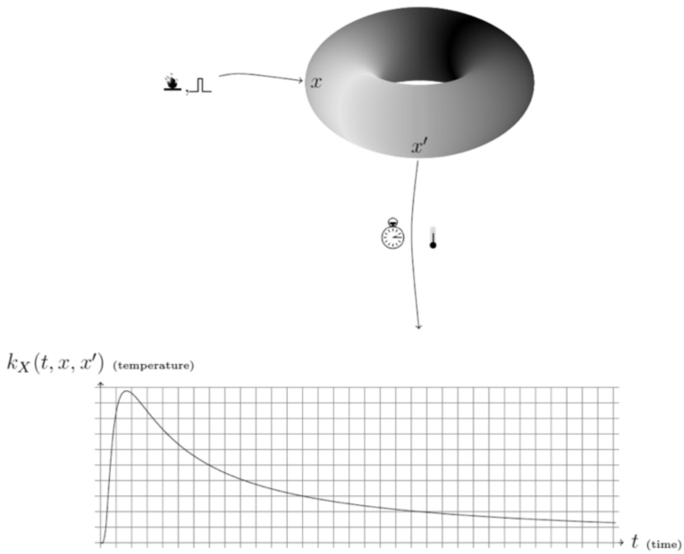


Fig. 12. A physics based way of characterizing/measuring a shape. For each pair of points x and x' on the shape X, one heats a tiny area around point x to a very high temperature in a very short interval of time around $t = 0$. Then, one measures the temperature at point x' for all later times and plots the resulting graph of the heat kernel $k_X(t, x, x')$ as a function of t. The knowledge of these graphs for all $x, x' \in X$ and $t > 0$ translates into knowledge of the heat kernel of X (the plot in the figure corresponds to $x \neq x'$). In contrast, one can think that a geometer's way of characterizing the shape would be via the use of a geodesic ruler that can be used for measuring distances between all pairs of points on X, see Figure 2. According to Varadhan's Lemma, both approaches are equivalent in the sense that they both capture the same information about X.

Example 1. Let X be any finite set, $Y = \{a, b\}$ a set with two elements, and P_X a partition of X. Assume first that $P_Y = \{\{a\}, \{b\}\}$ and let $f : X \to Y$ be any map. Then, in order for f to be a morphism in $\mathrm{Mor}_{\underline{\mathcal{C}}}\big((X, P_X), (Y, P_Y)\big)$ it is necessary that x and x' be in different blocks of P_X whenever $f(x) \neq f(x')$. Assume now that $P_Y = \{a, b\}$ and $g : Y \to X$. Then, the condition that $g \in \mathrm{Mor}_{\underline{\mathcal{C}}}\big((Y, P_Y), (X, P_X)\big)$ requires that $g(a)$ and $g(b)$ be in the same block of P_X.

We will also construct a category of *persistent sets*, which will constitute the output of hierarchical clustering functors.

Definition 8 ($\underline{\mathcal{P}}$, a category of outputs of hierarchical clustering schemes). *Let $(X, \theta_X), (Y, \theta_Y)$ be persistent sets. A map of sets $f : X \to Y$ is said to be* persistence preserving *if for each $r \in \mathbb{R}$, we have that $\theta_X(r)$ is a refinement of $f^*(\theta_Y(r))$. We define a category $\underline{\mathcal{P}}$ whose objects are persistent sets, and where $\mathrm{Mor}_{\underline{\mathcal{P}}}((X, \theta_X), (Y, \theta_Y))$ consists of the set maps from X to Y which are persistence preserving.*

Three categories of finite metric spaces. We will describe three categories $\underline{\mathcal{M}}^{iso}$, $\underline{\mathcal{M}}^{inj}$, and $\underline{\mathcal{M}}^{gen}$, whose collections of objects will all consist of the collection of finite metric spaces $\mathcal{M}$. For (X, d_X) and (Y, d_Y) in $\mathcal{M}$, a map $f : X \to Y$ is said to be **distance non increasing** if for all $x, x' \in X$, we have $d_Y(f(x), f(x')) \leq d_X(x, x')$. It is easy to check that composition of distance non-increasing maps are also distance non-increasing, and it is also clear that id_X is always distance non-increasing. We therefore have the category $\underline{\mathcal{M}}^{gen}$, whose objects are finite metric spaces, and so that for any objects X and Y, $\mathrm{Mor}_{\underline{\mathcal{M}}^{gen}}(X, Y)$ is the set of distance non-increasing maps from X to Y. It is clear that compositions of injective maps are injective, and that all identity maps are injective, so we have the new category $\underline{\mathcal{M}}^{inj}$, in which $\mathrm{Mor}_{\underline{\mathcal{M}}^{inj}}(X, Y)$ consists of the **injective distance non-increasing maps**. Finally, if (X, d_X) and (Y, d_Y) are finite metric spaces, $f : X \to Y$ is an **isometry** if f is bijective and $d_Y(f(x), f(x')) = d_X(x, x')$ for all x and x'. It is clear that as above, one can form a category $\underline{\mathcal{M}}^{iso}$ whose objects are finite metric spaces and whose morphisms are the isometries. Furthermore, one has inclusions

$$\underline{\mathcal{M}}^{iso} \subseteq \underline{\mathcal{M}}^{inj} \subseteq \underline{\mathcal{M}}^{gen} \tag{4}$$

of subcategories (defined as in [57]). Note that although the inclusions are bijections on object sets, they are proper inclusions on morphism sets.

Remark 3. The category $\underline{\mathcal{M}}^{gen}$ is special in that for any pair of finite metric spaces X and Y, $\mathrm{Mor}_{\underline{\mathcal{M}}^{gen}}(X, Y) \neq \emptyset$. Indeed, pick $y_0 \in Y$ and define $\phi : X \to Y$ by $x \mapsto y_0$ for all $x \in X$. Clearly, $\phi \in \mathrm{Mor}_{\underline{\mathcal{M}}^{gen}}(X, Y)$. This is not the case for $\underline{\mathcal{M}}^{inj}$ since in order for $\mathrm{Mor}_{\underline{\mathcal{M}}^{inj}}(X, Y) \neq \emptyset$ to hold it is necessary (but not sufficient in general) that $|Y| \geq |X|$.

Functors and functoriality. Next we introduce the key concept in our discussion, that of a *functor*. We give the formal definition first, and several examples will appear as different constructions that we use in the paper.

Definition 9 (Functor). *Let $\underline{C}$ and $\underline{D}$ be categories. Then a* functor *from $\underline{C}$ to $\underline{D}$ consists of:*

- *A map of sets $F : \mathrm{ob}(\underline{C}) \to \mathrm{ob}(\underline{D})$.*
- *For every pair of objects $X, Y \in \underline{C}$ a map of sets $\Phi(X, Y) : \mathrm{Mor}_{\underline{C}}(X, Y) \to \mathrm{Mor}_{\underline{D}}(FX, FY)$ so that*
  1. *$\Phi(X, X)(id_X) = id_{F(X)}$ for all $X \in \mathrm{ob}(\underline{C})$, and*
  2. *$\Phi(X, Z)(g \circ f) = \Phi(Y, Z)(g) \circ \Phi(X, Y)(f)$ for all $f \in \mathrm{Mor}_{\underline{C}}(X, Y)$ and $g \in \mathrm{Mor}_{\underline{C}}(Y, Z)$.*

Given a category $\underline{C}$, an ***endofunctor*** *on $\underline{C}$ is any functor $F : \underline{C} \to \underline{C}$.*

Remark 4. In the interest of clarity, we will always refer to the pair (F, Φ) with a single letter F. See diagram (6) below for an example.

Example 2 (Scaling functor). For any $\lambda > 0$ we define an endofunctor $\sigma_\lambda : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{M}}^{gen}$ on objects by $\sigma_\lambda(X, d_X) = (X, \lambda \cdot d_X)$ and on morphisms by $\sigma_\lambda(f) = f$. One easily verifies that if f satisfies the conditions for being a morphism in $\underline{\mathcal{M}}^{gen}$ from (X, d_X) to (Y, d_Y), then it readily satisfies the conditions of being a morphism from $(X, \lambda \cdot d_X)$ to $(Y, \lambda \cdot d_Y)$. Clearly, σ_λ can also be regarded as an endofunctor in $\underline{\mathcal{M}}^{iso}$ and $\underline{\mathcal{M}}^{inj}$.

Similarly, we define a functor $s_\lambda : \underline{\mathcal{P}} \to \underline{\mathcal{P}}$ by setting $s_\lambda(X, \theta_X) = (X, \theta_X^\lambda)$, where $\theta_X^\lambda(r) = \theta_X(\frac{r}{\lambda})$.

5.2 Clustering Algorithms as Functors

The notion of categories, functors and functoriality provide useful framework for studying algorithms. One first defines a class of input objects $\mathcal{I}$ and also a class of output objects $\mathcal{O}$. Moreover, one associates to each of these classes a class of natural maps, the morphisms, between objects, making them into *categories* $\underline{\mathcal{I}}$ and $\underline{\mathcal{O}}$. For the problem of HC for example, the input class is the set of finite metric spaces and the output class is that of dendrograms. An algorithm is to be regarded as a functor between a category of input objects and a category of output objects.

An algorithm will therefore be a procedure that assigns to each $I \in \underline{\mathcal{I}}$ an output $O_I \in \underline{\mathcal{O}}$ with the further property that it respects relations between objects in the following sense. Assume $I, I' \in \underline{\mathcal{I}}$ such that there is a "natural map" $f : I \to I'$. Then, the algorithm has to have the property that the relation between O_I and $O_{I'}$ has to be represented by a natural map for output objects.

Remark 5. Assume that $\underline{\mathcal{I}}$ is such that $\mathrm{Mor}_{\underline{\mathcal{I}}}(X, Y) = \emptyset$ for all $X, Y \in \underline{\mathcal{I}}$ with $X \neq Y$. In this case, since there are no morphisms between input objects any functor $\mathfrak{A} : \underline{\mathcal{I}} \to \underline{\mathcal{O}}$ can be specified arbitrarily on each $X \in \mathrm{ob}(\underline{\mathcal{O}})$. It is much more interesting and arguably more useful to consider categories with non-empty morphism sets.

More precisely. We view any given clustering scheme as a procedure which takes as input a finite metric space (X, d_X), and delivers as output either an object in $\underline{\mathcal{C}}$ or $\underline{\mathcal{P}}$:

- **Standard clustering**: a pair (X, P_X) where P_X is a partition of X. Such a pair is an object in the category $\underline{\mathcal{C}}$.
- **Hierarchical clustering**: a pair (X, θ_X) where θ_X is a persistent set over X. Such a pair is an object in the category $\underline{\mathcal{P}}$.

The concept of **functoriality** refers to the additional condition that the clustering procedure should map a pair of input objects into a pair of output objects in a manner which is consistent with respect to the morphisms attached to the input and output spaces. When this happens, we say that the clustering scheme is **functorial**. This notion of consistency is made precise in Definition 9 and described by diagram (6). Let $\underline{\mathcal{M}}$ stand for any of $\underline{\mathcal{M}}^{gen}$, $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{iso}$.

According to Definition 9, in order to view a standard clustering scheme as a functor $\mathfrak{C} : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$ we need to specify:

(1) how it maps objects of $\underline{\mathcal{M}}$ (finite metric spaces) into objects of $\underline{\mathcal{C}}$, and
(2) how a morphism $f : (X, d_X) \to (Y, d_Y)$ between two objects (X, d_X) and (Y, d_Y) in the input category $\underline{\mathcal{M}}$ induces a map in the output category $\underline{\mathcal{C}}$, see diagram (6).

$$\begin{tikzcd}
(X,d_X) \arrow[r, "f"] \arrow[d, "\mathfrak{C}"] & (Y,d_Y) \arrow[d, "\mathfrak{C}"] \\
(X,P_X) \arrow[r, "\mathfrak{C}(f)"] & (Y,P_Y)
\end{tikzcd} \tag{5}$$

Similarly, in order to view a hierarchical clustering scheme as a functor $\mathfrak{H} : \underline{\mathcal{M}} \to \underline{\mathcal{P}}$ we need to specify:

(1) how it maps objects of $\underline{\mathcal{M}}$ (finite metric spaces) into objects of $\underline{\mathcal{P}}$, and
(2) how a morphism $f : (X, d_X) \to (Y, d_Y)$ between two objects (X, d_X) and (Y, d_Y) in the input category $\underline{\mathcal{M}}$ induces a map in the output category $\underline{\mathcal{P}}$, see diagram (6).

$$\begin{tikzcd}
(X,d_X) \arrow[r, "f"] \arrow[d, "\mathfrak{H}"] & (Y,d_Y) \arrow[d, "\mathfrak{H}"] \\
(X,\theta_X) \arrow[r, "\mathfrak{H}(f)"] & (Y,\theta_Y)
\end{tikzcd} \tag{6}$$

Precise constructions will be discussed ahead.

We have 3 possible "input" categories ordered by inclusion (4). The idea is that studying functoriality over a larger category will be more stringent/demanding than requiring functoriality over a smaller one. We will consider different clustering algorithms and study whether they are functorial over our choice of the input category. The least demanding one, $\underline{\mathcal{M}}^{iso}$ basically enforces that clustering schemes are not dependent on the way points are labeled.

We will describe uniqueness results for functoriality over the most stringent category $\underline{\mathcal{M}}^{gen}$, and also explain how relaxing the conditions imposed by the morphisms in $\underline{\mathcal{M}}^{gen}$, namely, by restricting ourselves to the smaller but intermediate category $\underline{\mathcal{M}}^{inj}$, one allows more functorial clustering algorithms.

5.3 Results for Standard Clustering

Let (X, d_X) be a finite metric space. For each $r \geq 0$ we define the equivalence relation $\sim_r$ on X given by $x \sim_r x'$ if and only if there exist $x_0, x_1, \ldots, x_k \in X$ with $x = x_0$, $x' = x_k$ and $d_X(x_i, x_{i+1}) \leq r$ for all $i = 0, 1, \ldots, k-1$.

Definition 10. *For each $\delta > 0$ we define the* ***Vietoris-Rips clustering functor***

$$\mathfrak{R}_\delta : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{C}}$$

as follows. For a finite metric space (X, d_X), we set $\mathfrak{R}_\delta(X, d_X)$ to be $(X, P_X(\delta))$, where $P_X(\delta)$ is the partition of X associated to the equivalence relation $\sim_\delta$. We define how $\mathfrak{R}_\delta$ acts on maps $f : (X, d_X) \to (Y, d_Y)$: $\mathfrak{R}_\delta(f)$ is simply the set map f regarded as a morphism from $(X, P_X(\delta))$ to $(Y, P_Y(\delta))$ in $\underline{\mathcal{C}}$.

The Vietoris-Rips functor is actually just **single linkage clustering** as it is well known, see [15,18].

By restricting $\mathfrak{R}_\delta$ to the subcategories $\underline{\mathcal{M}}^{iso}$ and $\underline{\mathcal{M}}^{inj}$, we obtain functors $\mathfrak{R}_\delta^{iso} : \underline{\mathcal{M}}^{iso} \to \underline{\mathcal{C}}$ and $\mathfrak{R}_\delta^{inj} : \underline{\mathcal{M}}^{inj} \to \underline{\mathcal{C}}$. We will denote all these functors by $\mathfrak{R}_\delta$ when there is no ambiguity.

It can be seen [19] that the Vietoris-Rips functor is surjective: Among the desirable conditions singled out by Kleinberg [51], one has that of *surjectivity* (which he referred to as "richness"). Given a finite set X and $P_X \in \mathcal{P}(X)$, surjectivity calls for the existence of a metric d_X on X such that $\mathfrak{R}_\delta(X, d_X) = (X, P_X)$.

For $\underline{\mathcal{M}}$ being any one of our choices $\underline{\mathcal{M}}^{iso}$, $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{gen}$, a clustering functor in this context will be denoted by $\mathfrak{C} : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$. **Excisiveness** of a clustering functor refers to the property that once a finite metric space has been partitioned by the clustering procedure, it should not be further split by subsequent applications of the same algorithm.

Definition 11 (Excisive clustering functors). *We say that a clustering functor $\mathfrak{C}$ is* ***excisive*** *if for all $(X, d_X) \in \mathrm{ob}(\underline{\mathcal{M}})$, if we write $\mathfrak{C}(X, d_X) = (X, \{X_\alpha\}_{\alpha \in A})$, then*

$$\mathfrak{C}\left(X_\alpha, d_{X|_{X_\alpha \times X_\alpha}}\right) = (X_\alpha, \{X_\alpha\}) \textit{ for all } \alpha \in A.$$

It can be seen that the Vietoris-Rips functor is excisive.

However, there exist non-excisive clustering functors in $\underline{\mathcal{M}}^{inj}$.

*Example 3 (***A non-excisive functor in $\underline{\mathcal{M}}^{inj}$***).* For each finite metric space X let $\eta_X := (\mathbf{sep}\,(X))^{-1}$. Consider the clustering functor $\widehat{\mathfrak{R}} : \underline{\mathcal{M}}^{inj} \to \underline{\mathcal{C}}$ defined as follows: for a finite metric space (X, d_X), we define $\widehat{\mathfrak{R}}(X, d_X)$ to be $(X, \widehat{P}_X)$, where $\widehat{P}_X$ is the partition of X associated to the equivalence relation $\sim_{\eta_X}$ on X. That $\widehat{\mathfrak{R}}$ is a functor follows from the fact that whenever $\phi \in \mathrm{Mor}_{\underline{\mathcal{M}}^{inj}}(X, Y)$ and $x \sim_{\eta_X} x'$, then $\phi(x) \sim_{\eta_Y} \phi(x')$.

Now, the functor $\widehat{\mathfrak{R}}$ is **not excisive** in general. An explicit example is the following: Consider the metric space (X, d_X) depicted in Figure 13, where the metric is given by the graph metric on the underlying graph. Note that $\mathbf{sep}\,(X) = 1/2$ and thus $\eta_X = 2$. We then find that $\widehat{\mathfrak{R}}(X, d_X) = (X, \{\{A, B, C\}, \{D, E\}\})$. Let $(Y, d_Y) = \left(\{A, B, C\}, \left(\begin{smallmatrix} 0 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{smallmatrix}\right)\right)$. Then, $\mathbf{sep}\,(Y) = 1$ and hence $\eta_Y = 1$. Therefore,

$$\widehat{\mathfrak{R}}\left(\{A, B, C\}, \left(\begin{smallmatrix} 0 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & 0 \end{smallmatrix}\right)\right) = (\{A, B, C\}, \{A, \{B, C\}\}),$$

and we see that $\{A, B, C\}$ gets further partitioned by $\widehat{\mathfrak{R}}$.

It is interesting to point out that the similar constructions of a non-excisive functor in $\underline{\mathcal{M}}^{gen}$ would not work, see [19].

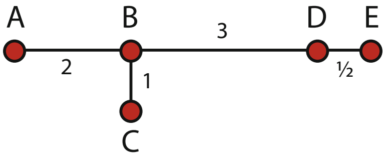


Fig. 13. Metric space used to prove that the functor $\widehat{\mathfrak{R}} : \underline{\mathcal{M}}^{inj} \to \underline{\mathcal{C}}$ is not excisive. The metric is given by the graph distance on the graph.

The case of $\underline{\mathcal{M}}^{iso}$ One can easily describe all $\underline{\mathcal{M}}^{iso}$-functorial clustering schemes. Let $\mathcal{I}$ denote the collection of all isometry classes of finite metric spaces. For each $\zeta \in \mathcal{I}$ let (X_ζ, d_{X_ζ}) denote an element of the class ζ, G_ζ the isometry group of (X_ζ, d_{X_ζ}), and Ξ_ζ the set of all fixed points of the action of G_ζ on $\mathcal{P}(X_\zeta)$.

Theorem 9 (Classification of $\underline{\mathcal{M}}^{iso}$-functorial clustering schemes, [19]). *Any $\underline{\mathcal{M}}^{iso}$-functorial clustering scheme determines a choice of $p_\zeta \in \Xi_\zeta$ for each $\zeta \in \mathcal{I}$, and conversely, a choice of p_ζ for each $\zeta \in \mathcal{I}$ determines an $\underline{\mathcal{M}}^{iso}$-functorial scheme.*

Representable Clustering Functors. In what follows, $\underline{\mathcal{M}}$ is either of $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{gen}$. For each $\delta > 0$ the Vietoris-Rips functor $\mathfrak{R}_\delta : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$ can be described in an alternative way. A first trivial observation is that the condition that $x, x' \in X$ satisfy $d_X(x, x') \leq \delta$ is equivalent to requiring the existence of a map $f \in \mathrm{Mor}_{\underline{\mathcal{M}}}(\Delta_2(\delta), X)$ with $\{x, x'\} \subset \mathrm{Im}(f)$. Using this, we can reformulate the condition that $x \sim_\delta x'$ by the requirement that there exist $z_0, z_1, \ldots, z_k \in X$ with $z_0 = x$, $z_k = x'$, and $f_1, f_2, \ldots, f_k \in \mathrm{Mor}_{\underline{\mathcal{M}}}(\Delta_2(\delta), X)$ with $\{x_{i-1}, x_i\} \subset \mathrm{Im}(f_i)$ $\forall i = 1, 2, \ldots, k$. Informally, this points to the interpretation that $\{\Delta_2(\delta)\}$ is the "parameter" in a "generative model" for $\mathfrak{R}_\delta$.

This suggests considering more general clustering functors constructed in the following manner. Let Ω be any fixed collection of finite metric spaces. Define a clustering functor

$$\mathfrak{C}^\Omega : \underline{\mathcal{M}} \to \underline{\mathcal{C}}$$

as follows: let $(X, d) \in \mathrm{ob}(\underline{\mathcal{M}})$ and write $\mathfrak{C}^\Omega(X, d) = (X, \{X_\alpha\}_{\alpha \in A})$. One declares that points x and x' belong to the same block X_α if and only if there exist

- a sequence of points $z_0, \ldots, z_k \in X$ with $z_0 = x$ and $z_k = x'$,
- a sequence of metric spaces $\omega_1, \ldots, \omega_k \in \Omega$ and
- for each $i = 1, \ldots, k$, pairs of points $(\alpha_i, \beta_i) \in \omega_i$ and morphisms $f_i \in \mathrm{Mor}_{\underline{\mathcal{M}}}(w_i, X)$ s.t. $f_i(\alpha_i) = z_{i-1}$ and $f_i(\beta_i) = z_i$.

Also, we declare that $\mathfrak{C}^\Omega(f) = f$ on morphisms f. Notice that above one can assume that $z_0, z_1, \ldots, z_k$ all belong to X_α.

Definition 12. *We say that a clustering functor $\mathfrak{C}$ is **representable** whenever there exists a collection of finite metric spaces Ω such that $\mathfrak{C} = \mathfrak{C}^\Omega$. In this case, we say that $\mathfrak{C}$ is **represented** by Ω. We say that $\mathfrak{C}$ is **finitely representable** whenever $\mathfrak{C} = \mathfrak{C}^\Omega$ for some* finite *collection of finite metric spaces Ω.*

As we saw above, the Vietoris-Rips functor $\mathfrak{R}_\delta$ is (finitely) represented by $\{\Delta_2(\delta)\}$.

Representability and excisiveness. Notice that excisiveness is an axiomatic statement whereas representability asserts existence of generative model for the clustering functor, and interestingly they are equivalent.

Theorem 10 ([19]). *Let $\underline{\mathcal{M}}$ be either of $\underline{\mathcal{M}}^{inj}$ or $\underline{\mathcal{M}}^{gen}$. Then any clustering functor on $\underline{\mathcal{M}}$ is excisive if and only if it is representable.*

A factorization theorem. For a given collection Ω of finite metric spaces let

$$\mathfrak{T}^{\Omega} : \underline{\mathcal{M}} \to \underline{\mathcal{M}} \tag{7}$$

be the endofunctor that assigns to each finite metric space (X, d_X) the metric space (X, d_X^{Ω}) with the same underlying set and metric d_X^{Ω} given by the maximal metric bounded above by W_X^{Ω}, where $W_X^{\Omega} : X \times X \to \mathbb{R}_+$ is given by

$$(x, x') \mapsto \inf\big\{\lambda > 0 |\, \exists w \in \Omega \,\text{and}\, \phi \in \mathrm{Mor}_{\underline{\mathcal{M}}}(\lambda \cdot \omega, X) \,\text{with}\, \{x, x'\} \subset \mathrm{Im}(\phi)\big\}, \tag{8}$$

for $x \neq x'$, and by 0 on $\mathrm{diag}(X \times X)$. Above we assume that the inf over the empty set equals $+\infty$. Note that $W_X^{\Omega}(x, x') < \infty$ for all $x, x' \in X$ as long as $|\omega| \leq |X|$ for some $\omega \in \Omega$. Also, $W_X^{\Omega}(x, x') = \infty$ for all $x \neq x'$ when $|X| < \inf\{|\omega|, \, \omega \in \Omega\}$.

Theorem 11 ([19]). *Let $\underline{\mathcal{M}}$ be either $\underline{\mathcal{M}}^{gen}$ or $\underline{\mathcal{M}}^{inj}$ and $\mathfrak{C}$ be any $\underline{\mathcal{M}}$-functorial finitely representable clustering functor represented by some $\Omega \subset \mathcal{M}$. Then, $\mathfrak{C} = \mathfrak{R}_1 \circ \mathfrak{T}^{\Omega}$.*

This theorem implies that all finitely representable clustering functors in $\underline{\mathcal{M}}^{gen}$ and $\underline{\mathcal{M}}^{inj}$ arise as the composition of the Vietoris-Rips functor with a functor that changes the metric.

A Uniqueness theorem for $\underline{\mathcal{M}}^{gen}$. In $\underline{\mathcal{M}}^{gen}$ clustering functors are very restricted, as reflected by the following theorem.

Theorem 12 ([19]). *Assume that $\mathfrak{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{C}}$ is a clustering functor for which there exists $\delta_{\mathfrak{C}} > 0$ with the property that*

- *$\mathfrak{C}(\Delta_2(\delta))$ is in one piece for all $\delta \in [0, \delta_{\mathfrak{C}}]$, and*
- *$\mathfrak{C}(\Delta_2(\delta))$ is in two pieces for all $\delta > \delta_{\mathfrak{C}}$.*

Then, $\mathfrak{C}$ is the Vietoris-Rips functor with parameter $\delta_{\mathfrak{C}}$. i.e. $\mathfrak{C} = \mathfrak{R}_{\delta_{\mathfrak{C}}}$.

Recall that the Vietoris-Rips functor is excisive.

Scale invariance in $\underline{\mathcal{M}}^{gen}$ and $\underline{\mathcal{M}}^{inj}$. It is interesting to consider the effect of imposing Kleinberg's scale invariance axiom on $\underline{\mathcal{M}}^{gen}$-functorial and $\underline{\mathcal{M}}^{inj}$-functorial clustering schemes. It turns out that in $\underline{\mathcal{M}}^{gen}$ there are only two possible clustering schemes enjoying scale invariance, which turn out to be the trivial ones:

Theorem 13 ([19]). *Let $\mathfrak{C} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{C}}$ be a clustering functor s.t. $\mathfrak{C} \circ \sigma_\lambda = \mathfrak{C}$ for all $\lambda > 0$. Then, either*

- *$\mathfrak{C}$ assigns to each finite metric space X the partition of X into singletons, or*
- *$\mathfrak{C}$ assigns to each finite metric the partition with only one block.*

By refining the proof of the previous theorem, we find that the behavior of any $\underline{\mathcal{M}}^{inj}$-functorial clustering functor is also severely restricted [19].

5.4 Results for Hierarchical Clustering

*Example 4 (**A hierarchical version of the Vietoris-Rips functor**).* We define a functor

$$\mathfrak{R} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{P}}$$

as follows. For a finite metric space (X, d_X), we define (X, d_X) to be the persistent set $(X, \theta_X^{\mathrm{VR}})$, where $\theta_X^{\mathrm{VR}}(r)$ is the partition associated to the equivalence relation $\sim_r$. This is clearly an object in $\underline{\mathcal{P}}$. We also define how $\mathfrak{R}$ acts on maps $f : (X, d_X) \to (Y, d_Y)$: The value of $\mathfrak{R}(f)$ is simply the set map f regarded as a morphism from $(X, \theta_X^{\mathrm{VR}})$ to $(Y, \theta_Y^{\mathrm{VR}})$ in $\underline{\mathcal{P}}$. That it is a morphism in $\underline{\mathcal{P}}$ is easy to check.

Clearly, this functor implements the hierarchical version of single linkage clustering in the sense that for each $\delta \geq 0$, if one writes $\mathfrak{R}_\delta(X, d_X) = (X, P_X(\delta))$, then $P_X(\delta) = \theta_X^{\mathrm{VR}}(\delta)$.

Functoriality over $\underline{\mathcal{M}}^{gen}$: A uniqueness theorem. We have a theorem of the same flavor as the main theorem of [51], except that one obtains existence and uniqueness on $\underline{\mathcal{M}}^{gen}$ instead of impossibility in our context.

Theorem 14 ([15]). *Let $\mathfrak{H} : \underline{\mathcal{M}}^{gen} \to \underline{\mathcal{P}}$ be a hierarchical clustering functor which satisfies the following conditions.*

(I) *Let $\alpha : \underline{\mathcal{M}}^{gen} \to \underline{Sets}$ and $\beta : \underline{\mathcal{P}} \to \underline{Sets}$ be the forgetful functors $(X, d_X) \to X$ and $(X, \theta_X) \to X$, which forget the metric and persistent set respectively, and only "remember" the underlying sets X. Then we assume that $\beta \circ \mathfrak{H} = \alpha$. This means that the underlying set of the persistent set associated to a metric space is just the underlying set of the metric space.*

(II) *For $\delta \geq 0$ let $\Delta_2(\delta) = (\{p, q\}, \left(\begin{smallmatrix} 0 & \delta \\ \delta & 0 \end{smallmatrix}\right))$ denote the two point metric space with underlying set $\{p, q\}$, and where dist$(p, q) = \delta$. Then $\mathfrak{H}(\Delta_2(\delta))$ is the persistent set $(\{p, q\}, \theta_{\Delta_2(\delta)})$ whose underlying set is $\{p, q\}$ and where $\theta_{\Delta_2(\delta)}(t)$ is the partition with one element blocks when $t < \delta$ and is the partition with a single two point block when $t \geq \delta$.*

(III) *Write $\mathfrak{H}(X, d_X) = (X, \theta^{\mathfrak{H}})$, then for any $t <$ **sep** (X), the partition $\theta^{\mathfrak{H}}(t)$ is the discrete partition with one element blocks.*

Then $\mathfrak{H}$ is equal to the functor $\mathfrak{R}$.

Extensions. There are extensions of the ideas described in previous sections that induce functorial clustering algorithms that are more sensitive to density, see [16,19].

6 Discussion

Imposing metric and or category structures on collections of datasets is useful. Doing this enables organizing the landscape composed by several algorithms commonly used in data analysis. With this in mind is possible to reason about the well posedness of some of these algorithms, and furthermore, one is able to infer new algorithms for solving data and shape analysis tasks.

Acknowledgements. This work was supported by DARPA grant number HR0011-05-1-0007 and ONR grant number N00014-09-1-0783.

References

1. Ankerst, M., Kastenmüller, G., Kriegel, H.-P., Seidl, T.: 3d shape histograms for similarity search and classification in spatial databases. In: Güting, R.H., Papadias, D., Lochovsky, F.H. (eds.) SSD 1999. LNCS, vol. 1651, pp. 207–226. Springer, Heidelberg (1999)
2. Asimov, D.: The grand tour: a tool for viewing multidimensional data. SIAM J. Sci. Stat. Comput. 6, 128–143 (1985)
3. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24(4), 509–522 (2002)
4. Berchtold, S.: Geometry-based Search of Similar Parts. PhD thesis. University of Munich, Germany (1998)
5. Boutin, M., Kemper, G.: On reconstructing n-point configurations from the distribution of distances or areas. Adv. in Appl. Math. 32(4), 709–735 (2004)
6. Bowman, G.R., Huang, X., Yao, Y., Sun, J., Carlsson, G., Guibas, L.J., Pande, V.S.: Structural insight into rna hairpin folding intermediates. Journal of the American Chemical Society (2008)
7. Brinkman, D., Olver, P.J.: Invariant histograms. University of Minnesota. Preprint (2010)
8. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Topology-invariant similarity of nonrigid shapes. Intl. Journal of Computer Vision (IJCV) 81(3), 281–301 (2009)
9. Bronstein, A.M., Bronstein, M.M., Kimmel, R., Mahmoudi, M., Sapiro, G.: A gromov-hausdorff framework with diffusion geometry for topologically-robust nonrigid shape matching (Submitted)
10. Bronstein, A., Bronstein, M., Bruckstein, A., Kimmel, R.: Partial similarity of objects, or how to compare a centaur to a horse. International Journal of Computer Vision
11. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Efficient computation of isometry-invariant distances between surfaces. SIAM Journal on Scientific Computing 28(5), 1812–1836 (2006)
12. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Calculus of nonrigid surfaces for geometry and texture manipulation. IEEE Trans. Vis. Comput. Graph. 13(5), 902–913 (2007)

13. Burago, D., Burago, Y., Ivanov, S.: A Course in Metric Geometry. AMS Graduate Studies in Math, vol. 33. American Mathematical Society, Providence (2001)
14. Bustos, B., Keim, D.A., Saupe, D., Schreck, T., Vranić, D.V.: Feature-based similarity search in 3d object databases. ACM Comput. Surv. 37(4), 345–387 (2005)
15. Carlsson, G., Mémoli, F.: Persistent Clustering and a Theorem of J. Kleinberg. ArXiv e-prints (August 2008)
16. Carlsson, G., Mémoli, F.: Multiparameter clustering methods. Technical report, technical report (2009)
17. Carlsson, G.: Topology and data. Bull. Amer. Math. Soc. 46, 255–308 (2009)
18. Carlsson, G., Mémoli, F.: Characterization, stability and convergence of hierarchical clustering methods. Journal of Machine Learning Research 11, 1425–1470 (2010)
19. Carlsson, G., Mémoli, F.: Classifying clustering schemes. CoRR, abs/1011.5270 (2010)
20. Chazal, F., Cohen-Steiner, D., Guibas, L., Mémoli, F., Oudot, S.: Gromov-Hausdorff stable signatures for shapes using persistence. In: Proc. of SGP (2009)
21. Clarenz, U., Rumpf, M., Telea, A.: Robust feature detection and local classification for surfaces based on moment analysis. IEEE Transactions on Visualization and Computer Graphics 10 (2004)
22. Coifman, R.R., Lafon, S.: Diffusion maps. Applied and Computational Harmonic Analysis 21(1), 5–30 (2006)
23. Cox, T.F., Cox, M.A.A.: Multidimensional scaling. Monographs on Statistics and Applied Probability, vol. 59. Chapman & Hall, London (1994) With 1 IBM-PC floppy disk (3.5 inch, HD)
24. d'Amico, M., Frosini, P., Landi, C.: Natural pseudo-distance and optimal matching between reduced size functions. Technical Report 66, DISMI, Univ. degli Studi di Modena e Reggio Emilia, Italy (2005)
25. d'Amico, M., Frosini, P., Landi, C.: Using matching distance in size theory: A survey. IJIST 16(5), 154–161 (2006)
26. Davies, E.B.: Heat kernels in one dimension. Quart. J. Math. Oxford Ser. (2) 44(175), 283–299 (1993)
27. Edelsbrunner, H., Harer, J.: Computational Topology - an Introduction. American Mathematical Society, Providence (2010)
28. Elad (Elbaz), A., Kimmel, R.: On bending invariant signatures for surfaces. IEEE Trans. Pattern Anal. Mach. Intell. 25(10), 1285–1295 (2003)
29. Frosini, P.: A distance for similarity classes of submanifolds of Euclidean space. Bull. Austral. Math. Soc. 42(3), 407–416 (1990)
30. Frosini, P.: Omotopie e invarianti metrici per sottovarieta di spazi euclidei (teoria della taglia). PhD thesis. University of Florence, Italy (1990)
31. Frosini, P., Mulazzani, M.: Size homotopy groups for computation of natural size distances. Bull. Belg. Math. Soc. Simon Stevin 6(3), 455–464 (1999)
32. Gelfand, N., Mitra, N.J., Guibas, L.J., Pottmann, H.: Robust global registration. In: SGP 2005: Proceedings of the Third Eurographics Symposium on Geometry Processing, p. 197. Eurographics Association, Aire-la-Ville (2005)
33. Ghrist, R.: Barcodes: The persistent topology of data. Bulletin-American Mathematical Society 45(1), 61 (2008)
34. Grigorescu, C., Petkov, N.: Distance sets for shape filters and shape recognition. IEEE Transactions on Image Processing 12(10), 1274–1286 (2003)
35. Gromov, M.: Metric structures for Riemannian and non-Riemannian spaces. Progress in Mathematics, vol. 152. Birkhäuser Boston Inc., Boston (1999)

36. Ben Hamza, A., Krim, H.: Geodesic object representation and recognition. In: Nyström, I., Sanniti di Baja, G., Svensson, S. (eds.) DGCI 2003. LNCS, vol. 2886, pp. 378–387. Springer, Heidelberg (2003)
37. Hartigan, J.A.: Statistical theory in clustering. J. Classification 2(1), 63–76 (1985)
38. Hastie, T., Stuetzle, W.: Principal curves. Journal of the American Statistical Association 84(406), 502–516 (1989)
39. Hilaga, M., Shinagawa, Y., Kohmura, T., Kunii, T.L.: Topology matching for fully automatic similarity estimation of 3d shapes. In: SIGGRAPH 2001: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 203–212. ACM, New York (2001)
40. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology 233(1), 123–138 (1993)
41. Huang, Q.-X., Adams, B., Wicke, M., Guibas, L.J.: Non-rigid registration under isometric deformations. Comput. Graph. Forum 27(5), 1449–1457 (2008)
42. Huber, P.J.: Projection pursuit. The Annals of Statistics 13(2), 435–525 (1985)
43. Huttenlocher, D.P., Klanderman, G.A., Rucklidge, W.J.: Comparing images using the Hausdorff distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(9) (1993)
44. Inselberg, A.: Parallel Coordinates: Visual Multidimensional Geometry and Its Applications. Springer-Verlag New York, Inc., Secaucus (2009)
45. Ion, A., Artner, N.M., Peyre, G., Marmol, S.B.L., Kropatsch, W.G., Cohen, L.: 3d shape matching by geodesic eccentricity. In: IEEE Computer Society Conference on, Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2008, pp. 1–8 (June 2008)
46. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice Hall Advanced Reference Series. Prentice Hall Inc., Englewood Cliffs (1988)
47. Janowitz, M.F.: An order theoretic model for cluster analysis. SIAM Journal on Applied Mathematics 34(1), 55–72 (1978)
48. Jardine, N., Sibson, R.: Mathematical taxonomy. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Ltd., London (1971)
49. Johnson, A.: Spin-Images: A Representation for 3-D Surface Matching. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA (August 1997)
50. Kastenmüller, G., Kriegel, H.P., Seidl, T.: Similarity search in 3d protein databases. In: Proc. GCB (1998)
51. Kleinberg, J.M.: An impossibility theorem for clustering. In: Becker, S., Thrun, S., Obermayer, K. (eds.) NIPS, pp. 446–453. MIT Press, Cambridge (2002)
52. Koppensteiner, W.A., Lackner, P., Wiederstein, M., Sippl, M.J.: Characterization of novel proteins based on known protein structures. Journal of Molecular Biology 296(4), 1139–1152 (2000)
53. Lafon, S.: Diffusion Maps and Geometric Harmonics. PhD thesis, Yale University (2004)
54. Le, T.M., Mémoli, F.: Local scales of embedded curves and surfaces. preprint (2010)
55. Ling, H., Jacobs, D.W.: Using the inner-distance for classification of articulated shapes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 719–726 (2005)
56. Lu, C.E., Latecki, L.J., Adluru, N., Yang, X., Ling, H.: Shape guided contour grouping with particle filters. In: IEEE 12th International Conference on, Computer Vision 2009, pp. 2288–2295. IEEE, Los Alamitos (2009)
57. Lane, S.M.: Categories for the working mathematician, 2nd edn. Graduate Texts in Mathematics, vol. 5. Springer, New York (1998)

58. Manay, S., Cremers, D., Hong, B.W., Yezzi, A.J., Soatto, S.: Integral invariants for shape matching 28(10), 1602–1618 (2006)
59. Mémoli, F.: Gromov-Hausdorff distances in Euclidean spaces. In: IEEE Computer Society Conference on, Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2008, pp. 1–8 (June 2008)
60. Mémoli, F.: Gromov-wasserstein distances and the metric approach to object matching. In: Foundations of Computational Mathematics, pp. 1–71 (2011) 10.1007/s10208-011-9093-5
61. Mémoli, F.: Some properties of gromov-hausdorff distances. Technical report, Department of Mathematics. Stanford University (March 2011)
62. Mémoli, F.: A spectral notion of Gromov-Wasserstein distances and related methods. Applied and Computational Mathematics 30, 363–401 (2011)
63. Mémoli, F., Sapiro, G.: Comparing point clouds. In: SGP 2004: Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing, pp. 32–40. ACM, New York (2004)
64. Mémoli, F., Sapiro, G.: A theoretical and computational framework for isometry invariant recognition of point cloud data. Found. Comput. Math. 5(3), 313–347 (2005)
65. Nicolau, M., Levine, A.J., Carlsson, G.: Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Proceedings of the National Academy of Sciences 108(17), 7265–7270 (2011)
66. Norris, J.R.: Heat kernel asymptotics and the distance function in Lipschitz Riemannian manifolds. Acta. Math. 179(1), 79–103 (1997)
67. Olver, P.J.: Joint invariant signatures. Foundations of computational mathematics 1(1), 3–68 (2001)
68. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Shape distributions. ACM Trans. Graph. 21(4), 807–832 (2002)
69. Pottmann, H., Wallner, J., Huang, Q., Yang, Y.-L.: Integral invariants for robust geometry processing. Comput. Aided Geom. Design (2008) (to appear)
70. Raviv, D., Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Symmetries of non-rigid shapes. In: IEEE 11th International Conference on, Computer Vision, ICCV 2007, October 14-21, pp. 1–7 (2007)
71. Reeb, G.: Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numérique. C. R. Acad. Sci. Paris 222, 847–849 (1946)
72. Reuter, M., Wolter, F.-E., Peinecke, N.: Laplace-spectra as fingerprints for shape matching. In: SPM 2005: Proceedings of the 2005 ACM Symposium on Solid and Physical Modeling, pp. 101–106. ACM Press, New York (2005)
73. Reuter, M., Wolter, F.-E., Peinecke, N.: Laplace-Beltrami spectra as "Shape-DNA" of surfaces and solids. Computer-Aided Design 38(4), 342–366 (2006)
74. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290(5500), 2323–2326 (2000)
75. Ruggeri, M., Saupe, D.: Isometry-invariant matching of point set surfaces. In: Proceedings Eurographics 2008 Workshop on 3D Object Retrieval (2008)
76. Rustamov, R.M.: Laplace-beltrami eigenfunctions for deformation invariant shape representation. In: Symposium on Geometry Processing, pp. 225–233 (2007)
77. Sakai, T.: Riemannian geometry. Translations of Mathematical Monographs, vol. 149. American Mathematical Society, Providence (1996)
78. Semple, C., Steel, M.: Phylogenetics. Oxford Lecture Series in Mathematics and its Applications, vol. 24. Oxford University Press, Oxford (2003)

79. Shi, Y., Thompson, P.M., de Zubicaray, G.I., Rose, S.E., Tu, Z., Dinov, I., Toga, A.W.: Direct mapping of hippocampal surfaces with intrinsic shape context. NeuroImage 37(3), 792–807 (2007)
80. Singh, G., Mémoli, F., Carlsson, G.: Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, pp. 91–100. Eurographics Association, Prague (2007)
81. Singh, G., Memoli, F., Ishkhanov, T., Sapiro, G., Carlsson, G., Ringach, D.L.: Topological analysis of population activity in visual cortex. J. Vis. 8(8), 1–18 (2008)
82. Stuetzle, W.: Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. J. Classification 20(1), 25–47 (2003)
83. Sturm, K.-T.: On the geometry of metric measure spaces. I. Acta. Math. 196(1), 65–131 (2006)
84. Sun, J., Ovsjanikov, M., Guibas, L.: A concise and provably informative multi-scale signature based on heat diffusion. In: SGP (2009)
85. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2323 (2000)
86. Thureson, J., Carlsson, S.: Appearance based qualitative image description for object class recognition. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3022, pp. 518–529. Springer, Heidelberg (2004)
87. Tsuchida, T.: Long-time asymptotics of heat kernels for one-dimensional elliptic operators with periodic coefficients. Proc. Lond. Math. Soc (3) 97(2), 450–476 (2008)
88. Verri, A., Uras, C., Frosini, P., Ferri, M.: On the use of size functions for shape analysis. Biological cybernetics 70(2), 99–107 (1993)
89. Villani, C.: Topics in optimal transportation. Graduate Studies in Mathematics, vol. 58. American Mathematical Society, Providence (2003)
90. von Luxburg, U., Ben-David, S.: Towards a statistical theory of clustering. presented at the pascal workshop on clustering, london. Technical report, Presented at the PASCAL Workshop on Clustering, London (2005)
91. Zomorodian, A., Carlsson, G.: Computing persistent homology. In: SCG 2004: Proceedings of the Twentieth Annual Symposium on Computational Geometry, pp. 347–356. ACM, New York (2004)

Semi-fragile Watermarking in Biometric Systems: Template Self-Embedding*

Reinhard Huber[1], Herbert Stögner[1], and Andreas Uhl[1,2,**]

[1] School of CEIT, Carinthia University of Applied Sciences, Austria
[2] Department of Computer Sciences, University of Salzburg, Austria
uhl@cosy.sbg.ac.at

Abstract. Embedding biometric templates as image-dependent watermark information in semi-fragile watermark embedding is proposed. Experiments in an iris recognition environment show that the embedded templates can be used to verify sample data integrity and may serve additionally to increase robustness in the biometric recognition process.

1 Introduction

There has been a lot of work done during the last years proposing watermarking techniques to enhance biometric systems security in some way (see [4] for our recent survey on the topic). Major application scenarios include biometric watermarking (where biometric templates are embedded as "message" as opposed to classical copyright information), sample data replay prevention (by robustly watermarking once acquired sample data), covert biometric data communication (by steganographic techniques), and employing WM is a means of tightly coupled transport of sample data and embedded (template or general purpose authentication) data for multibiometric or two-factor authentication schemes, respectively.

In this work we consider the application scenario where the aim of WM is to ensure the integrity and authenticity of the sample data acquisition and transmission process. During data acquisition, the sensor (i.e. camera) embeds a watermark into the acquired sample image before transmitting it to the feature extraction module. The feature extraction module only proceeds with its tasks if the WM can be extracted correctly (which means that (a) the data has not been tampered with and (b) the origin of the data is the correct sensor).

Attack. An attacker aims at *inserting* the WM in order to mimic correctly acquired sensor data or to *manipulate* sample data without affecting the WM.

WM properties and content. The WM needs to be unique in the sense that it has to uniquely identify the sensor. Resistance against a WM insertion

* This work has been partially supported by the Austrian Science Fund, project no. L554-N15.

** Correspondig author.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 34–41, 2011.

attack can be achieved by sensor-key dependent embedding. Since the watermarking scheme has to be able to detect image manipulations, (semi-)fragile embedding techniques are the method of choice. Especially in semi-fragile watermarking it was found to be highly advantageous to embed image-dependent watermark data in order to prevent copy attacks. WM extraction should be blind.

Crypto alternative. Classical authentication protocols can be used to secure the communication between sensor and feature extraction module – a digital signature signed with the private key of the acquisition device can ensure the authenticity of the sensor and the integrity of the image data. However, this approach cannot provide robustness and no information about tampering locations is obtained.

Yeung et al. [9] propose a fragile watermarking technique to add the ability for integrity verification of the captured fingerprint images against altering during transmission or in a database. Ratha et al. [8] propose to embed a response to an authentication challenge sent out by a server into a WSQ compressed fingerprint image in order to authenticate the sensor capturing the fingerprint image. If the (fragile) watermark cannot be extracted, either the image has been tampered with or the image does not come from the correct sensing device.

Also, semi-fragile watermarking has been suggested to verify authenticity of biometric sample data. PCA features are used as embedded data in [7], while [1] proposes the embedding of robust signatures into fingerprint images.

Finally, dual WM techniques have been proposed applying two different embedding techniques concurrently. The first technique in [6] is used for checking integrity on a block level using CRC checks, the second provides reversible watermarking in case the first technique rates the sample as being authentic. Two different embedding techniques (a semi-fragile and a robust one) for embedding both, a sample image dependent signature as well as a template of a different modality are proposed by Komninos et al. [5].

In this paper we focus on protecting the transmission of sample data from the sensor to the feature extraction module employing a specific semi-fragile watermarking technique. In particular, we propose to embed biometric template data instead of general purpose watermark information which can then be used in the matching process in addition to checking integrity. In Section 2, we introduce the template-embedding based semi-fragile watermarking approach and discuss its properties. Section 3 presents experiments where the proposed concept is evaluated in the context of iris recognition using a variant of a well known watermark embedding scheme. Section 4 concludes the paper.

2 Semi-fragile Watermarking by Template Self Embedding

In the context of biometrics, we propose to embed template data as semi-fragile WM information instead of general purpose image descriptors as used in classical semi-fragile WM schemes [2]. This is sensible since on the one hand template data are of course image dependent data and therefore are able to prevent WM copy attacks or similar. On the other hand, in case of tampering or other significant data manipulations, the aim is not to reconstruct the sample data at first hand, but to be able to generate template data from the sample data required for matching. So data for reconstructing the sample data is suggested to be replaced by data for directly generating template data. In the following, we describe the WM embedding and extraction processes:

1. From the acquired sample data, a template is extracted.
2. The template is embedded into the sample data employing a semi-fragile embedding technique (this template is referred to as "template watermark" subsequently).
3. The data is sent to the feature extraction and matching module.
4. At the feature extraction module, the template watermark template is extracted, and is compared to the template extracted from the sample (denoted simply as "template" in the following). In this way, the integrity of the transmitted sample data is ensured when there is sufficient correspondence between the two templates. In case of a biometric system operating in verification mode the template watermark can also be compared to the template in the database corresponding to the claimed identity (denoted "database template" in the following).
5. Finally, in case the integrity of the data has been proven, the watermark template and the template are used in the matching process, granting access if the similarity to the database template(s) is high enough.

When comparing this approach to previous techniques proposed in literature, we notice the following differences / advantages: As opposed to techniques employing robust template embedding watermarking (e.g. as proposed for enabling tightly coupled transport of sample and template data of different modalities), the proposed scheme can ensure sample data integrity. The importance of this property has been recently demonstrated [3] in an attack against robust embedding schemes used in the multibiometric and two-factor authentication scenarios. As opposed to techniques employing arbitrary (semi-)fragile watermarks for integrity protection (instead of the template watermark used here), the template watermark data can be used to provide a more robust matching process after data integrity has been assured.

However, some issues need to be investigated with respect to the proposed scheme (which will be done in the experiments):

- Does integrity verification indeed work in a robust manner ?
- What is the impact of the embedded template watermark on the recognition performance using the template for matching only ?
- Can a combination of template watermark and template result in more robustness in an actual matching process ?

3 Experiments in the Case of Iris Recognition

3.1 Iris Recognition and Iris Databases

The employed iris recognition system is Libor Masek's Matlab implementation[1] of a 1-D version of the Daugman iris recognition algorithm. First, this algorithm segments the eye image into the iris and the remainder of the image. Iris image texture is mapped to polar coordinates resulting in a rectangular patch which is denoted "polar image". For feature extraction, a row-wise convolution with a complex Log-Gabor filter is performed on the polar image pixels. The phase angle of the resulting complex value for each pixel is discretized into 2 bits. The 2 bit of phase information are used to generate a binary code. After extracting the features of the iris, considering translation, rotations, and disturbed regions in the iris (a noise mask is generated), the algorithm outputs the similarity score by giving the Hamming distance between two extracted templates.

The following three datasets are used in the experiments:

CASIAv3 Interval database[2] consists of 2639 images with 320 × 280 pixels in 8 bit grayscale .jpeg format, out of which 500 images have been used in the experiments.

MMU database[3] consists of 450 images with 320 × 240 pixels in 24 bit grayscale .bmp format, all images have been used in the experiments.

UBIRIS database[4] consists of 1876 images with 200 × 150 pixels in 24 bit colour .jpeg format, out of which 318 images have been used in the experiments.

3.2 The Watermarking Scheme

As the baseline system, we employ the fragile watermarking scheme as developed by Yeung et. al and investigated in the context of fingerprint recognition [9]. For this algorithm, the watermark embedded is binary and padded to the size of the host image. Subsequently, the WM is embedded into each pixel according to some key information. As a consequence, the WM capacity is 89600, 76800, and 30000 bits for CASIAv3, MMU, and UBIRIS, respectively.

Since this technique is a fragile WM scheme, no robustness against any image manipulations can be expected of course. However, the usually smaller size of biometric templates can be exploited to embed the template in redundant manner, i.e. we embed the template several times. After the extraction process, all template watermarks are used in a majority voting scheme which constructs a "master" template watermark. We expect to result in higher robustness as compared to the original algorithm due to redundant embedding leading to an overall quasi semi-fragile WM scheme for the watermark templates. In our implementation, the iris code consists of 9600 bits, therefore, we can embed 9, 8,

[1] http://www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html

[2] http://www.cbsr.ia.ac.cn/IrisDatabase.htm/

[3] http://pesona.mmu.edu.my/~ccteo/

[4] http://www.di.ubi.pt/~hugomcp/investigacao.htm

and 3 templates into images from the CASIAv3, MMU, and UBIRIS databases, respectively.

Note that instead of this embedding scheme, any semi-fragile WM scheme [2] with sufficient capacity to embed template information can be employed.

3.3 Experimental Results

As first topic, we investigate integrity verification under conditions which require robustness properties. As "attacks" against the sample data with embedded WM, we consider mean filtering, noise addition, and JPEG compression. As a first scenario S1 (restricted to the verification scenario), comparison between extracted template WM and database (DB) template is covered. We consider the case that 5 different templates are stored in the database out of which a single database template is generated by majority coding like explained before in the case of the template WM. Table 1 (left) shows the bit error rate (BER) for the different attacks considered. The second scenario S2 is the comparison between extracted template WM and the template extracted from the watermarked sample data the results of which are shown in Table 1 (right).

Table 1. BER for seven different attacks

	DB template vs. template			template WM vs. template		
Attack	CASIAv3	MMU	UBIRIS	CASIAv3	MMU	UBIRIS
No attack	0.21	0.23	0.19	0.14	0.06	0.07
Mean filtering	0.49	0.50	0.50	0.49	0.50	0.50
Gaussian Noise $N = 0.0005$	0.21	0.23	0.19	0.14	0.06	0.07
Gaussian Noise $N = 0.001$	0.21	0.23	0.19	0.14	0.06	0.07
JPEG Q100	0.21	0.23	0.19	0.14	0.06	0.08
JPEG Q99	0.21	0.24	0.22	0.14	0.07	0.11
JPEG Q98	0.25	0.30	0.32	0.20	0.18	0.26
JPEG Q95	0.41	0.45	0.45	0.39	0.41	0.44

The first thing to note is that even without attack, BER is clearly above zero. For S2 this effect is solely due to the influence the embedded WM has on the extracted template - obviously the WM changes the sample in a way that about 10% of the bits are altered. For S1 the differences are higher which is clear since the DB template is constructed from several distinct templates. We have to consider that a typical decision threshold value for the iris recognition system in use is at a BER in $[0.3, 0.35]$. When taking this into account, the extent of template similarity is of course enough to decide on proven sample integrity. For both S1 and S2, adding noise and applying JPEG compression with quality set to 100 (Q100) does not change the BER. When decreasing JPEG quality to 98, BER starts to increase slightly. The situation changes drastically when applying JPEG Q95 and mean filtering: BER is up to 0.4 - 0.5 which means that integrity cannot be verified successfully. We realize that integrity verification in our technique is indeed robust against moderate JPEG compression and noise. On the other hand, mean filtering and JPEG compression at quality 95% destroys the template WM and indicates modification. The distribution of incorrect bits can be used to differentiate between malicious attacks (where an accumulation of incorrect bits can be observed in certain regions) and significant global distortions like compression where incorrect bits are spread across the entire data.

S1 and S2 can be combined into a single integrity verification scheme. The idea is to combine the single templates extracted from the watermark and the template extracted from the watermarked sample into a weighted "fused template": in our example, we use 4 copies of the template and the embedded number of templates from the template WM in a majority voting scheme to generate the fused template. Table 2 shows the corresponding BER when comparing the fused template to the DB template.

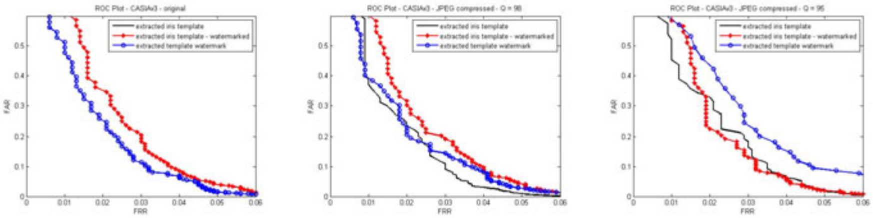
Table 2. BER for the fused template under seven different attacks

Attack	CASIAv3	MMU	UBIRIS
No attack	0.21	0.21	0.21
Mean filtering	0.30	0.27	0.21
Gaussian Noise $N = 0.0005$	0.21	0.21	0.21
Gaussian Noise $N = 0.001$	0.21	0.21	0.21
JPEG Q100	0.21	0.21	0.21
JPEG Q99	0.21	0.21	0.21
JPEG Q98	0.23	0.23	0.21
JPEG Q95	0.27	0.26	0.21

It can be clearly seen that while the BER without attack and applying moderate attacks is higher as compared to S2, we get much better robustness against JPEG Q95 and even mean filtering. With the fusing strategy, robustness even against those two types of attacks can be obtained. Of course, the fusion scheme does only make sense in a biometric system in verification mode, since integrity verification is done against templates stored in the template database.

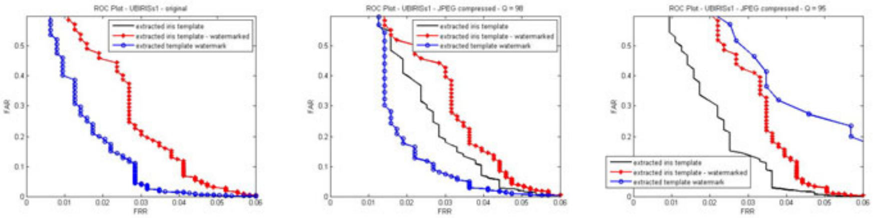
As a second topic, we investigate iris recognition performance using the template extracted from the watermarked sample (W1) and the extracted template WM (W2), and compare the behavior to the "original" results using templates extracted from the original sample data (without embedded WM, W0). For this purpose, we compare ROC curves of the three cases with and without attacks (i.e. JPEG compression, noise insertion, and mean filtering) conducted against the sample data.

In both Figs. 1.a and 2.a the curve W0 is hidden by W2 and we clearly note that the embedded WM impacts on recognition performance since W1 shows clearly inferior ROC (note that this contrasts to the case of fingerprint matching reported in [9]). So without attack, using the template WM is beneficial over the template. This situation is also typical for moderate attacks being conducted as shown in Figs. 1.b and 2.b as an example for the case of JPEG compression with Q98. While for the CASIAv3 data W0 and W2 are close, both being superior to



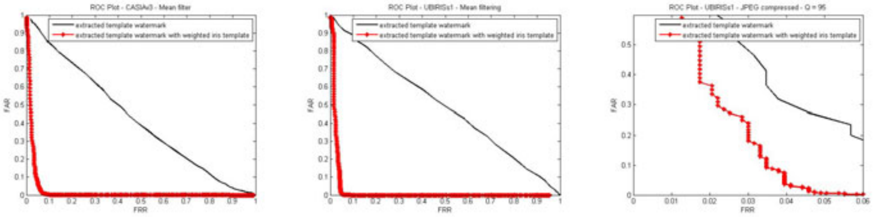
(a) no attack (b) JPEG Q98 (c) JPEG Q95

Fig. 1. ROC curves of the CASIAv3 data



(a) no attack (b) JPEG Q98 (c) JPEG Q95

Fig. 2. ROC curves of the UBIRIS data



(a) CASIAv3, mean (b) UBIRIS, mean (c) UBIRIS, Q95

Fig. 3. ROC curves for fused templates

W1, for the UBIRIS data W2 is the best option. W0 is clearly inferior to W2, while W1 is the worst option. Obviously, the embedded template watermark is not yet severely impacted by the compression artifacts.

The situation changes when the attacks get more severe. As shown in Figs. 1.c and 2.c, under JPEG compression with Q95 W2 is the worst option now since the robustness of the WM is not sufficient any more. While for the CASIAv3 data W0 and W1 are close (so the impact of the WM is negligible), for UBIRIS the impact of the WM is quite significant (which can be explained by the fact that the UBIRIS data is of already quite low quality without any further degradation, the additional WM complicates template extraction). For mean filtering the result for W2 is even worse as shown in Figs. 3.a and 3.b, no recognition can be performed at all with the extracted template WM after this attack.

Finally, the strategy of combining W1 and W2 into a fused template for integrity verification (results given in Table 2) can also be applied for matching. Fig. 3 shows examples where the ROC behavior of W2 can be significantly improved by using this approach. In particular, in the case of mean filtering the fused template can be used for recognition purposes as shown in Figs. 3.a and 3.b.

4 Conclusion

We have introduced the concept of embedding biometric templates as image-dependent watermark information in semi-fragile watermark embedding which serves the purpose of verifying the integrity and authenticity of the sensor - feature extraction communication. Experiments in an iris recognition environment show the feasibility of the approach and demonstrate, that the embedded templates can be used to verify integrity and may serve additionally as a means to increase robustness in the biometric recognition process.

References

[1] Ahmed, F., Moskowitz, I.S.: Composite signature based watermarking for fingerprint authentication. In: Proceedings of the ACM Workshop on Multimedia and Security (MMSEC 2005), pp. 799–802 (2005)

[2] Ekici, Ö., Sankur, B., Akcay, M.: A comparative evaluation of semi-fragile watermarking algorithms. Journal of Electronic Imaging 13(1), 209–216 (2003)

[3] Hämmerle-Uhl, J., Raab, K., Uhl, A.: Attack against robust watermarking-based multimodal biometric recognition systems. In: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N.C., Fairhurst, M.C. (eds.) BioID 2011. LNCS, vol. 6583, pp. 25–36. Springer, Heidelberg (2011)

[4] Hämmerle-Uhl, J., Raab, K., Uhl, A.: Watermarking as a means to enhance biometric systems: A critical survey. In: Ker, A., Craver, S., Filler, T. (eds.) Proceedings of the 2011 Information Hiding Conference (IH 2011), Prague, Czech Republic. LNCS. Springer, Heidelberg (to appear, 2011)

[5] Komninos, N., Dimitriou, T.: Protecting biometric templates with image watermarking techniques. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 114–123. Springer, Heidelberg (2007)

[6] Lee, H., Lim, J., Yu, S., Kim, S., Lee, S.: Biometric image authentication using watermarking. In: Proceedings of the International Joint Conference SICE-ICASE, 2006, pp. 3950–3953 (2006)

[7] Li, C., Ma, B., Wang, Y., Zhang, Z.: Protecting biometric templates using authentication watermarking. In: Qiu, G., Lam, K.M., Kiya, H., Xue, X.-Y., Kuo, C.-C.J., Lew, M.S. (eds.) PCM 2010. LNCS, vol. 6297, pp. 709–718. Springer, Heidelberg (2010)

[8] Ratha, N.K., Figueroa-Villanueva, M.A., Connell, J.H., Bolle, R.M.: A secure protocol for data hiding in compressed fingerprint images. In: Maltoni, D., Jain, A.K. (eds.) BioAW 2004. LNCS, vol. 3087, pp. 205–216. Springer, Heidelberg (2004)

[9] Yeung, M.M., Pankanti, S.: Verification watermarks on fingerprint recognition and retrieval. Journal of Electronal Imaging, Special Issue on Image Security and Digital Watermarking 9(4), 468–476 (2000)

The Weighted Landmark-Based Algorithm for Skull Identification

Jingbo Huang, Mingquan Zhou, Fuqing Duan, Qingqong Deng,
Zhongke Wu, and Yun Tian

College of Information Science and Technology, Beijing Normal University,
Beijing, 100875, P.R. China
huangjingbo@mail.bnu.edu.com, {mqzhou,fqduan}@bnu.edu.cn,
qqdeng@nlpr.ia.ac.cn, {zwu,tianyun}@bnu.edu.cn

Abstract. Computer aided craniofacial reconstruction plays an important role in criminal investigation. By comparing the 3D facial model produced by this technology with the picture database of missing persons, the identity of an unknown skull can be determined. In this paper, we propose a method to quantitatively analyze the quality of the facial landmarks for skull identification. Based on the quality analysis of landmarks, a new landmark-based algorithm, which takes fully into account the different weights of the landmarks in the recognition, is proposed. Moreover, we can select an optimal recognition subset of landmarks to boost the recognition rate according to the recognition quality of landmarks. Experiments validate the proposed method.

Keywords: Skull identification, landmark quality, 3D-2D face recognition, optimal recognition subset, Q-weighted algorithm.

1 Introduction

When an unknown skull is found at a crime scene, enforcement officials usually compare it against a gallery of facial images of missing persons in order to determine its identity. There are two kinds of methods for this. One is the craniofacial superimposition [1], which directly compares the skull against the photos, the other is to reconstruct the victim's face model by the craniofacial reconstruction technology [2, 3] and then compare the 3D face model with facial images of missing persons. Although some successful cases using the craniofacial superimposition have been reported, many researchers are still suspicious of its scientific and validity.

This paper employs the latter method. The basic flow is shown in Figure.1. The probe object is the 3D face model reconstructed from the skull, and the recognized result is the 2D facial image.

The recognition here is to compare the 3D face model against the 2D face image, and the 3D face model has only the shape information but no texture information (Figure 2). In the realm of 3D vs. 2D face recognition, Blanz. and Vetter employed a Morphable Model(3DMM) [4, 5]and G.Toderici et al. employed Annotated Face

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 42–48, 2011.

Model (AFM)[6] to compare 2D and 3D faces. However, the 3DMM and AFM both are complex to be developed and computationally expensive. Similar to 2D face recognition, several subspace-based algorithms to compare the 3D faces were also proposed. This kind of algorithms includes Canonical Correlation Analysis (CCA) based algorithm [7, 8] and Partial Principal Component Analysis [9] and so on. Those algorithms can achieve good results, but need a great number of training samples. In [10], D.Riccio et al. propose a particular 2D-3D face recognition method based on 16 geometric invariants, which are calculated from a number of control points. The main problem is the sensitiveness of the algorithm with respect to the pose variations and inaccuracy in the detection of the control points. These algorithms all have not been used in skull identification.

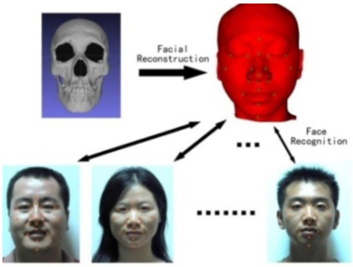


Fig. 1. The procedure of skull identification with craniofacial reconstruction

For skull identification, only Peter Tu etc. propose a landmark-based recognition algorithm [3]. The algorithm recognizes the face by extracting the landmarks in the 3D and 2D face, and then calculates the reprojection errors. This method is simple and practicable. However, it fails to take into account of the varying quality of landmarks.

In this paper, we propose a novel measurement for recognition quality of the landmarks, and based on this measurement, a new landmark-based algorithm, which takes fully into account the different weights of the landmarks in the recognition, is proposed. Moreover, we can select an optimal recognition subset of landmarks to improve the recognition rate according to the recognition quality of landmarks.

2 Landmark Definition and Analysis

To define the facial landmarks, it is necessary to meet two requirements: 1) landmarks must be obvious and easily to be labeled; 2) landmarks also need to be stable and do not vary greatly with the change of expression, age and weight.

After determining the 3D and 2D landmarks, we project the 3D landmarks to 2D image, and then calculate the distance of the projection results to their corresponding 2D landmarks. The distance can be used to measure the similarity of 3D-2D faces.

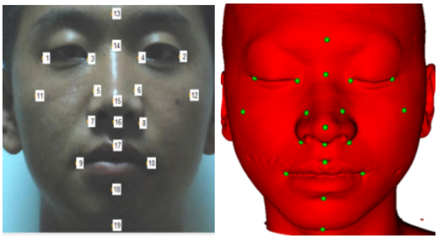


Fig. 2. Facial landmarks in 2D and 3D faces

Let $x = \left\{ (x_i, y_i, 1)^T \mid i = 1, 2, ..., n \right\}$ and $X = \left\{ (X_i, Y_i, Z_i, 1)^T \mid i = 1, 2, ..., n \right\}$ denote the 2D image landmarks and the 3D model landmarks respectively. According to the pinhole camera model, the 3D landmarks can be projected to the 2D image plane as follows:

$$\mathbf{x}_{3\times n} = \mathbf{P}_{3\times 4} * \mathbf{X}_{4\times n}, \tag{1}$$

where $\mathbf{P}_{3\times 4}$ is the camera projection matrix, which can be estimated from a group of 2D-3D correspondences by minimizing the following objective function:

$$RMS^2 = \frac{1}{n}\sum_{i=1}^{n} \| \mathbf{x}_i - \mathbf{P}\mathbf{X}_i \|_2^2 \cdot \tag{2}$$

This optimization problem can be solved by the least squares method. The optimal value of the objective function is defined as the disparity value of the 3D face and 2D image. Smaller the disparity value is, more similar the two faces are. Thus a landmark based method to recognize the 3D face model is defined.

In the following, we analyze the recognition quality of landmarks with given samples. Suppose we have m pairs of 3D-2D faces and n landmarks per face. The 3D face and the 2D face of a same person have a same index. Let $\mathbf{P}_{kl}$ be the optimal camera matrix which gives the minimal reprojection error for projecting landmarks of the *kth* 3D face model to the *lth* 2D face image. $\mathbf{X}_{ij}$ is the *jth* landmark in the *ith* 3D face model, and $\mathbf{x}_{ij}$ is the *jth* landmark in the *ith* 2D face image.

1) The reliability. The reliability of *kth* landmark can be defined as:

$$R_k = \sum_{i=1}^{m} \| \mathbf{x}_{ik} - \mathbf{P}_{ii}\mathbf{X}_{ik} \|_2^2 \quad , \quad k=1,2,...,\text{n} \tag{3}$$

A smaller *R* value means a less reprojection error of the landmark under the optimal projection, and the *kth* landmark can be located accurately and is not disturbed greatly by the change of expressions, ages and weights.

2) The discrimination. The discrimination of *kth* landmark can be calculated by the following formula:

$$D_k = \sum_{i=1}^{m} \sum_{j=1,i\neq j}^{m} \| \mathbf{x}_{jk} - \mathbf{P}_{ij}\mathbf{X}_{ik} \|_2^2 \quad , \quad k=1,2,...,\mathrm{n} \cdot \tag{4}$$

Contrary to the *R* value, the larger the *D* value is, the better the discriminative quality of the landmark is.

Through the above analysis, we can see that the quality of landmark is directly proportional to its *D* value and inversely proportional to its *R* value. Therefore, the two indications can be combined to describe the quality of landmarks as follows:

$$Q_k = \frac{D_k}{R_k} = \frac{\sum_{i=1}^{m} \sum_{j=1,i\neq j}^{m} \| \mathbf{x}_{jk} - \mathbf{P}_{ij}\mathbf{X}_{ik} \|_2^2}{\sum_{i=1}^{m} \| \mathbf{x}_{ik} - \mathbf{P}_{ii}\mathbf{X}_{ik} \|_2^2} \quad , \quad k=1,2,...,\mathrm{n} \tag{5}$$

3 A Recognition Algorithm Based on Landmark Quality Analysis

3.1 A Weighted Similarity Measurement of a 3D-2D Face Pair

The procedure of the weighted similarity measurement is as follows:

1) To calculate the R value, D value and the Q value of each landmark.
2) The landmarks of unknown 3D face are projected onto all pictures of missing people, and then to calculate the reprojection errors weighting by Q values. Specifically,

2.1) estimate the optimal projection matrix **P** of the 3D-2D face pair.

2.2) the final distance of the 3D-2D face pair can be calculated by the following formula:

$$dist = \frac{1}{n} \sum_{i=1}^{n} Q_i \| \mathbf{x}_i - \mathbf{P}\mathbf{X}_i \|_2^2 \cdot \tag{6}$$

3) From all the candidates of the face image, choose the face image of the smallest distance as the recognition result of the unknown skull.

3.2 Searching an Optimal Recognition Subset of Landmarks

Experiments have shown that the best recognition result is not achieved by using all the landmarks. In fact, selecting a part of landmarks with high quality can obtain a better recognition performance than using the all landmarks.

With above quantitative analysis of the quality of landmarks, the landmarks can be ordered by the Q values in a descending order. Let AR_i be the average rank of recognition using the top i landmarks and $ARi_0 = \max\{AR_i \,|\, i = 1..n\}$. Then the top i_0 landmarks compose the optimal recognition subset of landmarks, and ARi_0 is the best average rank of recognition.

4 Experimental Results

In the experiment, we have a total 41 3D- 2D face pairs. We get 2D landmarks by manually labeling. Firstly we compute the *R* values and the *D* values of all landmarks through the samples in the training database, and then get the *Q* values of those landmarks. The detailed results are shown in Table 2, and Figure 3 is the visualization of the results.

We compare four methods. Table 1 shows the results. The attribute of each method is also shown in the table. Less the average rank of a method is, better its

Table 1. The four methods in the experiment

Methods	Using the optimal subset of landmarks	Weighting the reprojection errors	The average rank
Method 1	N	N	12.34146
Method 2	Y	N	10.34146
Method 3	N	Y	10.41463
Method 4	Y	Y	10.02439

Table 2. The results of recognition

Rank	ID	R value	D value	Q value	AR(No weighting)	AR-Weighting
1	7	0.212852	1	1	0	0
2	4	0.148425	0.542458	0.777926	0	0
3	3	0.393937	0.104859	0.056658	23.26829268	22.90243902
4	2	0.775259	0.155122	0.04259	21.80487805	21.75609756
5	8	0.239004	0.032522	0.028964	20	19.19512195
6	1	0.383031	0.049366	0.027433	21.53658537	21.43902439
7	17	0.322997	0.037264	0.024557	18.2195122	18.87804878
8	9	0.564052	0.062562	0.023609	14.14634146	14.43902439
9	16	0.149526	0.01615	0.02299	12.43902439	13.41463415
10	15	1	0.107021	0.02278	11.04878049	12.92682927
11	12	0.606698	0.061926	0.021726	11.82926829	13.02439024
12	18	0.18197	0.018084	0.021153	11.17073171	12.87804878
13	5	0.281467	0.0274	0.020721	10.58536585	12.12195122
14	19	0.22985	0.021787	0.020176	10.87804878	11.65853659
15	10	0.53388	0.047192	0.018815	10.34146341	10.46341463
16	13	0.5345	0.035885	0.01429	11.07317073	11.12195122
17	6	0.353525	0.02368	0.014257	11.58536585	10.63414634
18	14	0.983102	0.063833	0.013821	12.36585366	10.02439024
19	11	0.69372	0.039972	0.012265	12.34146341	10.41463415

performance. It can be seen that the average rank of the method 4 is decreased about 18.78% compared with the method 1. The detailed information can be seen in Table 2. In Table 2, the landmarks were arranged by descending order of their *Q* values. Each column of the *ith* row of the table includes the rank of the Q values, the landmark number corresponding to this rank (1-19), and the *R, D, Q* values of the landmark. The last two columns represent the average rank using the top *i* landmarks in the table with weighting by the *Q* values or not. We use different colors in Table 2 to represent the average ranks of the four methods.

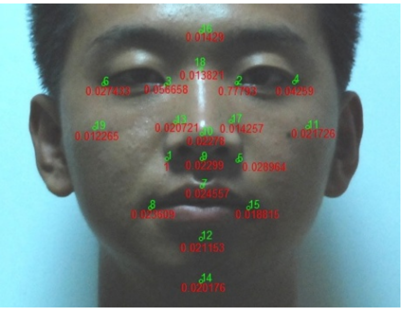


Fig. 3. Landmarks ranking by *Q* values (in descending order). The red numbers are the *Q* values of landmarks, and the green are the landmark ranks.

The Cumulative Match Characteristic (CMC) graphs of the four methods are shown in Figure 4. From Figure 4, we can see that using merely the optimal subset of the landmarks (method 2), and only weighting the reprojection errors (method 3), or using the both measures (method 4), all have better performance than the algorithm (method 1) without any improving measure for nearly all sizes of the candidate list (except that the method 3 is slightly worse than method 1 when the size is 35 or 36).

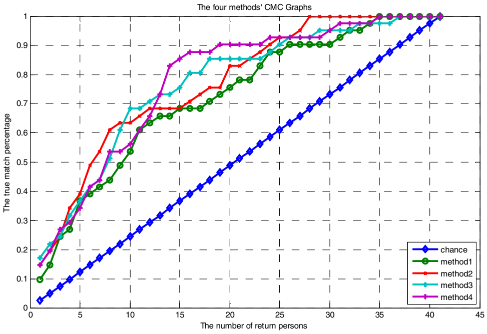


Fig. 4. The CMC graphs of the four method

5 Conclusions

In this paper, we propose a new method to quantitatively analyze the qualities of the facial landmarks for skull identification. Based on this, a new landmark-based algorithm, which takes fully into account the different weights of the landmarks in the recognition, is proposed. In addition, we can select an optimal recognition subset of landmarks to improve the recognition rate according to the recognition quality of landmarks. Experiments show that the improving measures are effective.

Acknowledgements. This work was partially supported by the National Natural Science Foundation of China (No. 60872127, No.60736008, No. 61003133).

References

[1] Damas, S., Cordón, O., Ibáñez, O.: Forensic identification by computer-aided craniofacial superimposition: a survey (2011)

[2] Claes, P., et al.: Computerized craniofacial reconstruction: Conceptual framework and review. Forensic Science International 201(1-3), 138–145 (2010)

[3] Tu, P., et al.: Automatic Face Recognition from Skeletal Remains (2007)

[4] Blanz, V., Scherbaum, K., Seidel, H.P.: Fitting a morphable model to 3D scans of faces. In: Computer vision, pp. 1–8 (2007)

[5] Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH 1999. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (1999)

[6] Toderici, G., et al.: Bidirectional relighting for 3D-aided 2D face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA (2010)

[7] Di, H., et al.: Automatic Asymmetric 3D-2D Face Recognition. In: The 20th International Conference on Pattern Recognition, IEEE Computer Society, Washington, DC, USA (2010)

[8] Yang, W., et al.: 2D-3D face matching using CCA. In: 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam (2008)

[9] Rama, A., et al.: Mixed 2D-3D Information for Pose Estimation and Face Recognition. In: Proceedings of Acoustics, Speech and Signal Processing, ICASSP 2006, Toulouse (2006)

[10] Riccio, D., Dugelay, J.: Geometric invariants for 2D/3D face recognition. Pattern Recognition Letters 28(4), 1907–1914 (2007)

Sequential Fusion Using Correlated Decisions for Controlled Verification Errors

Vishnu Priya Nallagatla and Vinod Chandran

School of Engineering Systems, Queensland University of Technology, Brisbane, Australia
{v.nallagatla,v.chandran}@qut.edu.au

Abstract. Fusion techniques have received considerable attention for achieving lower error rates with biometrics. A fused classifier architecture based on sequential integration of multi-instance and multi-sample fusion schemes allows controlled trade-off between false alarms and false rejects. Expressions for each type of error for the fused system have previously been derived for the case of statistically independent classifier decisions. It is shown in this paper that the performance of this architecture can be improved by modelling the correlation between classifier decisions. Correlation modelling also enables better tuning of fusion model parameters, '*N*', the number of classifiers and '*M*', the number of attempts/samples, and facilitates the determination of error bounds for false rejects and false accepts for each specific user. Error trade-off performance of the architecture is evaluated using HMM based speaker verification on utterances of individual digits. Results show that performance is improved for the case of favourable correlated decisions. The architecture investigated here is directly applicable to speaker verification from spoken digit strings such as credit card numbers in telephone or voice over internet protocol based applications. It is also applicable to other biometric modalities such as finger prints and handwriting samples.

Keywords: Multi-instance fusion, multi-sample fusion, verification error trade-off, sequential decision fusion, correlation, verification error bounds.

1 Introduction

Reliability of the performance of biometric identity verification systems remains a significant challenge. Performance degradation arises from intra-class variability and inter-class similarity. Intra-class variability is caused when individual samples of the same person are not identical for each presentation and inter-class similarity arises from high degree of identicalness of the same biometric trait between different persons. These limitations may lead to misclassification of the verification claims resulting in false alarms and false rejects. These two errors are dependent and in general, it is difficult to reduce the rate of one type of error without increasing the other. Fusion techniques attempt to reduce both.

Fusion techniques have been classified into the 6 categories: multi-instance, multi-sample, multi-sensor, multi-algorithm, multi-modal and hybrid. A system that integrates multi-instance and multi-sample fusion proposed in [1] is analytically

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 49–56, 2011.

shown to improve performance and allow a controlled trade-off between false rejection rate (FRR) and false acceptance rate (FAR) when the classifier decisions are assumed to be statistically independent. A statistical analysis of the problem of fusing independent decisions from classifiers has also been addressed in the context of writer identification from different handwritten words in [2]. This analysis did not consider multiple instances in a sequential scheme as used in [1].

Attempts [3] have also been made to model the correlation between classifiers and incorporate the statistical dependence information into the fusion scheme in order to improve performance. There have been claims of marginal improvement [4] in performance when correlation is considered, but a systematic analysis of this problem has not yet been presented. In this paper, we analyse the effect of modelling the statistical dependence between classifier decisions for multi-instance and multi-sample fused biometric identity verification system.

Section 2 and section 3 explain the methodology and theoretical analysis of the proposed sequential decision fusion scheme in the context of text-dependent speaker verification system. Section 4 develops the equations required for modelling the correlation between the classifier decisions and section 5 provides a brief conclusion with suggestions for possible future work.

2 Experimental Setup

Speech data from the CSLU Speaker Recognition Version 1.1 database is used for evaluating performance of the proposed fusion scheme. The data comprise of spoken digit strings that are manually segmented into individual digits. The methodology used is the same as explained in [1]. Mel Frequency Cepstral Coefficient features are extracted by processing utterances in 26 ms frames. Left - Right HMM models with five states per phoneme and three mixtures per state are created for each digit. The digit models are trained separately for each speaker. A universal background model is used for speaker normalization and this model is adapted using MAP and MLLR.

Data from 11 male speakers is used for performance evaluation. Each speaker data is divided into train, tune and test subsets that are kept disjoint. Impostor testing for a client is the done using data from the 10 speakers other than the client. Several combinations are used to obtain reliable estimates of error rates. A training set (21 client utterances) is first chosen for creating speaker specific digit dependent HMM models. Once the models are trained, the remaining data are divided into 5 different tune and test data subset combinations. Each tune set (35 client and 140 impostor utterances) is used to set appropriate digit dependent threshold and evaluate individual classifier error rates and finally the test set (70 client and 420 impostor utterances) is used to evaluate the performance of the proposed fusion.

In text-dependent speaker verification (TDSV) mode, the digit is known and the speaker is unknown. If the claimed speaker's model for the digit matches the utterance, it is accepted. This may be a true or false acceptance depending on whether the utterance came from the claimed speaker or an impostor. Impostor testing is done using utterances of the same (known) digit, resulting in true rejections or false acceptances. An instance in the context of TDSV by the proposed architecture refers to the text or digits which form the decision stages. A sample represents any single utterance of a digit from a speaker. If a sample is rejected at a decision stage, the next sample is randomly picked from the remaining utterances.

3 Multi-biometric Fusion for Speaker Verification

As explained in [1], the combination of multi-instance and multi-sample fusion schemes allows control of the verification error trade-off. It is desirable in most of the speaker verification applications such as remote authentication, telephone and internet shopping applications to serve both security and user convenience requirements which can be achieved by setting the parameters of the architecture, the number of attempts at each decision stage (samples) and the number of decision stages (instances), to be used for verification of a specific speaker.

In the proposed architecture (Fig. 1), the maximum permissible number of repeated samples, '*M*', and the number of instances *'N'* are fixed prior based on the error rates obtained from the tune dataset. In this system, the speaker presents an input test utterance $X_{m,n}$ *(m=1, 2 ...M, n =1, 2 ...N)* and the classifier C_n (here HMM) makes a decision to either accept or reject the claimed identity.

For a speaker to be declared genuine for a particular instance (or spoken text), it is considered sufficient if any one sample (or utterance) presented to the system gets accepted. Acceptance decisions are logical 'OR' for multiple samples. However if the speaker is accepted by 'i^{th} sample' ($1<i<m$) then the subsequent samples need not be verified. The speaker is considered to be an impostor when all the '*m*' samples are rejected. Rejection decisions are logical 'AND' for multiple samples. Conversely, it is considered necessary in the sequential decision framework that a speaker be accepted by all instances in the sequence of decision stages. Acceptance is thus logical 'AND' for multiple instances. If the speaker is rejected by any decision stage, the sequence terminates and thus rejection decisions are logical 'OR' for multiple instances. Considering false acceptance rate or FAR (α) and false rejection rate or FRR (ρ) to be independent for each instance, the fusion scheme equations are:

$$\text{Multiple Samples}: \quad \alpha_{(m)} = m\alpha; \rho_{(m)} = \rho^m \tag{1}$$

$$\text{Multiple Instances}: \quad \alpha_{(n)} = \alpha^n; \rho_{(n)} \approx n\rho \;\; (when \;\; \rho << 1) \tag{2}$$

$$\text{Multi - Instance \& Multi - Sample Fusion}: \; \alpha_{(m,n)} = (m\alpha)^n; \rho_{(m,n)} \approx n(\rho^m) \tag{3}$$

From the above equations it is clear that while the FRR decreases (since ρ is less than 1) multiplicatively with the number of attempts *'m'*, the FAR increases additively with *'m'* and the reduction in the FAR is multiplicative (Equation 2) with the number of instances *'n'*, while the increase in the FRR is approximately additive with *'n'*. The facts to be noted here are (a) the behaviour with respect to *'m'* and *'n'* are complementary and (b) multiplicative changes are faster than additive ones and this enables control of the errors through these parameters in the architecture.

With the above equations, it is possible to design a fused system that has lower errors of both types compared to a single verification stage using a single sample. It is

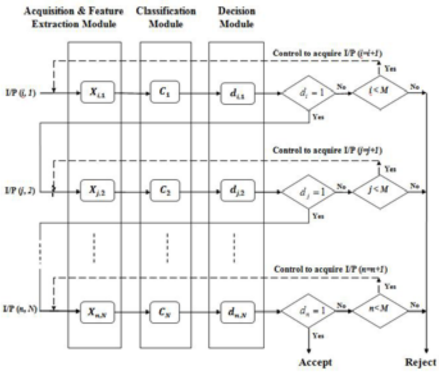


Fig. 1. Architecture for a multi-instance and multi-sample fusion scheme with '*M*' repetition of samples and '*N*' classifiers arranged sequentially

also possible to keep both errors within reasonable bounds – without false rejections rising quickly to nearly 100% when the false acceptance reduces or the other way around. The trade-off in achieving this is the time for computations required to perform multiple matches and make decisions with every sample and instance in the architecture. It will indeed be so if the decisions were statistically independent as assumed, for multiple samples as well as for multiple instances.

In the above analysis, it is assumed that the FAR and FRR are the same for all stages (instances) for the purpose of simplicity. This can be relaxed and more complicated and exact formulae obtained as:

$$\alpha_{Ideal} = m\alpha_1 * m\alpha_2 ... * m\alpha_n \tag{4}$$

$$\rho_{Ideal} = \rho_1^m + (1-\rho_1^m)\rho_2^m + ... + (1-\rho_1^m)..(1-\rho_{n-1}^m)\rho_n^m \tag{5}$$

Verification error rates for this fusion architecture can be estimated using the above equations and substituting the error rates for individual digits from the tune dataset (Ideal Error Rates). Digit models with reasonably lower error rates need to be used for fusion, otherwise ideal error rates may sometimes reach 100%. In case of statistically independent decisions, these ideal error rates are the same as the experimental error rates. However, statistical independence between decisions may not be always valid. Ideal error rates may be different from the experimentally obtained error rates and the difference can be statistically significant as demonstrated in [1]. This most likely cause of the difference is statistical dependence (correlation) between classifier decisions, resulting in error rates that are larger or smaller than the ideal values obtained under independence assumption [5, 6]. The input data presented at each classifier may also be correlated even though the text is different [7]. Taking classifier decision correlations into account is a further refinement of the statistical analysis as done in this work. In the next section, the effect of correlation modelling for the sequential decision fusion scheme is analysed.

4 Fusion of Correlated Decisions

A limitation of the analysis presented in [1] is the assumption that the decisions made on each instance are independent, which in general is not true for several words/digits spoken by the same individual. For modelling correlation, it is important to express the degree of dependence between the decisions and then to derive the appropriate decision fusion rule for fusing these decisions. The degree of dependence between the classifier decisions can be estimated based on the Bahadur-Lazarsfeld Expansion (BLE) [7]. The expansion begins with the ideal error rates (calculated assuming statistical independence), and then multiplies them by a correction factor. The ideal error rates for multi-instance fusion are obtained by using equations 4 & 5 with *'m'* equal to one where as for a multi-sample system the *'n'* is equal to one. For the proposed sequential fusion, the decision fusion rules used are 'AND' and 'OR' logics.

The equations to calculate the estimated false acceptance rate (α) and false rejection rate (ρ), for multiple instances using the BLE [7] and error rates for individual instances can be given as:

$$\alpha_{Est} = \alpha_{Ideal}\left(1+\sum_{i<j}\gamma_{ij}^{1}\sqrt{\frac{(1-\alpha_i)(1-\alpha_j)}{\alpha_i\alpha_j}}+\sum_{i<j<k}\gamma_{ijk}^{1}\sqrt{\frac{(1-\alpha_i)(1-\alpha_j)(1-\alpha_k)}{\alpha_i\alpha_j\alpha_k}}....\right) \tag{6}$$

$$\rho_{Est} = 1-\rho_{Ideal}\left(1+\sum_{i<j}\gamma_{ij}^{0}\sqrt{\frac{\rho_i\rho_j}{(1-\rho_i)(1-\rho_j)}}+\sum_{i<j<k}\gamma_{ijk}^{0}\sqrt{\frac{\rho_i\rho_j\rho_k}{(1-\rho_i)(1-\rho_j)(1-\rho_k)}}...\right) \tag{7}$$

For multi-sample system, the estimated values of true rejection rate ($\beta = 1-\alpha$) and false rejection rate (ρ) for correlated decisions can be given as:

$$\beta_{Est} = \beta_{Ideal}\left(1+\sum_{i<j}\gamma_{ij}^{1}\sqrt{\frac{\alpha_i\alpha_j}{(1-\alpha_i)(1-\alpha_j)}}+\sum_{i<j<k}\gamma_{ijk}^{1}\sqrt{\frac{\alpha_i\alpha_j\alpha_k}{(1-\alpha_i)(1-\alpha_j)(1-\alpha_k)}}....\right) \tag{8}$$

$$\rho_{Est} = \rho_{Ideal}\left(1+\sum_{i<j}\gamma_{ij}^{0}\sqrt{\frac{(1-\rho_i)(1-\rho_j)}{\rho_i\rho_j}}+\sum_{i<j<k}\gamma_{ijk}^{0}\sqrt{\frac{(1-\rho_i)(1-\rho_j)(1-\rho_k)}{\rho_i\rho_j\rho_k}}...\right) \tag{9}$$

Here γ^k (k=0, 1) are the correlation coefficients for client and impostor decisions and are defined using z_i's, variables that are orthogonal with respect to the independence model with zero mean and unit variance,

$$\gamma_{12...n} = \sum[z_1\, z_2....z_n],\ \text{and}\ z_i = \left(\frac{d_i - p_i}{\sqrt{p_i(1-p_i)}}\right), p_i = P(d_i = 1);\ 1-p_i = P(d_i = 0) \tag{10}$$

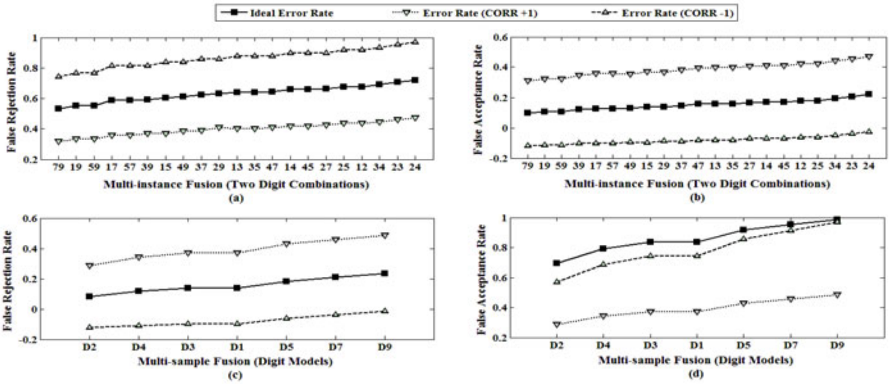


Fig. 2. Comparison of Ideal Error Rates and Estimated Error Rates calculated using positive and negative correlation coefficients. (a) FRR for Multi-instance Fusion (b) FAR for Multi-instance Fusion (c) FRR for Multi-sample Fusion (d) FAR for Multi-sample Fusion.

Figure 2 demonstrates the effect of 2nd order correlation coefficients on the performance of multi-instance and multi-sample fusion schemes. The error rates plotted in the figure are calculated using the dataset 1 for ‘speaker 0241’. The lines plotted for multi-instance fusion refers to the different two digit combinations whereas multi-sample fusion lines represent the error rates for a single instance (digit model) verified on two samples. It is evident that for multi-instance fusion the reduction in the estimated FRR is proportional to the increase in decision correlation (Fig. 2(a)) and whereas the estimated FAR is inversely proportional to the correlation (Fig 2(b)). However for multi-sample fusion, the reduction in experimental FRR is because of lower correlation values for a client (Fig. 2(c)) and experimental FAR decreases with higher decision correlation for an impostor (Fig 2(d)). The comparison of ideal error rates with the estimated values (Fig. 2) represents the same conclusion regarding the favourable dependence for fusion as explained in [5, 6] using Q values between classifier decisions. The favourable conditional dependence for OR fusion [5] is negative for clients and positive for impostors. However, for AND fusion [6] the favourable dependence is positive for clients and negative for impostors.

Favourable dependence between individual digits enables determination of the set of favourable digit combinations for a specific speaker. Table 1 represent the decrease in mean error rates for three random speakers for all possible digit sequences and the set of digit sequences/combinations with favourable correlation. It can be said from the results that favourable digit combinations are similar across different datasets for a given speaker and differ slightly between different speakers. So verifying a speaker using his/her favourable digit sequence can result in lower error rates. Selecting the optimal set of digit models specific for performance enhancement can be further based on phoneme correlation which will be explored in future.

It can also be noted from table 1 that trade-off between security and user convenience can be achieved by selecting the parameter set (*n*D, *m*S), *‘n’* - number of ‘Digits’ and *‘m’* - number of ‘Samples’, required for verification. For example, the FRR and FAR values reduce from initial mean error rates (1S-1D) by 4.9% and 27% respectively for (4D, 2S) and 18.8% and 8.7% respectively for (4D, 3S) in the case of

Table 1. Mean Error Rates for proposed fusion (1D-1S: One Digit-One Sample Combination...)

Speaker		1D-1S	4D-2S	4D-2S(γ)	4D-3S	4D-3S(γ)
0047	FRR	$0.233^{\pm 0.08}$	$0.241^{\pm 0.04}$	$0.222^{\pm 0.03}$	$0.069^{\pm 0.02}$	$0.053^{\pm 0.01}$
	FAR	$0.231^{\pm 0.09}$	$0.047^{\pm 0.02}$	$0.029^{\pm 0.02}$	$0.083^{\pm 0.03}$	$0.055^{\pm 0.02}$
0176	FRR	$0.314^{\pm 0.09}$	$0.364^{\pm 0.06}$	$0.300^{\pm 0.06}$	$0.159^{\pm 0.06}$	$0.143^{\pm 0.06}$
	FAR	$0.295^{\pm 0.08}$	$0.079^{\pm 0.02}$	$0.062^{\pm 0.01}$	$0.179^{\pm 0.04}$	$0.157^{\pm 0.04}$
0241	FRR	$0.392^{\pm 0.07}$	$0.377^{\pm 0.08}$	$0.343^{\pm 0.07}$	$0.229^{\pm 0.05}$	$0.204^{\pm 0.05}$
	FAR	$0.392^{\pm 0.07}$	$0.146^{\pm 0.04}$	$0.122^{\pm 0.03}$	$0.311^{\pm 0.05}$	$0.304^{\pm 0.04}$

favorable correlation (speaker 0241). This performance can be further improved by increasing the number of instances and samples used for verification.

The equations derived above can thus be used to tune the parameters, such as number of instances, number of samples and favourable set of digit sequences, required to determine the performance of the fusion method on test data set. For tune dataset, the correlation between decisions are known and so the experimental values obtained are equal to the estimated values obtained using Equations (6-9). However in real world applications (test dataset), the correlation values are unknown. In order to estimate the error rates for the test set, the correlation coefficient for a speaker across different tune datasets can be used. Figure 3(a) & 3(b) show the overlap of 2nd order correlation coefficient values between the tune and test datasets for 'speaker 9'. By ensuring that the tune set considers all the (prior) conditions under which a speaker may be tested, the overlap between the correlation sets can be maximised. Thereby using the variance of correlation values for a specific speaker, the maximum and minimum error rates (i.e., error bounds) can be calculated for fixed '*M*' and '*N*' values using equations 6-9 with individual error rates for each instance from tune set. The error bounds obtained using the correlation coefficients for 2 digit combinations are shown in figure 3(c) & 3(d). It is evident that most of the experimental error rates obtained from the test set fall within the bounds of error rates estimated using the tune set parameters.

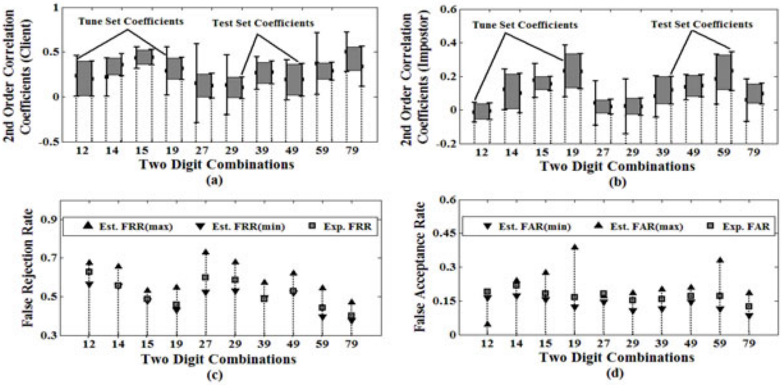


Fig. 3. Comparison of tune and test dataset parameters (a) Correlation for Client (b) Correlation for an Impostor (c) Estimated and Experimental FRR (d) Estimated and Expected FAR

In real world applications, the verification system may set initial acceptable values for FRR and FAR. These error rates can be easily estimated using the mathematical formulae discussed and the fusion parameters, i.e., the number of digits and samples, the particular digit sequence and variance of correlation. This fusion method can be applied to biometric systems used for remote authentication with modalities such as voice, handwriting, fingerprints, and keyboard strokes.

5 Conclusion and Future Work

A sequential decision fusion architecture with multiple attempts can be effectively used to control the trade-off between false accepts and false rejects. It was shown in [1] that there is potential to improve the performance of weaker classifiers by combining decisions under the assumption of statistical independence. This work demonstrates that superior performance can be obtained by considering the correlation values that are favourable in the multi-instance and multi-sample components. Correlation modelling also enables prediction of verification errors using parameters adjusted using a tune data set. Future work possible in this direction includes (a) the modelling of user adaptation in repetitive samples and (b) optimal classifier selection in this architecture amongst many possible instances or digit combinations.

Acknowledgement. The authors acknowledge the contributions of Dr. Eddie Wong and D. Lawson in the QUT Vocal Access Speaker Verification software package and assistance from Prof. Sridharan and members of the speech and audio research group at QUT.

References

[1] Nallagatla, V.P., Chandran, V.: Sequential decision fusion for controlled detection errors. In: 13th International Conference on Information Fusion (FUSION), Edinburgh (2010)

[2] Zois, E.N., Anastassopoulos, V.: Decision Fusion for Writer Identification. In: Proc. Int. Conf. DSP 1997, Santorini, Greece (1997)

[3] Karthik, N., et al.: Biometric Fusion: Does Modeling Correlation Really Matter? In: Proceedings of IEEE Third International Conference on BTAS, Washington DC, pp. 271–276 (2009)

[4] Ushmaev, O., Novikov, S.: Biometric Fusion: Robust Approach. In: Proc. of MMUA, Toulouse, France (2006)

[5] Venkataramani, K., Kumar, B.V.K.V.: OR rule fusion of conditionally dependent correlation filter based classifiers for improved biometric verification. In: Proceedings of SPIE, p. 62450A (2006)

[6] Venkataramani, K., Vijaya Kumar, B.V.K.: Conditionally-dependent classifier fusion using AND rule for improved biometric verification. In: Int.Conf. on Advances in Pattern Recognition, pp. 277–286 (August 2005)

[7] Zois, E.N., Anastassopoulos, V.: Fusion of correlated decisions for writer verification. Pattern Recognition 34, 47–61 (2001)

An Online Three-Stage Method for Facial Point Localization

Weiyuan Ni[1], Ngoc-Son Vu[2], and Alice Caplier[2]

[1] ICA-ACROE
[2] GIPSA-lab, Grenoble, France
ni.weiyuan@imag.fr

Abstract. Finding facial features respectively under expression and illumination variations is always a difficult problem. One popular solution for improving the performance of facial point localization is to use the spatial relation between facial feature positions. While existing algorithms mostly rely on the priori knowledge of facial structure and on a training phase, this paper presents an online approach without requirements of pre-defined constraints on feature distributions. Instead of training specific detectors for each facial feature, a generic method is first used to extract a set of interest points from test images. With a robust feature descriptor named Patterns Oriented Edge Magnitude (POEM) histogram, a smaller set of these points are picked as candidates. Then we apply a game-theoretic technique to select facial points from the candidates, while the global geometric properties of face are well preserved. The experimental results demonstrate that our method achieves satisfactory performance for face images under expression and lighting variations.

Keywords: facial point localization, game-theoretic matching, POEM.

1 Introduction

Although there exists some reliable face detection methods, e.g. Viola-Jones detector [12], the output faces are still not error-free. Hence, localization of facial points is an important step for many tasks such as face recognition and face alignment. Finding facial features respectively under expression and illumination variations is always a difficult problem. One popular solution for improving the localization performance is to use the spatial relation between facial feature positions. Existing algorithms mostly rely on the priori knowledge of facial structure and on a training phase. In [2,11], pairwise spatial relations between facial point positions are learned for detection. With the knowledge of facial feature distributions, [14] divides faces into several regions of interest(ROI), then individual feature patch templates are used to detect points in the relevant ROI. Ding et al.[4] first localize two eyes and estimate the approximate positions of other features with a priori knowledge about face.

Inspired by the work of [1], where the game-theoretic technique is used for 3D image registration and where the global consistency between correspondences is

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 57–64, 2011.

well preserved, we propose here an online, three-stage method for facial point localization. While [1] matches the features of images for the *same* scene/object, we try to find the correspondences between feature points of two different face images with *different* identities and even of *different* expressions and illuminations. As can be seen in Figure 1, we cast the feature point localization problem in a coarse-to-fine matching task. In our model, the template ($\boldsymbol{T}$) is an image with manually labeled *target points* and for each test image ($\boldsymbol{I}$), we aim at finding the corresponding feature points. In the first step, instead of training specific detectors for each facial feature, as commonly used in other algorithms [2,14], a generic method is applied to extract a set of interest points from $\boldsymbol{I}$. Then, for each target point in $\boldsymbol{T}$, a smaller set of these interest points are picked as candidates, using a robust feature descriptor named Patterns Oriented Edge Magnitude (POEM) histogram [13]. Finally, we apply the game-theoretic technique to select desired facial points from candidates, without requirements of pre-defined constraints on feature distributions.

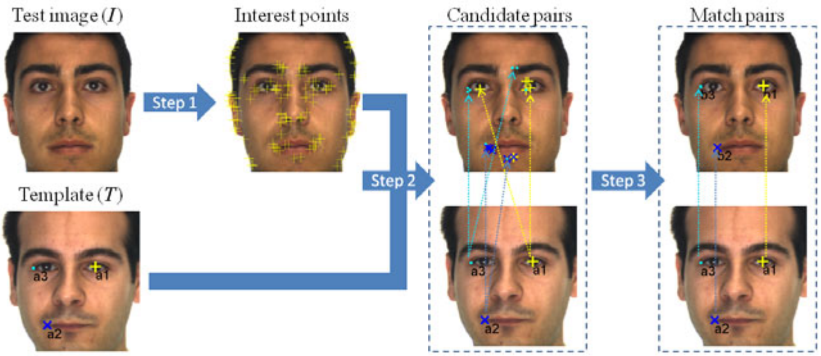


Fig. 1. Overview of our method. For clarity, only 3 facial points are located as examples. In Step 1, interest points are found by a generic detector. For each point in the template, a small set of points are picked as candidates in Step 2. The desired facial points are selected from candidates in Step 3.

2 Methodology

2.1 Step 1: Detection of Interest Points

Unlike some approaches requiring trained detectors for specific facial features [2,14], we first use a more generic method to extract a set of interest points which are invariant to scale, rotation and translation and which are also robust to illumination changes. A smaller set of these points will be picked as candidates in the following step. The fundamental idea behind this is that we believe some facial features, e.g. eye corners, mouth corners and nostrils are invariant to similarity transformations with respect to the change of identity, expression and illumination. We have tested several interest point detectors which are commonly used, including Difference of Gaussian(DoG)[6], Laplacian-of-Gaussian(LoG) [5], Hessian-Laplacian and Harris-Laplacian [9]. According to the visual results on several images (see an example in Figure 2), we adopt Harris-Laplacian detector in this paper, since it can find more facial feature points.

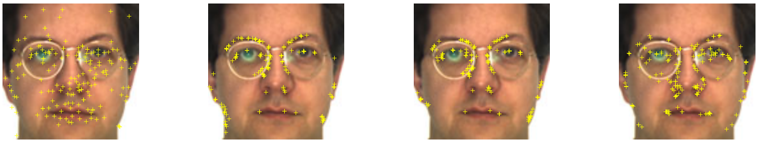


Fig. 2. Interest points detected by different methods. From left to right: DoG, LoG, Hessian-Laplacian and Harris-Laplacian detector.

2.2 Step 2: Candidate Points Screening with POEM Descriptor

After the extraction of interest points, the localization of facial points turns into a matching problem between the target points from $\boldsymbol{T}$ and the interest points from ($\boldsymbol{I}$). Considering the efficiency of matching, for each target point, only K (e.g. $K \leq 10$) points in $\boldsymbol{I}$ with the nearest descriptor are picked as candidates.

Since facial features are not stable under identity, expression and lighting variations, we need a robust descriptor to distinguish facial points. We propose here to use the recent feature descriptor called Patterns Oriented Edge Magnitude (POEM), which has been successfully applied for face representation with very strong quality results [10,13]. The main steps of calculating POEM histogram are (for more details see [13]):

(1) Calculation of image gradient and quantification of orientations.

(2) Magnitude Accumulation. For a pixel p, a local histogram of gradients over all pixels within the cell, centered on p, is calculated and assigned to p.

(3) Computation of self-similarity-based operator. In each orientation θ_i, the magnitude at pixel p is compared with l surrounding pixels in a radius r:

$$POEM^{\theta_i}(p) = \sum_{j=1}^{l}(I_p^{\theta_i} - I_{c_j}^{\theta_i} > \tau)2^j, \tag{1}$$

where $I_p^{\theta_i}$, $I_{c_j}^{\theta_i}$ are the magnitudes of central and surrounding pixels p, c_j, the threshold τ is 0.2.

So for each pixel, there will be a set of m values:

$$POEM(p) = \left\{POEM^{\theta_1}(p), ..., POEM^{\theta_m}(p)\right\}, \tag{2}$$

where m equals to the number of defined orientations.

(4) Finally, for a pixel, we calculate m histograms of POEM (one for each orientation) over a small window, centered on that pixel. These m histograms are concatenated and used as the feature descriptor of the considered pixel.

Depending on the distances between histograms, K interest points with the nearest descriptor are picked as candidates for each target point.

2.3 Step 3: Multi-template Game-Theoretic Matching

Up to this point, there are several candidate points in $\boldsymbol{I}$ for each target point in $\boldsymbol{T}$. Let $O_1 = \{\boldsymbol{a}_1, ..., \boldsymbol{a}_N\}$ and $O_2 = \{\boldsymbol{b}_1, ..., \boldsymbol{b}_L\}$ be the target and candidate point sets respectively, where $\boldsymbol{a}_i$,$\boldsymbol{b}_j$ represent the coordinates. Thus a target point $\boldsymbol{a}_i$ corresponds to K candidate point pairs: $(\boldsymbol{a}_i, \boldsymbol{b}_1)$,...,$(\boldsymbol{a}_i, \boldsymbol{b}_K)$. In this stage, we aim at finding the match pairs for every target point, e.g. $(\boldsymbol{a}_1, \boldsymbol{b}_1)$,$(\boldsymbol{a}_2, \boldsymbol{b}_2)$ and $(\boldsymbol{a}_3, \boldsymbol{b}_3)$ in Figure 1. As facial features have certain geometric structure, there exists a compatible transformation for all these match pairs. The selection process can be seen as a matching game [1], in which candidate pairs $(\boldsymbol{a}_i, \boldsymbol{b}_j)$ are defined as pure strategies available to players and the payoffs for every combination of strategies are calculated as:

$$\pi((\boldsymbol{a}_1, \boldsymbol{b}_1), (\boldsymbol{a}_2, \boldsymbol{b}_2)) = \frac{min(\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|, \|\boldsymbol{b}_1 - \boldsymbol{b}_2)\|)}{max(\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|, \|\boldsymbol{b}_1 - \boldsymbol{b}_2)\|)}, \tag{3}$$

where $\|\cdot\|$ represents the Euclidian distance.

With Equation 3, strategies that correspond to rigid transformation have high payoff values, while less compatible pairs get lower scores. Take Figure 1 for example, $\pi((\boldsymbol{a}_1, \boldsymbol{b}_1)$,$(\boldsymbol{a}_2, \boldsymbol{b}_2))$ and $\pi((\boldsymbol{a}_1, \boldsymbol{b}_1)$,$(\boldsymbol{a}_3, \boldsymbol{b}_3))$ are higher than $\pi((\boldsymbol{a}_1, \boldsymbol{b}_2)$, $(\boldsymbol{a}_2, \boldsymbol{b}_3))$. Since players always want to get higher payoffs, they prefer to pick strategies that are compatible with their opponents' choices. As the game is repeated by a large population of players, a set of strategies with high mutual compatibility will be assigned to high weights. The compatible set of strategies can be obtained by calculating evolutionary stable states (ESS's), see Appendix for details. Finally, the point pairs with high weights are taken as match pairs.

Since facial features in test images vary with the change of identity and expression, the matching problem will suffer from the error of candidate screening. More precisely, the correspondence $\boldsymbol{b}_i$ of one target point $\boldsymbol{a}_i$ may not be involved in the candidate set of $\boldsymbol{a}_i$. In that case, all pairs that contain $\boldsymbol{a}_i$ will get low weights after the matching game, i.e. this facial point is *miss-located.* To increase the robustness of game-theoretic matching, we apply multiple templates to match with test images. Only if one of these templates gives a match point of target point $\boldsymbol{a}_i$, this facial point can be successfully located. Hence, the probability of "miss-located" is very low. If a facial point is located by several templates, the average location is used as the final result.

3 Experimental Results

3.1 Experiment Settings

Database. We use images from AR-face database [8] which contains over 4,000 color images corresponding to 126 people's faces. Images feature frontal view faces with different facial expressions and illumination conditions (see Figure 3).

Evaluation criterion. Let $\boldsymbol{b}_i$ and $\boldsymbol{b}_i^+$ be the predicted and manually labeled locations (ground truth), the localization error is calculated as: $m_i = \left\|\boldsymbol{b}_i - \boldsymbol{b}_i^+\right\| /d_{eye}$, where d_{eye} is the average distance between two eye pupils in ground truth.

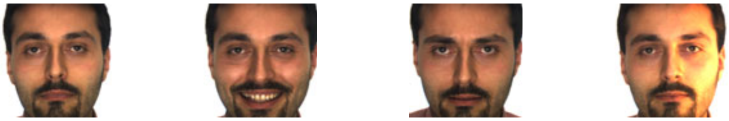


Fig. 3. Examples of test images. From left to right: neutral, smile, anger and side light.

If we choose a threshold c, the correct localization rate will be:

$$rate = \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} \left(m_i^j < c \right)}{M \times N}, \tag{4}$$

where M is the number of test images and N is the number of target points per template.

3.2 Matching of Labeled Points

In order to verify the effectiveness of our method for facial features, with the assumption of perfect selection of candidates, we first applied our method to match two sets of labeled points from two different images. We randomly selected 20 images of different individuals with neutral expression from AR-face database and ran game-theoretic matching between every two images, i.e. 190 image pairs. Original images with the resolution 768×576 are used directly in this experiment.

For each image pair, we take one image as template ($\boldsymbol{T}$) and calculate feature descriptors for all labeled points in both images. For each point in $\boldsymbol{T}$, 5 points with nearest descriptor in another image ($\boldsymbol{I}$) are used as candidates. AR-face images have been manually labeled with 22 landmarks, so there are 110 point pairs which are then regarded as strategies in the matching game.

A point in image $\boldsymbol{I}$ assigned to the corresponding point in $\boldsymbol{T}$, means a correct match. We adopted different POEM parameters to determine the closest neighbors, the average match rate is about 98% and the results are not sensitive to parameter selection. Hence, our method works well for the matching of facial points.

3.3 Facial Features Localization

Here, we aim at locating 10 facial points in test images (Figure 8). We form two image sets for evaluation: Data 1 consists of frontal faces with neutral, smile and anger expression and Data 2 is a set of face images under side illumination (Figure 3). All the face images are extracted by Viola-Jones detector [12].

1. Using different number of templates
We first evaluated the impact of adopting different numbers of templates. The templates and 350 test images were randomly selected from Data 1. The localization results can be seen in Figure 4. It is clear that matching with single template gets lower accuracy than multiple templates, due to the high probability of "miss-located". While the results with 10, 15 and 20 templates are very

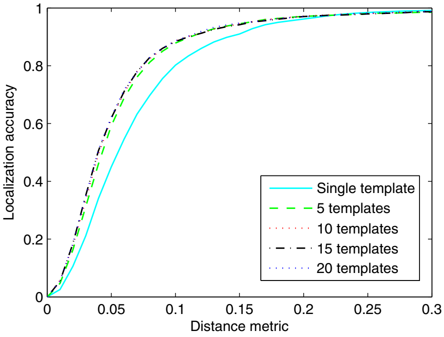


Fig. 4. Different numbers of templates

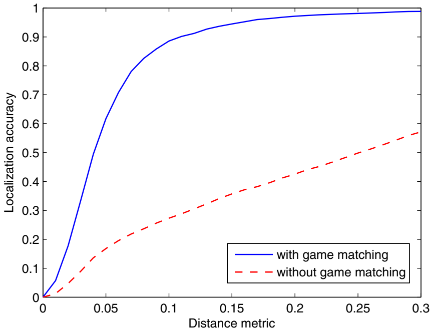


Fig. 5. With and without game matching

similar, the accuracy of using 5 templates is slightly worse. For efficiency, we adopt 10 templates in the following experiments, which have no overlap with test images.

2. Verification of the importance of game-theoretic matching
To show the importance of game-theoretic matching, we also tried to localize points without this step, i.e. we directly picked the points with closest descriptor in $\boldsymbol{I}$ as the correspondence of a target point in $\boldsymbol{T}$. Suffering from the variation of facial features, the closest-feature-based method is more like a random selection from detected interest points (Figure 5, Data 1), while game-matching-based method achieves a good performance. Hence the game-theoretic technique, which carries the information of face structure, is very important to facial point localization.

3. Using different feature descriptors under neutral condition
This section compares the performance of our method when different feature descriptors are used: intensity, SIFT[7] and POEM descriptor. When using intensity values, the sum of squared differences(SSD) between two sub-regions is computed as the measure of distance. We calculated the three feature descriptors with the same window size, and the localization results can be seen in Figure 6 (Data 1). The facial point localization method works better with POEM than with SIFT, and SSD does not seem to be suitable in this case. Using a threshold $m < 0.15$, our approach is successful in 95% of points (see some examples in Figure 8), while localization accuracy with SIFT only reaches 82%. The rates of other methods, e.g. 96% for PRFR [2] and TST [3], 95% for [11], are very close to our result. Considering that our approach runs without specific trained detectors nor face models, the localization performance is satisfactory.

4. Using different feature descriptors under lighting changes
Few evaluations have been done specifically for locating facial points under lighting changes. Here, 100 images were selected randomly as test images from Data 2, and template set consists of 5 images from Data 1 and 5 images from Data 2.

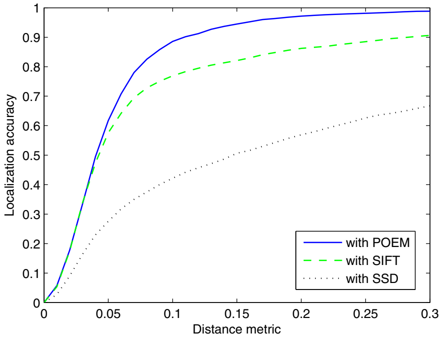


Fig. 6. Result of using Data 1

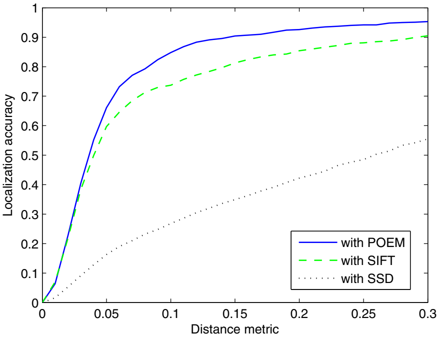


Fig. 7. Result of using Data 2

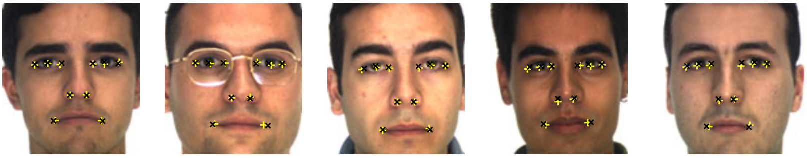


Fig. 8. Examples of localized facial points, where "+" is the output of our method and "×" is the manually labeled location

Three kinds of features are also compared in this case, and the results are shown in Figure 7. The game-theoretic method with POEM still gives better result than with other two features. For $m < 0.15$, our method reaches a success rate of 90% and the method with SIFT gets 81%. The accuracies are slightly lower than in neutral condition but still acceptable.

4 Conclusion

This paper presents an online approach to locating facial points, requiring no pre-defined constraints on feature distributions. We cast the localization problem in a matching game which preserves global geometric consistency of facial points. The experimental results demonstrate that the game-theoretic technique works well for facial point localization with a combination of a generic interest point detector. Besides, POEM descriptor is adopted in our method, and it shows better ability to represent facial features than SIFT.

References

1. Albarelli, A., Rodola, E., Torsello, A.: A game-theoretic approach to fine surface registration without initial motion estimation. In: CVPR (2010)
2. Cristinacce, D., Cootes, T., Scott, I.: A multi-stage approach to facial feature detection. In: 15th BMVC, pp. 277–286 (2004)

3. Cristinacce, D., Cootes, T.: Facial Feature Detection and Tracking with Automatic Template Selection. In: 7th FG (2006)
4. Ding, L., Martinez, A.: Precise detailed detection of faces and facial features. In: CVPR, pp. 1–7 (2008)
5. Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision 30(2), 79–116 (1998)
6. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999)
7. Lowe, D.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
8. Martinez, A., Benavente, R.: The AR face database. Tech. rep., CVC (1998)
9. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision 60(1), 63–86 (2004)
10. Ni, W., Caplier, A.: Newton optimization based Congealing for facial image alignment. In: ICIP (2011)
11. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. In: CVPR (2010)
12. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
13. Vu, N., Caplier, A.: Face Recognition with Patterns of Oriented Edge Magnitudes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 313–326. Springer, Heidelberg (2010)
14. Vukadinovic, D., Pantic, M.: Fully automatic facial feature point detection using Gabor feature based boosted classifiers. In: IEEE International Conference on Systems, Man and Cybernetics (2005)

Appendix: Basic Knowledge of Game Theory

Let $O = \{1, 2, ..., n\}$ be the *pure strategies* set and $C = (C_{ij})$ stands for the *payoff matrix*. A *mixed strategy* is a probability distribution $\boldsymbol{x} = (x_1, ..., x_n)^T$ over O, and belongs to $\Delta = \{\boldsymbol{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1 \textit{ and } x_i \geq 0, i = 1, ..., n\}$. The *support* of a mixed strategy $\sigma(\boldsymbol{x})$ defines the set of elements with non-zero probability.

If a player plays pure strategy i against a mixed strategy x, the payoff will be $(C\boldsymbol{x})_i = \sum_j c_{ij}x_j$. Hence, the expected payoff by adopting a mixed strategy $\boldsymbol{y}$ against $\boldsymbol{x}$ is $\boldsymbol{y}^TC\boldsymbol{x}$. The *best replies* against a mixed strategy x are $\beta(\boldsymbol{x}) = \{y \in \Delta : y^TCx = max_z z^TCx\}$. A mixed strategy $\boldsymbol{x}$ is a *Nash equilibrium* if it is the best reply to itself, i.e. $\forall \boldsymbol{y} \in \Delta, \boldsymbol{y}^TC\boldsymbol{x} \leq \boldsymbol{x}^TC\boldsymbol{x}$. A strategy is said to be an *evolutionary stable strategy*(ESS) if it is a Nash equilibrium and $\forall \boldsymbol{y} \in \Delta, \boldsymbol{x}^TC\boldsymbol{x} = \boldsymbol{y}^TC\boldsymbol{x} \Rightarrow \boldsymbol{x}^TC\boldsymbol{y} > \boldsymbol{y}^TC\boldsymbol{y}$. An ESS can be estimated iteratively by:

$$\boldsymbol{x}_i(t+1) = x_i(t)\frac{(C\boldsymbol{x}(t))_i}{\boldsymbol{x}(t)^TC\boldsymbol{x}(t)} \tag{5}$$

where t is the number of iteration.

Extraction of Teeth Shapes from Orthopantomograms for Forensic Human Identification

Dariusz Frejlichowski and Robert Wanat

West Pomeranian University of Technology, Szczecin,
Faculty of Computer Science and Information Technology,
Zolnierska 52, 71-210, Szczecin, Poland
{dfrejlichowski,rwanat}@wi.zut.edu.pl

Abstract. Dental biometrics are commonly used in the process of forensic human identification. In order to automatize the identification, a method of extracting and comparing dental features from digital radiograms was developed by the creators of Automated Dental Identification System (ADIS). In this paper, a novel method of extracting teeth shapes from extraoral radiograms, known as orthopantomograms, is proposed. The method segments the image using the watershed algorithm and classifies every resulting region as belonging either to the tooth or the background. Example results obtained by means of the proposed method are also presented.

Keywords: dental biometrics, image processing, forensic identification, ADIS.

1 Introduction

Forensic human identification is the process of establishing the identity of an individual, to be later used in judicial proceedings. Various biometrics are applied for this purpose, e.g. fingerprints, DNA or dental records. After the successful implementation of the Automatic Fingerprint Identification System (AFIS), other biometrics have received similar scrutiny from researchers in the hope of automatization of the process of identification. Existing dental identification systems, such as WinID, compare dental records previously codified by an expert. Another approach, presented by the creators of the Automated Dental Identification System (ADIS), consists in the automatic extraction of dental biometrics from a radiographic image, thus minimizing the participation of an expert, which results in speeding up the whole process ([1]). Whereas systems like WinID utilize dental works as a basis for the comparison, teeth contour shapes extracted from radiograms are used in ADIS in the process of matching ([2]).

In this paper, a method for extracting the shapes of teeth from orthopantomograms is proposed. As opposed to intraoral radiograms, which are taken with

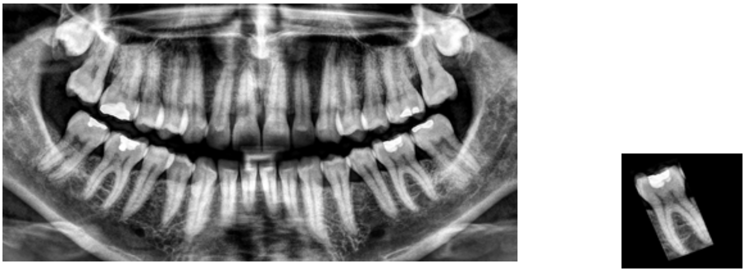
A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 65–72, 2011.

the film situated inside the patient's mouth, showing only a fragment of the dentition, pantomograms are taken with the film outside the patient's mouth (extraoral imaging) and show the full dentition on a single image. This type of radiogram is considered to be of poorer quality than intraoral images, because of the relatively lower dose of radiation used in the process of developing the film. The representation of semi-circular geometry of the jaw on a 2-dimensional image also results in neighboring teeth occluding with each other more frequently than in intraoral images.

Before the teeth contours can be extracted, firstly an image needs to be contrast-enhanced and segmented into areas containing only a single tooth. The image enhancement method preceding the algorithm presented in this paper consists in decomposing the radiogram into a set of smaller images containing a subset of information from the original image, called the Laplacian pyramid ([3]). The decomposed images, also known as the pyramid layers, are then filtered and recomposed, creating as a result an enhanced version of the original image.

The segmentation method proposed in the paper is a combination of an existing method created for intraoral images and a new approach utilizing dental features easily localizable on a pantomogram. After using the integral projections method described in [4] to determine a line separating the upper and lower jaw, the resulting curve is translated vertically in order to find a position where it passes through the soft tissue in the center of a tooth known as dental pulp. Once the location of such a line is established for both upper and lower jaw, a new image is created by combining a range filtered original image and the negative of the original image, which helps in emphasizing the gaps between teeth. Lastly, the values of the pixels on the new image through which the dental pulp curve passes are grouped in an array, which is then searched for sharp spikes in values. These spikes occur in points where the original image is dark (negative component) and surrounded by pixels of high and low intensity values (range filtering component), indicating a gap between teeth. After finding all the necessary gaps, for each molar tooth an additional search is performed in order to find the slope of the line separating neighboring teeth. This is caused by the fact that molars have a higher probability of malalignment, which makes the use of a vertical line passing through the detected gap between teeth insufficient to properly separate them. To determine the location of the second point, a greedy algorithm is used, moving iteratively one pixel vertically towards the root of the tooth and selecting the darkest pixel in its horizontal vicinity. After the amount of iterations equal to an average length of a tooth on the image, the last selected point becomes the second segmentation point and a line passing through the aforementioned gap position and this second point becomes the segmentation line.

All pantomograms presented in this paper are used courtesy of Pomeranian University of Medicine in Szczecin, Poland. A sample pantomogram with a single segmented tooth is displayed on Fig. 1.



(a) (b)

Fig. 1. A sample pantomogram (1(a)) and a single segment from which a tooth shape is extracted (1(b))

2 Methods Developed for Intraoral Images

Several approaches for extracting the shapes of teeth from dental radiograms have been presented in scientific literature so far. These algorithms are usually developed with intraoral images in mind and do not address the problems typical for pantomograms. The first method, described in [5], utilizes the active contour model (so-called 'snakes') to extract the shapes from a previously segmented image. Active contours, first described in [6], are a model of parametrized curves that, while under the influence of an external driving force (usually derived from the image), attempt to minimize the sum of their external and internal energy by moving in the spatial domain in accordance with limitations imposed on their shape. The external driving force needs to be chosen in such a way that the function takes low values in the points belonging to the contour and high values outside of the contour; in [5] the assumed external energy function is represented by the formula:

$$E_{ext} = -|\nabla[G_\sigma(x,y) * I(x,y)]|^2, \tag{1}$$

where G_σ is a Gaussian with the standard deviation σ and ∇ is a Laplacian. The Laplacian of Gaussian is commonly utilized in image processing for edge detection if the image background is noisy. It results in the reduction of false edge detection.

A modified approach based on snakes — the active contour without edges, was used in [7]. Instead of minimizing the energy of the contour, a model fit error term was applied to guide the contour. Thus, the minimized energy function becomes:

$$E(C) = \int_{inside(C)} |I_0(x,y) - c_1|^2 dxdy + \int_{outside(C)} |I_0(x,y) - c_2|^2 dxdy, \tag{2}$$

where C is the current contour, I_0 is the original image, c_1 and c_2 are the mean intensities of the pixels inside and outside C, respectively. The energy of a given contour is minimized when the difference between c_1 and c_2 is maximized, i.e.

when the contour C contains an homogeneous area with high-intensity pixels and the area outside C contains low-intensity pixels. The contour shape can still be controlled by the limitations imposed on its curvature.

Another method, applied by Chen and Jain ([8]), consists in the use of active shape models in the process of dental shape extraction. Presented in [9] active shape models are "used to extract eigen-shapes from aligned training tooth contours, which include tooth contours and their scaled and rotated variations" ([8]). After the resulting contour and the tooth on the image are aligned, splines are used to represent the extracted shape.

The last described approach was presented in [4]. In this method, it is assumed that the center of the crown (the uncovered part of the tooth) is located in the segment. A radial scan is then performed with the angle ranging between 0 and π, from the crown center to the edge of the image. Along every scan line, a single point with the highest bayesian probability of belonging to the contour (determined by its intensity and the intensity of the next pixel in the given direction) is accepted and connected to the previously selected point to form the crown contour. To extract the shape of the root (part of the tooth covered by gums) an iterative algorithm is used, starting from both ends of the crown contours, i.e. points selected for the angles 0 and π, moving towards the horizontal edge of the image and choosing a single point in the horizontal vicinity maximizing the intensity difference between the points inside and outside the contour. Which points are considered to be inside depends which side of root's shape is being extracted, e.g. for the left side of the root the pixels to the right of the selected contour point are considered to be inside the contour. When the horizontal edge of the image is reached on both sides of the tooth, the contour is complete.

3 Description of the Proposed Method

While the approaches presented in the previous section provide good results for intraoral images, frequent occlusions appearing in pantomograms require a different solution, one that does not require high contrast between the pixels of the tooth and the background. For instance, if two neighboring teeth occlude with each other, their edes will have higher intensities than the pixels in the center of the tooth.

It is assumed that before the proposed algorithm starts, a detection step is performed to decide whether a tooth is present in a given segment of the radiogram. Then, the image is morphologically opened in order to reduce the noise and to create larger areas of similar intensity range. Afterwards, the image is entropy filtered in order to detect the edges of similarly colored areas and then segmented into small fragments using the watershed method. Because the image was previously morphologically opened, the resulting segments are larger than on a watershed-segmented original image, thus reducing the number of segments and, as a result, speeding up the later stages of the presented method. The size of the resulting segments depends on the structuring element used in morphological opening — the larger the structuring element, the larger the segments on the image.

For every thus achieved segment, a set of features is calculated from the original image: segment's centroid, normalized mean value of the intensities of its pixels and the normalized vertical distance from the centroid to the curve separating the upper and lower jaw. The Euclidean distance between the centroids is also calculated and for each segment, 50 segments with the closest centroids are chosen to calculate the distinction of its mean intensity. The distinction value of segment i is calculated as:

$$D(i) = \sum_{j=1}^{N} \max(\bar{I}(i) - \bar{I}(j), 0), \tag{3}$$

where $\bar{I}(i)$ and $\bar{I}(j)$ are respectively the mean intensities of segments i and j. The distinction values are later normalized and are used as an indication whether the chosen segment is brighter than its surrounding segments. Finally, a mean intensity is calculated for all the non-zero pixels to serve as a reference of the image exposure.

To determine which segments belong to the tooth, a fitness function is calculated. The values used in the calculation of the fitness function depend on the type of tooth being segmented: for the first two teeth from the center of the jaw (incissors) only the distinction function and vertical distance from the curve separating upper and lower jaws are used, with the weights of 0.7 and 0.3 respectively, for all the other teeth the mean intensity is added, with the weights of 0.4 (mean intensity), 0.4 (distinction function) and 0.2 (vertical distance from the curve separating jaws). Once every region has an assigned fitness function, the segments with the fitness above a preselected threshold are considered to belong to the tooth and have their pixel values set to 1, and all other regions are excluded and set to 0. The thresholds used in this study were: 0.4 for the incissors, 0.5 for the third and fourth tooth from the center (the canine and the first premolar) and the mean intensity of the whole image multiplied by 0.8 for all the other teeth. All the values were established experimentally and did not require scaling for different images, as all the radiograms used in this study were the same size and subject to the same contrast enhancement.

Finally, after all rejected regions are set to 0, the remaining regions are morphologically dilated in order to remove the borders between them. Exterior boundaries of the objects on the resulting image are then traced and the longest contour that also lies close to the jaws separating curve is selected as the tooth contour. In order to smooth the contour, all border pixel positions are then Gaussian filtered. The resulting list of points is the final contour of the tooth. The result of consecutive stages of the proposed algorithm on the sample tooth is presented on Fig. 2.

4 Experimental Results

The algorithm was tested on a database containing 218 digital pantomograms belonging to 176 different people and the result for the exemplary tooth is presented on Fig. 3(a). A contour of the same tooth extracted using the active

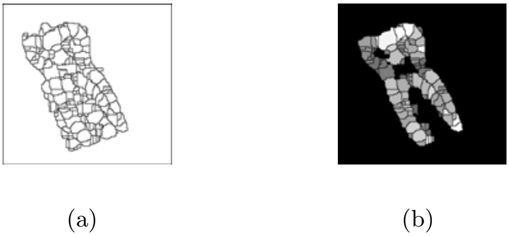

(a) (b)

Fig. 2. Results of consecutive stages of the algorithm: 2(a) watershed segmentation, 2(b) regions remaining after thresholding, with brightness equal to their fitness value

contours without edges ([7]) can be seen on Fig. 3(b). The contour shown on 3(b) was achieved after 650 iterations and a further increase of this parameter did not result in the inclusion of the area of dental pulp inside the contour. This is caused by the fact that the dental pulp is a soft tissue and appears darker than the surrounding tooth on the radiogram. Because of this, the inclusion of the area inside the contour results in the increase of its energy. This problem is also evident on Fig. 2(b), but it has no impact on the final result.

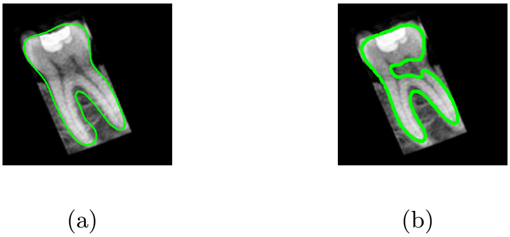
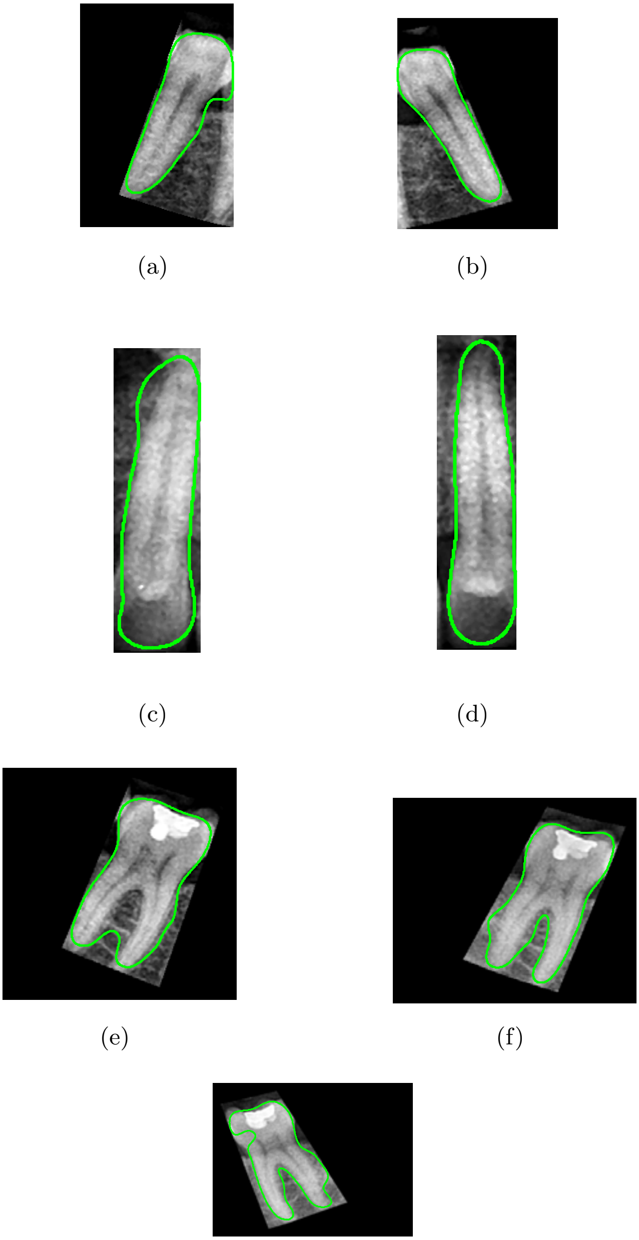

(a) (b)

Fig. 3. A comparison of the results of shape extraction using the proposed method (3(a)) and active contour without edges ([7], 3(b))

More test results of the presented method are shown on Fig. 4. The extracted contours are repeatable across different images of the same person, as seen on Fig. 3(a) and Fig. 4(g). The incorrect results are often caused by the bone formation known as trabecula. Other problems might be caused by incorrect segmentation that results in a fragment of neighboring tooth visible on the image segment, like on Fig. 4(a). Incorrectly excluded regions do not affect the resulting contour as they do in the case of the active contours without edges (Fig. 4(g)).

It should be noted that because the extracted shapes are later used for identification, the best way to compare the presented methods is to assess their influence on the successful retrieval rate. Repeatability in similar conditions is more important than the correctness of the result, because two incorrectly extracted shapes (for example because of the presence of dental braces on the image) could still lead to a successful retrieval, if the error is similarly reflected on both of them. Such a comparison should be the basis of a future study.



(a) (b)

(c) (d)

(e) (f)

(g)

Fig. 4. Exemplary results of the proposed method. Teeth on figures 4(a)-4(e) come from the image shown on Fig. 1(a). Teeth on figures 4(f) and 4(g) are the same teeth as those shown on 4(e) and 3(a), respectively. They were extracted from an another pantomogram belonging to the same person.

5 Conclusions and Future Work

In this paper, a novel method of extracting teeth contours from orthopantomograms was presented. The method works fully automatically and provides acceptable results, which can be later used in the process of forensic human identification. The proposed algorithm was compared with another approach — active contour model without edges.

Further development of the method could include the use of artificial neural networks instead of the fitness function in the process of deciding which regions on the image belong to the tooth. Another improvement could be achieved if a post-processing step is added, removing protrusions from the contour that do not fit the general shape of a tooth.

References

1. Fahmy, G., et al.: Toward an Automated Dental Identification System (ADIS). In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 789–796. Springer, Heidelberg (2004)
2. Nassar, D., Ammar, H.H.: A Prototype Automated Dental Identification System (ADIS). In: Proc. of the 2003 Annual National Conference on Digital Government Research, pp. 1–4 (2003)
3. Frejlichowski, D., Wanat, R.: Application of the Laplacian Pyramid Decomposition to the Enhancement of Digital Dental Radiographic Images for the Automatic Person Identification. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6112, pp. 151–160. Springer, Heidelberg (2010)
4. Jain, A.K., Chen, H.: Matching of Dental X-ray Images for Human Identification. Pattern Recognition 37(7), 1519–1532 (2004)
5. Zhou, J., Abdel-Mottaleb, M.: A Content-based System for Human Identification Based on Bitewing Dental X-ray Images. Pattern Recognition 38(11), 2132–2142 (2005)
6. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active Contour Models. International Journal of Computer Vision 1(4), 321–331 (1988)
7. Shah, S., Abaza, A., Ross, A., Ammar, H.: Automatic Tooth Segmentation Using Active Contour Without Edges. In: Proc. of Biometrics Symposium, pp. 1–6 (2006)
8. Chen, H., Jain, A.K.: Automatic Forensic Dental Identification. In: Jain, A.K., Flynn, P., Ross, A.A. (eds.) Handbook of Biometrics, pp. 231–251 (2008)
9. Cootes, T.F., Taylor, C.J.: Active Shape Models — "Smart Snakes". In: Proc. of the British Machine Vision Conf., pp. 266–275 (1992)

Effects of JPEG XR Compression Settings on Iris Recognition Systems*

Kurt Horvath[1], Herbert Stögner[1], and Andreas Uhl[1,2,**]

[1] School of CEIT, Carinthia University of Applied Sciences, Austria
[2] Department of Computer Sciences, University of Salzburg, Austria
uhl@cosy.sbg.ac.at

Abstract. JPEG XR is considered as a lossy sample data compression scheme in the context of iris recognition techniques. It is shown that apart from low-bitrate scenarios, JPEG XR is competitive to the current standard JPEG2000 while exhibiting significantly lower computational demands.

1 Introduction

With the increasing usage of biometric systems the question arises naturally how to store and handle the acquired sensor data (denoted as sample data subsequently). In this context, the compression of these data may become imperative under certain circumstances due to the large amounts of data involved. Among other possibilities (e.g. like compressed template storage on IC cards and optional storage of (encrypted) reference data in template databases), compression technology is applied to sample data in distributed biometric systems, where the data acquisition stage is often dislocated from the feature extraction and matching stage (this is true for the enrolment phase as well as for authentication). In such environments the sample data have to be transferred via a network link to the respective location, often over wireless channels with low bandwidth and high latency. Therefore, a minimisation of the amount of data to be transferred is highly desirable, which is achieved by compressing the data before transmission and any further processing. As an alternative, the application of feature extraction before transmission looks promising due to the small size of template data but cannot be done under most circumstances due to the prohibitive computational demand of these operations (current sensor devices are typically far too weak to support this while compression can be done e.g. in dedicated low power hardware).

While current international standards define the application of JPEG2000 for lossy iris sample data compression, we focus in this paper on the corresponding application of the recent JPEG XR still image coding standard. We experimentally compare the achieved results to a JPEG2000 based (and therefore standard

* This work has been partially supported by the Austrian Science Fund, project no. L554-N15.

** Corresponding author.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 73–80, 2011.

conformant) environment. In particular, we investigate the effects of applying different settings concerning the use of the optional Photo Overlap Transform (POT) as a part of JPEG XR's Lapped Biorthogonal Transform (LBT) with respect to iris recognition accuracy. In Section 2, we review related standards and literature in the area of lossy iris sample data compression. Section 3 presents experiments where we first shortly review the four different iris recognition systems employed in this study. Subsequently, JPEG XR basics and the investigated transform settings are briefly explained. Experimental results comparing JPEG XR and JPEG2000 are shown with respect to PSNR (image quality), execution speed, and iris recognition accuracy in terms of EER. Section 4 concludes the paper.

2 Biometric Iris Sample Compression

During the last decade, several algorithms and standards for compressing image data relevant in biometric systems have evolved. The certainly most relevant one is the ISO/IEC 19794 standard on Biometric Data Interchange Formats, where in its former version (ISO/IEC 19794-6:2005), JPEG and JPEG2000 (and WSQ for fingerprints) were defined as admissible formats for lossy compression, whereas for lossless and nearly lossless compression JPEG-LS as defined in ISO/IEC 14495 was suggested. In the most recently published version (ISO/IEC FDIS 19794-6 as of August 2010), only JPEG2000 is included for lossy compression while the PNG format serves as lossless compressor. These formats have also been recommended for various application scenarios and standardised iris images (IREX records) by the NIST Iris Exchange (IREX `http://iris.nist.gov/irex/`) program.

The ANSI/NIST-ITL 1-2011 standard on "Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information" (2nd draft as of February 2011, former ANSI/NIST-ITL 1-2007) supports both PNG and JPEG2000 for the lossless case and JPEG2000 only for applications tolerating lossy compression.

In literature on compressing iris imagery, rectangular as well as polar iris sample data has been considered. With respect to employed compression technology, we find JPEG [1, 8], JPEG2000 [4, 1, 8], and other general purpose compression techniques [8] being investigated. Superior compression performance of JPEG2000 over JPEG is seen especially for low bitrates (thus confirming the choice of the above-referenced standards), however, for high and medium quality JPEG is found still to be competitive in terms of impacting recognition accuracy. Apart from applying the respective algorithms with their default settings and standard configurations, work has been done to optimise the compression algorithms to the application domain: For JPEG2000, we have proposed to invoke RoI coding for the iris texture area [3] whereas the removal of the image background before compression has also been suggested (i.e. parts of the image not being part of the eye like eye-lids are replaced by constant average gray [1]). For JPEG, we have demonstrated an optimisation of quantisation matrices to achieve better matching accuracy compared to the standard values for rectangular iris image data [7] as well as for polar iris images [6].

The recent JPEG XR standard has not yet been investigated in the context of biometric systems. It might represent an interesting alternative to JPEG2000 due to its simpler structure and less demanding implementations in terms of memory and CPU resources.

3 Experiments on Compressing Iris Sample Data

3.1 Iris Recognition and Iris Database

It is crucial to assess the effects of compressing iris samples using a set of different iris recognition schemes since it can be expected that different feature extraction strategies will react differently when being confronted with compression artefacts and reduced image quality in general.

Many iris recognition methods follow a quite common scheme close to the well known and commercially most successful approach by Daugman. In our pre-processing approach (following e.g. Ma et al. [9]) we assume the texture to be the area between the two almost concentric circles of the pupil and the outer iris. These two circles are found by contrast adjustment, followed by Canny edge detection and Hough transformation. After the circles are detected, unwrapping along polar coordinates is done to obtain a rectangular texture of the iris. In our case, we always re-sample the texture to a size of 512x64 pixels. Subsequently, features are extracted from this iris texture (which has also been termed polar iris image), we consider the following four techniques in this work:

1. A wavelet-based approach proposed by Ma et al. [9] is used to extract a bit-code. The texture is divided into N stripes to obtain N one-dimensional signals, each one averaged from the pixels of M adjacent rows. We used $N = 10$ and $M = 5$ for our 512x64 pixel textures (only the 50 rows close to the pupil are used from the 64 rows, as suggested in [9]). A dyadic wavelet transform is then performed on each of the resulting 10 signals, and two fixed subbands are selected from each transform. This leads to a total of 20 subbands. In each subband we then locate all local minima and maxima above some threshold, and write a bitcode alternating between 0 and 1 at each extreme point. Using 512 bits per signal, the final code is then 512x20 bit. Matching different codes is done by computing the Hamming Distance.
2. Again restricting the texture to the same $N = 10$ stripes as described before, we use a custom C implementation similar to Libor Masek's Matlab implementation[1] of a 1-D version of the Daugman iris recognition algorithm as the second feature extraction technique. A row-wise convolution with a complex Log-Gabor filter is performed on the texture pixels. The phase angle of the resulting complex value for each pixel is discretized into 2 bits. Those 2 bit of phase information are used to generate a binary code, which therefore is 512x20 bit (again, Hamming Distance can be used for similarity determination).

[1] http://www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html

3. The third algorithm has been proposed by Ko et al. [5]. Here feature extraction is performed by applying cumulative-sum-based change analysis. The algorithm discards parts of the iris texture, from the right side [45^o to 315^o] and the left side [135^o to 225^o], since the top and bottom of the iris are often hidden by eyelashes or eyelids. Subsequently, the resulting texture is divided into basic cell regions (these cell regions are of size 8×3 pixels). For each basic cell region an average gray scale value is calculated. Then basic cell regions are grouped horizontally and vertically. It is recommended that one group should consist of five basic cell regions. Finally, cumulative sums over each group are calculated to generate an iris-code. If cumulative sums are on an upward slope or on a downward slope these are encoded with 1s and 2s, respectively, otherwise 0s are assigned to the code. In order to obtain a binary feature vector (to enable Hamming Distance computation for comparison) we rearrange the resulting iris-code such that the first half contains all upward slopes and the second half contains all downward slopes. With respect to the above settings the final iris-code consists of 2400 bits.
4. Finally, we employ the feature extraction algorithm of Zhu et al. [10] which applies a 2-D wavelet transform to the polar image first. Subsequently, first order statistical measures are computed from the wavelet subbands (i.e. mean and variance) and concatenated into a feature vector. The similarity between two of these real-valued feature vectors is determined by computing the corresponding l^2-Norm.

The following dataset is used in the experiments:

CASIAv3 Interval database[2] consists of NIR images with 320×280 pixels in 8 bit grayscale .jpeg format (high quality) of 249 persons, where for many persons both eyes are available which leads to 391 (image) classes overall.

For intra-class matches (genuine user matches), we consider all possible template pairs for each class (overall 8882 matches), while for inter-class matches (impostor matches) the first two templates of the first person are matched against all templates of the other classes (overall 2601 matches).

3.2 Compression Techniques: JPEG XR and JPEG2000

Originally developed by Microsoft and termed “HD Photo”, JPEG XR got standardized by ITU-T and ISO in 2009 [2], which makes it the most recent still image coding standard. The original scope was to develop a coding scheme targeting “extended range” applications which involves higher bit-depths as currently supported. However, much more than 10 years after JPEG2000 development and 10 years after its standardisation it seems to be reasonable to look for a new coding standard to eventually employ “lessons learnt” in JPEG2000 standardisation. In particular, the focus is on a simpler scheme which should offer only the amount of scalability actually required for most applications (as opposed

[2] http://www.cbsr.ia.ac.cn/IrisDatabase.htm/

to JPEG2000 which is a rather complex scheme offering almost unconstraint scalability). JPEG XR shares many properties with JPEG and JPEG2000 but exhibits also elements of the recent H.264 video standardisation [2].

JPEG XR is a transform coding scheme showing the classical three-stage design: transform, quantisation, and entropy encoding. JPEG XR supports lossless to lossy compression of up to 32 bits per colour channel. The transform operates on macroblocks consisting of 16 (arranged in 4 by 4) 4 × 4 pixel blocks. The first stage of the integer-based transform allowing for perfect reconstruction is applied to all 4×4 pixel blocks of a macroblock. Subsequently, the resulting coefficients are partitioned into 240 "high pass (HP) coefficients" and 16 coefficients corresponding to the lowest frequency in each block. The latter are aggregated into a square data layout (4 x 4 coefficients) onto which the transform is applied for a second time. The result are 15 "low pass (LP) coefficients" and a single "DC" coefficient (per macroblock). It is interesting to note that the concept of recursively applying a filtering operation is "borrowed" from the wavelet transform. Obviously, this also corresponds to three scalability layers: DC, LP, and HP coefficients, similar to the scans being built in the spectral selection JPEG progressive mode.

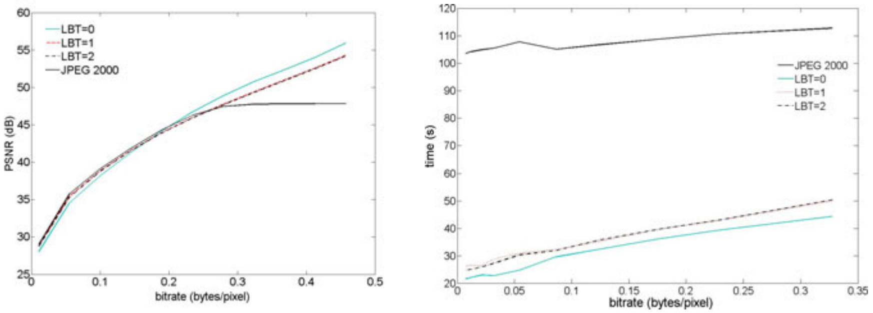
In fact, the transform used in JPEG XR is more complicated as compared to JPEG, it is a so-called "two-stage lapped biorthogonal transform (LBT)" which is actually composed of two distinct transforms: The Photo Core Transform (PCT) and the Photo Overlap Transform (POT). The PCT is similar to the widely used DCT and exploits spatial correlation within the 4 x 4 pixels block, however, it suffers from the inability to exploit inter-block correlations due to its small support and from blocking artifacts at low bitrates. The POT is designed to exploit correlations across block boundaries as well as mitigate blocking artifacts.

Each stage of the transform can be viewed as a flexible concatenation of POT and PCT since the POT is functionally independent of the PCT and can be switched on or off, as chosen by the encoder (this is signalled by the encoder in the bitstream). There are three options: disabled for both PCT stages, enabled for the first PCT stage but disabled for the second PCT stage, or enabled for both PCT stages.

Since our experiments are focused on the evaluation of those three options concerning POT employment, we do not describe the subsequent JPEG XR stages in the following, please consult the standard or related publications with respect to this issue [2]. For experimentation, we use the official JPEG-XR reference software 1.8 (as of September 2009) and for JPEG2000 compression, imagemagick 8.6.6.0.4-3 (employing libJASPER 1.900.1-7+b1) is used with standard settings.

3.3 Experimental Results

For enabling a fair comparison in the experiments, the same bitrate has to be set in JPEG XR and JPEG2000. While this is straightforward in JPEG2000, JPEG XR suffers from the same weakness as JPEG being unable to explicitly specify a target bitrate. Therefore we have employed a wrapper-program, continuously adapting the JPEG XR quantisation factors (set to identical values for DC, LP,



(a) PSNR (b) Speed

Fig. 1. Comparing JPEG XR and JPEG2000 in terms of PSNR and Execution Speed

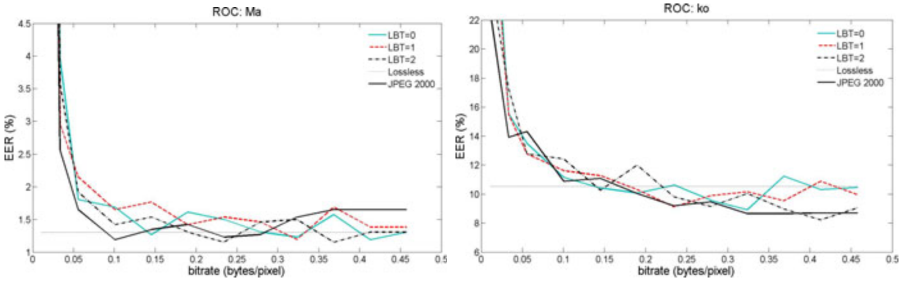
and HP band as used in the default settings) to achieve a certain target bitrate (given in bytes per pixel bpp).

In Fig. 1.a we compare PSNR performance averaged over all images in the considered dataset. Up to 0.2 bpp, JPEG2000 provides the highest values. In this bitrate range, applying no POT (LBT= 0) clearly gives the worst results (PSNR is about 1dB reduced as compared to JPEG2000). Applying POT for the first (LBT= 1) or both transform stages (LBT= 2) leads to almost identical results across the entire bitrate range, up to 0.2bpp PSNR quality is only slightly below that of JPEG2000.

The situation is different for higher bitrates. JPEG2000 saturates from 0.3bpp upwards due to the employed irreversible 9/7 transform and is clearly outperformed by all JPEG XR settings. Interestingly, for bitrates larger than 0.2bpp, applying no POT gives the best PSNR values, which is explained by the fact that POT application is targeted to optimise data for human perception but not for numerical error minimisation.

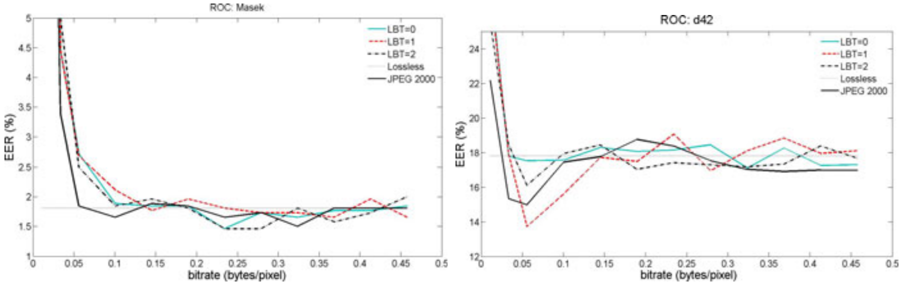
Fig. 1.b shows a comparison of execution timings for compressing the entire dataset. We note that depending on the target bitrate considered, JPEG XR is faster by a factor of 2-5 as compared to JPEG2000 (target bitrate optimisation is disabled for this evaluation). This result underlines that JPEG XR could be an interesting alternative to JPEG2000 in biometric environments, especially in cases with limited CPU resources at the compressing site.

In the following, we will investigate the impact compression of one template involved in matching has on the recognition performance of the four iris recognition systems considered (e.g., the sample data acquired by the sensor is compressed and sent to the feature extraction / matching site). For this purpose, we plot equal error rate (EER, on the vertical axis) for applying compression in an entire range of target bitrates (in bpp, on the horizontal axis) and compare JPEG2000 to the three JPEG XR POT employment variants. For reference, also the "Lossless" case (i.e. recognition accuracy in EER without any compression applied) is indicated as a horizontal line in Figs. 2 and 3.



(a) Ma et al. (b) Ko et al.

Fig. 2. EER for varying bitrates and JEPG XR compression settings



(a) Masek (b) Zhu et al.

Fig. 3. EER for varying bitrates and JEPG XR compression settings

For the algorithms of Ma and Masek, JPEG2000 provides the lowest (i.e. best) EER up to a bitrate of 0.15bpp, while for the other two recognition algorithms, no clear tendency can be observed. In particular, for no algorithm there is a clear indication whether application of POT would be beneficial or not. Further, it is interesting to see that for some algorithms and bitranges, the results involving a compressed template are superior to the uncompressed case (e.g. Ko and Masek for bitrates $>$ 0.2bpp, Zhu for bitrates between 0.04 and 0.15). This can be explained by the fact that compression acts as a denoising filter and has been observed in earlier studies as well [6].

What is especially interesting to observe, is that PSNR behaviour as shown in Fig. 1.a does not directly propagate to recognition accuracy. While the better PSNR behaviour of JPEG2000 at low bitrates is at least reflected by the results of two algorithms, we do not find any superiority of JPEG XR for higher bitrates. On the other hand it is interesting to see that except for two recognition algorithms at low bitrates, JPEG XR compressed sample data perform almost equivalent to JPEG2000 compressed one. Given the significantly reduced computational demand as shown in Fig. 1.b, JPEG XR can be considered a promising alternative to JPEG2000 in this application scenario and should be considered in future standardisation efforts in the area.

4 Conclusion

We have found that in the context of biometric systems, JPEG XR can be an interesting alternative to the current standard JPEG2000, especially due to its significantly lower computational demand. A minor decrease in EER as compared to JPEG2000 can be seen only for lower bitrates for two out of four iris recognition systems only. For most iris recognition scenarios, compression with JPEG XR has been identified to be quite competitive to compression with JPEG2000.

References

[1] Daugman, J., Downing, C.: Effect of severe image compression on iris recognition performance. IEEE Transactions on Information Forensics and Security 3(1), 52–61 (2008)

[2] Dufaux, F., Sullivan, G.J., Ebrahimi, T.: The JPEG XR image coding standard. IEEE Signal Processing Magazine 26(6), 195–199 (2009)

[3] Hämmerle-Uhl, J., Prähauser, C., Starzacher, T., Uhl, A.: Improving compressed iris recognition accuracy using JPEG2000 RoI coding. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1102–1111. Springer, Heidelberg (2009)

[4] Ives, R.W., Broussard, R.P., Kennell, L.R., Soldan, D.L.: Effects of image compression on iris recognition system performance. Journal of Electronic Imaging 17, 11015 (2008), doi:10.1117/1.2891313

[5] Ko, J.-G., Gil, Y.-H., Yoo, J.-H., Chung, K.-I.: A novel and efficient feature extraction method for iris recognition. ETRI Journal 29(3), 399–401 (2007)

[6] Konrad, M., Stögner, H., Uhl, A.: Custom design of JPEG quantization tables for compressing iris polar images to improve recognition accuracy. In: Tistarelli, M., Nixon, M.S. (eds.) ICB 2009. LNCS, vol. 5558, pp. 1091–1101. Springer, Heidelberg (2009)

[7] Kostmajer, G.S., Stögner, H., Uhl, A.: Custom JPEG quantization for improved iris recognition accuracy. In: Gritzalis, D., Lopez, J. (eds.) SEC 2009. IFIP AICT, vol. 297, pp. 76–86. Springer, Heidelberg (2009)

[8] Matschitsch, S., Tschinder, M., Uhl, A.: Comparison of compression algorithms' impact on iris recognition accuracy. In: Lee, S.-W., Li, S.Z. (eds.) ICB 2007. LNCS, vol. 4642, pp. 232–241. Springer, Heidelberg (2007)

[9] Ma, L., Tan, T., Wang, Y., Zhang, D.: Efficient iris recognition by characterizing key local variations. IEEE Transactions on Image Processing 13(6), 739–750 (2004)

[10] Zhu, Y., Tan, T., Wang, Y.: Biometric personal identification based on iris patterns. In: Proceedings of the 15th International Conference on Pattern Recognition (ICPR 2000), vol. 2, pp. 2801–2804. IEEE Computer Society, Los Alamitos (2000)

A Recursive Sparse Blind Source Separation Method for Nonnegative and Correlated Data in NMR Spectroscopy

Yuanchang Sun and Jack Xin

Math Department, Univ of California Irvine, Irvine, CA 92697, USA
yuanchas@uci.edu, jxin@math.uci.edu

Abstract. Motivated by the nuclear magnetic resonance (NMR) spectroscopy of biofluids (urine and blood serum), we present a recursive blind source separation (rBSS) method for nonnegative and correlated data. A major approach to non-negative BSS relies on a strict non-overlap condition (also known as the pixel purity assumption in hyperspectral imaging) of source signals which is not always guaranteed in the NMR spectra of chemical compounds. A new dominant interval condition is proposed. Each source signal dominates some of the other source signals in a hierarchical manner. The rBSS method then reduces the BSS problem into a series of sub-BSS problems by a combination of data clustering, linear programming, and successive elimination of variables. In each sub-BSS problem, an ℓ_1 minimization problem is formulated for recovering the source signals in a sparse transformed domain. The method is substantiated by NMR data.

Keywords: NMR spectroscopy, non-negative correlated sources, recursive blind separation.

1 Introduction

Blind source separation (BSS) aims to recover source signals from their mixtures without detailed knowledge of the mixing process. Nonnegative BSS has received much attention in various fields lately, such as image processing, analytical chemistry, metabolic fingerprinting, and disease diagnosis [1, 2, 5–8, 10–14] where nonnegative constraints are imposed on the mixing process and source signals. The nonnegative BSS problem is defined by the following matrix model:

$$X = A\,S, \quad \text{with} \quad A_{ij} \geq 0,\;\; S_{ij} \geq 0, \tag{1.1}$$

where $X \in \mathbb{R}^{m\times p}$ is the mixture matrix containing known mixture signals as its rows, $S \in \mathbb{R}^{n\times p}$ is the unknown source matrix, $A \in \mathbb{R}^{m\times n}$ is the unknown mixing matrix. The dimensions of the matrices are expressed in terms of three numbers: (1) p is the number of available samples, (2) m is the number of mixture signals, and (3) n is the number of source signals. Both X and S are sampled functions of an acquisition variable (time, frequency, position, or wavenumber). The problem

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 81–88, 2011.

is to estimate nonnegative A and S from X, also known as nonnegative matrix factorization (NMF [5]).

Naanaa and Nuzillard (NN) proposed a nonnegative BSS method [6] based on the sparseness assumption (NNA) that the source signals be strictly non-overlapping at some locations of acquisition variable. Each source signal must have a stand-alone peak where other sources are strictly zero. Such a strict sparseness condition leads to a dramatic mathematical simplification of a general *nonconvex* NMF problem (1.1). Geometrically speaking, the problem of finding the mixing matrix A reduces to the identification of a minimal cone containing the column vectors of X. The latter can be done by linear programming. Similar assumption and geometric construction were known earlier [2, 12] in blind hyper-spectral unmixing. The analogue of NNA is called pixel purity assumption. The resulting geometric (cone) method is the so called N-findr [12]. However, certain class of NMR data may not satisfy NNA as seen in the following two examples.

Example 1: Consider the NMR spectra of two chemical compounds β-sitosterol and menthol in Fig. 1. The β-sitosterol (blue) has a stand-alone peak (circled) however menthol (red) does not have such a peak. Hence NNA does not hold. However, β-sitosterol (blue) has a dominant interval ($[420, 600]$) over menthol (red), though spectral overlap occurs on $[0, 420]$.

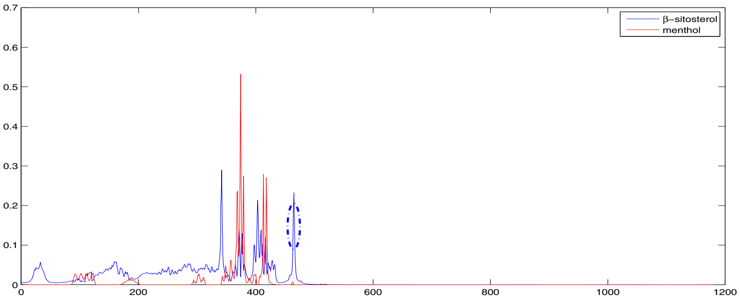


Fig. 1. NMR spectra of two chemical compounds. In the circled region, β-sitosterol (blue) has a stand-alone peak, while menthol (red) does not have such region.

Example 2: The data in Fig. 2 are from NMR spectroscopy of urine and blood serum. The complicated NMR spectra contain both wide-peak source signals and narrow-peak source signals. The blood serum has constituents with wide spectral peaks which overlap others over the whole acquisition region. The urine NMR spectrum is similar. NNA does not hold for this type of data.

The above two examples show that new BSS methods should be developed for these non-NNA signals where wide spectral peaks exist and violates NNA. Our work is motivated by NMR spectroscopy of biofluids such as urine and blood serum (example 2) which provide important information for metabolic fingerprinting and disease diagnosis [1, 11, 13, 14]. The main challenge of the

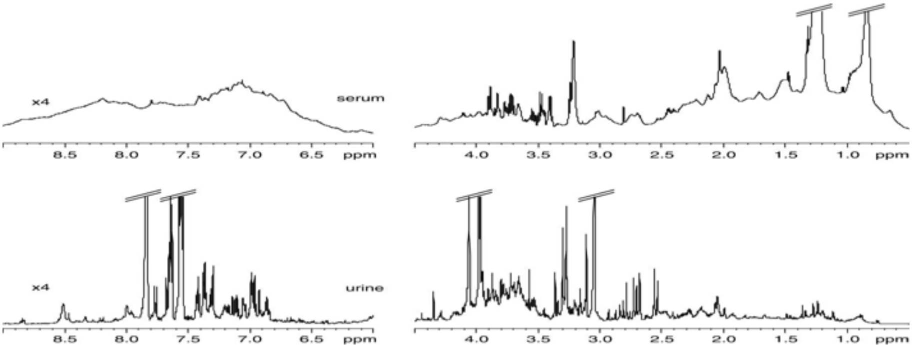


Fig. 2. Standard NMR spectra of serum and urine, showing representative structural complexity produced by multiple metabolite signals (plot from [1])

non-NNA problem is that the mixing matrix A cannot be recovered from data matrix X independently of S as in [6]. Our method breaks the source separation process into two stages. In the first stage, clustering and linear programming techniques are employed to recursively identify columns of the mixing matrix while simultaneously eliminating source variables. The first stage also serves to convexify the orginal non-convex matrix factorization problem because half of the unknowns are estimated. The second stage is to solve a sequence of ℓ_1 regularized convex optimization problems to recover the source signals.

The paper is organized as follows. In section 2, we propose a new condition on the source signals motivated by NMR spectroscopy data of biofluids. Then we present our recursive BSS method, and illustrate it with a numerical example. Section 3 is the conclusion. The following notations will be used throughout the paper. The notation A^j (X^j) denotes the j-th column of matrix A (X); S_j (X_j) is the j-th row of matrix S (X).

This work was partially supported by NSF-ADT grant DMS-0911277.

2 Source Assumption and Recursive Method

Let us consider the determined case ($m = n$) for simplicity. Each column in X of model (1.1) represents data collected at a particular value of the acquisition variable, and each row represents a mixture spectrum. Motivated by the NMR spectra of urine and blood serum, we propose here a more general and relaxed condition on the source signals. Rows $S_1, S_2, \ldots, S_n$ of S, i.e. the source signals, satisfy: for $i = 2, 3, \ldots, n$, the source signal S_i has a dominant interval over $S_{i-1}, \ldots, S_2, S_1$, while the other part of S_i may overlap with $S_{i-1}, \ldots, S_2, S_1$. More precisely, the source matrix S satisfies the hierarchical dominant interval (DI) condition:

- For each $k \in \{2, 3, \ldots, n\}$, there is a set $\mathcal{I}_k \subset \{1, 2, \ldots, p\}$ such that for each $l \in \mathcal{I}_k$ $s_{il} \gg s_{jl}, i = k, k+1, \ldots, n, j = 1, 2, \ldots, k-1$.

The recursive method consists of the backward and forward steps. In the backward step (elimination of variables from S_n to S_1), the original BSS problem is reduced to a series of smaller BSS problems. The DI condition implies that there are columns of X such that $X^k = s_{n,k}A^n + \sum_{i=1}^{n-1} o_{i,k}A^i$, where $s_{n,k}$ dominate $o_{i,k}(i = 1,\ldots,n-1)$, i.e., $s_{n,k} \gg o_{i,k}$. The A^n is found inside a cluster formed by these X^k's in $\mathbb{R}^n$. All X's column vectors form a set of points $\mathcal{P} = \{X^1, X^2, \ldots, X^p\}$ in n dimensional space. The convex hull of $\mathcal{P}$ is a polytope, $\mathcal{A}$ in $\mathbb{R}^n$. The frame $\mathcal{F}$ of these points is the set of extreme points of the convex hull. To determine if the element X^k of $\mathcal{P}$ constitutes an element of $\mathcal{F}$, the following constraint is examined: $\sum_{j=1,j\neq k}^{p} X^j\lambda_j = X^k,\ \lambda_j \geq 0,\ k = 1,\ldots,p.$ X^k belongs to $\mathcal{F}$ if it cannot be written as a linear combination of other points of $\mathcal{P}$. The above constraint is solved by linear programming. Among the elements of $\mathcal{F}$, A^n is the one attracting a cluster or most number of data points in its neighborhood.

After A^n is obtained, we reduce the model by eliminating S_n from X. Let row vectors of X be $X_1,\ldots,X_n$. Using A^n, we eliminate S_n by performing $X_i \to X_i - \frac{A_{in}}{A_{nn}}X_n$, $i = 1, 2, \ldots, n-1$. The reduced mixture matrix is: $X_{(1,2,\ldots,n-1)}$ consisting of rows: $X_1 - \frac{A_{1n}}{A_{nn}}X_n,\ X_2 - \frac{A_{2n}}{A_{nn}}X_n,\ \cdots,\ X_{n-1} - \frac{A_{n-1,n}}{A_{nn}}X_n$. which contains $n-1$ mixtures from source signals $S_1,\ldots,S_{n-1}$. The reduced BSS system is: $X_{(1,2,\ldots,n-1)} = A^{(1,2,\ldots,n-1)}\, S_{(1,2,\ldots,n-1)}$, where $A^{(1,2,\ldots,n-1)}$ is the mixing matrix of sources $S_1,\ldots,S_{n-1}$. In $X_{(1,2,\ldots,n-1)}$, the source S_{n-1} has dominant regions over other sources. So data clustering and linear programming can be used again to recover the mixing coefficients of S_{n-1} from $X_{(1,2,\ldots,n-1)}$. Then we reduce the mixture matrix further to $X_{(1,2,\ldots,n-2)}$ containing $S_1,\ldots,S_{n-2}$. The procedure iterates until the source S_1 is obtained.

In summary, the backward step not only extracts source signal S_1, but also generates reduced mixtures $X_{(1,2)}, X_{(1,2,3)}, \ldots, X_{(1,2,\ldots,k)}, \ldots, X_{(1,2,\ldots,n-1)}$. Although the original model (1.1) contains nonnegative A, S and X, the reduced mixtures and mixing matrices may have negative entries from variable eliminations. The geometric cone method is still applicable, only that the cone may not lie in the sector consisting of nonnegative vectors.

The forward step (recovery of sources from S_2 to S_n) is as follows. With S_1 recovered by the end of the forward step, we continue to separate out the source signals $S_2,\ldots,S_n$. We shall use sparseness property in a transformed domain. Analytical chemistry [4] says that an NMR spectrum is represented as a sum of symmetric, positive Lorentzian-shaped peaks. An NMR spectrum can be viewed as a linear convolution of Lorentzian kernel with some sparse function or $S = \hat{S} * \mathcal{L}(x,w)$, where $\mathcal{L}(x,w) = \frac{1}{\pi}\frac{\frac{1}{2}w}{x^2+(\frac{1}{2}w)^2}$, w specifies its width, and $\hat{S}$ is a sparse function. The sparsity under the Lorentzian kernel suggests an ℓ_1 minimization problem to recover the source signals. To estimate S_k $(k = 2,\ldots,n-1)$ with S_j $(j = 1,\cdots,k-1)$ known, we solve:

$$\min_{\substack{A^{(1,2\ldots,k-1)} \in \mathbb{R}^{k\times(k-1)}\ , \\ \hat{S}\in\mathbb{R}^{k\times p},\ \hat{S}\geq 0}} \mu\|\hat{S}\|_1 + \|X_{(1,2,\ldots,k)} - A^{(1,2\ldots,k-1)}\, S_{(1,2,\ldots,k-1)} - \hat{S} * \mathcal{L}(w_k)\|_2^2\ , \tag{2.1}$$

where $X_{(1,2,\ldots,k)} \in \mathbb{R}^{k\times p}$ is the mixture matrix that contains source $S_1,\ldots,S_k$, the columns of $A^{(1,2\ldots,k-1)} \in \mathbb{R}^{k\times(k-1)}$ correspond to the mixing coefficients of sources $S_1,\ldots,S_{k-1}$ in $X_{(1,2,\ldots,k)}$. The rows of $\hat{S} * \mathcal{L}(w_k)$ represent source S_k in $X_{(1,2,\ldots,k)}$, w_k is the peak width of S_k. Because (2.1) allows the constraint $A^{(1,2\ldots,k-1)}\, S_{(1,2,\ldots,k-1)} + \hat{S} * \mathcal{L}(w_k) = X_{(1,2,\ldots,k)}$ to be relaxed, it is applicable when the mixtures are contaminated by measurement errors. The l_2 norm in (2.1) models the unknown measurement error as Gaussian. When there is minimal measurement error, one assigns a tiny value to μ to heavily weigh the fidelity term. The widths w_k's may be estimated from peaks in the mixture. An upper bound often suffices. The convex optimization (2.1) is solved by a projected gradient descent method which converges to a global minimum. At this point, we have retrieved $S_1,\ldots,S_{n-1}$. Finally, we extract the last source signal S_n from the original mixture matrix X. We solve the ℓ_1 minimization problem:

$$\min_{\substack{0\le A^{(1,\ldots,n-1)}\in\mathbb{R}^{n\times(n-1)},\ , \\ \hat{S}\in\mathbb{R}^{n\times p},\ \hat{S}\ge 0}} \mu\|\hat{S}\|_1 + \|X - A^{(1,\ldots,n-1)}\, S_{(1,\ldots,n-1)} - \hat{S} * \mathcal{L}(w_n)\|_2^2\ , \quad (2.2)$$

where rows of $X \in \mathbb{R}^{n\times p}$ represent the n mixture signals, the columns of $A_{(1,\ldots,n-1)}$ correspond to the mixing coefficients of $S_1,\ldots,S_{n-1}$ in X. The rows of $\hat{S} * \mathcal{L}(w_n)$ are the multiples of S_n in X. Again, we use projected gradient descent approach to solve (2.2). The difference is that, in (2.1) the nonnegativity constraint is only imposed on the source signals, while in (2.2) both the mixing matrix and sources are required to be nonnegative.

A brief pseudo-code is: (B1) recover last column A_n of mixing matrix A by clustering columns of data matrix X; eliminate S_n from mixing equation. (B2) repeat (B1) and eliminate S_k ($k = n-1, \cdots, 2$) till S_1 is recovered. (F1) Recover

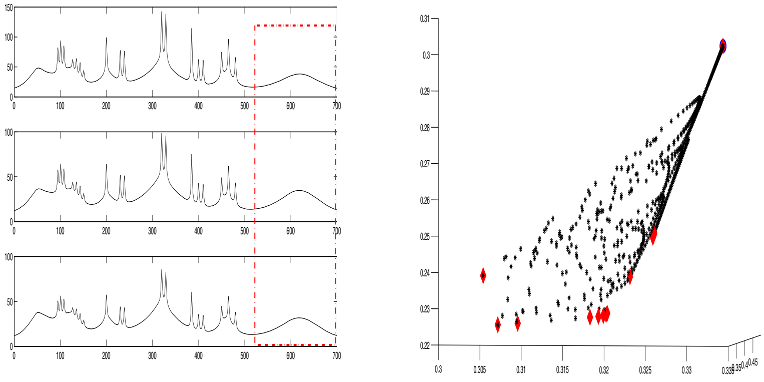


Fig. 3. Backward step 1. Left: the three mixtures. Right: the geometry of the mixture and the recovery of A^3 (the one in the blue circle). A dominant region containing the widest spectral peak is in the red rectangle. An estimate $w_3 = 130$ for the peak width of source S_3 can be read off.

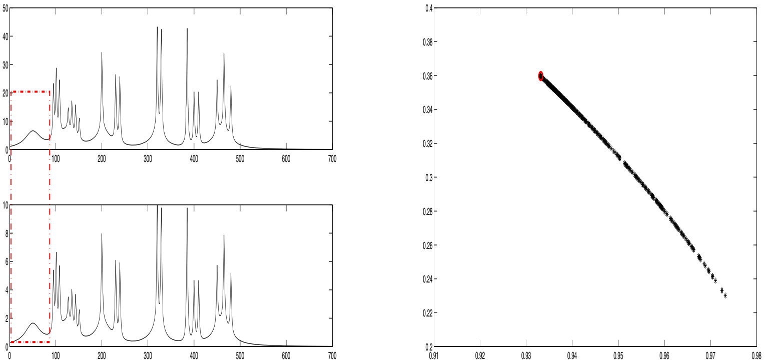


Fig. 4. Backward step 2. Model reduction via eliminating S_3. The two mixtures are on the left. The geometrical visualization is on the right. The mixing coefficient vector (red spot in the right plot) of source S_2 in $X_{(1,2)}$ attracts a dense cluster of planar points. An estimate $w_2 = 60$ (peak width of S_2) is read off from the peaks in the rectangular region.

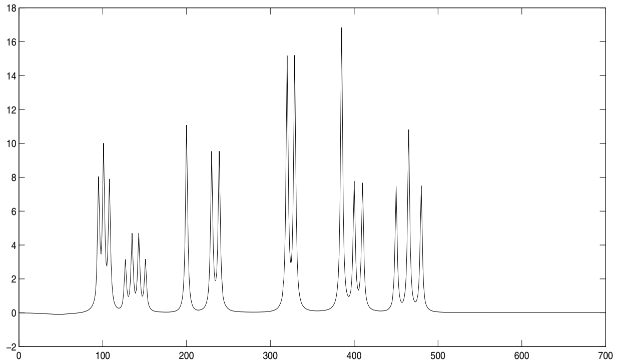


Fig. 5. Backward step 3. The recovery of S_1 by eliminating S_2 from the reduced mixture $X_{(1,2)}$.

$(S_2, \cdots, S_{n-1})$ successively from S_1 up by solving (2.1) based on reduced mixing equations in (B1)-(B2). (F2) Recover S_n and $(A_1, \cdots, A_{n-1})$ by solving (2.2).

We illustrate our method by a computational example where three sources are to be separated from three mixtures. One source has narrow peaks, one has wider peaks, and the last one has very wide peaks. The results are presented in a series of plots. Fig. 3 to Fig. 5 illustrate the backward step, and Fig. 6 presents the recovered source signals by ℓ_1 minimization in the forward step. In the step of recovering the source signal S_2 and S_3 via ℓ_1 minimization, the peak widths $w_2 = 60, w_3 = 130$ are read off from the mixture signals. Compared to ground truth, the separation results by our method are accurate. We also applied our method to separate mixture NMR spectra of menthol and β-sitosterol, and

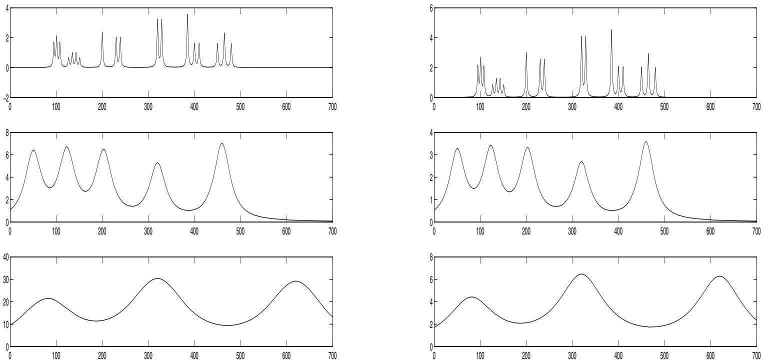


Fig. 6. Forward step. Left is the recovered sources by ℓ_1 minimization. Right is the reference spectra.

urine mixture data. More evaluation results based on experimental NMR data, and complexity analysis of algorithms are being reported in a comprehensive companion paper [9].

3 Concluding Remarks

A new source condition (the hierarchical dominant interval condition) is proposed for non-negative BSS of NMR mixtures. Though well-known minimal cone method does not work for such data, we found a recursive method integrating data clustering, successive elimination of variables, convex source recovery, and l_1 norm regularized minimization in transformed domains. A large non-convex NMF problem eventually boils down to smaller convex optimization problems. In future work, we shall further study NMR data of biofluids with our method.

References

1. Barton, R., Nicholson, J., Elliot, P., Holmes, E.: High-throughput 1H NMR-based metabolic analysis of human serum and urine for large-scale epidemiological studies: validation study. Int. J. Epidemiol. 37(suppl 1), i31–i40 (2008)
2. Boardman, J.: Automated spectral unmixing of AVRIS data using convex geometry concepts. In: Summaries of the IV Annual JPL Airborne Geoscience Workshop, vol. 1, pp. 11–14. JPL Pub., 93-26 (1993)
3. Chang, C.-I. (ed.): Hyperspectral Data Exploitation: Theory and Applications. Wiley-Interscience, Hoboken (2007)
4. Ernst, R., Bodenhausen, G., Wokaun, A.: Principles of Nuclear Magnetic Resonance in One and Two Dimensions. Oxford University Press, Oxford (1987)
5. Lee, D.D., Seung, H.S.: Learning of the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
6. Naanaa, W., Nuzillard, J.–M.: Blind source separation of positive and partially correlated data. Signal Processing 85(9), 1711–1722 (2005)

7. Nuzillard, D., Bourgb, S., Nuzillard, J.–M.: Model-Free analysis of mixtures by NMR using blind source separation. J. Magn. Reson. 133, 358–363 (1998)
8. Sun, Y., Ridge, C., del Rio, F., Shaka, A.J., Xin, J.: Postprocessing and Sparse Blind Source Separation of Positive and Partially Overlapped Data. Signal Processing 91(8), 1838–1851 (2011)
9. Sun, Y., Xin, J.: Nonnegative Sparse Blind Source Separation for NMR Spectroscopy by Data Clustering, Model Reduction, and ℓ_1 Minimization, preprint (2011); under review and revision for publication
10. Stadlthanner, K., Tom, A., Theis, F., Gronwald, W., Kalbitzer, H.-R., Lang, E.: On the use of independent analysis to remove water artifacts of 2D NMR Protein Spectra. In: Proc. Bioeng 2003 (2003)
11. Vitols, C., Weljie, A.: Identifying and Quantifying Metabolites in Blood Serum and Plasma. Chenomx Inc., (2006)
12. Winter, M.E.: N-findr: an algorithm for fast autonomous spectral endmember determination in hyperspectral data. In: Proc. of the SPIE, vol. 3753, pp. 266–275 (1999)
13. Wu, W., Daszykowski, M., Walczak, B., Sweatman, B.C., Connor, S., Haselden, J., Crowther, D., Gill, R., Lutz, M.: Peak alignment of urine NMR spectra using fuzzy warping. J. Chem. Inf. Model. 46, 863–875 (2006)
14. Yang, W., Wang, Y., Zhou, Q., Tang, H.: Analysis of human urine metabolites using SPE and NMR spectroscopy. Sci. China. Ser. B-Chem. 51, 218–225 (2008)

A Novel Face Recognition Approach under Illumination Variations Based on Local Binary Pattern

Zhichao Lian, Meng Joo Er, and Juekun Li

School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore 639798
{LIAN0069,EMJER,LI0009UN}@ntu.edu.sg

Abstract. Local Binary Pattern (LBP) is one of the most important facial texture features in face recognition. In this paper, a novel approach based on the LBP is proposed for face recognition under different illumination conditions. The proposed approach applies Difference of Gaussian (DoG) filter in the logarithm domain of face images. LBPs are extracted from the filtered images and used for recognition. A novel measurement is also proposed to calculate distances between different LBPs. The experimental results on the Yale B and Extended Yale B prove superior performances of the proposed method and measurement compared to other existing methods and measurements.

Keywords: Face Recognition, Illumination Variation, Local Binary Pattern.

1 Introduction

Illumination variation is one of the most challenging issues in face recognition. In [1], differences between varying illumination conditions are proven to be more significant than differences between individuals. A number of approaches have been proposed to address the issue, which can be classified into three categories: illumination modeling, illumination normalization and illumination invariant feature extraction.

Among all existing illumination invariant features, local binary pattern (LBP) [2-3] has gained much attention. The LBP operator is one of the best local texture descriptors. Besides the robustness against pose and expression variations as common texture features, the LBP is also robust to monotonic gray-level variations caused by illumination variations. The main idea in the LBP is to compare the gray value of central point with the gray values of other points in the neighborhood, and set a binary value to each point based on the comparison. After that, a binary string is transformed to a decimal label. A histogram of the labels is used for further recognition task. However, the labels are not stable when small changes occur such as noise. To overcome the problem, local directional pattern (LDP) [4] is proposed. The LDP is obtained by computing the edge response values in all eight directions at each pixel position and generating a binary code based on their edge response magnitudes. Tan and Triggs [5] proposed local ternary pattern (LTP) which extended the LBP to 3-valued codes. It is more discriminant and less sensitive to noise in uniform region. All the LBP and its several extensions mentioned above are not robust enough against large illumination variations.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 89–96, 2011.

In this paper, different from existing methods, we propose a novel distance measurement that can provide a stable distance based on the LBPs, instead of making the labels (patterns) stable when noise exists. The idea is more direct and easier to implement. In the new measurement, a distance based on pixel-level information is calculated besides a distance between histograms in a global level. A tolerance parameter is involved which can take two patterns as the same even if they have slight difference.

Besides, we also propose a novel face recognition method under varying illuminations based on the LBP. The proposed method applies Difference of Gaussians (DoG) filter in the logarithm domain of face images firstly and extracts the LBPs from the filtered images. The experimental results on the Yale B and Extended Yale B prove superior performances of the proposed method and measurement compared to other existing methods and measurements.

The rest of this paper is organized as follows. In Section 2, we introduce the proposed novel method and measurement in details. Experimental results and discussions are presented in Section 3. Finally, conclusions are drawn in Section 4.

2 Proposed Illumination Invariant Approach

2.1 Difference of Gaussians (DoG) Filter

The Difference of Gaussians (DoG) filter can enhance edge information, which is important for illumination invariant face recognition. In this paper, the image F is processed by a DoG filter:

$$F^{'} = DoG * F \tag{1}$$

where the DoG is given by

$$DoG(x,y) = \frac{1}{2\pi\sigma_1^2} e^{-\frac{x^2+y^2}{2\sigma_1^2}} - \frac{1}{2\pi\sigma_2^2} e^{-\frac{x^2+y^2}{2\sigma_2^2}} \tag{2}$$

$\sigma 1$ and $\sigma 2$ are standard deviations of two low-pass filters, and ω is the size of the DoG filter.

In this paper, the DoG filter is applied in the logarithm domain of face images firstly. The logarithm transform can compress the light pixel values and expand the dark pixel ones [1]. As a result, the transform can partially reduce the effects caused by illumination variations. For the DoG, we set $\sigma 1$=2.5, $\sigma 2$=2 and ω=6. After the DoG filter, the LBPs are extracted from the filtered images. The experiments shown in the following section will prove that the performance of the LBP in the DoG filtered images can be improved significantly.

2.2 Local Binary Pattern

The LBP used in face recognition was firstly proposed by Ahonen et al. [2]. The main idea in LBP is to compare the gray value of central point with the gray values of other points in the neighborhood, and set a binary value to each point based on the comparison. After that, a binary string is transformed to a decimal label as shown in the following equation

$$LBP_{P,R}(x,y)=\sum_{i=0}^{P-1} s(g_i - g_c)2^i \tag{3}$$

where $LBP_{P,R}(x,y)$ is the decimal label of point (x, y), P is the number of sampling points, R is the radius of a circle neighborhood, g_c is the gray level of central point (x, y), g_i is the gray level of neighborhood sampling point around central point (x, y) and

$$s(x)=\begin{cases}1, x>0\\ 0, x\le 0.\end{cases} \tag{4}$$

A histogram of the decimal label is calculated and can be used as a texture feature. The histogram is defined as

$$H_i=\sum_{x,y} I\{LBP_{P,R}(x,y)=i\}, i=0,1,..,...,l \tag{5}$$

where l is the number of different labels produced by the LBP operator and

$$I\{A\}=\begin{cases}1, A \quad is \quad true\\ 0, A \quad is \quad false.\end{cases} \tag{6}$$

After an image is divided into non-overlapped blocks, the LBP operator is applied to each block and a histogram of different labels is calculated for each block. All the histograms of blocks are concatenated to an entire histogram to build a global description of the image. The details of the LBP can be referred to [2-3].

The LBP descriptor contains three levels information: the labels for the histogram contain information about the patterns on a pixel-level, the labels are summed over a small block to produce information on a regional level and an entire histogram concatenated by regional histograms presents a global description of the image [2].

In this paper, we divide the images into blocks of 24×24. After that, the histograms of $LBP_{8,1}$ uniform patterns of blocks are calculated and concatenated into a global histogram. The global histogram will be used in the global level distance measurement.

2.3 Proposed Distance Measurement

Most popular methods [2-5] use histogram intersection or Chi Square statistic as means of distance measurement, and they are defined as follows:
Histogram intersection:

$$D(Q,S)=\sum_i \min(Q_i,S_i) \tag{7}$$

Chi square statistic:

$$\chi^2(Q,S)=\sum_i \frac{(Q_i-S_i)^2}{Q_i+S_i} \tag{8}$$

where *Q* and *S* are two concatenated LBP histograms of image A and B respectively. When the image is divided into blocks, some blocks may contain more discriminant information than others. Therefore, it is reasonable to set weights for different blocks based on the importance of the information they contain. The weighted Chi square statistic is defined as

$$\chi_w^2(Q,S)=\sum_{i,j} w_j \frac{(Q_{i,j}-S_{i,j})^2}{Q_{i,j}+S_{i,j}} \tag{9}$$

where *Q* and *S* are the concatenated histograms to be compared, indices *i* and *j* refer to *i*th bin in histogram corresponding to the *j*th local block and w_j is the weight for block *j*.

No matter either non-weighted or weighted measurements is used, it is obvious that they only use histogram information, which means that pattern information on a pixel-level is ignored.

In this paper, a novel measurement is proposed which considers both the differences between images on a pixel-level and the differences between images on a global level. The distance between images on a global level D_1 is defined as the differences between concatenated histograms, using histogram intersection distance as Eq. (7). The distance between images on a pixel-level D_2 is defined as the percentage of pixels which have different patterns in two images, shown as follows.

$$D_2(A,B)=\sum_{i=1}^{m}\sum_{j=1}^{n} Z(label_A(i,j),label_B(i,j)) \Big/ m\cdot n \tag{10}$$

$$Z(x,y)=\begin{cases}0,\text{if the hamming distance between x and y } i\text{s smaller than } T\\ 1, otherwise\end{cases} \tag{11}$$

where the size of images is *m*×*n*, $label_A(i,j)$ is the decimal label of point (*i, j*) in the image A, $label_B(i,j)$ is the decimal label of point (*i, j*) in the image B, and *T* is a threshold parameter.

As mentioned before, the labels of pixels in the LBP are not stable when some noise exists. Several extensions of the LBP have been proposed to obtain stable labels to address the problem [4-5]. However, no matter which extension is applied, the

labels are still not stable in some cases. In this paper, we study the problem in another view. We propose a novel measurement to obtain a stable distance even if the labels have some small changes instead of obtaining stable labels. In our measurement, two binary pattern labels are regarded the same if the number of bits having different values is not greater than T. When T is set to 0, it means that two binary pattern labels are taken as the same only when the values of each bit in two labels are the same.
The final distance between two images A and B is defined as

$$D(A,B) = \alpha D_1(Q,S) + (1-\alpha)D_2(A,B) \tag{13}$$

where α is the ratio between the global level distance and the total distance. With the proposed novel measurement, the performance of the LBP is improved significantly, shown in the experiment section. The effects of parameters will also be discussed in the experiment section.

3 Experimental Results and Discussions

3.1 Database

In the experiments, we use the Yale Face database B [6] and Extended Yale Face database B [7] as the test database. In total there are 38 persons with 64 different illumination conditions for nine poses per person. Because the main concern in this paper is on illumination variations, only 64 frontal face images per person under different illumination conditions are chosen. After combining these two databases except 18 corrupted images, there are 2414 images of 38 subjects named as the Completed Yale B. The images are divided into 5 subsets based on the angle between the light direction and the camera axis shown in Table 1. All the images are cropped with the size of 192×168 and aligned by the database [7]. In the following experiments, only one frontal image per person with normal illumination (0°light angle) is applied as a training sample, which increases the difficulty of recognition. Recognition is performed with the nearest neighbor classifier.

Table 1. Subsets divided based on light source direction

	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
Light angle	0~12	13~25	26~50	51~77	>77
Number of images in Completed Yale B	263	456	455	526	714

3.2 Performance Comparisons for Different Distance Measurements

In this section, we compare the proposed distance measurement with other existing distance measurements. The LBP and LDP are implemented with our proposed measurement and histogram intersection as distance measurement. For comparison, the DCT [8] and the LN [9], two of the most representative illumination invariant recognition approach, are also implemented. All the results are shown in Table 2. Please note that the DoG filter is not involved in the process.

From the table, it is clear that although the performances of the LBP under small illumination variations as Subsets 2 and 3 are acceptable, the LBP is not robust against larger illumination variation such as Subsets 4 and 5. The extension LDP is also not robust against larger illumination variations and it obtains even worse performances compared to the LBP in Subsets 3 and 4. The reason is that histogram intersection distance only makes use of histograms that represent global level information but ignores pixel-level information. The results prove that histograms could represent facial features well under small illumination variations but histograms cannot provide sufficient discriminant information under larger illumination variations. In our proposed measurement, pixel-level information is taken into consideration. With the proposed measurement, the performance of the LBP has been improved significantly. In Subset 3, the LBP with our measurement even achieves a better performance compared to the DCT and the LN. In Subset 4, the performance of the LBP is acceptable. In the most difficult cases as Subset 5, although the proposed distance measurement improves the performance of the LBP obviously compared to other distance measurements, the error rate still much higher than that of the DCT and the LN. Similarly, our proposed measurement also improves the performance of the LDP a lot. Thus, our distance measurement outperforms other distance measurements, especially in the cases with larger illumination variations.

Table 2. Performance comparisons of different measurements

Method	Error rate (%)				
	Subset 2	Subset3	Subset 4	Subset 5	Total
The LBP with histogram intersection	0.4	9.0	64.3	90.5	42.5
The LDP with histogram intersection	0.2	23.1	73.4	83.6	45.1
The LBP with our distance measurement	0.2	3.7	17.1	50.7	19.5
The LDP with our distance measurement	0	4.2	20.5	43.1	18.0
The DCT	0	10.5	10.8	12.6	8.1
The LN	0	12.3	6.3	8.4	6.2

3.3 Performance Comparisons for Different Parameter Values

Here, we evaluate the effects of parameters on our proposed measurement. The DoG filter is still not involved in the process and a better result of the proposed method with the DoG filter will be presented in the next section.

The results for different values of tolerance T are shown in Table 3. As mentioned before, tolerance T is the number of different value bits in two pattern labels, below which these two patterns are still taken as the same. In the strictest case where T is set to 0, two binary pattern labels are taken as the same if and only if they are completely the same. For the cases where small noise may exist, T can be set to 1 or 2. From Table 3, we can notice that with the increase of the T value, the performances of the LBP are also improved in the cases with larger illumination variations such as Subsets 4 and 5. Please also note that even with the strictest measurement ($T=0$), our proposed measurement still improves the performance of the LBP significantly compared to histogram intersection distance, especially in Subsets 3 and 4.

The results of different values of α are shown in Table 4. The parameter α reflects the ratio between the global level distance and the total distance. From the table, it is clear that it is better to select smaller value of α under larger illumination variations. This is because histograms cannot provide sufficient discriminant information in the case with larger illumination variations and pixel-level information could provide more useful information in such cases. Therefore, the weight of the global level distance should be decreased.

Table 3. Performance comparisons for different values of *T*

Parameter *T*	Error rate (%)				
	Subset 2	Subset3	Subset 4	Subset 5	Total
0	0.4	3.3	32.3	82.2	32.1
1	0.4	3.1	18.3	54.6	20.8
2	0.2	3.7	17.1	50.7	19.5
3	0.2	4.8	18.8	52.2	20.5

Table 4. Performance comparisons for different values of α

	Error rate (%)				
	Subset 2	Subset3	Subset 4	Subset 5	Total
0.4 (T=1)	0.2	3.1	14.5	43.0	16.5
0.5 (T=1)	0.4	3.1	18.3	54.6	20.8
0.4 (T=2)	0.2	4.4	14.6	42.9	16.7
0.5 (T=2)	0.2	3.7	17.1	50.7	19.5

3.4 Performance Comparisons for Different Methods

Furthermore, we compare our proposed method involving the DoG filter with other existing methods, including the DCT, the LN and the LBP. All the results are listed in Table 5. From the table, we can see that our method significantly outperforms other methods and our method can achieve a very satisfactory performance either in small illumination variations or large illumination variations.

Table 5. Comparison of different methods

Method	Error rate (%)				
	Subset 2	Subset3	Subset 4	Subset 5	Total
The DCT	0	10.5	10.8	12.6	8.1
The LN	0	12.3	6.3	8.4	6.2
The LBP	0.4	9.0	64.3	90.5	42.5
Our Proposed Method	0	2.6	1.3	6.6	2.7

4 Conclusions

In this paper, a novel distance measurement for the LBP and its extensions is proposed. Different from existing distance measurements, pixel-level information are fused in the measurement besides the common histogram differences in a global level. To overcome the problem that pattern labels of the LBP are not stable when noise exists, a tolerant parameter is considered in the measurement. Therefore the proposed measurement can provide a stable distance between the LBPs even if the labels have small changes due to some noise. Besides, we proposed a novel approach based on the LBP to improve face recognition performance under illumination variations. The proposed approach applies DoG filter in the logarithm domain of face images. LBPs are extracted from the filtered images and used for recognition. The experimental results on the Yale B and Extended Yale B prove the superior performances of our proposed method and measurement compared to other existing methods and measurements. Further research will focus on face recognition under other variations based on the proposed distance measurement.

References

1. Adini, Y., Moses, Y., Ullman, S.: Face recognition: the problem of compensating for changes in illumination direction. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 721–732 (1997)
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: Proceedings of 8th European Conf. Computer Vision, Prague, Czech Republic, pp. 469–481 (2004)
3. 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, 2037–2040 (2006)
4. Jabid, T., Kabir, M.H., Chae, O.: Local Directional Pattern (LDP) for face recognition. In: Proceedings of 2010 Digest of Technical Papers International Conference on Consumer Electronics, Las Vegas, NV, USA, pp. 329–330 (2010)
5. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. on Image Processing 19, 1635–1650 (2010)
6. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Transactions on Pattern Analysis and Machine Intelligence 23, 643–660 (2001)
7. Lee, K.C., Ho, J., Kriegman, D.: Acquiring Linear Subspaces for Face Recognition under Variable Lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence 27, 684–698 (2005)
8. Chen, W., Er, M.J., Wu, S.: Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 36, 458–466 (2006)
9. Xie, X., Lam, K.M.: An efficient illumination normalization method for face recognition. Pattern Recognition Letters 27, 609–617 (2006)

A New Pedestrian Detection Descriptor Based on the Use of Spatial Recurrences

Carlos Serra-Toro and V. Javier Traver

Departamento de Lenguajes y Sistemas Informáticos &
Institute of New Imaging Technologies,
Universitat Jaume I, 12071 Castellón, Spain
{cserra,vtraver}@uji.es

Abstract. Recent work on pedestrian detection has relied on the concept of local co-occurences of features to propose higher-order, richer descriptors. While this idea has proven to be benefitial for this detection task, it fails to properly account for a more general and/or holistic representation. In this paper, a novel, flexible, and modular descriptor is proposed which is based on the alternative concept of visual recurrence and, in particular, on a mathematically sound tool, the recurrence plot. The experimental work conducted provides evidence on the discriminatory power of the descriptor, with results comparable to recent similar approaches. Furthermore, since its degree of locality, its visual compactness, and the pair-wise feature similarity can be easily changed, it holds promise to account for characterizations of other descriptors, as well as for a range of accuracy-computational trade-offs for pedestrian detection and, possibly, also for other object detection problems.

Keywords: Pedestrian detection, Recurrence plot, Oriented gradients, Feature descriptor.

1 Introduction

Human detection is the base from which other more specialized recognition tasks can be performed, e.g. identification, categorization according to some criteria, or body silhouette extraction. Human detection itself is even more difficult when no temporal information is available. This problem is often addressed using images of standing people which are usually obtained from urban scenes and that is why the term *pedestrian* is so widely used the literature, as is the case for this paper.

Several approaches, in terms of both feature descriptors and classifiers, have been proposed in the literature to solve the problem of the pedestrian detection in static images. According to the conclusions of a recent survey [5], features based on local edge orientation seem to be useful to encode a human figure when no processing constraints are imposed. Similarly, [13] shows that gradient-based features are majoritary used to solve this task in state-of-the-art approaches.

Besides the improvement of the detection accuracy, some other problems related to pedestrian detection are of current interest. On the one hand, speeding

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 97–104, 2011.

up the detection is important, and has been tackled both with ad-hoc cascades [12] or with a more generic approach aimed at reducing the number of sliding windows required [14]. On the other hand, the output of the detection can be refined beyond the bounding box by detecting subparts of the human figure [11] or by getting the bounding box closer to the person [10].

Some recent approaches to pedestrian detection rely on co-occurrences of neighboring low-level features [7,8] so that richer descriptors are obtained by encoding this higher-order information. While this kind of representation has proven useful, the global structure of the human figure is not characterized explicitly and it is therefore left to the subsequent classifier to implicitly discover this information. In this paper, an alternative idea is explored, where the pairwise relationships of local features are captured. Experimental evidence is provided on the usefulness of this recurrence for pedestrian detection. To the best of our knowledge, this pedestrian representation, based on spatial *recurrences*, rather than *co-occurences*, has not been proposed before.

This paper is organized as follows: Section 2 presents our new method to detect pedestrians using a descriptor based on recurrences. First experimental results using our descriptor are presented in Section 3. Section 4 concludes the paper by summarizing some of our prelimilary findings about this technique.

2 Recurrence-Based Descriptor

Our recurrence-based descriptor is inspired by recurrence plots [2], a mathematically sound concept which is useful for visualizing or describing dynamical systems. Given a system represented by a sequence of states, $S_1, S_2, \ldots, S_\xi$, the recurrence plot $\rho \in \{0,1\}^{\xi \times \xi}$ is defined as this 2D binary matrix:

$$\rho_{i,j} = \begin{cases} 1 & \text{if } d(S_i, S_j) \leq \theta, \\ 0 & \text{otherwise ,} \end{cases} \tag{1}$$

where d is a similarity or distance measure, θ is a threshold on this distance, and $1 \leq i, j \leq \xi$. Because of d acting as a distance function, its symmetrical property implies $\rho_{i,j} = \rho_{j,i}$, and therefore $\rho_{i,i} = 1$.

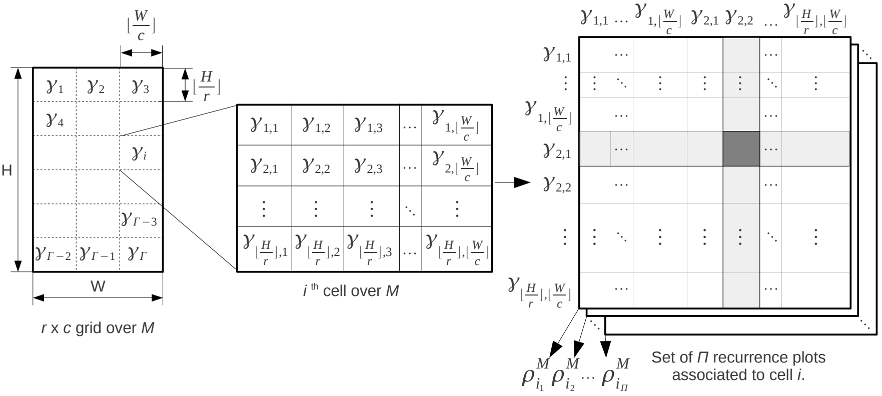
While the most immediate use of the recurrence plot is for states describing the temporal behavior of a system, the extension to spatial data is also possible [2]. We propose a representation with one or several recurrence plots, each of them still 2D, simply by choosing the "states" to represent certain visual information at a given spatial location, and the sequence of states resulting from an arbitrary ordering of these location.

More formally, let I be the original (2D) image, and M a (2D) map obtained from I with some visual information (edges, gradients, colour channels, etc.). Let the size of M be $H \times W$ (height × width). A $r \times c$ uniform cartesian grid is superimposed over M, resulting in $\Gamma = r \cdot c$ non-overlapping cells over the image (r cells across and c cells down, discarding the elements of the last rows and columns of M not covered by the cells when H or W are not divisible by r or c, respectively) with $\lfloor \frac{H}{r} \rfloor \times \lfloor \frac{W}{c} \rfloor = \xi$ elements per cell.

Based on this, a general descriptor is proposed in which the information covered by the region delimited by the cell γ of M, $1 \leq \gamma \leq \Gamma$, is encoded by a vector of Π recurrence plots $P_\gamma^M = \rho_{\gamma_1}^M \rho_{\gamma_2}^M \dots \rho_{\gamma_\Pi}^M$, with $\rho_{\gamma_i}^M$ being the i-th recurrence plot corresponding to cell γ over the information map M. Each $\rho_{\gamma_i}^M$ therefore encodes the visual information at a given spatial location, and may have its own threshold θ and similarity function d (see (1)). Having defined this, the feature vector v_M associated to the map M is

$$v_M = \underbrace{\rho_{1_1}^M \rho_{1_2}^M \dots \rho_{1_\Pi}^M}_{P_1^M} \underbrace{\rho_{2_1}^M \rho_{2_2}^M \dots \rho_{2_\Pi}^M}_{P_2^M} \cdots \underbrace{\rho_{\Gamma_1}^M \dots \rho_{\Gamma_\Pi}^M}_{P_\Gamma^M} , \qquad (2)$$

i.e. all the recurrence plots of a given cell, and all those vectors obtained from all the cells, are concatenated to form the resulting feature vector. Figure 1 illustrates the proposed method.



$\lfloor \frac{W}{c} \rfloor$
$\lfloor \frac{H}{r} \rfloor$
γ_1
γ_2
γ_3
γ_4
γ_i
$\gamma_{\Gamma-3}$
$\gamma_{\Gamma-2}$
$\gamma_{\Gamma-1}$
γ_Γ
H
W
r x c grid over M
$\gamma_{1,1}$ $\gamma_{1,2}$ $\gamma_{1,3}$ … $\gamma_{1,\lfloor \frac{W}{c} \rfloor}$
$\gamma_{2,1}$ $\gamma_{2,2}$ $\gamma_{2,3}$ … $\gamma_{2,\lfloor \frac{W}{c} \rfloor}$
$\gamma_{\lfloor \frac{H}{r} \rfloor,1}$ $\gamma_{\lfloor \frac{H}{r} \rfloor,2}$ $\gamma_{\lfloor \frac{H}{r} \rfloor,3}$ … $\gamma_{\lfloor \frac{H}{r} \rfloor,\lfloor \frac{W}{c} \rfloor}$
i^{th} cell over M
$\gamma_{1,1}$ … $\gamma_{1,\lfloor \frac{W}{c} \rfloor}$ $\gamma_{2,1}$ $\gamma_{2,2}$ … $\gamma_{\lfloor \frac{H}{r} \rfloor,\lfloor \frac{W}{c} \rfloor}$
$\gamma_{1,1}$
$\gamma_{1,\lfloor \frac{W}{c} \rfloor}$
$\gamma_{2,1}$
$\gamma_{2,2}$
$\gamma_{\lfloor \frac{H}{r} \rfloor,\lfloor \frac{W}{c} \rfloor}$
$\rho_{i_1}^M \rho_{i_2}^M \cdots \rho_{i_\Pi}^M$
Set of Π recurrence plots associated to cell i.

Fig. 1. Illustration of the proposed method: given an information map M, a cartesian grid $r \times c$ is applied over it, resulting in Γ non overlapping cells. Each cell γ, $1 \leq \gamma \leq \Gamma$, has $\lfloor \frac{H}{r} \rfloor \times \lfloor \frac{W}{c} \rfloor = \xi$ elements which are used to create Π recurrence plots. In each plot, elements $\gamma_{k,l}$ and $\gamma_{m,n}$ of cell γ, $1 \leq k, l, m, n \leq \xi$ are compared using a distance function. In the figure, the emphasized cell will store the result of the comparison between states $\gamma_{2,1}$ and $\gamma_{2,2}$. The feature vector is the concatenation of all the recurrence plots obtained for all the cells, as defined in (2).

Some comments on the properties of this descriptor and its implications follow. The complexity and generality of the proposed approach depends on two factors: the density of the grid, determined by the parameters r and c, and the number of recurrence plots, Π, for each cell.

The grid configuration (parameters r and c) determines how local or global the descriptor is with respect to the whole information map M. Denser grids imply smaller cells and, therefore, smaller recurrence plots, each capturing the recurrence on a small local neighborhood. In the other extreme, $r = c = 1$, results in a *single* larger recurrence plot covering the whole M and therefore capturing all pair-wise relationships of the elements of M.

Also, the number of recurrence plots per cell, Π, has an impact on how variate is the spatial information encoded for each cell. A large Π can result in a richer and heterogeneous descriptor, while a small Π (even $\Pi = 1$) can result in a more homogeneous descriptor per cell, independtly of its complexity.

These parameters also determine the dimensionality of the features vector. Since each cell in M has ξ elements and each element has to be compared with all the other elements of the cell, each recurrence plot needs to store ξ^2 elements. Since recurrence plots are symetrical by definition, only half of those elements are required to represent the information, $\frac{\xi^2+\xi}{2}$. Since a cell has associated Π recurrence plots, the dimensionality of the feature vector v_M is:

$$dim(v_M) = \frac{\xi^2 + \xi}{2} \cdot \Gamma \cdot \Pi \ . \tag{3}$$

3 Experimentation

We performed our experiments with the dataset described in Section 3.1. Section 3.2 specifies the implementation details and the evaluation method followed to obtain the results of our experiments, presented and commented in Section 3.3.

3.1 Image Dataset

We use the DaimlerChrysler Pedestrian Classification Benchmark Dataset[1] [1], which consists of five disjoint sets, each containing 4,800 pedestrian pictures and 5,000 non-pedestrian pictures. Three of those sets are marked as training sets while the other two are intended for testing purposes. Each picture is a manually labelled 36 × 18 pixels gray-scale image.

3.2 Implementation Details and Evaluation Method

For every image in the dataset, its oriented gradients map is obtained by using the Sobel operator. Each gradient orientation is discretized into eight possible orientations. If the magnitude of the gradient is below a given threshold δ, it is marked as no-gradient. Therefore, each image results in a discretized gradient map M whose elements may take nine possible values (one per each orientation plus one more for the no-gradient case).

We use an evaluation scheme where the three different training sets are merged into one single training set of 29,400 instances, and the two different testing sets are merged into a single testing set of 19,600 instances. Performance is given both in terms of accuracy rate, Receiver Operating Characteristics (ROC) curves and, to summarize some results, the Area Under the Curve (AUC) [9].

[1] http://www.science.uva.nl/research/isla/downloads/pedestrians/

We use a Suport Vector Machine (SVM) [3] to classify our data. We use a linear SVM since our method yields very high dimensional feature vectors. We use the LIBLINEAR 1.7 library [4] and, due to the exploratory nature of this work, the linear SVM penalty parameter C was fixed to $C = 1$, so no grid search was done to find the optimal parameter when classifying.

3.3 Experimental Results

We study the impact of the number of recurrence plots per cell (determined by parameter the Π). The results are summarized in Figure 2 and in Table 1. Later, we study the locality/globality impact (determined by parameters r and c) showing the results in Table 2. In all the experiments we set $\delta = 0.5$.

Experiment 0: First of all, to prove the expresivity of our proposed recurrence-based feature descriptor we compared its performance against a naive descriptor consisting of simple concatenating all the elements of M into one raw vector. This approach resulted in an accuracy of 67.06% with an AUC = 0.819 (see Table 1).

Study of the Impact of the Number of Recurrence Plots per Cell. Then, we focused on the impact of the number of recurrence plots per cell, Π, to determine if it is convenient to use a small or large set of recurrence plots per cell. In all the following experiments, we set $r = 6, c = 3$ as the grid configuration parameters. Although the grid configuration determines the size of each cell, and thus restricts the locality/globality of each recurrence plot, it is possible to create a complex distance function d (see (1)) independently of the area of each cell determined by the grid.

Experiment 1: First, we selected a distance function based on the equality of all the discretized gradients across each cell, including the no-gradient value. So, the distance function between two discretized orientations S_i and S_j was:

$$d(S_i, S_j) = |S_i - S_j| \ , \tag{4}$$

i.e. the absolute value of the difference between the discretized orientations. A threshold $\theta = 0$ was chosen so that only if both (discretized) orientations are equal then $\rho_{i,j} = 1$ according to (1), otherwise $\rho_{i,j} = 0$.

Experiment 2: We divided the information encoded by each recurrence plot in the previous experiment in such a way that there were a recurrence plot per possible discretized orientation, and each recurrence plot encoded the similarities of a single discretized orientation (including the no-gradient value).

We used a function distance d_k in which each discretized gradient k (including the no-gradient) was compared only with its same gradient. Therefore, there was a recurrence plot for each orientation in each cell (i.e. $\Pi = 9$),

$$d_k(S_i, S_j) = \begin{cases} 0 & \text{if } S_i = S_j \text{ and } S_i = k, \\ 1 & \text{otherwise} \ , \end{cases} \tag{5}$$

with threshold $\theta = 0$. The accuracy was 90.80% and the AUC = 0.969. As it can be seen, when comparing the experimental results obtained here with those obtained with the previous experiment (see Table 1), it seems that a large amount of recurrence plots per cell, each focusing on a different aspect of the information, is prefereable to a more complex, unique recurrence plot that stores all that information in a common place. This is because, although a single recurrence plot globally captures all the relations between pairs of states, it really does not consider what those states represent. Splitting a complex information in pieces of simpler information allow each recurrence plot to be more concise about the information which it is representing, and thus more discriminative.

Experiment 3: To confirm the last hipothesis, we performed another experiment halfway between the two previous ones. We created four recurrence plots per cell, each with a distance function similar to that defined by (5) but, in this case, opposite orientations (i.e. with a difference of 180° between them) were considered equivalent and thus were encoded in the same recurrence plot. As was expected, the accuracy obtained was halfway between the two previous approaches (see Table 1), since, as stated before, splitting complex information between different recurrence plots allow them to be more meaningful.

Table 1. Accuracies (%) and AUC obtained for the experiments performed to study the impact of the number of recurrence plots per cell (see Section 3.3). The dimensionality per feature vector (determined by (3)) and the execution time relative to Experiment 0 are also shown. When meassuring the time, the portions of code related with I/O and with the computation of the discretized gradients map were not measured.

Exp.	Description		No. features	Time Factor	Accuracy	AUC
0	No recurrent information used		648	×1	67.06	0.819
1	$r = 6, c = 3$	$\Pi = 1$	11,988	×9	76.37	0.848
2		$\Pi = 9$	107,892	×15	90.80	0.969
3		$\Pi = 4$	47,952	×23	86.46	0.943

Study of the Impact of the Grid Configuration. We tested several $r \times c$ uniform cartesian grids over the image, all uniformly sampled and without overlap between them. So, we tested $r \in \{3, 4, \ldots, 8\}$ and $c \in \{2, 3, 4, 5\}$ values. We chose a number of recurrence plots defined by Experiment 2 in this section since it was the distance that yielded a higher accuracy. However, the purpose of this study is not to reach a high accuracy but to do a first exploration of the behaviour of the proposed descriptor as the grid varies.

Since accuracies tend to increase with smaller grids (Table 2), it seems that global approaches (i.e., lower number of cells, Γ) perform better than local ones, which suggests the benefit of holistic approaches accounting for the global human shape. While further investigation is required, a global approach with bigger recurrence plots covering larger cells seem preferable. For instance, with a 3×3 grid, our approach results in an AUC which is only 0.017 below of the obtained with our implementation of a similar method to ours [7].

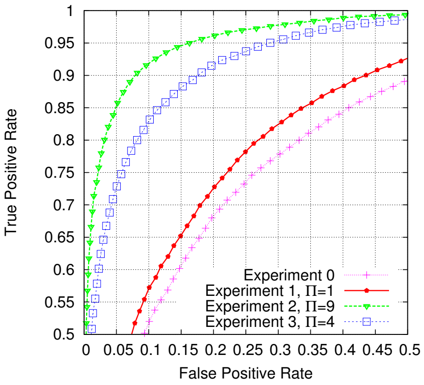


Fig. 2. ROC curves comparing the experiments done to investigate the impact of the number of recurrence plots per cell (Π parameter)

Table 2. Accuracies (%) and AUC obtained using the approach described in Experiment 2 using several grids of sizes $r \times c$

		***r* parameter**											
		3		4		5		6		7		8	
		Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
***c* param.**	2	91.20	0.972	91.03	0.972	90.27	0.968	90.57	0.969	89.86	0.964	87.36	0.949
	3	91.34	0.972	91.12	0.971	90.51	0.967	90.80	0.969	89.73	0.963	87.48	0.948
	4	90.91	0.970	90.73	0.969	89.82	0.963	90.33	0.966	89.10	0.959	86.38	0.939
	5	90.26	0.966	89.92	0.964	89.05	0.958	89.65	0.962	88.29	0.953	85.21	0.930

4 Conclusions

A novel descriptor based on recurrences of visual features has been proposed and its properties and possibilities have been explored. The flexibility of the descriptor and the modularity of its design and implementation facilitates its experimental validation: the local-to-global character of the descriptor, as well as the compactness of the visual representation can be easily varied. Due to its generality, the proposed approach can be considered to subsume a co-occurrence-like concept. The results for pedestrian detection are comparable to the state of the art and suggest that best results can be obtained by splitting the raw features into several recurrences, each focusing on a different piece of information, and considering a global approach instead of a local one. Future work is aimed at further improving the expresiveness of the descriptor and reducing its current computational requirements.

Acknowledgments. The authors acknowledge the Spanish research programme Consolider Ingenio-2010 CSD2007-00018, and Fundació Caixa-Castelló Bancaixa under project P1·1A2010-11. Carlos Serra-Toro is funded by Generalitat Valenciana under the "VALi+d program for research personnel in training" with grant code ACIF/2010/135.

References

1. Munder, S., Gavrila, D.M.: An Experimental Study on Pedestrian Classification. IEEE Trans. on PAMI 28(11), 1863–1868 (2006)
2. Marwan, N., Romano, M.C., Thiel, M., Kurths, J.: Recurrence Plots for the Analysis of Complex Systems. Physics Reports 438(5–6), 237–329 (2007)
3. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
4. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9, 1871–1874 (2008), Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear
5. Enzweiler, M., Gavrila, D.M.: Monocular Pedestrian Detection: Survey and Experiments. IEEE Trans. on PAMI 31(12), 2179–2195 (2009)
6. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Conf. on CVPR, vol. 1, pp. 886–893 (2005)
7. Watanabe, T., Ito, S., Yokoi, K.: Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection. In: Wada, T., Huang, F., Lin, S. (eds.) PSIVT 2009. LNCS, vol. 5414, pp. 37–47. Springer, Heidelberg (2009)
8. Ito, S., Kubota, S.: Object Classification Using Heterogeneous Co-occurrence Features. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6312, pp. 209–222. Springer, Heidelberg (2010)
9. Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
10. Pedersoli, M., Gonzàlez, J., Bagdanov, A.D., Villanueva, J.J.: Recursive Coarse-to-Fine Localization for Fast Object Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 280–293. Springer, Heidelberg (2010)
11. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object Detection with Discriminatively Trained Part-Based Models. IEEE Trans. on PAMI 32(9), 1627–1645 (2010)
12. Zhu, Q., Avidan, S., Yeh, M.-C., Cheng, K.-T.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: IEEE Conf. on CVPR, vol. 2, pp. 1491–1498 (2006)
13. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: A Benchmark. In: IEEE Conf. on CVPR, pp. 304–311 (2009)
14. Lampert, C.H., Blaschko, M.B., Hofmann, T.: Beyond Sliding Windows: Object Localization by Efficient Subwindow Search. In: IEEE Conf. on CVPR, pp. 1–8 (2008)

Facial Expression Recognition Using Nonrigid Motion Parameters and Shape-from-Shading

Fang Liu[1], Edwin R. Hancock[2], and William A.P. Smith[2]

[1] School of Computer Sci. and Tech., Huazhong University of Sci. and Tech.
fang.liu@hust.edu.cn,
[2] Department of Computer Science, The University of York
{erh,wsmith}@cs.york.ac.uk

Abstract. This paper presents a 3D motion based approach to facial expression recognition from video sequences. A non-Lambertian shape-from-shading (SFS) framework is used to recover 3D facial surfaces. The SFS technique avoids heavy computational requirements normally encountered by using a 3D face model. Then, a parametric motion model and optical flow are employed to obtain the nonrigid motion parameters of surface patches. At first, we obtain uniform motion parameters under the assumptions that motion due to change in expressions is temporally consistent. Then we relax the uniform motion constraint, and obtain temporal motion parameters. The two types of motion parameters are used to train and classify using Adaboost and HMM-based classifier. Experimental results show that temporal motion parameters perform much better than uniform motion parameters, and can be used to efficiently recognize facial expression.

Keywords: Facial Expression Recognition, SFS, Nonrigid Motion.

1 Introduction

Over the past two decades automatic facial expression recognition has become an active area of research. A large number of methods have been proposed to extract and represent features associated with facial expressions. These methods can be categorized according to whether they focus on the motion or deformation of faces associated with forming an expression [1]. Deformation-based features [2,3,4] depict facial actions by capturing shape and texture changes, and these are good indicators for facial actions. Compared with the deformation based and indirect approach of representing facial actions, motion-based methods offer the advantage of directly focussing on the physical action needed to form an expression. Dense optical flow [5] and feature point tracking [6] have been used as the basis of motion-based methods. Focussing in more detail on [5], Black and Yacoob have shown that local parameterized models of image motion are effective on recovering the nonrigid motion of human faces, and also for recognizing facial expressions within localized space-time intervals. Their work focuses on the motion of intransient facial features such as the eyes, mouth, eye-brows

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 105–113, 2011.

which are involved in the formation of facial expressions. Optical flow and an eight-parameter motion model are used to estimate the motion parameters from 2D image data. Fasel and Luettin [1] show that the motion estimations can be significantly improved if they are recovered using a 3D facial model. However, such 3D models often require complex mapping procedures and these in turn place significant computational overheads on the method. In this paper, we aim to estimate the motion parameters from a 3D facial surface which is recovered by using a non-Lambertian shape-from-shading (SFS) method. By utilising SFS, we avoid some of the computational overheads. In addition, we are able to extract the motion of small patches over the complete facial surface. As a result we can detect transient facial features.

Nonrigid motion recovery methods generally assume that the motion within a small patch must be spatially consistent. To overcome the ill-posedness of non-rigid montion recovery, Zhou and Kambhamettu [7] assumes that the motion should be not only be spatially consistent but also temporally consistent. This means that the motion associated with facial expressions is uniform. Unfortunately, this does not agree well with the real world conditions. Here we relax the uniform motion constraint and obtain temporal motion parameters by enlarging the size of the patches. Potentially large patch size may cause inaccurate parameter estimation. However, by comparing the temporal parameters to those obtained under the uniform motion assumption, we observe that when facial expression recognition is attempted then the temporal properties of the parameters are more useful.

For the purpose of expression classification we use the Adaboost algorithm for classifying the non-temporal features, while a HMM-based classification method is used for temporal motion. The HMM-based classification method has been successful used in the field of speech recognition, and the method has also been used to recognize facial expressions [8,9]. In this paper, we use multiple HMM-based classifiers instead of a single classifier. A set of voting rules are used to combine the classifier outputs and reach a decision. Experiments indicate that this method gives good performance.

The oultine of this paper is as follows: In Section 2 we describe the elements of our method namely a non-Lambertian SFS technqiue and a local parameterized motion model for 3D objects. Section 3 and 4 describes the details of implementation for our method. Section 5 presents and discusses our experimental results. Finally, we conclude the paper and offer directions for future investigation.

2 Background

Non-Lambertian SFS. In this paper, a SFS method is employed to derive a 3D face description from a single 2D brightness image. SFS aims to recover 3D shape from the gradual variations of shading in an image. To solve the SFS problem, it is important to consider the image formation process. A commonly assumed model of image formation is based on Lambertian reflectance, which assumes that the surface reflectance is from a matte surface of uniform albedo.

However, many types of surface, including those of faces, do not always follow Lambert's law. To overcome this problem, here we use the SFS framework for non-Lambertian surfaces proposed by Smith and Hancock [10]. The aim is to recover the surface normal $\overline{n}(x,\, y)$ and the facial depth map $Z(x{,}y)$ which gives the relative surface height above the point x-y on the image plane.

Smith and Hancock's non-Lambertian SFS algorithm first obtains the surface normal $\overline{n}$ by minimising the brightness error which is defined as a function of a point on the manifold S^2 for unit surface normals.

$$f(\overline{n}) = (g(\phi(n), \mathbf{L}, \mathbf{V}, P) - I)^2 \tag{1}$$

where $\phi(n){=}\overline{n}$ and ϕ: $S^2 \mapsto R^3$ is an embedding of the unit surface normal as a sphere, I is the measured image intensity, $\mathbf{L}$, $\mathbf{V}$ are the unit vectors in the direction of the lightsource and viewer respectively, the function $g()$ is the radiance function of the assumed (non-Lambertian) reflectance model, and P denotes the set of additional parameters specific to the particular reflectance model adopted. The radiance function of the algorithm employs the Torrance and Sparrow model. A minimisation method applicable to functions defined over a Riemannian manifolds is used to minimise Eq. (1). The local gradient of the error function f can be approximated in terms of a vector on the tangent plane to the manifold T_nS^2 using finite differences.

In addition to the brightness constraint, the algorithm also satisfies a statistical regularisation constraint. A surface in the 3D space can be expressed in terms of a linear combination of K surface basis functions Ψ_i (or modes of variation), and the height function is:

$$z_{\mathbf{b}}(x, y) = \sum_{i=1}^{K} b_i \Psi_i(x, y) \tag{2}$$

where $\mathbf{b} = (b_1,...,b_K)^T$ are the surface parameters. A surface height basis set is learnt from a set of exemplar face surfaces and the modes of variation are found by applying PCA to a representative sample of face surfaces. Here Ψ_i is the eigenvector of the covariance matrix of the training samples corresponding to the ith largest eigenvalue. In terms of the parameter vector $\mathbf{b}$, the surface normals are given by:

$$\overline{n}_{\mathbf{b}}(x, y) = (\sum_{i=1}^{K} b_i \partial_x \Psi_i(x, y), \sum_{i=1}^{K} b_i \partial_y \Psi_i(x, y), -1)^T \tag{3}$$

In order to apply this constraint to the field of surface normals $\overline{n}(x,\, y)$ satisfying the brightness contraint, the parameter vector $\mathbf{b}^*$ which minimises the distance between $\overline{n}(x,\, y)$ and $\overline{n}_{\mathbf{b}^*}(x,\, y)$, must be found. Once $\mathbf{b}^*$ is found, the surface normal and surface depth can be respectively obtained according to Eq. (3) and Eq. (2). An iterative scheme is used to compute surface normal, surface depth and the parameters of the reflection model. Details of the algorithm can be found in [10].

3 3D Motion Model

In this paper, we choose to use an affine motion model since it has proven effective in describing nonrigid motion within a small region, and has been successfully used in the application of facial motion recovery [7]. Consider a point $p_{i,l}$ in a 3D space with position $(x_{i,l},\, y_{i,l},\, z_{i,l})$ at time i. If we assume that the image is formed under perspective projection and f=1, then

$$X_{i,l} = \frac{x_{i,l}}{z_{i,l}}, Y_{i,l} = \frac{y_{i,l}}{z_{i,l}} \tag{4}$$

where $(X_{i,l},\, Y_{i,l})$ is the 2D projection of $p_{i,l}$ in frame i (i.e. at time i).

At time $i+1$, the point moves to position $(x_{i+1,l},\, y_{i+1,l},\, z_{i+1,l})$ under a nonrigid motion. Under the affine motion model, we have

$$(x_{i+1,l}, y_{i+1,l}, z_{i+1,l}, 1)^T = M_i * (x_{i,l}, y_{i,l}, z_{i,l}, 1)^T \tag{5}$$

where M_i is the affine transformation matrix given by Eq. (7). From Eq. (4), Eq. (5) and $z_{i,l}{*}Z_{i,l}{=}k$. we have

$$\begin{aligned} X_{(i+1),l} &= \frac{a_{1i}X_{i,l} + a_{2i}Y_{i,l} + a_{3i} + (a_{4i}/k)Z_{i,l}}{a_{9i}X_{i,l} + a_{10i}Y_{i,l} + a_{11i} + (a_{12i}/k)Z_{i,l}} \\ Y_{(i+1),l} &= \frac{a_{5i}X_{i,l} + a_{6i}Y_{i,l} + a_{7i} + (a_{8i}/k)Z_{i,l}}{a_{9i}X_{i,l} + a_{10i}Y_{i,l} + a_{11i} + (a_{12i}/k)Z_{i,l}} \\ Z_{(i+1),l} &= \frac{Z_{i,l}}{a_{9i}X_{i,l} + a_{10i}Y_{i,l} + a_{11i} + (a_{12i}/k)Z_{i,l}} \end{aligned} \tag{6}$$

$$M_i = \begin{pmatrix} a_{1i} & a_{2i} & a_{3i} & a_{4i} \\ a_{5i} & a_{6i} & a_{7i} & a_{8i} \\ a_{9i} & a_{10i} & a_{11i} & a_{12i} \\ 0 & 0 & 0 & 1 \end{pmatrix} M^i = \begin{pmatrix} A_{1i} & A_{2i} & A_{3i} & A_{4i} \\ A_{5i} & A_{6i} & A_{7i} & A_{8i} \\ A_{9i} & A_{10i} & A_{11i} & A_{12i} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{7}$$

If the non-rigid motion is uniform, the affine matrices are constant i.e. $M_1{=}M_2{=}...{=}\,M$, and can be learnt from the frames of the video sequence. However, using the information provided by SFS, the height information $Z_{(i+1),l}$ for each frame is known in advance. Here we aim to use the facial surface recovered from the first frame using SFS, i.e. $Z_{1,l}$ to seed the estimation of the remainder using the affine motion model. The motion model is

$$(x_{(i+1),l}, y_{(i+1),l}, z_{(i+1),l}, 1)^T = M^i(x_{1,l}y_{1,l}z_{1,l}1)^T \tag{8}$$

where M^i is given by Eq. (7). The derived equations are as follows,

$$\begin{aligned} X_{(i+1),l} &= \frac{A_{1i}X_{1,l} + A_{2i}Y_{1,l} + A_{3i} + (A_{4i}/k)Z_{1,l}}{A_{9i}X_{1,l} + A_{10i}Y_{1,l} + A_{11i} + (A_{12i}/k)Z_{1,l}} \\ Y_{(i+1),l} &= \frac{A_{5i}X_{1,l} + A_{6i}Y_{1,l} + A_{7i} + (A_{8i}/k)Z_{1,l}}{A_{9i}X_{1,l} + A_{10i}Y_{1,l} + A_{11i} + (A_{12i}/k)Z_{1,l}} \\ Z_{(i+1),l} &= \frac{Z_{1,l}}{A_{9i}X_{1,l} + A_{10i}Y_{1,l} + A_{11i} + (A_{12i}/k)Z_{1,l}} \end{aligned} \tag{9}$$

Assuming M is diagonalisable, there exists an invertible matrix P:

$$M = P\lambda P^{-1}; M^i = P\lambda^i P^{-1} \tag{10}$$

$$P = \begin{pmatrix} p_1 & p_2 & p_3 & p_4 \\ p_5 & p_6 & p_7 & p_8 \\ p_9 & p_{10} & p_{11} & p_{12} \\ p_{13} & p_{14} & p_{15} & p_{16} \end{pmatrix} P^{-1} = \begin{pmatrix} p'_1 & p'_2 & p'_3 & p'_4 \\ p'_5 & p'_6 & p'_7 & p'_8 \\ p'_9 & p'_{10} & p'_{11} & p'_{12} \\ p'_{13} & p'_{14} & p'_{15} & p'_{16} \end{pmatrix} \lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}$$

According to Eq. (10), each element in M^i is a function of i, and elements in P, P^-1, λ. We take A_{1i} as an example: $A_{1i} = B_{1i} - B_{2i}B_{3i}/B_{4i}$, where $B_{1i} = p_1p'_1\lambda_1^i + p_2p'_5\lambda_2^i + p_3p'_9\lambda_3^i + p_4p'_{13}\lambda_4^i$; $B_{2i} = p_{13}p'_1\lambda_1^i + p_{14}p'_5\lambda_2^i + p_{15}p'_9\lambda_3^i + p_{16}p'_{13}\lambda_4^i$; $B_{3i} = p_1p'_4\lambda_1^i + p_2p'_8\lambda_2^i + p_3p'_{12}\lambda_3^i + p_4p'_{16}\lambda_4^i$; $B_{4i} = p_{13}p'_4\lambda_1^i + p_{14}p'_8\lambda_2^i + p_{15}p'_{12}\lambda_3^i + p_{16}p'_{16}\lambda_4^i$. Once $(p_1, p_2, ..., p_{12})$, $(p'_1, p'_2, ..., p'_{12})$ and $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ are obtained, $(a_1, a_2, ..., a_{12})$ can be computed according to Eq. (10).

If the nonrigid motion is non-uniform, then M_i can be learnt from two consecutive frames so that $Z_{(i+1),l}$ can be straighforwardly calculated using Eq. (6). However, this surface recovery method is rather unreliable, due to the lack of a regularization constraint. As a result, the subsequent motion and facial surface estimation would potentially become more and more inaccurate. In a manner similar to that for uniform motion, we use Eq. (11) and an iteration scheme to obtain $(X_{i+1,l}, Y_{i+1,l}, Z_{i+1,l})$:

$$(x_{(i+1),l}, y_{(i+1),l}, z_{(i+1),l}, 1)^T = M_i(M_{i-1}...(M_1(x_{1,l}y_{1,l}z_{1,l}1)^T)) \tag{11}$$

The motion equations appearing in Eq. (11) are almost identical to those appearing in Eq. (9). The difference lies in the expressions $(A_{1i}, A_{2i}, ..., A_{12i})$.

4 Motion Parameters for Recognition

Facial Surface recovery. The non-Lambertian SFS algorithm describes in Section 2 is used to obtain the depth map $Z(x, y)$ which can be used in estimating nonrigid motion parameters. This SFS algorithm satisfies not only the brightness constraint but also a statistical regularisation constraint to ensure accurate recovery of the facial surface. This statistical constraint is learnt from a set of face surfaces with neutral expression. In this paper, the algorithm is used to recover the facial surface from the first image in the expression sequence since this is usually a face with a neutral expression.

Local Motion Estimation. The non-linear least-square optimization method is utilized to estimate the motion parameters. The 2D optical flow algorithm of Horn and Schunck [11] is used to compute the required flow. Let $(U_{i,l}\ V_{i,l})$ be the observed optical flow vectors. Then $(a_{1i}, a_{2i}, ..., a_{12i})$ are estimated by minimizing Eq. (12) or Eq. (13) for uniform motion and non-uniform motion respectively.

$$\chi_1^2 = \sum_{i=2}^{numf} \sum_{l=1}^{nump} (U_{i,l} - X_{i,l} - X_{i-1,l})^2 + (V_{i,l} - Y_{i,l} - Y_{i-1,l})^2 \tag{12}$$

$$\chi_2^2 = \sum_{l=1}^{nump} (U_{i,l} - X_{i,l} - X_{i-1,l})^2 + (V_{i,l} - Y_{i,l} - Y_{i-1,l})^2 \tag{13}$$

where $numf$ is the number of frames of the video sequence, and $nump$ is the number of pixels within a small patch which undergoes nonrigid motion.

Voting Rules for Classification. The motion parameters for all of the regions are used to recognize facial expression. The Adaboost algorithm is utilized to handle uniform nonrigid motion parameters, and a HMM-based classification approach is employed to train and classify non-uniform motion parameters.

We use multiple HMM-based classifiers to avoid the computational overheads associated with high-dimensional data and problems caused by uniform scaling. Each classifier is based on a HMM model for a particular expression class. The overall classification decision depends on the log-likelihood given by each HMM with the test sequence. Since we have multiple classifier we need to combine the outputs to make a final decision. Here we use the plurality rule: 1)If there is one winner according to the plurality rule, then the final decision is the only winner. 2)If there are multiple winners, then the sum of the log-likelihood for each winner is compared. The final decision is the most likely one. 3)If rules 2 still can not make the filnal decision, then randomly selects one of them as the final winner.

5 Experiments

We carry out experiments on 38 selected image sequences from the Cohn-Kanade AU-Coded Facial Expression Database [12]. The data set contains 19 subjects, and each subject has two expressions i.e. smile and surprise. The length of the sequences varies from 7 to 29 frames, and the average length is 15 frames.

Non-temporal classification. Performance comparisons between the uniform and non-uniform motion features are shown in Table 1. The accuracy rates for n-fold tests are the averages of three running iterations.

Table 1. Accuracy rate for uniform and non-uniform features

	LOO	5-fold	3-fold		LOO	5-fold	3-fold
uniform	86.84%	80.70%	78.13%	nonuniform	73.68%	72.81%	72.80%
smile	*100%*	*95.24%*	*88.33%*	*smile*	*78.95%*	*91.83%*	*87.90%*
surprise	*73.68%*	*66.67%*	*68.37%*	*surprise*	*68.42%*	*55.61%*	*59.30%*

From the results, we observe that in non-temporal classification, the uniform motion features perform better than the non-uniform motion features. Moreover, the accuracy rates for the smile expression are much higher than those for the surprise expression.

Temporal classification. Each face image is divided into 12*14 patches, and motion parameters for the patches in one row are modeled by one HMM. So 14

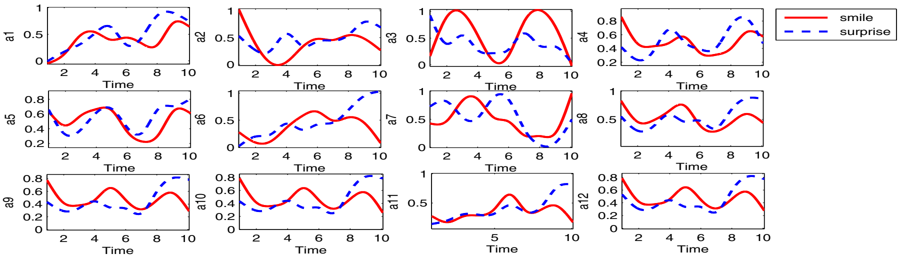


Fig. 1. Temporal non-uniform parameters

pairs of HMM models are used for classification. Each model is a 3 state left-to-right HMM with Gaussian observation symbols. The model parameters are generated randomly. The training and classification steps using the HMM-based classifier are implemented using Kevin Murphy's HMM toolbox.

Table 2 shows the classification results for the combined decision from the complete set of classifiers. It shows that the temporal features perform better than the non-temporal features. Fig. 1 shows the temporal non-uniform parameters for the two expressions.

Table 2. Accuracy rate for temporal non-uniform features

LOO	*smile*	*surprise*	5-fold	*smile*	*surprise*	3-fold	*smile*	*surprise*
92.11%	*94.74%*	*89.47%*	92.11%	*94.74%*	*89.47%*	86.84%	*84.21%*	*89.47%*

Fig. 2 gives the number of votes for each sample in a leave-one-out test. There is one incorrect classification (No.10) in the smile samples and two incorrect classifications (No.1 and No.10) in the surprise samples. In the surprise samples, No.6, No.10, and No.19 make use of voting rule 2 to make the final decision. Voting rule 3 has not been invoked by the data used in our experiments since the equality of the sum of log-likelihoods rarely occurs. Fig. 3 shows the performance

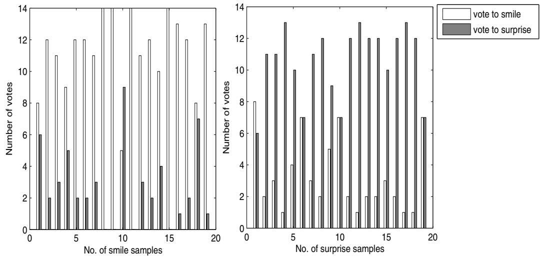


Fig. 2. votes in LOO test

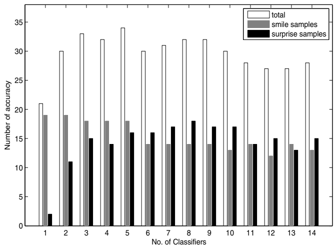


Fig. 3. accurate classification in LOO test

of each classifier. We observe that the classifiers which handle parameters from the middle face or the middle and upper face performs better than the remaining classifiers.

6 Conclusion

This paper has explored the use of local parameterized non-rigid motion recovered from 3D facial surfaces in recognizing facial expressions from video sequences. A SFS method is used to recover the 3D facial surface. An affine non-rigid motion model and an optical flow technique are used to estimate motion parameters from the estimated 3D facial surface. Finally, Adaboost and multiple HMM-based classifiers are employed to recognize expressions. We observe that the recovered non-rigid motion parameters are efficient in discriminating smile and surprise expressions.

Our future work will revolve around seeking a robust method to obtain non-rigid motion parameters. We also aim to find a means to initialize the HMM parameters and select models according to the motion data. Of course, more subjects and more expressions will be used in further research.

Acknowledgement. Edwin R. Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

References

1. Fasel, B., Luettin, J.: Automatic facial expression analysis: A survey. Pattern Recognition 36, 259–275 (2003)
2. Bartlett, M.S., Littlewort, G., Frank, M., et al.: Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In: CVPR, pp. 568–573 (2005)
3. Lucey, S., Ashraf, A.B., Cohn, J.F.: Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face. In: Delac, K., Grgic, M. (eds.) Face Recognition, Vienna, Austria, pp. 275–286 (2007)
4. Chang, Y., Hu, C., Feris, R., Turk, M.: Manifold based Analysis of Facial Expression. J. Image and Vision Computing 24(6), 605–614 (2006)
5. Black, M.J., Yacoob, Y.: Recognizing facial expressions in image sequences using local parameterized models of image motion. International Journal of Computer Vision 25, 23–48 (1997)
6. Zhu, Z., Ji, Q.: Robust pose invariant facial feature detection and tracking in real-time. In: Proceedings of the 18th ICP Recognition, pp. 1092–1095 (2006)
7. Zhou, L., Kambhamettu, C.: Hierarchical structure and nonrigid motion recovery from 2d monocular views. In: CVPR, pp. 752–759 (2000)
8. Otsuka, T., Ohya, J.: Spotting segments displaying facial expression from image sequences using hmm. In: Proceedings of the 2nd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 442–447 (1998)
9. Aleksic, P.S., Katsaggelos, A.K.: Automatic facial expression recognition using facial animation parameters and multistream hmms. IEEE Transactions on Information Forensics and Security (2006)

10. Smith, W.A.P., Hancock, E.R.: A new framework for grayscale and colour non-lambertian shape-from-shading. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 869–880. Springer, Heidelberg (2007)
11. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artifical Intelligence 17, 185–203 (1981)
12. Kanade, T., Tian, Y., Cohn, J.F.: Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53 (2000)

TIR/VIS Correlation for Liveness Detection in Face Recognition

Lin Sun*, WaiBin Huang, and MingHui Wu

Department of Computer Science, Zhejiang University City College, China
{sunl,huangwb,mhwu}@zucc.edu.cn

Abstract. Face liveness detection in visible light (VIS) spectrum is facing great challenges. Beyond visible light spectrum, thermal IR (TIR) has intrinsic live signal itself. In this paper, we present a novel liveness detection approach based on thermal IR spectrum. Live face is modeled in the cross-modality of thermal IR and visible light spectrum. In our model, canonical correlation analysis between visible and thermal IR face is exploited. The correlation of different face parts is also investigated to illustrate more correlative features and be helpful to improve live face detection ability. An extensive set of liveness detection experiments are presented to show effectiveness of our approach and other correlation methods are also tested for comparison.

Keywords: liveness detection, thermal IR, correlation analysis.

1 Introduction

Biometrics is an emerging technology that recognizes human identities based upon one or more intrinsic physiological or behavioral characteristics, e.g. faces, fingerprints, irises, voice. However, spoofing attack (or copy attack) is still a fatal threat for biometric authentication systems [1,2]. Liveness detection, which aims at recognition of human physiological activities as the liveness indicator to prevent spoofing attack, is becoming very active in fields of fingerprint recognition and iris recognition [1,3].

In the face recognition community, numerous recognition approaches have been presented, but the effort on anti-spoofing is still limited [4]. The most common faking way is to use a facial photograph of the valid user to spoof the face recognition system, since usually one's facial image is very easily available for the public, for example, downloaded from the web, captured unknowingly by the camera. Photo attack is one of the cheapest and easiest spoofing approaches. The another face spoofing way is video of valid user. It is not difficult to get and display nowadays thanks to high quality pinhole camera and tablet PC. The spoofing video has more physiological clues than photos, such as eyeblink, facial expression, and head movement. The difficulty of detecting spoofing video is that it is a re-imaging of the original live face. High quality spoofing videos are almost the same as live faces in a non-intrusive scenario.

* Corresponding author.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 114–121, 2011.

1.1 Analysis of Face Liveness

The definition of liveness in biometrics is to determine whether the biometric being captured is an actual measurement from live person who is present at the time of capture [3]. The live signals in face can be investigated from physiology, psychology, physics, etc. For example, Eyeblink is a physiological form of conditioning reflex, Q&A (question and answer) is a kind of intelligence test and thermal IR is a physical phenomenon.

Most face liveness researches are based on visible light images. From the static view, the essential difference between the live face and photograph is that a live face is a fully three dimensional object while a photograph could be considered as a two dimensional planar structure. With this natural trait, Choudhary et al. [5] employed the structure from motion yielding the depth information of the face to distinguish live person and still photo. Kollreider et al. [6] applied optical flow to obtain the movement of different parts in face for liveness judgment. Some researchers used the Q&A approaches to against spoofing, e.g. exploiting the lip movement during speech [7,8], requiring user to act an obvious response of head movement [9], reading the numbers which are hinted [10]. This kind of method needs user collaboration. Besides, the imaging difference between photo and live face in visible light spectrum is investigated for liveness detection. Li et al. [11] presented Fourier spectra to classify live faces or the faked images, based on the assumption that the high frequency components of the photo is less than those of live face images. Li et al. also stated that it would be defeated if a clear and big size photo was used to fool the system. Tan et al. [12] and Bai et al. [13] modeled the imaging difference according to Lambertian model and BRDF respectively. The weakness of these methods is unstable and effected by the quality of photos, cameras, illumination, etc. Eye's blink [14] and movement [15,16] detection methods were proposed to find the physiological clue in live face.

All these methods above are facing great challenges, e.g. depth estimation, face movement and eyeblink can be fooled by videos, Q&A approaches can be fooled by photograph cut out mouth region and placed in font of attacker's face or using photograph-pasted head mold to respond head movement, a high quality and big size photo can spoof the imaging difference of live face and photograph. Strictly speaking, only Q&A intelligence test can be considered as the live signal detection in visible light spectrum. However, computer intelligence is not high

Table 1. Spoofing live face detection examples

Live Detection methods	Spoofing methods
Depth information	Video face
Facial movement	Video face
Eyeblink	Video face
Head movement	Photograph-pasted head mold
Question & Answer	Photograph by cutting mouth out
Imaging difference in visible spectrum	High quality images

enough, therefore, it is easily fooled by human tricks. Table.1 lists the ways to spoof above methods.

1.2 Thermal IR for Live Detection

Beyond visible light spectrum, IR is also used to solve face problems, e.g. face recognition in near IR [17]. The human body emits thermal radiation in the bands of thermal IR spectrum, typically high emissions in long-wave infrared (LWIR) from 8.0-14.0μm. The thermal image of face is an intrinsic characteristic for human beings while energy metabolism is operating. Thermal face image itself indicates that it is a live face [1]. Different from visible light, thermal radiation of a live face depends on temperature, blood vessels patten, organ shape, etc. [18], therefore it is very hard to be simulated.

The efforts on anti-spoofing in thermal IR spectrum are still very limited. The most common ways are thermal IR face detection [19] and recognition [20,23]. The shortcoming of thermal IR face detection is that simple hand-drawn face can be detected as face in the classical face detection methods [21,22], therefore, thermal imaging of attacker tightly pasted by the photograph of valid user can also be possibly considered as thermal IR face. The other way is to use thermal IR for identification. However, the ability of thermal face recognition is still limited nowadays[23].

In this paper, our contributions are as follows, firstly, we present a novel live face detection method which utilizes thermal IR live signals for high security in face recognition system. To tackle this problem, we model the cross-modality of TIR/VIS face pairs in correlation analysis framework. Secondly, the canonical correlation analysis of the whole face and different face parts is illustrated and patch correlation coefficient based weighting is also presented to improve liveness detection. Thirdly, experiments show that effectiveness of our approach to detect live face and reject thermal IR spoofing.

The paper is organized as follows: in Section 2, we describe the proposed method in detail. The experiments and results are illustrated in Section 3. Section 4 gives the conclusion of this paper.

2 The Approach

2.1 Cross-Modality of TIR/VIS Face Pair

We model cross-modality of TIR/VIS face pair by canonical correlation analysis. TIR and VIS face images are captured at the same time in the authentication stage. Let (x, y) is a visible and thermal IR face pair variable with zero mean. The canonical correlation analysis [24] between x and y maximizes the correlation coefficient ρ by choosing projection directions ω_x and ω_y,

$$\rho = max_{\omega_x,\omega_y} \frac{\omega_x^T \varSigma_{xy} \omega_y}{\sqrt{\omega_x^T \varSigma_{xx} \omega_x \omega_y^T \varSigma_{yy} \omega_y}}, \tag{1}$$

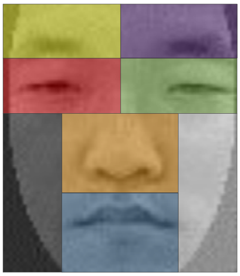


Fig. 1. Eight face patches illustration. Left eyebrow(yellow), left eye(red), left cheek(black), nose(orange), mouth(blue), right eyebrow(purple), right eye(green), right cheek(white).

where Σ_{xy} is covariance matrix between x and y. ω_x can be solved by following eigen problem,

$$\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}\omega_x = \lambda^2\omega_x, \tag{2}$$

where λ^2 is eigenvalue and ω_x is eigenvector. Then

$$\omega_y = \frac{\Sigma_{yy}^{-1}\Sigma_{yx}\omega_x}{\lambda}. \tag{3}$$

To control overfitting and avoid singular matrix Σ_{xx} and Σ_{yy}, regularization term τI are added to Σ_{xx} and Σ_{yy} , where τ is regularization parameter and I is identity matrix.

Given a new visible-thermal IR face pair (X, Y), the liveness confidence Ψ is defined as

$$\Psi(X, Y) = \frac{(\omega_x^T X) \bullet (\omega_y^T Y)}{\|\omega_x^T X\|\|\omega_y^T Y\|}, \tag{4}$$

where $\bullet$ is dot product of two vectors. A live face can be verified by comparing the liveness confidence Ψ to a predefined threshold.

2.2 Correlation Analysis of Face Patches

We divide whole face into eight patches, shown in Fig.1, and investigate the correlation between visible and thermal IR images on these patches individually. Let (U, V) are n TIR/VIS face image pairs. We use leave-one-out cross validation [25] to calculate the kth patch correlation coefficient ρ_k which is defined as follows,

$$\rho_k = corr(\left\{\omega_{U_k^{(-i)}}^T U_k^{(-i)}\right\}_{i=1}^n, \left\{\omega_{V_k^{(-i)}}^T V_k^{(-i)}\right\}_{i=1}^n), \tag{5}$$

where $\omega_{U_k^{(-i)}}^T, \omega_{V_k^{(-i)}}^T$ are the first eigen vectors trained by CCA on $(U_k^{(-i)}, V_k^{(-i)})$ with ith pair removed. (U_k, V_k) are the kth patch of visible light images U and thermal infrared images V respectively.

To emphasize higher correlated patches, we assign weight w_k to kth patch and the final face liveness confidence is a weighted sum,

$$\Psi_w(X,Y) = \Sigma_{k=1}^{K} w_k \Psi(X_k, Y_k). \tag{6}$$

The value of weight w_k is referred to following equations,

$$\begin{aligned} w_k &= ln(\frac{\rho_k}{1-\rho_k}) \\ w_k &\longleftarrow \frac{w_k}{\Sigma_{k=1}^{K} w_k}. \end{aligned} \tag{7}$$

3 Experiments

To evaluate performance of the proposed approach, we use public IRIS Thermal/Visible Face Database in OTCBVS benchmark database [26]. We select 120 visible-thermal IR image pairs of 30 persons with surprise, smile, angry and neutral expressions from this database. 8 persons wear glasses and other 22 persons do not.The eye and mouth centers are labelled manually. Faces are aligned with eyes and mouth's coordinates, resized to 70 × 80 pixels and cropped by an elliptical mask. Visible and thermal IR face image examples are shown in Fig.2. The pixels in the image are simply arranged into a vector in raster-scan manner. The dimension of correlation subspace is set to 60.

Patch correlation coefficients are calculated on data exclude glasses wearing. The second column of Tab.2 shows the correlation coefficient ρ_k of eight face patches. Eyes and mouth patches give the highest correlation in all.

Visible light and thermal IR images from the same person are considered as live face. In the live face detection experiment, one person is leaved out for test and others for training and totally 120 live tests are done. Thermal IR spoofing is carried out using valid user's photograph and attacker's thermal IR face to spoof our liveness detection approach. Any two persons are picked out from the database to spoof each other and the remaining persons are used to train correlation subspace.

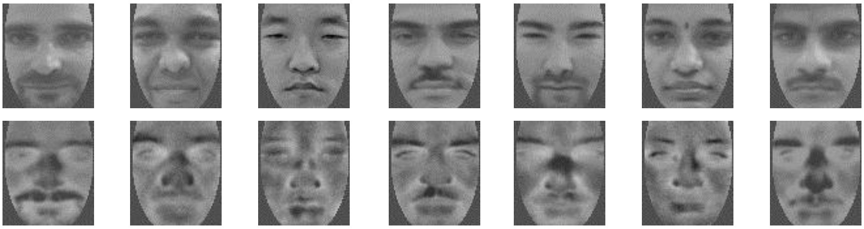


Fig. 2. Visible and thermal IR face image examples. Top row is visible light face and bottom row is thermal IR face.

Table 2. Correlation coefficient of eight patches in face

Patch	ρ_k
left eyebrow	0.861
left eye	0.938
left cheek	0.785
nose	0.831
mouth	0.903
right eyebrow	0.842
right eye	0.921
right cheek	0.791

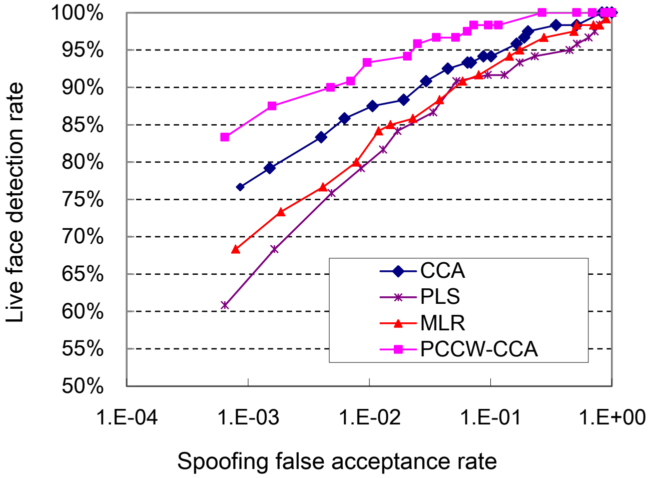


Fig. 3. Comparison of MLR, PLS, CCA and PCCW-CCA

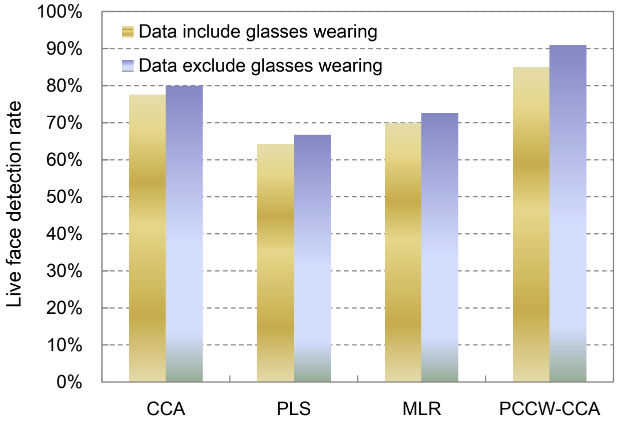


Fig. 4. Effects of glasses wearing for correlation analysis(FAR=0.1%)

Fig. 3 shows performance of our patch correlation coefficient based weighting CCA (PCCW-CCA) and the comparison results with other multivariate analysis methods, multivariate linear regression (MLR) and partial least squares (PLS). We do experiments on data include wearing glasses and exclude wearing glasses respectively. Fig. 4 shows that glasses wearing produce small bad effect because thermal IR is blocked by glasses. The results show that our PCCW-CCA outperforms than others and live detection rate achieve 85.1% and 90.8% on data include and exclude glasses wearing repectively when spoofing false acceptance rate is 0.1%.

4 Conclusions

This paper has given a novel approach to detect live face in multiple spectrum. Characteristics of live face in the thermal IR spectrum are intrinsic live signals. To against thermal IR spoofing, the cross-modality of thermal IR and visible light face is modeled by correlation analysis. The experiments shows that it is feasible to detect live face by TIR/VIS face pair and it is really highly secure for the impossibility of spoofing valid user's thermal IR face. Glasses wearing will produce small bad effect because of the blocked thermal IR and the difficulty of eye alignment. The comparison experimental results show that CCA outperforms other multivariate analysis methods, MLR and PLS, in our problem.

Acknowledgements. This work was partly supported by National Science Foundation of China and we would like to thank Dr. Gang Pan in Zhejiang University for valuable advice.

References

1. Schuckers, S.: Spoofing and Anti-Spoofing Measures. Information Security Technical Report, 7(4), 56-62 (2002)
2. Jain, A.K., Pankanti, S., Prabhakar, S., Hong, L., Ross, A.: Biometrics: A Grand Challenge. In: ICPR, pp. 935–942 (2004)
3. Li, S.Z., Jain, A.K.: Encyclopedia of Biometrics (2009)
4. Zhao, W., Chellappa, R., Phillips, J., Rosenfeld, A.: Face Recognition: A Literature Survey. ACM Computing Surveys, 399-458 (2003)
5. Choudhury, T., Clarkson, B., Jebara, T., Pentland, A.: Multimodal person recognition using unconstrained audio and video. In: AVBPA 1999, pp. 176–181 (1999)
6. Kollreider, K., Fronthaler, H., Bigun, J.: Evaluating liveness by face images and the structure tensor. In: Fourth IEEE Workshop on Automatic Identification Advanced Technologies, pp. 75–80 (2005)
7. Frischholz, R.W., Dieckmann, U.: BioID: A Multimodal Biometric Identification System. IEEE Computer 33(2), 64–68 (2000)
8. Chetty, G., Wagner, M.: Multi-level Liveness Verification for Face-Voice Biometric Authentication. In: Biometrics Symposium (2006)
9. Frischholz, R.W., Werner, A.: Avoiding Replay-Attacks in a Face Recognition System using Head-Pose Estimation. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2003), pp. 234–235 (2003)

10. Kollreider, K., Fronthaler, H., Faraj, M.I., Bigun, J.: Real-Time Face Detection and Motion Analysis With Application in "Liveness" Assessment. IEEE Transactions on Information Forensics and Security 2(3), 548–558 (2007)
11. Li, J., Wang, Y., Tan, T., Jain, A.: Live Face Detection Based on the Analysis of Fourier Spectra. In: Proc. SPIE of Biometric Technology for Human Identification, vol. 5404, pp. 296–303 (2004)
12. Tan, X.Y., Li, Y., Liu, J., Jiang, L.: Face Liveness Detection from A Single Image with Sparse Low Rank Bilinear Discriminative Model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 504–517. Springer, Heidelberg (2010)
13. Bai, J., Ng, T.T., Gao, X.T., Shi, Y.Q.: Is physics-based liveness detection truly possible with a single image? In: Proceedings of the 2010 IEEE International Symposium on Circuits and Systems, pp. 3425–3428 (2010)
14. Pan, G., Sun, L., Wu, Z.H., Lao, S.H.: Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera. In: ICCV, pp. 1–8 (2007)
15. Jee, H.K., Jung, S.U., Yoo, J.H.: Liveness detection for embedded face recognition system. International Journal of Biomedical Sciences 1(4), 235–238 (2006)
16. Wang, L.T., Ding, X.Q., Chi, F.: Face Live Detection Method Based on Physiological Motion Analysis. Tsinghua Science & Technology 14(6), 685–690 (2009)
17. Li, S.Z., Chu, R.F., Liao, S.C., Zhang, L.: Illumination Invariant Face Recognition Using Near-infrared Images. IEEE Transactions on PAMI (Special issue on Biometrics: Progress and Directions) 29(4), 627–639 (2007)
18. Buddharaju, P., Pavlidis, I., Tsiamyrtzis, P., Bazakos, M.: Physiology-Based Face Recognition in the Thermal Infrared Spectrum. IEEE Transactions on PAMI 29(4), 613–626 (2007)
19. Wang, X.Y., Chen, J.H., Wang, P.J., Huang, Z.H.: Infrared Human Face Auto Locating Based on SVM and A Smart Thermal Biometrics System. In: Proceedings of ISDA, pp. 1066–1069 (2006)
20. Socolinsky, D.A., Selinger, A., Neuheisel, J.D.: Face Recognition with Visible and Thermal Infrared Imagery. Computer vision and image understanding 91(1-2), 72–114 (2003)
21. Rowley, H., Baluja, S., Kanade, T.: Neural Network-Based Face Detection. IEEE Trans. on PAMI 20(1), 23–38 (1998)
22. Viola, P., Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: CVPR 2001, pp. 511–518 (2001)
23. Chen, X., Flynn, P.J., Bowyer, K.W.: IR and Visible Light Face Recognition. Computer Vision and Image Understanding 99, 332–358 (2005)
24. Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis, 5th edn. Prentice Hall, Englewood Cliffs (2002)
25. Leurgans, S.E., Moyeed, R.A., Silverman, B.W.: Canonical Correlation Analysis when the Data are Curves. Journal of the Royal Statistical Society 55(3), 725–740 (1993)
26. http://www.cse.ohio-state.edu/otcbvs-bench/

Person Localization and Soft Authentication Using an Infrared Ceiling Sensor Network

Shuai Tao, Mineichi Kudo, Hidetoshi Nonaka, and Jun Toyama

Division of Computer Science, Hokkaido University, Japan
{taoshuai,mine,nonaka,jun}@main.ist.hokudai.ac.jp

Abstract. Person localization and identification are indispensable to provide various personalized services in an intelligent environment. We propose a novel method for person localization and developed a system for identifying up to ten persons in an office room to realize soft authentication. Our system consists of forty-three infrared ceiling sensors with low cost and easy installation. In experiments, the average distance error of person localization was 31.6cm that is an acceptable error for sensors with 1.5m distance to each other. We also confirmed that walking path and speed gives sufficient information for authenticating the user. Through the experiments, we obtained the correct recognition rates of 98%, 95% and 86% for any pair, any three people and all ten people to identify individuals.

Keywords: localization, soft authentication, sensor network, infrared sensors.

1 Introduction

In recent years, along with the rapid development of network devices and person authentication technology, it has become possible to provide many kinds of personalized services in response to the implicit/explicit demands of the users. In such an intelligent environment, people can use voice, face, gait and other physical features to realize the person localization and authentication. In this situation, we need localization to know where the users are and need authentication to know those who want services.

Commercial authentication systems using various biometric evidences, such as fingerprint, iris, speech and palm vain, can maintain a high level of security, but such a high-level security is not necessary in daily life. In daily life situation, misidentification dose not cause a serious problem. Rather, psychological/physical disturbance should be seriously considered.

For distinguishing our motivation from the motivation for security, we call the authentication for personalized services *soft authentication*[1] and call the authentication for security *hard authentication.*

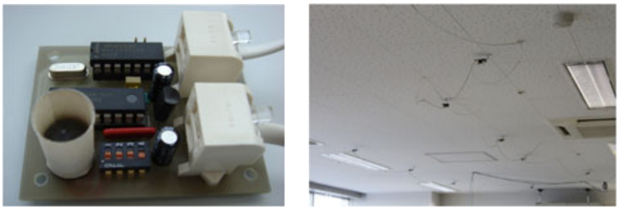
Video cameras have been used in some studies for person localization [2]. However, cameras might violate privacy. Schulz et al. [3] tried to use an ID badge for person localization and authentication. However, many people, especially elderly people, would be unwilling to wear such sensing devices. Shankar et al. [4] tried to use pyroelectric infrared sensors for human identification and localization in a relatively small room. On the contrary, our first ceiling-sensor system is applicable for a large room situation [1].

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 122–129, 2011.

Recently, in order to increase the sampling rate and reduce the noise in the previous system [1], we have improved the ceiling sensor system [5]. In this renovated system, sampling rate of 80 Hz for up to 128 nodes using 250 kbps equilibrium line has been realized. However, due to the characteristics of infrared sensors, the information we obtain is still only the fact that someone is under or just passed under the active sensor. In this study, we propose a novel method for person localization to bring a finer precision (31.6cm) than that of the geometrical precision (1.5m) of sensor placement. By using the location information, the performance of soft authentication has also been improved in discrimination rate.

2 Infrared Sensing System

In the improved system, ”pyroelectric infrared sensor”, sometimes called ”infrared motion sensor”, are attached to the ceiling [5]. This sensor detects an object with a different temperature from the surrounding temperature. We used NaPiOn (AMN11111, Panasonic Denko Co. Ltd.) as the sensor module. There are 16 lenses for gathering infrared radiation to 4 quadrants on the surface of the pyroelectric infrared detector. Then, 64 detection zones are formed in front of the sensor module. The detection area is up to 7.42 m × 5.66 m on a plane at a distance of 2.5 m from the sensor. In our system, a hand-made cylindrical lens hood was used to narrow the detection area of each sensor. The photographs of the sensor module and the interconnection of sensor nodes with cables are shown in Fig. 1. Such infrared sensors are easy to set up at a low cost ($20/unit). Light conditions and movable obstacles do not affect the performance.



(a) A sensor module (b) Connection of sensors

Fig. 1. The sensor module and the interconnection of sensor nodes with cables

Forty-three sensors were attached to the ceiling of our research room (15.0 m × 8.5 m) so as to cover all the area and not to produce any dead space. The average distance between each other is 1.5m. Figure 2 shows the layout of the room and the arrangement of the sensors. A binary response from each sensor can be read at the sampling rate from 1 Hz to 80 Hz.

Users are not required any cooperation for authentication and they do not feel that they are being observed. These are necessary requirements for soft authentication.

In our sensor network, motions of one person often make multiple sensors active. There is also a *get-out delay* of sensors in response to motions, that is, an active sensor keeps the active status for a few seconds after a person left the sensing area. There is no

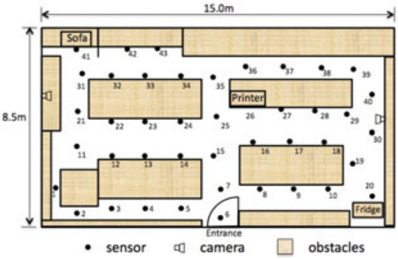


Fig. 2. Layout of infrared sensors

get-in delay. Another important fact is that the sensor sometimes cannot be active if the person moves slightly, such as, keyboard typing or browsing with a mouse.

3 Person Localization

3.1 The Method for Person Localization

In indoor environments, person localization has the requirements: (1) Estimate the location of the person at each time frame with an acceptable distance error; (2) Show the short-term walking trajectory of the person; (3) Evaluate the speed of moving.

In the ceiling sensor system of our laboratory, we can assume that: (1) The walking speed v of a person is known; (2) Detection area is a circle of radius R; (3) Active status will be kept for D_{delay} (sec.) after the person getting off the detection area and D_{delay} does not depend on the speed v; (4) In Fig. 3, we assume that the person enters the detection area with an angle α and the duration of active status is decomposed as $D = t_e - t_s = D_{detect} + D_{delay}$ if the person gets out of the detection area at time frame t $(> t_e)$.

From the sensor model in Fig. 3, we see that there are four cases to be considered: (1) At position P_0 (at time frame t_0 before detection), the distance from the sensor is $r_0 > R$. (2) At position P_1 (at time frame t_1 under detection), $r_1^2 = D^2v^2 + R^2 - 2RDv\cos\alpha$ ($D = t - t_s < \frac{2R\cos\alpha}{v}$). (3) At position P_2 (at time frame t_2 out of detection area but the sensor is still active), $r_2^2 = D^2v^2 + R^2 - 2RDv\cos\alpha$ ($\frac{2R\cos\alpha}{v} < D < \frac{2R\cos\alpha}{v} + D_{delay}$). (4) At position P_3 (at time frame t_3), the sensor becomes inactive again, and the distance from the sensor is $r_3 > R$.

For situations (2) and (3), with the expected value $\frac{2}{\pi}$ of $\cos\alpha$ in range $-\frac{\pi}{2} < \alpha < \frac{\pi}{2}$, we use the expected value of squared distance as $E(r^2) = D^2v^2 + R^2 - \frac{4}{\pi}RDv$.

Algorithm

(1) If a sensor S_i has already been active for duration D_i, we estimate the distance to the person by $r_i = \sqrt{D_i^2v^2 + R^2 - \frac{4}{\pi}RD_iv} = \sqrt{(D_iv - \frac{2}{\pi}R)^2 + (1 - \frac{4}{\pi^2})R^2}$

(2) Gathering all the information D_i and thus r_i $(i = 1, \cdots, n)$ from all active sensors, estimate the position $P_t = (x_t^*, y_t^*)$ at time frame t by solving

$$\min_{P_t} \sum_{i=1}^{n} (r_i - \|S_i - P_t\|)^2 = \min_{(x,y)} \sum_{i=1}^{n} \{r_i - \sqrt{(x_i - x)^2 + (y_i - y)^2}\}^2$$

The solution (x_t^*, y_t^*) satisfies:

$$\begin{cases} x = \sum w_i x_i / \sum w_i \\ y = \sum w_i y_i / \sum w_i \end{cases}, \quad w_i = \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2} - r_i}{\sqrt{(x_i - x)^2 + (y_i - y)^2}}.$$

Therefore, with appropriate initial values, we can find the solution P_t by iteration.

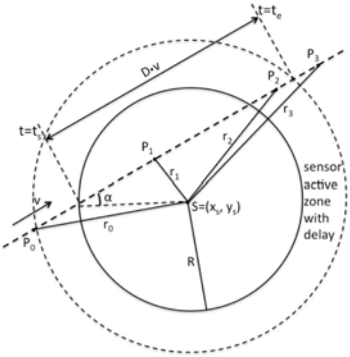


Fig. 3. The sensor model that contains four cases when a person passes by. Without generality, we may assume that he/she enters at the left end of x-axis.

3.2 Basic Evaluation of the Person Localization Method

We examined our localization method for the case that three subjects walked along the same route (from the entrance to the sofa in Fig. 2). The initial values of the estimated location were set to the average location of all the active sensors ($x_0 = \sum_{i=1}^{n} x_i/n, y_0 = \sum_{i=1}^{n} y_i/n$) and iteration was repeated 10 times. The moving speed was set to 1.3m/s that is an average speed of a person in our laboratory. The radius R of the detection area is 0.75m. The true positions of the three subjects at each time frame were determined subjectively from the image sequences taken from two video cameras established for the purpose of evaluation (the locations are shown in Fig. 2). The true positions and estimated positions of three subjects are both shown in Fig. 4. We can see that the trajectories of three subjects are successfully estimated. The average distance errors of estimated positions of three subjects was 31.6cm.

4 Identification Information

In an office environment, the motions of a person usually switch back and forth between walking and sitting. By observing people's behavior in our laboratory, we noticed the possibility to identify individuals from the different speeds of sitting up and starting walking. So, we measured the starting speeds of walking on the basis of our localization method in each time frame.

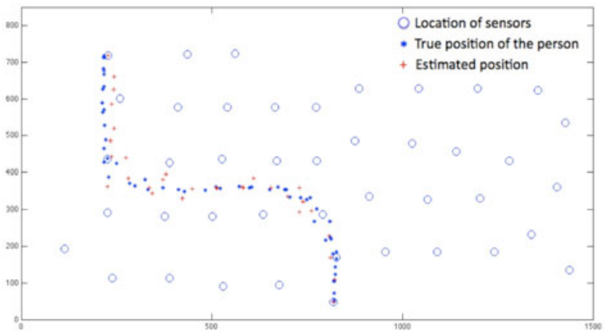


Fig. 4. The true positions and estimated positions of three subjects

For investigating the varying regularity of the speed, we examined the speed of a subject during the period in which the subject starts walking. A subject was asked to stay below a sensor for a while, and then to move to another position for 20 times. The sampling rate was 2 Hz. The speed was calculated after localization in every time frame. The average speed of the 20 times is shown in Fig. 5.

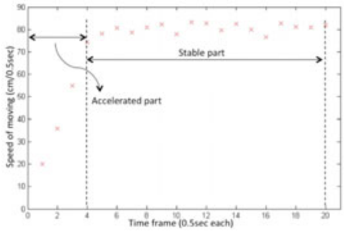
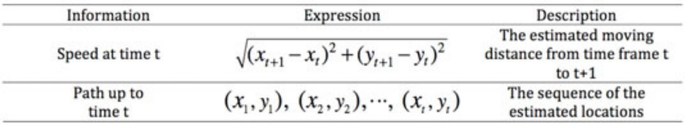


Fig. 5. The speed of a subject during the period of starting to move

We noticed first that the time required for accelerating is about two seconds. After that, the speed of the subject becomes stable with slight fluctuation. Therefore, we can separate the period into two parts: an ”accelerate part” and a ”stable part” (Fig. 5). We expected that there is a cue for indentifying the users in both of the two parts.

Each person in an office room has his/her own living habits and tends to linger at some certain areas. Therefore, the walking paths are also expected to hold information

Table 1. The description of the speed and path information



Information	Expression	Description
Speed at time t	$\sqrt{(x_{t+1}-x_t)^2+(y_{t+1}-y_t)^2}$	The estimated moving distance from time frame t to t+1
Path up to time t	$(x_1,y_1),\ (x_2,y_2),\cdots,\ (x_t,y_t)$	The sequence of the estimated locations

for recognizing multiple persons. In our experiment, the walking speed and path information for a short period (3 sec.) are both used for identifying multiple persons. Here, we use only a short path because we want to identify entering users as soon as possible. The descriptions of the speed and path information are given in Table 1.

5 Authentication Experiments

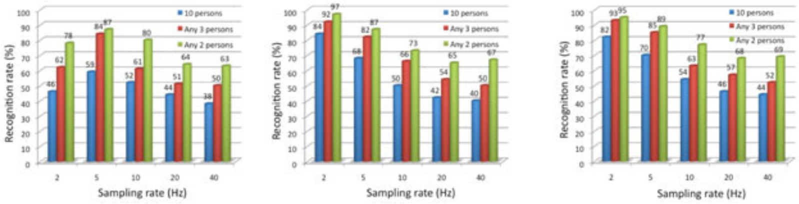
We distinguished the ”stable part” and ”accelerate part” to examine the potential power of the speed and path information in two cases of ”continuous soft authentication” and ”immediate soft authentication.” To do this, we prepared two kinds of datasets.

A. Ten subjects (laboratory students) were asked to enter the room from outside and then to go forward to their own desks directly without stopping for twenty times. For realizing a fast authentication, we consider the first 3 seconds to measure the speed and path.

B. The same subjects were given different instructions. They were asked to move into the room from outside, stay a while for changing shoes (2-3s with a strong motion), then move to their own desks for twenty times. When they finished changing shoes and started moving, we began to measure the speed and path for 3 seconds.

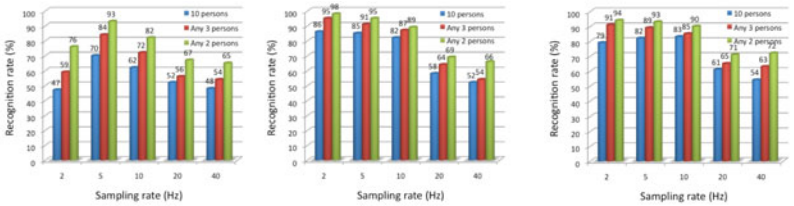
Dataset A provided the information of the stable part, and dataset B provided that of the accelerate part. We used five kinds of sampling rate: 2Hz, 5Hz, 10Hz, 20Hz and 40Hz. The recognition rate was calculated by 20-fold cross-validation. The classifier was a support vector machine (SVM) with a radial basic kernel with default parameter values, the soft margin parameter was 1.0 and the variance parameter was taken as the dimensionality (the number of features). Here, the number of full features is 3T of $(v_1, x_1, y_1), (v_2, x_2, y_2), \cdots, (v_T, x_T, y_T)$ for time period T (T=6 for 2Hz and 3s measurements). The numbers of features are T for speed only and 2T for path only. The results are shown in Fig. 6 and Fig. 7.

We can find that the sampling rates of 2Hz, 5Hz and 10Hz brought better results. In the stable part (Fig. 6), we see that the best performance with speed information only is 59% for ten subjects, 84% for any three subjects and 87% for any two subjects. With path information only, the best performance is 84% for ten subjects, 92% for any three subjects and 97% for any two subjects. With both of the speed and path information, the best performance is 82% for ten subjects, 93% for any three subjects and 95% for any two subjects. In the accelerate part (Fig. 7), we obtained 70%, 84%, 93% for ten,



(a) With speed (b) With path (c) With both speed and path

Fig. 6. Identification rates in the stable part



(a) With speed (b) With path (c) With both speed and path

Fig. 7. Identification rates in the accelerated part

three, two subjects with speed information, 86%, 95%, 98% for ten, three, two subjects with path information. We also obtained 83%, 91%, 94% for ten, three, two subjects with both of the speed and path information. The recognition rate is a little higher in the accelerate part than in the stable part, which means that people show their personalities more when they start walking than keep walking.

6 Discussion

In our previous system [1], finger vein was used for identification at the entrance as a necessary information of individual tracking. That, however, needs one time cooperation of users and is against the purpose of soft authentication. In addition, after a while, our system would lose the users because of the characteristics of this system. Then we need another evidence for recovering the identification precision to a required level. Therefore, to realize a highly-reliable system for soft authentication, we need to collect as many pieces of evidence as possible. In the previous system, we used the enter/leave information as strong pieces of evidence and the long-stay information at a certain desk as a weak piece of evidence. In addition, a chair system measuring *hip-print* was developed for authentication [6]. In this paper, we added two more pieces of evidence: the walking speed and path. Through the experiments we knew that both have information for authentication to some extent, especially short path after entrance gives sufficient information. In the next phase, combination of all the pieces of evidence would be gathered to improve this system.

Nowadays, there are many researches about the Activities of Daily Living (ADL) [7-8]. Precise person localization of our study has the potential application to automatically monitoring the ADL of single living elder. As described in Introduction, preserving privacy is necessary for this goal and thus our system is one of the promising ways.

7 Conclusion

We have tried to localize and identify subjects in an office room using a ceiling sensor network in order to realize soft authentication for providing personalized services.

In the experiments, the average error of person localization was 31.6cm. With the localization technique, the speed and path information were measured and exploited for identification. From the measurement of constantly walking persons, we obtained 59%, 84%, 87% for ten, three, two subjects with speed information, 84%, 92%, 97% for ten, three, two subjects with path information, 82%, 93%, 95% for ten, three, two subjects with both of the speed and path information. From the measurement of starting walking persons after shoes exchange, we obtained 70%, 84%, 93% for ten, three, two subjects with speed information, 86%, 95%, 98% for ten, three, two subjects with path information, 83%, 91%, 94% for ten, three, two subjects with both of the speed and path information.

Acknowledgement. This work was supported by JSPS Grant-in-Aid for Scientific Research (C) 10213216.

References

1. Hosokawa, T., Kudo, M., Nonaka, H., Toyama, J.: Soft authentication using an infrared ceiling sensor network. Pattern Anal. Applic. 12, 237–249 (2009)
2. Tung Ying, L., Tsung Yu, L., Szu Hao, H., Shang Hong, L., Shang Chih, H.: People localization in a camera network combining background subtraction and scene-aware human detection. In: Lee, K.-T., Tsai, W.-H., Liao, H.-Y.M., Chen, T., Hsieh, J.-W., Tseng, C.-C. (eds.) MMM 2011 Part I. LNCS, vol. 6523, pp. 151–160. Springer, Heidelberg (2011)
3. Schulz, D., Fox, D., Hightower, J.: People tracking with anonymous and id-sensors using rao-blackwellised particle lters. In: Proceedings of International Joint Conference on Articial Intelligence, IJCAI (2003)
4. Shankar, M., et al.: Human tracking systems using pyroelectric infrared detectors. Opt. Eng. 45(10), 106401–106410 (2006)
5. Nonaka, H., Tao, S., Toyama, J., Kudo, M.: Ceiling sensor network for soft authentication and person tracking using equilibrium line. In: The 1st International Conference of Pervasive and Embedded Computing and Communication Systems (PECCS), pp. 218–223 (2011)
6. Yamada, M., Kamiya, K., Kudo, M., Nonaka, H., Toyama, J.: Soft authentication and behavior analysis using a chair withsensors attached: hipprint authentication. Pattern Anal. Applic. 12, 251–260 (2009)
7. Bruno, L., et al.: What is happening now? Detection of activities of daily living from simple visual features. Personal and Ubiquitous Computing 14, 749–766 (2010)
8. Anthony, F., Norbert, N., Michel, V.: Improving Supervised Classification of Activities of Daily Living Using Prior Knowledge. International Journal of E-Health and Medical Communications (IJEHMC) 2, 17–34 (2011)

Categorization of Camera Captured Documents Based on Logo Identification

Venkata Gopal Edupuganti[1], Frank Y. Shih[1], and Suryaprakash Kompalli[2]

[1] Dept. of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA-07102
{vge2,shih}@njit.edu
[2] Hewlett-Packard Labs, Bangalore, India
kompalli@hp.com

Abstract. In this paper, we present a methodology to categorize camera captured documents into pre-defined logo classes. Unlike scanned documents, camera captured documents suffer from intensity variations, partial occlusions, cluttering, and large scale variations. Furthermore, the existence of non-uniform folds and the lack of document being flat make this task more challenging. We present the selection of robust local features and the corresponding parameters by comparisons among SIFT, SURF, MSER, Hessian-affine, and Harris-affine. We evaluate the system not only with respect to amount of space required to store the local features information but also with respect to categorization accuracy. Moreover, the system handles the identification of multiple logos on the document at the same time. Experimental results on a challenging set of real images demonstrate the efficiency of our approach.

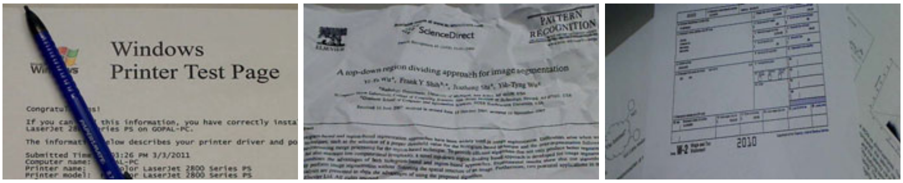
Keywords: Logo detection, affine-invariant features, clustering, hamming embedding.

1 Introduction

Logos [5,9,11] are interesting objects that serve enormous purposes ranging from ownership identification to document retrieval. There are three types of logos [11]: one with only graphics, the other with only text, and finally a mix of both. In this paper, we address categorization of camera captured documents based on logo detection. Unlike scanned documents, camera captured logo identification is more challenging as it encounters partial occlusions, background clutter, intensity variations, and crumpled documents as shown in Fig. 1. Furthermore, a single document might contain multiple logos as shown in Fig. 1(b).

In the recent years, logo identification and recognition have been addressed by a significant number of researchers. Majority of these approaches [5,11,12,14] rely on connected component extraction. A Bayesian approach by providing feedback between detection and recognition phases is specified in [11]. A method based on boundary extraction of feature rectangles to generate robust candidate logos is proposed in [12]. Geometric relationship among connected components is enforced in [5] to eliminate outliers. In [9], SIFT [6] features from a query image (image under observation) are matched against all the descriptors of logo models. Though the accuracies are good,

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 130–137, 2011.



(a) (b) (c)

Fig. 1. (a) partial occlusion, (b) crumpled document with multiple logos, and (c) background cluttering

matching against all the model descriptors is not a good choice. The main objective of all these methods is the logo identification on scanned documents. In the following sections, we introduce an efficient method to detect logos on camera captured documents under various deformations.

In order to address non trivial deformations such as partial occlusions, intensity variations, and view point changes, we adapt local affine-invariant features to represent the pre-defined logo models and the query image. Due to the availability of various local affine-invariant features such as SIFT [6], SURF [2], MSER [7], Hessian-Affine [7], and Harris-Affine [7], there is always a question of selecting the good feature type.

The rest of the paper is organized as follows: Section 2 presents the comparison of various local affine-invariant features and the selection of one for the logo detection task. We present the detailed methodology of camera captured document categorization in section 3. Section 4 presents the experimental results on a challenging data set, we also discuss the impact of dimensionality reduction and representation of the features. Finally, Section 5 concludes the paper.

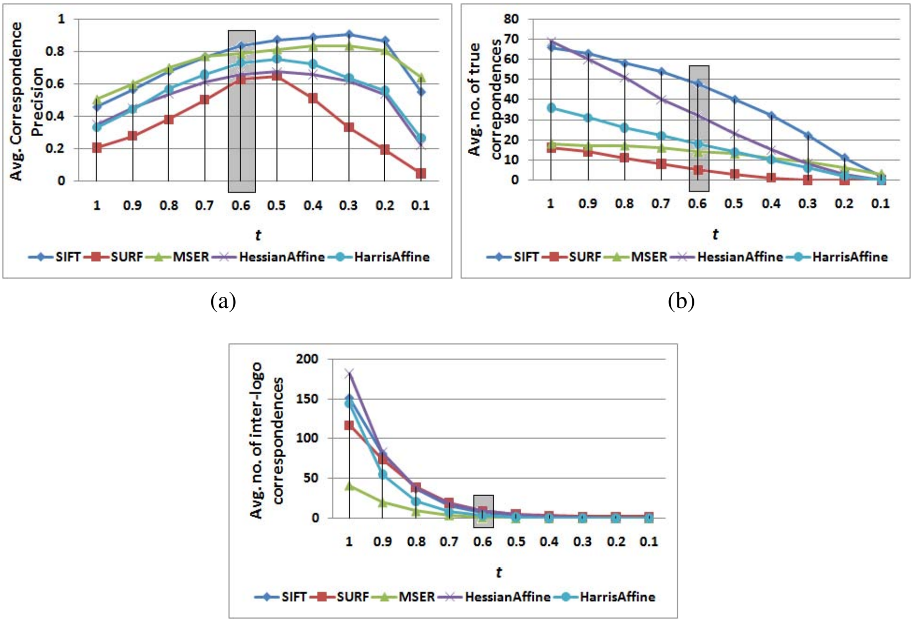
2 Comparative Analysis of Local Affine-Invariant Features

In this section, we present the selection of desired local feature by comparisons among various local affine-invariant features. The features in consideration are SIFT [6], SURF [2], MSER [7], Hessian-Affine [7], and Harris-Affine [7]. The comparison is done using 25 logo models and 125 camera captured documents with 5 documents under each logo model. Let $L = \{L_1, L_2, ..., L_m\}$ be a set of logo models, where m is the total number of logo models. Each logo model L_i is represented by using n_i feature points $L_i = \{(x^j, y^j, f^j)\}$ for $j \in \{1, 2, ..., n_i\}$, where n_i is the total number of feature points in the i^{th} logo model; (x^j, y^j) and f^j are the Cartesian coordinates and d-dimensional description of the j^{th} feature point respectively. Similarly, query image is represented as Q=$\{(x_q^j, y_q^j, f_q^j)\}$ for $j \in \{1, 2, ...n_q\}$, where n_q is the total number of features points extracted from the query document. We denote the j^{th} feature in L_i and Q as L_i^j and Q^j respectively, and the corresponding d-dimensional feature descriptors as f_i^j and f_q^j respectively.

We adapt Lowe's [6] threshold t for comparison (as defined in Eqn. 1), which is the ratio of distance between the logo descriptor and the first nearest neighbor among the query descriptors $f_q \in Q$ in the d-dimensional feature space (i.e.f_q^{nn1}) to that of the second nearest neighbor (i.e.f_q^{nn2}).

$$t = \frac{D(f_i^j, f_q^{nn1})}{D(f_i^j, f_q^{nn2})} \tag{1}$$

where $D()$ is the Euclidean distance in d-dimensional feature space, and $nn1, nn2 \in \{1, 2, ..., n_q\}$ are the indices of the first and second nearest neighbors to f_i^j in the feature space. A correspondence for each L_i^j is established with Q^{nn1} only if t is less than a pre-defined threshold, i.e. Q^{nn1} is the corresponding feature point to j^{th} feature of L_i in Q. As t goes down from 1 and approaches 0, the ambiguity in the correspondences decreases, and more discriminative correspondences will be established.



(a) (b)

(c)

Fig. 2. Comparison among various local affine-invariant features

We analyze the behavior of local affine-invariant features with respect to three important criteria: correspondence precision, number of true correspondences, and number of inter-logo correspondences. Correspondence precision (as defined in Eqn. 2) and the number of true correspondences are analyzed by establishing the correspondences between each logo model and the corresponding 5 camera captured documents. The number of true correspondences is counted with the help of established ground truth. Fig. 2(a) and Fig. 2(b) show the behavior of average correspondence precision and average number of true correspondences at different thresholds t respectively. An ideal feature type must have high average correspondence precision along with large number of feature points to support partial occlusions and non-rigid deformations in the logo. Fig. 2(c) shows the average number of inter-logo correspondences established with different feature types at various thresholds of t (for each logo model $L_i \in L$, we use the remaining models $L_{i'} \in L; i \neq i'$ as queries). As some of the local features are common among multiple logos, using all the features will reduce the discriminative power. One with lower number of average inter-logo correspondences should be preferred. From Fig. 2, SIFT features at the shaded threshold t, i.e. 0.6, is the desired choice compared to the remaining feature types and the thresholds. Section 3 presents how we use the derived feature type and the corresponding threshold t to build an efficient logo-based categorization system.

$$Correspondence\ Precision = \frac{Number\ of\ true\ correspondences}{Total\ number\ of\ correspondences} \tag{2}$$

3 Methodology

The system has two modes of operation: off-line and on-line. Off-line mode is responsible for feature extraction from logo models, representation, and storage of the extracted data. On-line mode works in two stages. In stage 1, features are extracted from a query document and are matched against the features in the database to determine the candidate logo models. In stage 2, top l candidate logo models are then subjected to cluster-based refinement process in the image space to eliminate false positives. Finally, the query document is categorized into the candidate logo models left after stage 2. Fig. 3 shows the overview of our system configuration. The following subsections briefly explain the individual components of the system. We discuss the significance of optional components of Fig. 3 in section 4.

3.1 Off-Line: Representation and Storage of Logo Model Features

Let $X = \{(x^j, y^j, f^j)\}$, $1 \leq j \leq n$ be the set of SIFT [6] features extracted from all the logo models $L_i \in L$; where n is the total number of logo model features.

1. **Dimensionality Reduction:** It reduces the dimensionality of SIFT [6] features. Generate a 128×128 dimensional matrix P with random numbers. Subject P to QR decomposition [3] to obtain an orthogonal matrix Q. The first r_d rows of the matrix Q form the projection matrix R. Project all the descriptors $f^j \in X$ onto R to reduce their dimensionality to r_d.
2. **Cluster Formation:** Form the clusters of descriptors $f^j \in X$ in r_d-dimensional space using k-means [4], and denote the cluster centroids as $C = \{c_i | 1 \leq i \leq k\}$.
3. **Hamming Embedding (HE):** The main objective of this step is to convert the feature $f^j \in X$ to a binary string b^j for efficient representation, storage, and matching. For each r_d-dimensional descriptor $f^j \in C_i; 1 \leq i \leq k$, we adapt hamming embedding [3] to convert it to a bit string b^j of length r_d as defined in Eqn. 3.

$$\begin{aligned} b^j(x) &= 1, if\ f^j(x) <= C_i(x);\ 1 \leq x \leq r_d \\ &= 0,\ otherwise; \end{aligned} \tag{3}$$

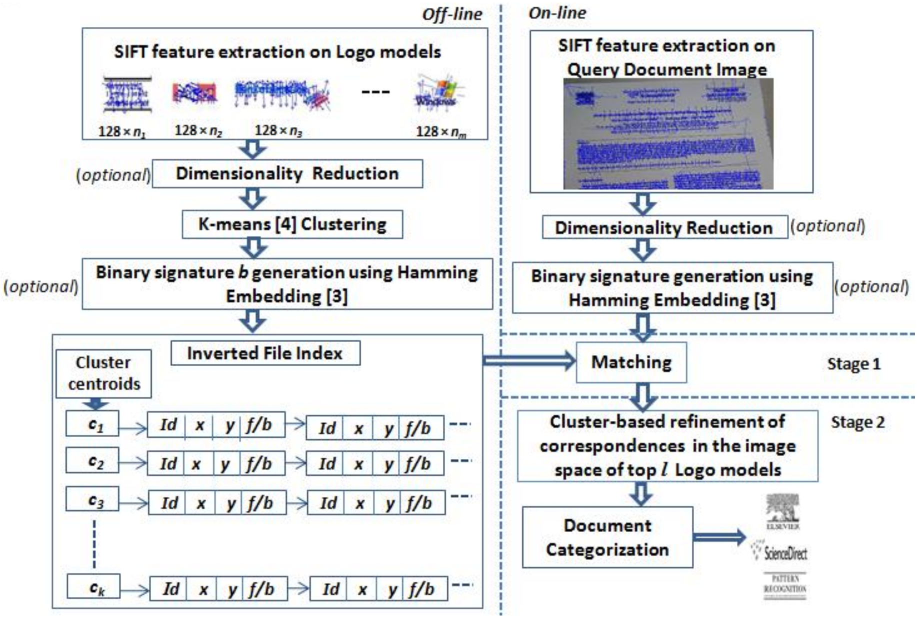


Fig. 3. Document categorization framework

4. **Inverted file index:** We adapt inverted file indexing [3,10] structure to store the logo models information. Only the cluster centroids $C_i \in C$ are indexed, and all the SIFT [6] features within each cluster are linked to their corresponding cluster centroid. The feature information attached is the logo model number(Id), Cartesian coordinates x^j, y^j, and the feature f^j (or) binary string b^j as shown in Fig. 3. Denote the established index structure as I.

3.2 On-line: Feature Extraction on Query Document and Matching

Let $Q = \{(x_q^j, y_q^j, f_q^j)\}$, $1 \le j \le n_q$ be the set of SIFT [6] features extracted from a query document image and represented in the similar manner as logo model features (section 3.1); where n_q is the total number of SIFT [6] features extracted from the query document. Algorithm 1 presents the mechanism of matching features in Q with the established inverted file index I of section3.1 and computation of scores $S_i \in S; 1 \le i \le m$ of logo models.

Stage 2 matching: Refinement of scores using neighborhood check. As the scores after stage 1 matching comprise lot of outliers, we refine the established correspondences in the top l candidate logo models using cluster-based neighborhood check in the image space. One can enforce the ordering among the local features [13], and check for the relative order consistency between query document and the candidate logo model, or refine the correspondences by fitting a transformation model [6] to the correspondences. Due to the non-rigid deformations (i.e. crumpled document), we apply a cluster-based

Algorithm 1. Stage 1 matching

```
Input: Inverted File Index $I$(section 3.1), Query features $Q$.
Output: Scores $S_i \in S; 1 \leq i \leq m$ of the logo models.
Initialize: All $S_i \in S$ to zero.
for all $Q^j \in Q$ do
    Determine the nearest cluster $C_i \in I$;
    Initialize: $D$ (Distance to all features $\in C_i$) to zero.
    for all $(b^z|f^z) \in C_i$ do
        Compute the distance $D^z$=D($b^z|f^z$,$Q^j$); where D() is xor() for $b^z$, and Euclidean distance in $r_d$-dimensional space for $f^z$;
    end for
    sort $D$ in decreasing order;
    Increment the score $S_{Id(nn1)}$ by 1 only if $(D^{nn1}/D^{nn2}) \leq t$; where $D^{nn1}$ and $D^{nn2}$ are the distances to the first and second nearest features of $Q^j$, and $t$ is Lowe's [6] threshold;
end for
sort $S$ in decreasing order;
```

Algorithm 2. Stage 2 matching: cluster-based neighborhood check

```
Input: Top $l$ candidate logo models $L' \in L$ after stage 1 matching, and the corresponding scores $S' \in S$.
Output: Refined scores $S'$ of the candidate logo models.
for all $L_i \in L'$ do
    Initialize: neighborhood cardinality $r_e$ to $\lceil$ sqrt($S'_i$)$\rceil$.
    repeat
        for all features $(x^j, y^j) \in L'_i$ do
            Let $N(x^j, y^j)$ and $N_q(x^j_q, y^j_q)$ be the $r_e$ neighborhood features of the $j^{th}$ correspondence between the logo model $L'_i$ and the query document $Q$ respectively;
            Determine the probability of $j^{th}$ correspondence being an inlier as $P^j = \frac{(N(x^j,y^j) \cap N_q(x^j_q,y^j_q))}{r_e}$;
            Mark the $j^{th}$ correspondence as inlier if $P^j \geq t_p$; where threshold $t_p$ is set to 0.5;
        end for
        Update the correspondences in $L_i$ with inliers, and refine the $S'_i$ with the cardinality of $L'_i$ i.e. $||L'_i||$;
        Update $r_e$ to minimum of $\lceil$ sqrt($S'_i$)$\rceil$ and ($r_e$-1);
    until $r_e \leq 3$
end for
sort $S'$ in decreasing order, and eliminate all the logo models $L'_i \in L'$ with the scores $S'_i \leq 3$;
```

neighborhood check in the image space to determine the outliers. Algorithm 2 presents the underlying mechanism.

4 Experimental Results and Discussion

Our test set consists of 375 camera captured query documents of resolution 1600×1200 belonging to 25 logo models. We adapt F_measure[8] as defined in Eqn. 4 to evaluate the system. The higher the F_measure, the better the categorization accuracy.

$$\begin{aligned} Recall &= \frac{Number\ of\ true\ categories\ identified}{Total\ number\ of\ true\ categories} \\ Precision &= \frac{Number\ of\ true\ categories\ identified}{Total\ number\ of\ identified\ categories} \\ F\_measure &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{aligned} \tag{4}$$

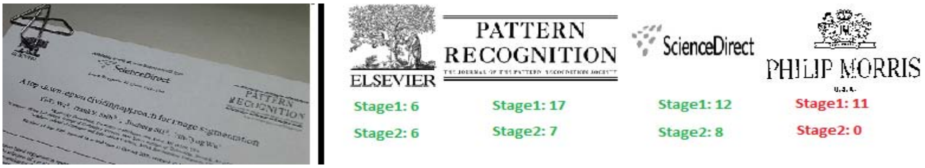


Fig. 4. Category identification: left:query document, right: predicted categories(true: scores in green, false: scores in red i.e. rightmost logo model)

Table 1. F_measure[8] at different stages of matching, and different feature representations

	Feature representation					
	128	64	32	16	HE-128	HE-64
Stage 1	64.03%	54.99%	58.21%	40.73%	44.93%	32.51%
Stage 2	77.95%	73.92%	72.52%	55.18%	68.24%	59.96%

Table 1 shows the accuracies at different stages, and different SIFT [6] feature representations with k=100, t=0.5, and l=5. HE-128 and HE-64 in the table 1 corresponds to feature representation with Hamming Embedding(HE) and bit string lengths of 128 and 64 respectively. From the table 1, as the dimension of the SIFT [6] features decreases from 128 to 16, the corresponding stage 2 F_measure decreases gradually, and stage 2 matching significantly improves the stage 1 matching F_measures. HE with 128-bit string representation achieves a reasonable F_measure of 68.24% with enormous savings in storage. We observe a similar kind of pattern at k=50 and k=200, with a minor change of 1 to 2% in F_measure, and slightly higher measures with increasing number of clusters k. We also empirically verified the derived threshold t=0.6 by a comparison among other threshold values, and observed higher F_measures at t=0.6. We achieved a F_measure of 36.54% by directly adapting the HE method of [3] with 128 bits and the specified parameters. Finally, we verified our method on Tobacco-800 [1] dataset and achieved a 95.14% F_measure as opposed to 92.5% using [5]. Finally, Fig. 4 shows the scores of identified categories of a query document at each stage. On an average, it takes 1 second to categorize the given query document on Intel core 2 duo machine using MATLAB.

5 Conclusions and Future Work

This paper presents a methodology to categorize camera captured documents based on logo identification. The selection of robust features is done by comparisons among various local affine-invariant features. The methodology not only categorize the document in the case of partial occlusions, intensity variations, and non-rigid deformations but also identify multiple categories if present. The system is evaluated with respect to different feature representations. Improving the categorization accuracies by adapting optimal representations constitutes the focus of our future work.

References

1. Agam, G., Argamon, S., Frieder, O., Grossman, D., Lewis, D.: The IIT Complex Document Image Processing (CDIP) Test Collection Project. Illinois Institute of Technology, USA (2006)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346–359 (2008)
3. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 304–317. Springer, Heidelberg (2008)
4. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C., Silverman, R., Wu, A.Y.: The analysis of a simple k-means clustering algorithm. In: Proc. the Sixteenth Annual Symposium on Computational Geometry, pp. 100–109. Hong Kong University of Science and Technology (June 2000)
5. Li, Z., Schulte-Austum, M., Neschen, M.: Fast logo detection and recognition in document images. In: Proc. 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp. 2716–2719 (August 2010)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
7. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision 65(1-2), 43–72 (2005)
8. Olsen, D.L., Delen, D.: Advanced Data Mining Techniques, 1st edn. Springer, Heidelberg (February 2008)
9. Rusinol, M., Llados, J.: Logo spotting by a bag-of-words approach for document categorization. In: Proc. 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, pp. 111–115 (July 2009)
10. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Proc. 9th International Conference on Computer Vision, Nice, France, pp. 1470–1477 (October 2003)
11. Wang, H.: Document logo detection and recognition using bayesian model. In: Proc. 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp. 1961–1964 (August 2010)
12. Wang, H., Chen, Y.: Logo detecion in document images based on boundary extension of feature rectangles. In: Proc. 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, pp. 1335–1339 (July 2009)
13. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: Proc. IEEE Intl. Conf. on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 25–32 (June 2009)
14. Zhu, G., Doermann, D.: Automatic document logo detection. In: Proc. 9th International Conference on Document Analysis and Recognition, Curitiba, Brazil, pp. 864–868 (September 2007)

Multiple Line Skew Estimation of Handwritten Images of Documents Based on a Visual Perception Approach

Carlos A.B.Mello[1], Ángel Sánchez[2], and George D.C. Cavalcanti[1]

[1] Center of Informatics, Federal University of Pernambuco, Recife, PE, Brazil
{cabm,gdcc}@cin.ufpe.br
http://www.cin.ufpe.br/~viisar
[2] Rey Juan Carlos University, Madrid, Spain
angel.sanchez@urjc.es

Abstract. This paper introduces Viskew: a new algorithm to estimate the skew of text lines in digitized documents. The algorithm is based on a visual perception approach where transition maps and morphological operators simulate human visual perception of documents. The algorithm was tested in a set of 19,500 synthetic text line images and 400 images of documents with multiple skew angles. The skew angles for the synthetic dataset are known and our algorithm achieved the lowest mean square error in average when compared with two other algorithms.

Keywords: Document processing, skew estimation, visual perception.

1 Introduction

Document analysis and recognition systems are subdivided into several modules in order to achieve high performance. In general, the scanned document is first thresholded into a bi-level image, i.e., an image where the paper is converted into white and the ink into black [13].

After thresholding, segmentation can be carried out in different ways [6]. Document segmentation identifies text and graphical areas in the image. The text areas are segmented into lines and the lines into words (or, even further, into characters in typewritten documents); this is called text segmentation. The final objects (words or characters) are submitted to classifiers for the final recognition phase. During segmentation phase, several factors can cause errors such as noise.

The source of errors can also come from the beginning of the process in the scanning phase. At this point, most part of the errors is due to rotation of the original document during digitization. When the document is typewritten, OCR (Optical Character Recognition) tools can correct this rotation with a small effect in the recognition rate. To correct the image it is common to use the Hough transform (HT) [5] for skew estimation and further skew correction [5][15].

However, rotation can appear in different and more difficult situations, some of which are not associated to errors in any phase. For example, it is common

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 138–145, 2011.

to find in handwritten documents skew angles that are related to the writing of the person who wrote that document; or inclinations that occur when someone writes in unlined paper. Hough transform is defined to find just one skew angle in an entire image; this and other methods are not suited for situations as the one shown in Fig. 1. Another problem associated to HT-like methods is the computational cost, although in [8], this cost is reduced by applying the HT on the horizontal decomposition generated by a wavelet transform (Haar family).

Handwritten documents are the most difficult type of images to process with an automatic recognition system [6]. In general, at every further step, it is harder to produce high quality responses. This is no different for skew estimation, even more in cases where there is more than one skew angle to be estimated.

In this paper, we present a new skew estimation algorithm which is developed for handwritten images with one or several, possibly different, line skew angles. The paper is divided as follows: next section reviews some skew estimation techniques; Section 3 presents the new method; experiments are described in Section 4, while, in Section 5, we conclude the paper.

2 Skew Estimation

As previously stated, most skew estimation algorithms consider that the document has a single skew angle. A rotated document will lead to a much harder segmentation process. Fig. 1 shows an example of the results of text line segmentation using Basu et al.'s method [1] in a rotated document with and without skew correction. A correct skew estimation and correction is clear with this example as without them several different lines are segmented as just one.

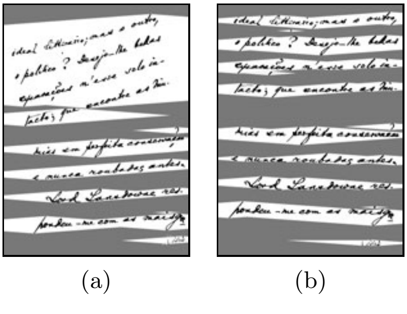

(a) (b)

Fig. 1. Basu et al.'s line segmentation algorithm [1] applied to a handwritten document (left) without and (right) with skew correction

A segmentation algorithm based on the idea that the text of a document can be put inside a bounding box is proposed in [11]. Drawing this bounding box by finding the extreme corners of the text image allows the evaluation of the skew angle. The same authors previously introduced another skew estimation algorithm based on histogram and connected component analysis [12]. In [2], an algorithm is proposed to segment text from complex background in video frames. After locating the text lines, the projection profile is found and the skew angle is the one that produces a minimum entropy histogram. A method based on morphological operators has been proposed in [16]. Skew estimation is achieved in [14] by counting the amount of black pixels in the rows, but it is not able to deal with rotated documents which do not have large black areas. A very robust algorithm was presented by Chou et al. [3] where the image is segmented into four slabs and parallelograms are formed in each slab as bounding boxes around the text lines. Different angles are tested to form these parallelograms. The skew angle that forms more white areas indicates the rotation of the image. An improvement to Chou et al.'s method was proposed in [10]. Not only the computational cost was highly reduced but also the original algorithm was improved, as it is now able to have a better response when applied to noisy images and documents with tables. In [9], it was presented an algorithm for skew estimation based on horizontal and vertical white runs counting.

As outlined, these algorithms are suited for documents with just one skew angle, which in general is caused by the scanning process.

3 Viskew: A New Algorithm for Skew Estimation

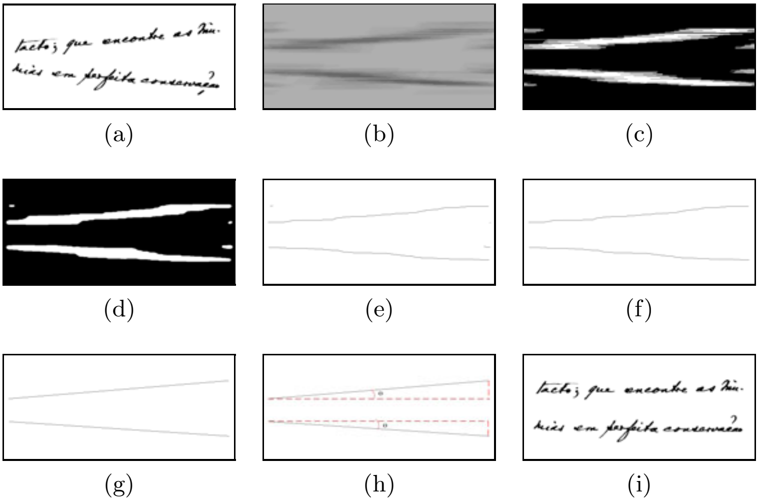
In order to illustrate the main steps of the proposed skew estimation method, a sample black-and-white image (Fig. 2a) was created with very different skew angles. This is not an expected real example but it is being used just to demonstrate the major steps of the proposed method.

The main idea of our proposal uses some aspects of our visual system and some theories from visual perception. The first time we see a document, we can perceive several characteristics of it without focusing our attention specifically on any of them. For example, in general, we can perceive if the document has figures or not, if it is handwritten or typewritten, or if its text is written with a zero degree skew angle in relation to the superior and inferior margins of the paper. Documents with text line with different skew angles is very common when the document is handwritten and the sheet of paper has no guide lines to help the writer to keep a straight text. Usually, the text lines change the skew angle along the document generating a document with multiple skew angles. After this initial perception of the document, we can concentrate in reading the document, focusing our attention on its contents. This aspect of our visual system can be understood as what is called pre-attentive vision [17] which involves the visual processes that operate before we attend to an object.

With this in mind, we can consider that several details of the document do not need to be observed to have some features perfectly perceived. Thus, the first idea is to loose details. This can be achieved by the application of a blur effect in the document image. However, this operation can group different text lines vertically. Then we opted to use transition maps to simulate this effect [7]. This map counts the number of transitions (from black to white or vice versa) which appears in a sliding window of height h and width w. For the evaluation of the transition map, it is proposed the use of 1-pixel height windows [7]. For our experiments, the width is set to 180 pixels as small widths can segment the lines into words. The window is centered in the pixel that is being processed. The result (the transition map) is a grayscale image where the zero values correspond to regions where there are no transitions, i.e., regions without text (Fig. 2b).

The result of the transition map gives an idea of the skew angles of the document. But as the behavior is still not completely clear, we try to decrease even more the amount of details. The transition map is binarized using Otsu thresholding algorithm [13] and a low pass filter is used to smooth the map, creating large homogeneous areas. These two steps are presented in Fig. 2 for the transition map of Fig. 2b. We also maintain the ideas of Gestalt grouping principles as established for document processing in [4]: good continuity is maintained as each text line is kept together; proximity is also applied as elements that are closer are merged together (words from the same text line).

Once we have this filtered image, the skeleton of each white area is computed using the algorithm of [18]. These skeletons represent the line axis of each text line. Small skeletons are removed as it is not expected that they represent real text lines. Based on our experiments, for 200 dpi resolution documents, the skeleton is considered small if it has area less than 250 pixels. This removal of small skeletons also makes our method suitable to deal with noisy images.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 2. (a) Sample document used to illustrate the main phases of the proposal and (b) transition map for the sample document of Fig. 2a (the contrast was enhanced for a better visualization), (c) binarization of the transition map of Fig. 2b, (d) the result after a low pass filtering on Fig. 2c, (e) skeletization of the filtered image of Fig. 2d, (f) final image with small skeletons removed, (g) best fit linear functions, (h) a zooming into the lines of Fig. 2g with the skew angle θ to be evaluated and (i) final image after skew estimation and correction

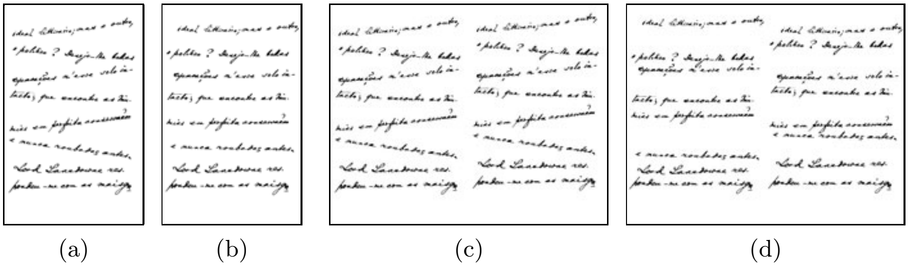
Fig. 2 shows the initial skeleton image (Fig. 2e) and the final image after the removal of the small skeletons (Fig. 2f). It is clear that these line axes have the skew angle information.

Using the coordinates of each separated line axis, a first degree function is calculated so that it fits best using least squares approximation (Fig. 2g). Now it is just necessary to compute the skew angle of each line and to proceed with the inverse rotation for skew correction (Fig. 2h). For example, for the sample document depicted in Fig. 2a, the algorithm estimates an angle of 4.73 degrees for the first text line and -3.81 degrees for the second one. The negative value indicates a clockwise rotation (which will thus require a counter-clockwise rotation for correction).

A search for the components of the original image that are connected to the line axis defines the text that is going to be rotated according to the specific skew angle of that axis. The text line that corresponds to each separate line axis receives a different label. Inverse rotations with the skew angles are then applied to each text line to generate the final image (Fig. 2i).

4 Experiments and Discussion

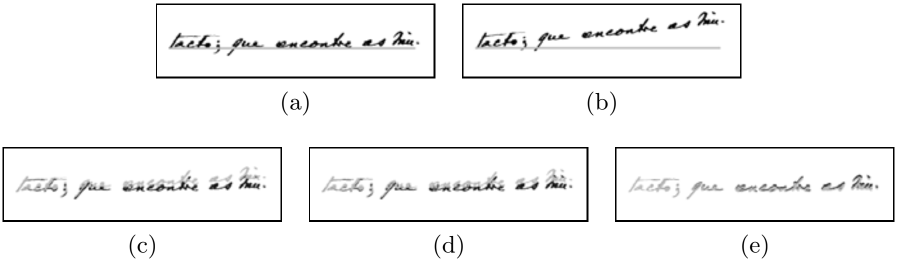
Fig. 3 presents a more complex synthetic sample document with very different skew angles and the final corrected image (for skew correction, we used [5]) after skew estimation by our algorithm. In Fig. 3c, we show a synthetic double column document, generated by duplicating the image of Fig. 3a. Fig. 3d presents the corrected image.



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

Fig. 3. (a) Original document and (b) the final image after skew estimation with Viskew and skew correction. (c) Original double column document and (d) the final image after skew estimation with Viskew and skew correction.

Another experiment was done with the first text line of the image depicted in Fig. 2i. We separated that line from the rest of the image and rotated it so that it becomes as straight as possible (Fig. 4a). This straight line was considered as the ground truth. Next, the text line was rotated 5 degrees counter-clockwise (Fig. 4b). We have applied Mascaro et al. [10], Chou et al. [3] and our algorithm to this rotated image and we compared the results to the gold standard. Fig. 4c, Fig. 4d and Fig. 4e present a comparison between the gold standard text line (in black) and the text lines obtained after skew correction of the rotated image with the skew angles detected by these three algorithms (in light gray). We also compared the results of skew estimation by the three different algorithms, as presented in Table 1. The proposed algorithm estimated the small angle for the gold standard image and it also estimated the closest slant for the 5 degrees rotated image.


(a) (b)

(c) (d) (e)

Fig. 4. (a) Gold standard, (b) its 5 degrees rotated version; a comparison of Fig. 4a and the skew correction with skew angles estimated by: (c) Chou et al., (d) Mascaro et al. and e) our proposed algorithm. In this figure the dark text is the gold standard and the light grey is the corrected text line (see quantitative results on Table 1).

Table 1. Skew estimation of Chou et al., Mascaro et al., and Viskew for the ground truth image and its 5 degrees rotated version (see Fig. 4)

Image	Chou et al.	Mascaro et al.	Viskew
Ground truth (0 degree)	-2.00	-1.50	0.42
5 degrees rotated	3.70	4.20	4.53

Other examples are shown in Fig. 5, which depicts real images of historical documents with multiple skew angles and the images generated after skew correction using the angles found by our proposed algorithm.

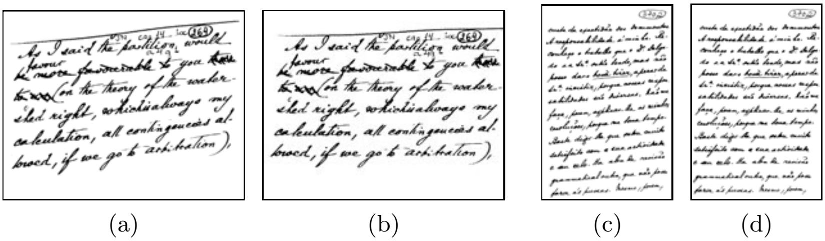

(a) (b) (c) (d)

Fig. 5. Two real handwritten historical documents: (a) and (c): the original documents; (b) and (d): the images after multiple skew estimation and correction

None of the tested algorithms achieved acceptable results, as they are suited for documents with just one skew angle. The algorithms introduced in [3] and [10] when applied to the sample document of Fig. 2a found unique angles of -0.1 and -0.5 degrees, respectively. These values are not even close to the angles found by our algorithm for each line (4.73 and -3.81 degrees) which resulted in the corrected image (Fig. 2i).

For evaluation of the proposed algorithm, two experiments were developed. In both of them, a set of 296 documents were segmented generating 1,500 text lines. These text lines were straightened in order to obtain gold standard text lines. Rotations are then imposed to these gold standard text lines from 0.5 to 6 degrees with a 0.5 step. This process produced 18,000 text lines (12 rotations for each one of the 1,500 gold standard lines).

In the first experiment, Mascaro et al. [10], Chou et al. [3] and our proposed algorithm were applied to the complete set of 19,500 text lines (the rotated and the gold standard images). As the angles are known a priori, it was possible to analyze the mean square error of the skew angle estimated for each technique. Our method achieved the lowest value in average (1.00) against 49.96 (Chou et al.) and 6.35 (Mascaro et al.). Other algorithms were not tested as they are not suited for this application.

For the second experiment, these rotated text lines were randomly grouped into 400 synthetic images of document (each one with ten text lines). Again, the skew angle of each text line is known. The proposed algorithm was applied to this set of 400 documents. The estimated skews were analyzed reaching an average mean square error of 0.896.

5 Conclusions

This paper presents Viskew: a robust algorithm for skew estimation of handwritten documents with different skew angles at each text line based on aspects of our visual system. It starts with the evaluation of a transition map to simulate pre-attentive aspects of vision removing details of the document image. The skeleton of this map is used to generate an axis line which allows the estimation of the skew angle. The components of the text that are connected to the skeleton are classified as part of that line and then rotated according to the defined skew angle for that line. The approach was applied to a set of 150 images of documents with a total amount of more than 1,700 text lines. Isolated rotated lines were also tested and compared with the results of Chou et al.'s and Mascaro et al.'s algorithms. In every experiment, our method achieved more precise skew angles estimation.

Different values of window width in the transition map phase could allow the skew estimation of different angles in the same text line. We will examine this issue in further studies. We will also try to attack the problem of non-linear text lines.

Acknowledgments. This research is partially sponsored by CNPq under grant 473926/2010-5, FACEPE APQ-0266-1.03/10 and Spanish Ministerio de Ciencia e Innovacin under project number TIN2008-06890-C02-02.

References

1. Basu, S., et al.: Text Line Extraction from Multi-Skewed Handwritten Documents. Pattern Recognition 40, 1825–1839 (2007)
2. Zhou, C., et al.: Skew Estimation and Segmentation of Text Line in Video Frames. In: International Symposiums on Information Processing, Moscow, pp. 379–383 (2008)
3. Chou, C., et al.: Estimation of skew angles for scanned documents based on piecewise covering by parallelograms. Pattern Recognition 40, 443–455 (2007)
4. Eglin, V., Bres, S.: Analysis and interpretation of visual saliency for document functional labeling. IJDAR 7, 28–43 (2004)
5. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Addison-Wesley, Reading (2007)
6. OGorman, L., Kasturi, R.: Document Image Analysis. IEEE Press, Los Alamitos (1996)
7. Kennard, D.J., Barrett, W.A.: Separating Lines of Text in Free-Form Handwritten Historical Documents. In: International Conference on Document Image Analysis for Libraries, France, pp. 12–23 (2006)
8. Khorissi, N., et al.: Application of the Wavelet and the Hough Transform for Detecting the Skew Angle in Arabic Printed Documents. In: ISSPA, United Arab Emirates, pp. 1–4 (2007)
9. Lu, S., Wang, J., Tan, C.L.: Fast and Accurate Detection of Document Skew and Orientation. In: International Conference on Document Analysis and Recognition, Brazil, pp. 684–688 (2007)
10. Mascaro, A.A., Cavalcanti, G.D.C., Mello, C.A.B.: Fast and robust skew estimation of scanned documents through background area information. Pattern Recognition Letters 31, 1403–1411 (2010)
11. Sarfraz, M., Rasheed, Z.: Skew Estimation and Correction of Text using Bounding Box. In: International Conference on Computer Graphics, Imaging and Visualisation, Malaysia, pp. 259–264 (2008)
12. Sarfraz, M., et al.: On Skew Estimation and Correction of Text. In: International Conference on Computer Graphics, Imaging and Visualisation, Thailand, pp. 308–313 (2007)
13. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging 1, 146–165 (2004)
14. Sharma, P.K., et al.: A Rule Based Approach for Skew Correction and Removal of Insignificant Data from Scanned Text Documents of Devanagari Script. In: SITIS, China, pp. 899–903 (2007)
15. Thanh, N.D.: A Rotation Method for Binary Document Images Using DDA Algorithm. In: ACM Document Engineering, Brazil, pp. 271–274 (2008)
16. Thanh, N.D., et al.: A Robust Document Skew Estimation Algorithm Using Mathematical Morphology. In: IEEE International Conference on Tools with Artificial Intelligence, Greece, pp. 496–503 (2007)
17. Wolfe, J.J., Kluender, K.R., Levi, D.M.: Sensation and Perception. Sinauer Associates Inc., (2009)
18. Zhang, T.Y., Suen, C.Y.: A Fast Parallel Algorithm for Thinning Digital Patterns. Communications of the ACM 27, 236–239 (1984)

Space Variant Representations for Mobile Platform Vision Applications

Naveen Onkarappa and Angel D. Sappa

Computer Vision Center, Edifici O, Campus UAB,
08193 Bellaterra, Barcelona, Spain
{naveen,asappa}@cvc.uab.es

Abstract. The log-polar space variant representation, motivated by biological vision, has been widely studied in the literature. Its data reduction and invariance properties made it useful in many vision applications. However, due to its nature, it fails in preserving features in the periphery. In the current work, as an attempt to overcome this problem, we propose a novel space-variant representation. It is evaluated and proved to be better than the log-polar representation in preserving the peripheral information, crucial for on-board mobile vision applications. The evaluation is performed by comparing log-polar and the proposed representation once they are used for estimating dense optical flow.

Keywords: log-polar mapping, space-variant representation, optical flow.

1 Introduction

Space variant representation schemes have been used in the computer vision field in order to improve the efficiency of proposed solutions. *Log-Polar Representation* (LPR) is one of the most widely used. It is inspired by the biological vision systems [1], [2] and has been exploited in the robotics and active vision communities for pattern recognition [3] and navigation [4] tasks. The LPR has many advantages with respect to the conventional cartesian representation of images [5]; the most important are the reduction in the data and invariance to scale and rotation. The data reduction due to the polar mapping and logarithmic sub-sampling leads to a high resolution in the fovea and a low resolution in the periphery, which is a desired feature for instance in the active vision community.

A review of log-polar imaging is presented in [6] for robotic vision applications such as: visual attention, target tracking and 3D perception. All these applications benefit from the high resolution of the fovea region. There have been also attempts to use LPRs for motion analysis [4] [7], mainly based on the estimation of optical flow. For instance, [5] presents the advantages of polar and log-polar mapping to the cartesian representation and proposes a technique to estimate time-to-impact using optical flow. In [8], a novel optical flow computation approach is proposed. It is based on the concept of variable window and generalized dynamic image model. The variable window adapts its size along the LP space. Also working in the LP space, [9] analyzes the polar deformation and proposes several local optical flow estimation techniques on log-polar plane.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 146–154, 2011.

In the particular contexts of robotics and advanced driver assistance systems (ADAS), LPR has attracted the attention of many researchers. In general, in these fields LPRs are obtained using the *vanishing point* (VP) as a center of the log-polar reference system. $VP_{(x,y,z)}$ corresponds to a point at $z \rightarrow \infty$ where two parallel lines of a road appear to converge in the image plane. Since LPR results in a high sampling in the fovea region, the periphery are under-sampled. It should be noted that the periphery corresponds to regions near to the camera reference system, hence are the most important areas for robotics navigation tasks and ADAS applications. Furthermore, features near to the camera are not only useful for detection tasks but also for an accurate calibration; note that the accuracy of 3D data decreases with the depth.

In the current work a new space variant representation scheme is proposed. It is intended to overcome the problem of LPR with respect to periphery in forward facing motion problems. The superiority of the proposed representation, to LPR, is analyzed using dense optical flow on these representations. The paper is organized as follows. Section 2 presents the proposed space variant representation and optical flow estimation. Then, experimental results and a comparative study are given in Section 3. Finally, the work is concluded in Section 4.

2 Proposed Approach

This section introduces first, the LPR and then the proposed space variant representation of cartesian images; next, the basic variational optical flow model is presented.

2.1 Space-Variant Representations

A log-polar representation is a polar mapping with logarithmic distance along the radial axis. For a given pixel (x, y), the log-polar (ρ, θ) are defined as:

$$\rho = log(\sqrt{(x-x_0)^2 + (y-y_0)^2}), \qquad \theta = \arctan((y-y_0)/(x-x_0)), \qquad (1)$$

where (x_0, y_0) is the origin of mapping; the current work focuses on the study of the particular scenario of forward facing moving platforms, hence the origin of the reference system corresponds to the vanishing point.

As mentioned above, LPR oversamples the fovea and undersamples the periphery. This leads to the non-preservation of vital information of the periphery useful for mobility applications. The latter motivates us to propose a better space variant representation, where a (x, y) pixel is mapped as:

$$\rho = log(r_{max} - \sqrt{(x-x_0)^2 + (y-y_0)^2}), \quad \theta = \arctan((y-y_0)/(x-x_0)), \quad (2)$$

where r_{max} is the radius of the largest inner circle around VP in the cartesian image. This is different from LPR in the sense that logarithmic subsampling is from the periphery towards the center and will be referred as *Reverse Log-Polar Representation* (RLPR). Figure 1 (*right*) shows LP (*top*) and RLP (*bottom*) representations of the same image Fig. 1(*left*). In both cases the images are sparsely

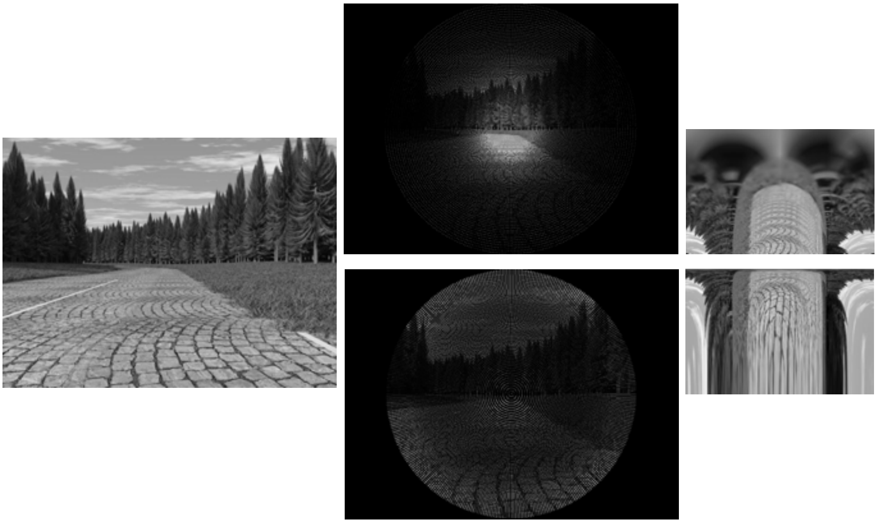


Fig. 1. (*top*) Log-Polar and (*bottom*) Reverse-Log-Polar representations of an image

sampled as depicted in Fig. 1(*middle*) correspondingly. Since the LP/RLP transformation involves both many-to-one and one-to-many mapping, the LP/RLP images cannot be straight forwardly dense. The dense images presented in the right column are obtained by querying for each (ρ, θ) to the cartesian and by bilinear interpolations—horizontal axis is angles $(\theta' s)$ and vertical axis is distances $(\rho' s)$. As can be seen in the grids in Fig. 1(*middle*), qualitatively, the RLPR image better preserves the periphery information, which covers most part of the road at the bottom in the scenario of a moving vehicle.

2.2 Variational Optical Flow

The aim in this paper is to evaluate the performance of LP and RLP representations once they are used to compute optical flow in the context of on-board vision systems. The variational optical flow [10] is based on two assumptions: *i*) the *brightness constancy* (BCA) and *ii*) the homogeneous regularization. The BCA, also called as *optical flow constraint*, assumes the grey value of objects remains constant over time. The homogeneous regularization assumes that the resulting flow field varies smoothly all over the image, necessary to overcome the aperture problem. The BCA can be formulated as: $I_1(\boldsymbol{x} + \boldsymbol{u}) - I_0(\boldsymbol{x}) = 0$, where I_0 and I_1 is the image pair, $\boldsymbol{x} = (x_1, x_2)$ is the pixel location within a rectangular image domain $\Omega \subseteq \mathbf{R}^2$; $\boldsymbol{u} = (u_1(\boldsymbol{x}), u_2(\boldsymbol{x}))$ is the two-dimensional displacement vector. Linearizing above equation using first-order Taylor expansion, and combining it with smoothness assumption in a single variational framework and squaring both constraints, the energy functional becomes:

$$E(\boldsymbol{u}) = \int_{\Omega} \{ \underbrace{(I_{x_1} u_1 + I_{x_2} u_2 + I_t)^2}_{Data\ Term} \quad + \alpha \underbrace{(|\nabla u_1|^2 + |\nabla u_2|^2)}_{Regularization} \} \, d\boldsymbol{x}, \qquad (3)$$

where α is a regularization parameter. Variational optical flow energy functions can be minimized in a number of ways. The most used way is to express and solve the set of Euler-Lagrange equations of the energy model. Another popular way of solving eq. (3) is by using a dual formulation based on iterative alternating steps [11]. In the current work a recent variational optical flow technique [12] is used. It explores the basic formulation and some concepts such as pre-processing, coarse-to-fine warping, graduated non-convexity, interpolation, derivatives, median filtering. [12] proposes an improved model underlying median filtering.

3 Experimental Results

As mentioned in Section 1, there have been many applications using LP represented images, some of them based on the optical flow estimation on those images. The current work aims to estimate the optical flow on RLP represented images and compare it with results from LPRs.

In LP/RLP representations of images the origin of mapping should be the vanishing point in the scenario of a forward facing moving vehicle, so that the mapped images better suit the applications. In the current work, vanishing points computed from a RANSAC based approach [13] are used. Then, the optical flow is computed on these LP and RLP represented images. The bottleneck to compare the flow fields from LP and RLP representations is that the flow field patches at a particular location in both representations correspond to different regions of the image in cartesian with varied resolution. Hence, the framework proposed to perform the comparison consists of inverse mapping the flow fields back to cartesian and compare them in the cartesian space. Figure 2 shows an image pair in cartesian (*top-left*), their ground-truth flow (*top-right*), LPR (*middle-left*) and RLPR (*bottom-left*), and their computed flow fields (*middle-right* and *bottom-right*). The color map used to display optical flow is shown in Fig. 2 bottom right corner. Since the image pairs correspond to a translation along the camera focal axis, the flow field in cartesian looks diverging. The computed flow field in both LP/RLP representations looks blue in color indicating all the vectors point downwards. Figure 3 depicts the inverse maps of both LP and RLP flow fields back to cartesian which are sparse. Hereinafter, the LP and RLP representations of flow fields refer to these mapped back to cartesian.

The well known error measures to compare flow fields are Average Angular Error (AAE) and Average End-Point error (AEP) [14] [15]. The AAE is chosen in the current work as the measure to compare flow fields. The angular error e between two vectors (u_1, v_1) and (u_2, v_2) is given by:

$$e((u_1, v_1), (u_2, v_2)) = \arccos\left(\frac{u_1 u_2 + v_1 v_2 + 1}{\sqrt{(u_1^2 + v_1^2 + 1)(u_2^2 + v_2^2 + 1)}}\right). \tag{4}$$

Since the flow fields from LP and RLP representations are sparse and of varied resolution, in order to do a fair comparison a common set of pixels (mask) is selected. Figure 4 shows the masks of LPR (*left*) and RLPR (*middle*) of flow

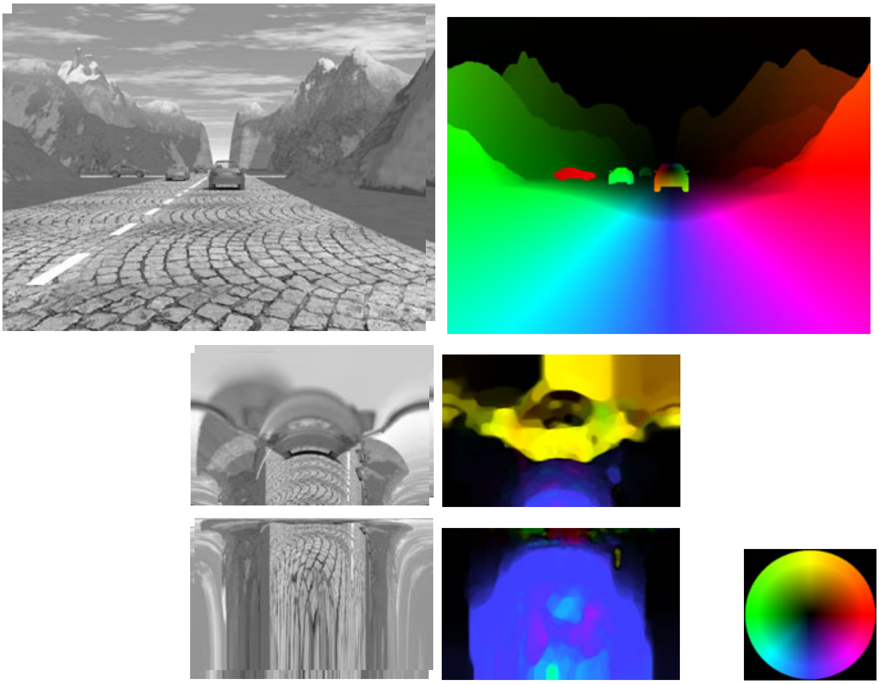


Fig. 2. (*top*) Flow fields in Cartesian, (*middle*) LP and (*bottom*) RLP representations

fields and the intersection mask (*right*) that is the set of positions those have flow values in both representations. This mask is used to compute the errors between LPR/RLPR and ground-truth flow fields. Table 1 shows AAE of ten different flow fields from sequence-1 of set-2 of [16]. The images in this dataset are of resolution 480×640. They are mapped to LP and RLP representations of resolution 230×360, placing the vanishing point at $(230, 340)$, computed from [13]. Then, optical flow is computed on these images using [12]. The flow fields are mapped back to cartesian and then, using the mask as shown in Fig. 4(*right*), the AAEs between LP and ground-truth flow fields, and between RLP and ground-truth flow fields, are computed. The AAEs in Table 1 show that flow fields estimated in RLP representations are more accurate than flow fields from LPRs. In all these experiments, the image region contained in the largest inner circle around the vanishing point is considered for mapping to LP/RLP.

Table 1. AAEs (deg.) for flow fields from sequences [16] in LPR and RLPR

		1	2	3	4	5	6	7	8	9	10
Seq. 1	LPR	24.38	24.35	23.99	23.95	23.92	23.80	23.63	23.78	23.53	23.42
	RLPR	20.98	19.15	19.04	19.43	18.30	18.47	18.47	18.00	17.92	18.94
Seq. 2	LPR	24.30	24.59	27.45	27.18	24.32	24.63	24.70	24.52	24.75	24.87
	RLPR	21.68	21.80	27.24	26.60	21.61	23.39	24.25	23.80	23.87	22.05

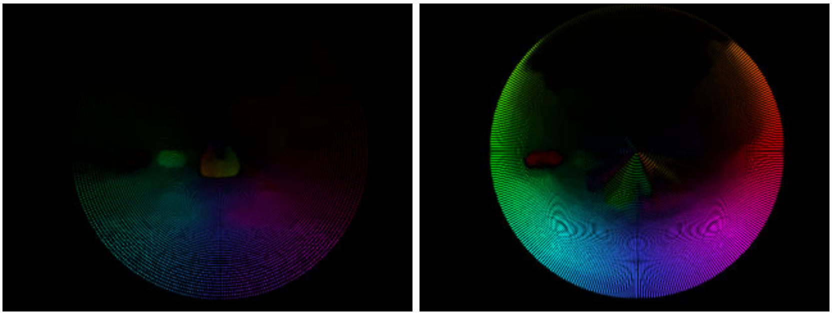


Fig. 3. Inverse mapped flow fields from (*left*) LP and (*right*) RLP

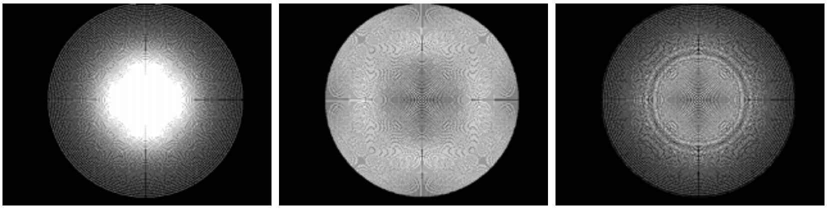


Fig. 4. (*left*) LP mask; (*middle*) RLP mask; (*right*) Mask from their intersection

A similar experiment on sequence-2 of set-2 of [16] is performed; results are presented in Table 1. Vanishing point for these 10 image pairs lies in $(240, 320)$, and the resolution of the mapped images is 240×360. In the results of sequence-2, the difference in AAEs between LP and RLP is smaller than the results of sequence-1 because the displacement between consecutive frames in sequence-2 is very high. These large displacements lead to more stretching in RLP represented images and hence more erroneous flow fields.

Further experiments are done to analyze how the error evolves along the space in these variant representations. Different circular regions around the vanishing point, with an increase in the radius of the circles within the flow field boundary, are considered. At each radius of the circle, the AAE is calculated inside the circle and outside the circle. This experiment is done on both LPR and RLPR. Since the radial axis for the flow fields of sequence-1 of set-2 is of length 230, nine circles with increasing radius from 23 till 207 in multiples of 23 are considered. Figure 5(*top-left*) and (*middle-left*) shows the AAEs in colormap for the region inside the circle at radius of 115 for LPR and RLPR. Figure 5(*top-right*) and (*middle-right*) show the AAEs in colormap for the region outside the circle at radius 115 for LPR and RLPR respectively. In Fig. 5(*bottom-left*), solid lines indicates AAEs (the average of ten flow fields' region inside the circle) with the increase in radius in LPR. The AAE increases as the inner area increases with the increase in radius. This proves that the flow field near the fovea is more accurate than in the periphery in LPR. The dashed lines correspond to AAEs (the average of ten flow fields' region inside the circle) in RLPRs with the increase in radius. In the plot Fig. 5(*bottom-left*) the AAE of RLPR decreases from radius 138 till the boundary. At radius 207, where most of the image area is covered inside the circle, the AAE of RLPR is less than the AAE of LPR. This shows RLPR is better at periphery than LPR.

Figure 5(*bottom-right*) shows the AAEs of LPR and RLPR, outside the circles, with the increase in radii of the circles. That means the outer area getting reduced with the increase in radius of the circle. The solid line indicating AAE of LPR increases as the outer area decreases, whereas the dashed line indicating AAE of RLPR decreases as the outer area decreases till the circle with radius 161. Then it increases due to some artifacts in the extreme periphery of RLPR flow field. Figure 5(*middle-right*) shows the artifact, thin band of circular arc on the top, whereas this band is absent in the LPR (*top-right*) flow field. This plot (*bottom-right*) gives the same conclusion obtained from the plot in (*bottom-left*).

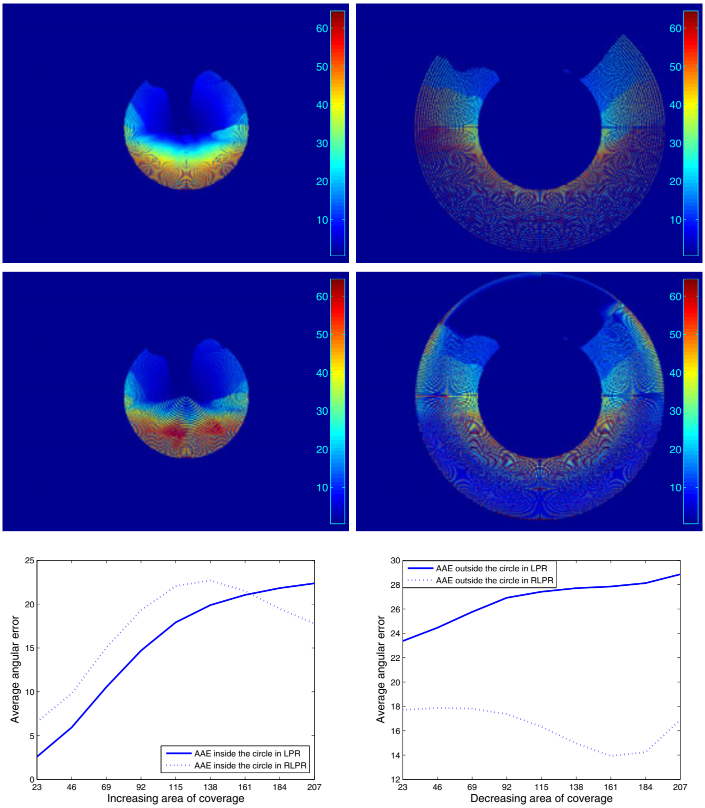


Fig. 5. Analysis of AAEs over space in LPR and RLPR (values in colormap scale computed from eq. 4). (*left*) Region inside circle. (*right*) Region outside circle.

4 Conclusion

The current paper shows that LPR, although inspired by biological vision systems, is not an appropriate representation for forward faced on-board vision systems, where translation in the optical axis is the predominant motion (e.g., mobile robotics, automotives). The previous statement is proved in a dense optical flow estimation framework, using as evaluation metric the average angular error. The optical flow is estimated on both, LP and RLP representations, and the results qualitative and quantitatively shows RLPR better preserves the peripheral information and hence more accurate flow field. The analysis of variance of errors along the space proves that the accuracy in flow field decreases along the distance from the fovea in LPR, whereas it increases along the distance from the fovea to periphery in RLPR. The possible future works are estimation of vanishing point along with the optical flow estimation in proposed representation, analysis of data reduction in RLPR to LPR and the cause of errors in space variant representations.

Acknowledgements. This work was supported by the Spanish govt. under Project TRA2010-21371-C03-01 and Research Program Consolider Ingenio 2010: MIPRCV (CSD2007-00018). The FI grant of AGAUR, Catalan govt. supports the first author.

References

1. Bolduc, M., Levine, M.D.: A review of biologically motivated space-variant data reduction models for robotic vision. Computer Vision and Image Understanding 69(2), 170–184 (1998)
2. Schwartz, E.L., Greve, D.N., Bonmassar, G.: Space-variant active vision: Definition, overview and examples. Neural Networks 8(7-8), 1297–1308 (1995)
3. Traver, V.J., Pla, F.: The log-polar image representation in pattern recognition tasks. In: Perales, F.J., Campilho, A.C., Pérez, N., Sanfeliu, A. (eds.) IbPRIA 2003. LNCS, vol. 2652, pp. 1032–1040. Springer, Heidelberg (2003)
4. Daniilidis, K.: Computation of 3-d-motion parameters using the log-polar transform. In: Hlaváč, V., Šára, R. (eds.) CAIP 1995. LNCS, vol. 970, pp. 82–89. Springer, Heidelberg (1995)
5. Tistarelli, M., Sandini, G.: On the advantages of polar and log-polar mapping for direct estimation of time-to-impact from optical flow. IEEE Trans. Pattern Anal. Mach. Intell. 15(4), 401–410 (1993)
6. Traver, V.J., Bernardino, A.: A review of log-polar imaging for visual perception in robotics. Robotics and Autonomous Systems 58(4), 378–398 (2010)
7. Traver, V.J., Pla, F.: Motion analysis with the radon transform on log-polar images. Journal of Mathematical Imaging and Vision 30(2), 147–165 (2008)
8. Yeasin, M.: Optical flow in log-mapped image plane-a new approach. IEEE Trans. Pattern Anal. Mach. Intell. 24(1), 125–131 (2002)
9. Daniilidis, K., Krüger, V.: Optical flow computation in the log-polar-plane. In: Hlaváč, V., Šára, R. (eds.) CAIP 1995. LNCS, vol. 970, pp. 65–72. Springer, Heidelberg (1995)

10. Horn, B.K.P., Schunk, B.G.: Determining optical flow. Artificial Intelligence 17, 185–203 (1981)
11. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L^1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 214–223. Springer, Heidelberg (2007)
12. Sun, D., Roth, S., Black, M.J.: Secrets of optical flow estimation and their principles. In: CVPR, pp. 2432–2439 (2010)
13. Onkarappa, N., Sappa, A.D.: On-board monocular vision system pose estimation through a dense optical flow. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6111, pp. 230–239. Springer, Heidelberg (2010)
14. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. In: ICCV, pp. 1–8 (2007)
15. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. International Journal of Computer Vision 12(1), 43–77 (1994)
16. Vaudrey, T., Rabe, C., Klette, R., Milburn, J.: Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In: Proc. Image and Vision Computing, Christchurch, New Zealand, pp. 1–6 (2008)

JBoost Optimization of Color Detectors for Autonomous Underwater Vehicle Navigation

Christopher Barngrover, Serge Belongie, and Ryan Kastner

University of California San Diego,
Department of Computer Science

Abstract. In the world of autonomous underwater vehicles (AUV) the prominent form of sensing is sonar due to cloudy water conditions and dispersion of light. Although underwater conditions are highly suitable for sonar, this does not mean that optical sensors should be completely ignored. There are situations where visibility is high, such as in calm waters, and where light dispersion is not significant, such as in shallow water or near the surface. In addition, even when visibility is low, once a certain proximity to an object exists, visibility can increase. The focus of this paper is this gap in capability for AUVs, with an emphasis on computer-aided detection through classifier optimization via machine learning. This paper describes the development of color-based classification algorithm and its application as a cost-sensitive alternative for navigation on the small Stingray AUV.

Keywords: Stingray, AUV, object detection, color, boosting.

1 Introduction

The goal of this paper is to use the Stingray platform to investigate object detection and classification as a basis for navigation. Reliable navigation on small AUVs is challenging in the absence of large and expensive sensors for estimating position. Using vision to detect and classify objects in the environment can be a source for estimating relative position. The target object can be used as a destination or could act as a path for the vehicle to follow [1]. The focus of this research is on developing robust object classifiers for specific targets based on color. The movement of the water and changes in lighting due to refraction and light dispersion cause colors to blur and change. In order to overcome these difficulties, we use a boosting algorithm to optimize the color classifier and improve the detector capability.

The target destination objects are three different colored buoys, anchored with relatively close proximity and varying depth. The buoy colors, chosen for their contrast with an underwater environment, are orange, yellow, and green in decreasing order of contrast. The green buoy should be more difficult to detect since it is most similar in color to the background. Once the algorithm can correctly detect and classify the target buoy, the vehicle demonstrates the

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 155–162, 2011.

navigation capability by approaching and touching the buoy. The path or bearing objects are orange pipes, which are anchored to the bottom. In some cases there are two pipes with different orientations in the same location. The vision algorithms detect and classify the pipe and then estimate the orientation. The vehicle demonstrates the vision-based navigation capability by centering over the pipe and altering its heading based on the estimated orientation. When there are multiple pipes, the vehicle must decide which direction to navigate. The two target types are shown in Figure 1 below.

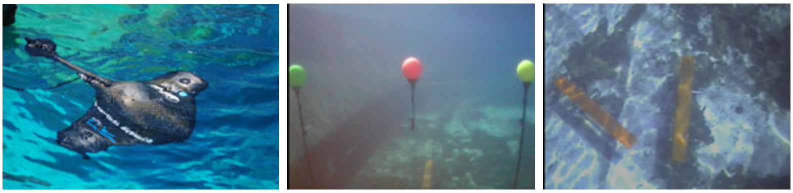


Fig. 1. (a) Stingray AUV. (b) Destination buoy objects. (c) Bearing pipe objects.

It turns out that the boosting of the classifiers for the buoys and pipes greatly improves the detectors. For the pipe, we show that the bearing estimation becomes extremely accurate as well. We implement the optimized detectors and bearing estimator on the Stingray, which is able to navigate to the correct buoy and change bearing based on the pipe with high reliability.

The remainder of this paper is organized as follows. In Section 2 we discuss related work, while in Section 3 we describe our process for developing a classification algorithm. In Sections 4 and 5 we focus on the specific targets of the buoy and pipe, providing results from the final algorithms for each. Finally, in Section 6 we conclude by discussing the aspects of this research that are novel and the promising directions for future work.

2 Related Work

There has been an increase of research in vision-based navigation for underwater vehicles in recent years. Most of the research focuses on avenues that do not parallel the work in this paper, but there are some similar efforts.

The papers that use landmarks as reference points for underwater navigation are most similar. The work of Yu *et al.* [7] uses yellow markers and colored cables for AUV navigation by thresholding the UV components of the YUV color space, which is similar to the baseline methods for this paper. Another method thresholds on the RG components of the RGB color space to detect yellow sensor nodes, as presented by Dunbabin *et al.* [3]. In the research by Soriano *et al.* [6] an average histogram is created for each target, which is compared to a region of interest for classification.

Cable or pipe tracking is another task, which is heavily researched in terms of vision-based systems. The work of Balasuriya *et al.* [1] shows a method of using Laplacian of Gaussian (LoG) filters to detect the edges of the pipe. Foresti *et al.* [4] use a trained neural network to recognize the pipeline borders, while Zingaretti and Zanoli [8] use vertical edge detection in horizontal strips and contour density within the strips to detect the pipe.

These papers avoid much of the underwater difficulties, which cause colors to change based on light absorption, by attaining proximity to the target. We show that without boosting, a simple color classifier is not sufficient on our test data set, which includes images of the targets at substantial distances and under varying lighting conditions.

3 Developing a Classification Algorithm

The process of developing the classification algorithm generally starts with choosing a feature set to describe the target. The feature chosen for these targets is color. The Hue-Saturation-Value (HSV) color model is used for its separation of brightness from the hue and saturation pair. Because of this isolation of the brightness element of a color, a single object is more reliably detectable under different lighting conditions. The more common Red-Green-Blue (RGB) color model is an additive model, which makes it difficult to identify the same color under different lighting conditions [2].

The boosting algorithm requires a large number of examples in order to optimize the decision tree. For the HSV color classifier, we labeled individual pixels as positive or negative in terms of the target. The examples, which number in the hundreds of thousands, are then inputs into the boosting algorithm.

For this research, the LogitBoost form of boosting is used via the JBoost software package. The JBoost application expects the input examples in a standard format with classifier data and a label. JBoost can output the resulting decision tree visually as well as in Java or C code.

4 Buoy Detection

The buoy targets have the same size and shape, only differing by color. To develop the algorithm, we focus first on the orange buoy. Once an algorithm is developed, including the pixel level optimized decision tree and post processing, we can train the classifier for the other buoys. The final algorithm will have a pixel decision tree for each color to create a binary image. The binary image will be post processed in the same way for each color. The goal is to accurately estimate the location of the designated buoy in the image and use the distance from the buoy to the center of the image as a heading offset for the Stingray vehicle.

4.1 Baseline

There must be a baseline algorithm in order to determine the improvements provided by using boosting to optimize the decision tree for the HSV classifier.

The baseline in this research is a simple HSV thresholding, which was previously implemented on the Stingray. An HSV estimation of the color orange in the buoy is extended to provide a range for each of hue, saturation and value, which was tuned over many iterations to achieve the best possible threshold range. The range is used to determine if a pixel is positive or negative, thus creating a binary image, which is used without post processing to estimate the center of the buoy based on the centroid of the positive pixels.

The metrics used to compare algorithms are the true positive rate (TPR) and false positive rate (FPR). There are two sets of images from two different environments. The first environment is a large anechoic pool, which is 300 ft by 200 ft by 38 ft deep. The other is a small above ground pool, which is 10 ft in diameter and 4 ft deep. Both pools are situated outside in natural lighting. For each environment there is a set of images for training the classifier and a set of images for testing the resulting classifier. Both image sets have examples of the buoy from different distances as well as images with no buoy present. To determine TPR and FPR, we label the center of the buoy in each test image, as well as the edge of the buoy. The distance between these points provides a threshold for the correctness of a center estimation. The baseline TPR is 0.45 and 0.18 for the Tank and the Pool respectively, while the FPR is 0.55 and 0.45.

4.2 Post Processing

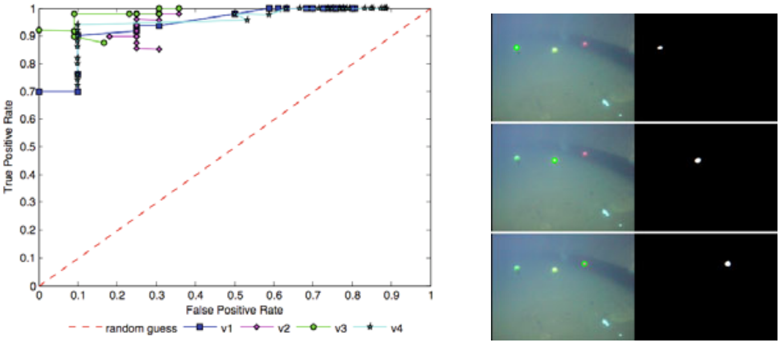
Since the boosted classification algorithm is for individual pixels, the output is a binary image without clearly defined object boundaries and with extraneous positive or negative pixel noise. The goal of the post processing techniques used in this research is to prepare the binary image for the best possible estimation of the location of the buoy.

We start by using one iteration of opening, which is erosion followed by dilation, to remove noise in the binary image. Next we use two iterations of closing, which is two dilations followed by two erosions, to fill binary objects containing gaps. Then the smoothing algorithm via Median blur with a 7x7 kernel creates smooth edges of binary objects in the image. Finally, we use the convex hull algorithm to approximate the shape of the binary object with only convex corners, which provides more complete binary objects in situations where part of the target is not correctly classified.

4.3 Boosting HSV

As described in Section 3, the first step to boosting the HSV classifier is labeling examples. The pixel examples are given as input to JBoost, which outputs a complex decision tree in a C code function. The function provides a score for a given pixel, which is labeled as a one or zero based on a threshold.

In order to determine the threshold that provides the best output, we look at the receiver operating characteristic (ROC) curve for thresholds from -2.0 to 5.0 over 0.1 increments. Since the threshold determines the status of a pixel and the performance of the classifier is determined by the accuracy of the center



(a) (b)

Fig. 2. (a) The ROC curves for four versions of the buoy classifier on the test image set from the tank environment. (b) Example of classifying specifically for different color buoys independently. The green circles show the estimated centers for each buoy.

estimation, the generated ROC curve is not a smooth curve. The tank is large and representative of an ocean environment in terms of acoustics and reflectivity, while the pool is small with reflective walls and bottom. The two environments are distinct enough that when we label extra examples for the pool, we ultimately overfit causing reduced performance for tank images. The simple solution is to develop target classifiers for the environments independently.

We start with the tank environment by generating a decision tree, which we use on our test image set to produce the ROC curve and choose the best threshold value. Based on the results at this threshold, additional labeling may improve the classifier. Figure 2 shows the ROC curves from four such iterations of the decision tree. The best results are at the threshold of 3.6, which gives a TPR of 0.98 and a FPR of 0.18, and the threshold of 4.2, which gives a TPR of 0.92 and a FPR of 0.0.

We follow the same iterative sequence for the pool environment, which is much more challenging because of its small size and shallow depth. The two best thresholds are 0.7, which gives a TPR of 0.68 and a FPR of 0.26, and the threshold 1.7, which gives a TPR of 0.61 and FPR of 0.05. These results are not as reliable as the tank results, but they are still a substantial improvement over the baseline.

4.4 Results

The same technique described in Section 4.3 can be applied to the other two buoy colors to create decision trees for classifying the pixels. The post processing techniques are the same for each color buoy. This means that the algorithm will switch between the decision trees based on the target buoy. Figure 2 shows the processing of the same image while looking for each of the different color buoys.

When combining the results of the three buoy classification algorithms on the test image set, we can calculate the total TPR and FPR for the overall algorithm as 0.84 and 0.16 respectively. The relatively low quality of the classifier for the green buoy reduces the overall result.

In practice the Stingray is able to reliably detect the designated target buoy at approximately six frames per second and the detection becomes more reliable as the Stingray approaches the buoy.

5 Pipe Detection

The pipe is an interesting target because it provides a bearing for navigation. There can be two pipes leading to different destinations, as shown in Figure 1, which means the algorithm needs to be able to classify multiple pipes in a single image. After determining that a binary object is a pipe, the algorithm must calculate the orientation. The goal is to use the orientation of the pipe as a target heading for the Stingray vehicle.

5.1 Baseline

The baseline for the pipe, similar to the buoy, is a simple HSV threshold used to create a binary image on which a custom algorithm, using least squares estimation, attempts to determine the orientation. This orientation estimation technique is not dependable and is only used in the baseline algorithm.

The same metrics are used for the pipe results as are used for the buoys. The main difference is that there are no examples from a secondary environment. This makes the classification problem slightly easier, so that the problem of estimating orientation can take focus. The baseline for the Tank is a TPR of 0.74 and a FPR of 0.16.

5.2 Classification

The pipe, like the buoy, has a unique color which makes for a useful classifier. The same process of labeling images and inputting the examples into JBoost to optimize a decision tree ultimately outputs a function for scoring individual pixels of the image. The same post processing techniques from Section 4.2 are applied to the pipe binary images to create smooth and closed binary objects.

The version of the decision tree that produces the best ROC results has two thresholds with a trade off between TPR and FPR. Both of these thresholds provide very reliable rates, -0.3 give a TPR of 0.97 and a FPR of 0.02, while the threshold 0.7 gives a TPR of 0.95 and a FPR of 0.01.

5.3 Bearing Estimation

The overall goal of the pipe detection is to determine the orientation of the pipe to be used as a bearing for navigation purposes. Therefore, with a binary object found, only the edges of the object are actually pertinent. The Canny edge detector, with threshold values of 50 and 150 pixels, is applied to the binary image and the output contains only the edges of all binary objects.

With only edges remaining, the Hough Transform can be used to easily estimate the straight lines in the image. We use the Probabilistic Hough Transform (PHT) due to its ability to combine similar lines with a gap between them [5]. We use a ρ of one pixel and a θ of $\frac{\pi}{120}$ or 1.5 degrees. Our threshold is set at 30 pixels, with an acceptable line segment length of 20 pixels and an acceptable gap of 20 pixels.

Often times the output from the PHT has extraneous line segments. The goal of the pruning portion of the algorithm is to reduce all the line segments from the Hough Transform down to the two per pipe that represent the long edges of the pipe. This is broken into two steps, starting with merging all line segments that are close to collinear. The next step is using the property of parallelism to remove extraneous line segments. Figure 3 shows three scenarios where different tests of parallelism remove extraneous line segments.

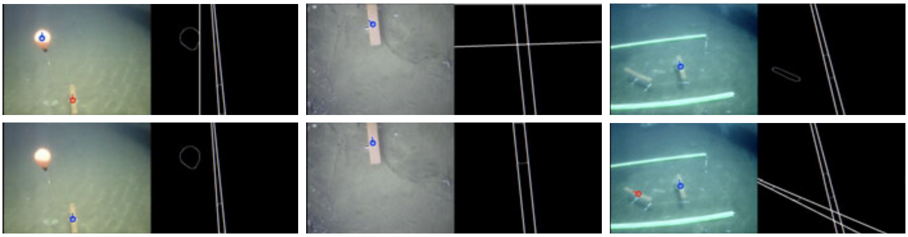


Fig. 3. Examples of the three algorithms of the pruning stage. The blue and red circles with lines show the estimated centers and orientations of the pipes.

5.4 Results

The important result of the pipe detection is the ability to estimate the orientation of the pipe with great precision, in order to provide the vehicle with useful bearing. Of course, detecting the location of the pipe is necessary to allow for the bearing estimation, which we have shown to be very reliable.

In order to quantify the accuracy of the bearing estimation, the edges of the pipes are labeled in the test image set and then compared to the algorithm's estimate. The average error with standard deviation for the baseline algorithm is $9.0° \pm 14.6°$ compared to $0.7° \pm 0.8°$ for the hough transform based algorithm.

In practice the Stingray is able to process the images at five frames per second, allowing the vehicle to center itself over the pipe and estimate the orientation. The vehicle then rotates to match its heading with the orientation of the pipe, and navigates in that direction.

6 Conclusion

This paper presents a method for using object detection and classification of target objects to aid in navigation for AUVs. The color classifier is one unique element of this research, as it is not common in underwater applications. Also, the use of boosting algorithms to optimize the classifier greatly improves on previous work. We incorporated the use of post processing techniques to make identifying the center of the target objects more reliable. We also showed a technique for calculating the orientation of up to two pipes simultaneously, and with high precision.

The result is two classification algorithms that are more efficient than the baseline algorithms of simple thresholding. We demonstrated these algorithms on the Stingray AUV, which navigates towards and touches a specific color of buoy and changes heading based on the pipe.

The process we presented for creating an optimized classifier via boosting can be applied to other targets and with classifiers other than color. The complex and dynamic properties of underwater environments cause these classifiers to be very specialized, which naturally leads this research towards efforts in adaptive learning to improve a classifier in real time for changing environments.

References

1. Balasuriya, B.A.A.P., Takai, M., Lam, W.C., Ura, T., Kuroda, Y.: Vision based autonomous underwater vehicle navigation: underwater cable tracking. In: OCEANS Proceedings, pp. 1418–1424 (1997)
2. Cheng, H. D., Jiang, X. H., Sun, Y., Wang, J. L.: Color image segmentation: Advances and prospects. Pattern Recognition 34, 2259–2281 (2001)
3. Dunbabin, M., Corke, P., Vasilescu, I., Rus, D.: Data muling over underwater wireless sensor networks using an autonomous underwater vehicle. In: IEEE Int. Conf. on Robotics and Automation, pp. 2091–2098 (2006)
4. Foresti, G.L., Gentili, S., Zampato, M.: A vision-based system for autonomous underwater vehicle navigation. In: OCEANS Proceedings, pp. 195–199 (1998)
5. Kiryati, N., Eldar, Y., Bruckstein, A. M.: A probabilistic Hough transform. Pattern Recognition 24, 303–316 (1991)
6. Soriano, M. and Marcos, S. and Saloma, C. and Quibilan, M. and Alino, P.: Image classification of coral reef components from underwater color video. In: OCEANS Proceedings, pp. 1008–1013 (2001)
7. Yu, S.C., Ura, T., Fujii, T., Kondo, H.: Navigation of autonomous underwater vehicles based on artificial underwater landmarks. In: OCEANS Proceedings, pp. 409–416 (2001)
8. Zingaretti, P., Zanoli, S.M.: Robust real-time detection of an underwater pipeline. Engineering Applications of Artificial Intelligence 11, 257–268 (1998)

Combining Structure and Appearance for Anomaly Detection in Wire Ropes

Esther-Sabrina Wacker and Joachim Denzler

Chair for Computer Vision, Friedrich Schiller University of Jena
Ernst-Abbe-Platz 2, 07743 Jena, Germany
{esther.wacker,joachim.denzler}@uni-jena.de

Abstract. We present a new approach for anomaly detection in the context of visual surface inspection. In contrast to existing, purely appearance-based approaches, we explicitly integrate information about the object geometry. The method is tested using the example of wire rope inspection as this is a very challenging problem.

A perfectly regular 3d model of the rope is aligned with a sequence of 2d rope images to establish a direct connection between object geometry and observed rope appearance. The surface appearance can be physically explained by the rendering equation. Without a need for knowledge about the illumination setting or the reflectance properties of the material we are able to sample the rendering equation. This results in a probabilistic appearance model. The density serves as description for normal surface variations and allows a robust localization of rope surface defects.

We evaluate our approach on real-world data from real ropeways. The accuracy of our approach is comparable to that of a human expert and outperforms all other existing approaches. It has an accuracy of 95% and a low false-alarm-rate of 1.5%, whereupon no single defect is missed.

Keywords: anomaly detection, image-based analysis, surface inspection.

1 Introduction

Automatic surface inspection is a research area of rising interest. It is an important problem as the inspection task is an exhausting and monotonous work for a human with high quality claims on the other hand. In addition, surface analysis in general is a difficult problem, as the visual appearance of surfaces is highly subjected to various kinds of noise and changing lighting conditions.

A good example for such a task is the visual inspection of wire ropes. This is a very important problem, as damaged ropes pose a risk for the human life. Furthermore, the long, heavy ropes cannot be unmounted, are often contaminated with *e.g* mud or oil and their material is highly reflective. In consequence, the surface appearance of an intact rope exhibits various characteristics. In contrast, defects in the surface structure are often very small and inconspicuous. Some examples for typical surface defects are displayed in the upper images of Fig. 5. Due to the high intra-class variability and the poor inter-class separability, a discrimination between defect and normal appearance variation is a difficult problem.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 163–170, 2011.

Furthermore, a common problem of visual inspection tasks is the limited amount of available defective samples which hinders a supervised learning. For this reason, anomaly detection techniques [1,4], also known as one-class classification [9] have been used in the past for defect detection in material surfaces [8,11]. In general, these approaches are highly dependent on their choice of features used to represent the intact class. Platzer *et al* [7] compared the performance of different textural features for the problem of defect detection in wire rope surfaces. Their results underline the importance of context information for the problem of surface defect detection, especially with respect to the complex structure of wire ropes. In [6] Platzer *et al* focused on contextual anomaly detection by modeling the intact class with help of Hidden Markov Models. Haase *et al* [2] diagnosed contextual anomalies in the rope surface with help of an autoregressive model which predicts the intact surface appearance given its neighborhood. Nevertheless, no approach achieves the accuracy of a human inspector.

We state that the main reason for this is the lack of *geometrical* context in these purely appearance-based approaches. Therefore, we present a model-based approach for visual surface inspection. By fusing a geometrical structure model with a statistical appearance model we achieve a much better discrimination between a real defect and normal appearance variations. In a first step the model geometry is estimated in an image-based manner with help of a perfectly regular 3d rope model introduced recently by Wacker and Denzler [10]. In contrast to our work, they used this model to monitor important rope parameters but they did not address the problem of rope surface defect detection. We introduce a statistical appearance model which is linked to the geometric constraints implied by the rope structure. This allows a description of the surface appearance dependent on the position in the rope. Our method is data-driven and purely image-based. Moreover, we have no need for calibration information with respect to camera positions or the illumination setting.

The remainder of this paper is structured as follows: in section 2 the 3d model and the geometry estimation are summarized. Section 3 explains how this structural model can be linked to an statistical appearance model based on the *rendering equation*, which gives a physical explanation for light transport. Finally, section 4 turns to the problem of anomaly detection for defect analysis. A special focus will be laid on a validation strategy, which normalizes the learned appearance model with respect to small inaccuracies, which result from the geometry estimation step. Our experimental evaluation on real-world rope data is provided in section 5. Finally, conclusions are given in section 6.

2 Geometric Rope Model

To estimate the rope geometry from 2d rope images, we use the framework described recently by Wacker and Denzler [10]. Their approach focuses on the image-based monitoring of important rope parameters and is not suitable for the automatic detection of surface defects.

A rope has a hierarchical structure composed of J strands $\mathbf{S}_j$ which comprise I wires $\mathbf{W}_i$. A wire centerline $\mathbf{W}_{i,j}$ of wire i in strand j for the time step t can

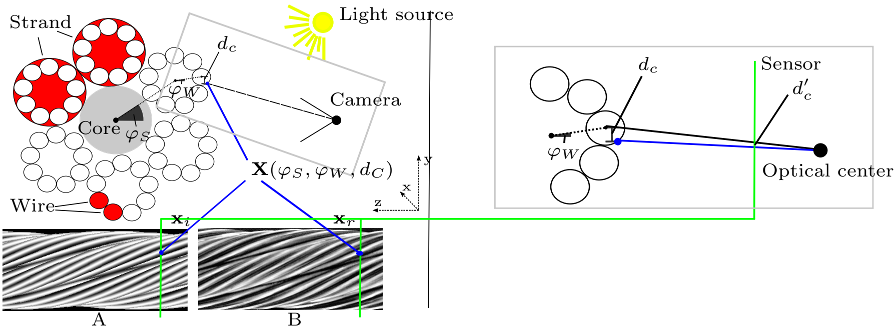


Fig. 1. Scenario sketch: given the point correspondence of a rope pixel $\mathbf{x}_r$ in the real rope image (B) and a rope pixel $\mathbf{x}_i$ in the aligned artificial model projection (A) a 3d surface point $\mathbf{X}$ of the rope can be parametrized by the two phase angles φ_S, φ_W and the 2d distance d'_c of $\mathbf{x}_i$ to its corresponding projected wire centerline. d'_c results from a 1:1 mapping of the unknown 3d distance d_c.

be described by a sum of two parametrized helices:

$$\mathbf{W}_{i,j}(\mathbf{p}, t) = \underbrace{\begin{pmatrix} t \\ r_S \sin(\varphi_S(\mathbf{p}, t)) \\ -r_S \cos(\varphi_S(\mathbf{p}, t)) \end{pmatrix}}_{\mathbf{S}_j} + \underbrace{\begin{pmatrix} 0 \\ r_W \sin(\varphi_W(\mathbf{p}, t)) \\ -r_W \cos(\varphi_W(\mathbf{p}, t)) \end{pmatrix}}_{\mathbf{W}_i} . \tag{1}$$

$\mathbf{p}$ is a vector of free model parameters and $\varphi_S(\mathbf{p}, t)$, $\varphi_W(\mathbf{p}, t)$ are the phase angles of the helices which are dependent on the model parametrization. The cross section through this model for one time step is shown in the top of Fig. 1.

By means of analysis-by-synthesis this parametric model is aligned with the digitally acquired 2d rope images. For that purpose an artificial 2d projection of the 3d rope model is computed. Real rope images and the artificial projections are then registered by optimizing the free model parameters in a non-linear fashion and these steps are repeated until convergence. We obtain a correspondence between a pixel $\mathbf{x}_i$ in the artificial projection and a pixel $\mathbf{x}_r$ in the real image.

In contrast to [10] we use this correspondence to form a parametric description of each surface point $\mathbf{X}$ in the rope. Fig. 1 clarifies that every 3d surface point can be described by the two phase angles φ_S and φ_W of the corresponding wire centerline and the 3d distance d_c to this surface point (time is neglected):

$$\mathbf{X}(\varphi_S(\mathbf{p}), \varphi_W(\mathbf{p}), d_c) = \mathbf{W}_{i,j}(\mathbf{p}) + \underbrace{\begin{pmatrix} 0 \\ d_c \\ -\sqrt{0.5\varnothing_W^2 - d_c^2} \end{pmatrix}}_{\mathbf{n}'} \tag{2}$$

Here $\varnothing_W$ is the known diameter of the wires and $\mathbf{n}'$ points into the direction of the surface normal. As the rope model reveals no volumetric information d_c is unknown, but there exists a 1:1 mapping to the measurable 2d distance d'_c of an image pixel $\mathbf{x}_i$ to its corresponding projected wire centerline. Therefore, we will use the parametric description $\theta = (\varphi_S(\mathbf{p}), \varphi_W(\mathbf{p}), d'_c)$ to characterize a surface point in the rope and to build a combined model for structure and appearance.

3 Combined Model for Structure and Appearance

The rendering equation is a physical model describing the observed radiance at a surface point of an geometric object. It was first introduced by Kajiya [3] in 1986 and is an integral equation describing the propagation of light. One of the most common formulations of the rendering equation is:

$$L_O(\mathbf{X}, \omega_o) = L_E(\mathbf{X}, \omega_o) + \int_{\Omega} f_r(\mathbf{X}, \omega_i, \omega_o)\ L_I(\mathbf{X}, \omega_i)\ (\omega_i \cdot \mathbf{n}) d\omega_i. \qquad (3)$$

The radiance which can be observed at a surface point $\mathbf{X}$ depends on the viewing direction ω_o, the emitted amount of light L_E and the reflected radiance which results from the incoming radiance L_I, the bidirectional reflectance distribution function f_r of the surface point and the inner product of surface normal $\mathbf{n}$ and the inward direction ω_i integrated over the hemisphere Ω.

Usually, in visual inspection scenarios we have neither calibration information nor knowledge about the illumination setting so that ω_o and ω_i are unknown. However, the relation between camera, object and position of the light source(s) typically stays fixed. This implies that the viewing direction and the incident angle of the incoming light depend only on the parametrization θ of the surface point $\mathbf{X}$, which we derived in section 2. Fig. 1 clarifies this scenario. In this case, the rendering equation can be re-parametrized and the emitting term L_E can be neglected for non-emitting objects like the rope:

$$\tilde{L}_O(\theta) = L_O(\mathbf{X}(\theta)) = \int_{\Omega} f_r(\mathbf{X}(\theta), \omega_i)\ L_I(\mathbf{X}(\theta), \omega_i)(\omega_i \cdot \mathbf{n}(\mathbf{X}(\theta)))d\omega_i, \qquad (4)$$

Now, we are able to sample the observed irradiance L_O at a surface point $\mathbf{X}$ of the rope only dependent on its parametrization θ without additional knowledge about the camera position or the illumination setting. As our goal is the estimation of a *representative* surface appearance model including normal appearance variations, we exploit the periodic structure of a rope to obtain several samples for the same surface point. We consider a whole sequence of rope images which are aligned with the rope model for this purpose.

The appearance model is learned from an images of an intact rope. We are interested in the likelihood of observing a gray value g_r at the position $\mathbf{x}_r$ in the real rope image given its corresponding 3d surface point $\mathbf{X}(\theta)$. This can be formulated as a density estimation problem. We estimate the joint distribution $p(g_r, \theta)$ for any parametrization θ and its corresponding observed gray values g_r in a non-parametric manner. To obtain a dense representation we apply a 4d Parzen estimator. This density constitutes a combined model for appearance and structure, which allows to describe the normal surface appearance of each surface point subjected to the underlying rope geometry.

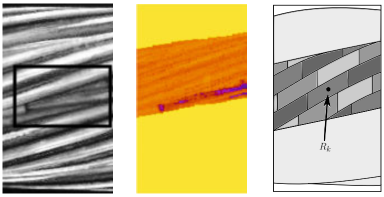


Fig. 2. Original rope image with defect (left), corresponding probability map (middle) for the strand with the defect and sketch of the rope regions (right)

4 Defect Analysis

Once having learned the rope surface appearance model, the defect diagnosis can be treated as anomaly detection problem. Again, the input rope images must be aligned with the rope model to obtain the parametrization θ of each surface point. Subsequently, the appearance representation is extracted from the density $p(g_r, \theta)$ as a function of the position in the rope. A probability map can be computed which contains the likelihood of observing gray value g_r for a pixel $\mathbf{x}_r$ in the real rope image given its corresponding parametrization θ

$$p(g_r \mid \theta) = \frac{p(g_r, \theta)}{p(\theta)}. \tag{5}$$

Fig. 2 shows a real rope image including a typical defect on the left and its corresponding probability map for the strand of interest in the middle. The darker the color in the probability map, the smaller the obtained likelihood.

Nevertheless, an alignment of a rigid rope model with the flexible structure of a real rope leads to systematic registration inaccuracies which arise mainly in the border areas between two strands. In these regions a robust estimation of the appearance model is hindered. Hence, we normalize the appearance model with respect to these stability variations.

Different regions in the rope can be encoded with help of the two phase angles φ_S, φ_W of the 3d model. This allows a separation into K discrete region classes R_k as sketched in the right hand side of Fig. 2. In order to increase the robustness of the appearance model with respect to systematic registration inaccuracies, we normalize the expectation of all rope regions. Hence, we compute the average likelihood $\overline{p}(R_k)$ for each rope region R_k and all N_k rope pixels belonging to R_k:

$$\overline{p}(R_k) = \frac{1}{N_k} \sum_{n=1}^{N_k} p(g_r^n \mid \theta^n). \tag{6}$$

This average is used to obtain a normalized likelihood according to (5):

$$\tilde{p}(g_r \mid \theta) = p(g_r \mid \theta) \, \frac{1}{\epsilon + \overline{p}(R_k)}. \tag{7}$$

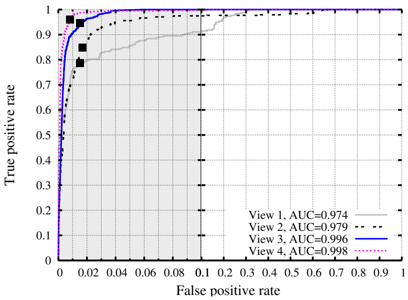


Fig. 3. Model-based approach: ROC curves and the 50% recovery of each defect marked by black squares

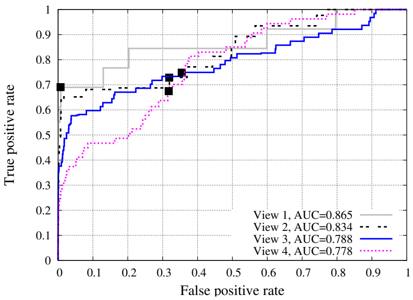


Fig. 4. Comparison to HMM-approach: ROC curves and 50% recovery of each defect marked by black squares

$\epsilon > 0$ is a stabilization factor. The validation compensates for a systematic problem caused by the alignment of a rigid model with flexible real-world data. Thus, the normalization is data-independent and can be performed on the training set.

Finally, the resulting probability map for the input rope image including the normalized likelihoods $\tilde{p}(g_r \mid \theta)$ is filtered along the wire course. To transfer this soft classification result into a hard discrimination between suspicious changes and normal variations in the rope surface, a thresholding operation can be used.

5 Experiments

We evaluate our approach on real-world data taken from real ropeways under realistic acquisition conditions. Our data set comprises 400 meters of rope in total which corresponds to 7.7 GB of data. It was carefully selected by a human expert to ensure, that a maximum amount of appearance variations and surface defects are contained. The used system [5] operates with four line cameras, which are equally placed around the rope. A concatenation of the four individual 1d measurements results in four different 2d image sequences which are referenced as `view 1 - 4` from now on. Thus the amount of rope meters is quadrupled and the set of natural variations which occur during the acquisition process is augmented. The reference labeling is also provided by a human expert. The appearance model is trained on 5 m of rope which are known to be defect free. The remaining 395 m were used for testing.

5.1 Overall Performance

In order to evaluate the overall performance of our approach, we compute Receiver Operating Characteristic (ROC) curves for each sequence. The results can be seen in Fig. 3. The Area Under the Curve (AUC) value for each curve is given in the legend. The True Positive Rate (TPR) represents the total area of recovered defects and the False Positive Rate (FPR) relates to the false alarm rate

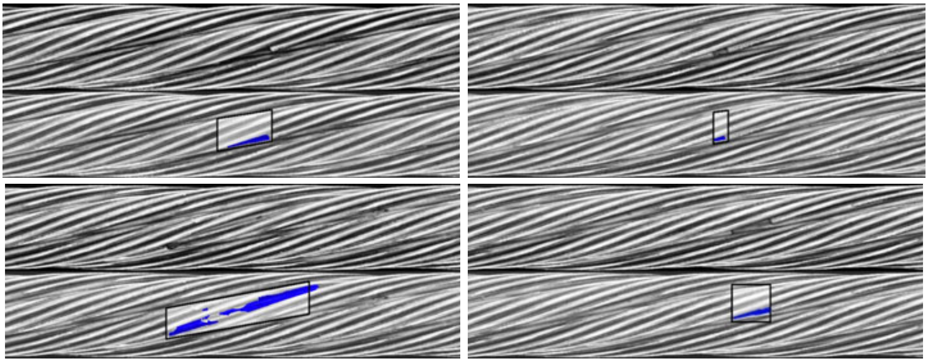


Fig. 5. Recovered defects: original rope image (upper image in each group) and result with recovered defect (blue) and ground truth labeling (black box)

(both measured in camera lines). As it is not sufficient to measure the error just as a function of the total length of detected anomalies we furthermore introduce the 50% recovery case. The black squares on each curve mark the recognition rates, which can be achieved if *every* known defect is recognized to at least 50% of its extent. Note, that these rates are bounded to the most inconspicuous defects in the sequence and the overall recognition rate is significantly higher than 50% in all cases. Keep in mind, that for the application it is not important to recover 100% of the defect area. But, it is crucial to recover *every single* defect to at least a certain extent while minimizing the FPR. In Fig. 5 some of our detection results are displayed. These results underline the high accuracy of the presented approach. As in most security relevant applications, the final decision must be made by a human expert who needs an image context of around 5 cm around each system alarm to judge weather it is a critical anomaly or a false alarm. With a false alarm rate of 1.5% for the 50% defect recovery case, a human expert would have to re-inspect only 103 m of the rope instead of 395 m.

5.2 Comparison to other Rope Defect Detection Approaches

We compare our results to the one obtained with the Hidden-Markov model (HMM) approach of Platzer *et al* [6] which leads to the best published results so far with regard to an individual analysis of each camera view.

Fig. 4 shows the ROC curves obtained on the same dataset with the HMM approach. Again the AUC values for each curve are given and the black squares mark the recognition rates obtained for the 50% recovery case of each defect.

It is obvious that our approach outperforms the HMM-based strategy. Particularly, in case of views 2–4 the HMM approach fails with an unfeasible high false alarm rate if the request is a detection of *every* single defect to at least 50%. But for a security-relevant task this claim is essential and this is not guaranteed by the existing approaches.

6 Summary and Conclusions

We presented a new approach for anomaly detection in wire ropes. The combination of a statistical appearance model with a parametric description of the object geometry leads to a position-dependent appearance representation. This combination allows a clearly enhanced discrimination between normal appearance variations and suspicious anomalies. One open question is the automatic determination of an optimal threshold. At the moment, the optimal threshold is evaluated with ROC curves, which always require a labeled data set.

Our results obtained on real-world rope data are very accurate and comparable to those of a human expert. We achieve low false alarm rates of 1.5% while fulfilling the claim that *every* single defect is recovered to a certain extent. This outperforms all existing approaches for automatic rope inspection and marks a clear improvement with respect to the practical applicability. Furthermore, our approach allows a precise localization of the defects.

References

1. Chandola, V., Banerjee, A., Kumar, V.: Anomaly Detection: A Survey. ACM Computing Surveys 41(3), 1–58 (2009)
2. Haase, D., Wacker, E.-S., Schukat-Talamazzini, E.-G., Denzler, J.: Analysis of Structural Dependencies for the Automatic Visual Inspection of Wire Ropes. In: VMV 2010: Vision, Modeling & Visualization, pp. 49–56 (2010)
3. Kajiya, J.T.: The rendering equation. ACM SIGGRAPH Computer Graphics 20(4), 143–150 (1986)
4. Markou, M., Singh, S.: Novelty detection: a review - part 1: statistical approaches. Signal Processing 83(12), 2481–2497 (2003)
5. Moll, D.: Innovative procedure for visual rope inspection. Lift Report 29(3), 10–14 (2003)
6. Platzer, E.-S., Nägele, J., Wehking, K.-H., Denzler, J.: HMM-Based Defect Localization in Wire Ropes - A New Approach to Unusual Subsequence Recognition. In: Denzler, J., Notni, G., Süße, H. (eds.) Pattern Recognition. LNCS, vol. 5748, pp. 442–451. Springer, Heidelberg (2009)
7. Platzer, E.-S., Süße, H., Nägele, J., Wehking, K.-H., Denzler, J.: On the Suitability of Different Features for Anomaly Detection in Wire Ropes. In: Ranchordas, A., Pereira, J.M., Araújo, H.J., Tavares, J.M.R.S. (eds.) VISIGRAPP 2009. CCIS, vol. 68, pp. 296–308. Springer, Heidelberg (2010)
8. Tajeripour, F., Kabir, E., Sheikhi, A.: Fabric Defect Detection Using Modified Local Binary Patterns. EURASIP Journal on Advances in Signal Processing 8(1), 12 (2008)
9. Tax, D.M.J.: One-class classification - Concept-learning in the absence of counter-examples. Phd thesis, Technische Universität Delft (2001)
10. Wacker, E.-S., Denzler, J.: An Analysis-by-Synthesis Approach to Rope Condition Monitoring. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Chung, R., Hammound, R., Hussain, M., Kar-Han, T., Crawfis, R., Thalmann, D., Kao, D., Avila, L. (eds.) ISVC 2010. LNCS, vol. 6454, pp. 459–468. Springer, Heidelberg (2010)
11. 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)

3D Cascade of Classifiers for Open and Closed Eye Detection in Driver Distraction Monitoring

Mahdi Rezaei and Reinhard Klette

The *.enpeda..* Project, The University of Auckland
Tamaki Innovation Campus, Auckland, New Zealand
mrez010@aucklanduni.ac.nz , r.klette@auckland.ac.nz

Abstract. Eye status detection and localization is a fundamental step for driver awareness detection. The efficiency of any learning-based object detection method highly depends on the training dataset as well as learning parameters. The research develops optimum values of Haar-training parameters to create a nested cascade of classifiers for real-time eye status detection. The detectors can detect eye-status of open, closed, or diverted not only from frontal faces but also for rotated or tilted head poses. We discuss the unique features of our robust training database that significantly influenced the detection performance. The system has been practically implemented and tested in real-world and real-time processing with satisfactory results on determining driver's level of vigilance.

1 Introduction

The automotive industries implements active safety systems into their top-end cars for lane departure warning, safe distance driving, stop and speed sign recognition, and currently also first systems for driver monitoring [Wardlaw 2011]. Stereo vision or pedestrian detection are further examples of components of a driver assistant system (DAS).

Any sort of driver distraction and drowsiness can lead to catastrophic cases of traffic crashes not only for the driver and passengers in the *ego-vehicle* (i.e. the car the DAS is operating in) but also for surrounding traffic participants. Face pose and eye status are two main features for evaluating a driver's level of fatigue, drowsiness, distraction or drunkenness. Successful methods for face detection emerged in the 2000s. Research is now focusing on real time eye detection. Concerns in eye detection still exist for non-forward looking face positions, tilted heads, occlusion by eye-glasses, or restricted lightening conditions.

According to [Zhang and Zhang 2010], research on eye localization can be classified into four categories. *Knowledge-based methods* include some predefined rules for eye detection. *Template-matching methods* generally judge the presence or absence of an eye based on a generic eye shape as a reference; a search for eyes can be in the whole image or in pre-selected windows. Since eye models vary for different people, the locating results are heavily affected by eye model initialization and image contrast. High computational cost also prevents a wide application for this method. *Feature-based approaches* are based on fundamental eye-structures; typically a method starts here with determining properties such

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 171–179, 2011.

as edges, intensity of the iris and sclera, plus colour distributions of the skin around eyes to identify ‘main features’ of eyes [Niu et al. 2006]. This approach is relatively robust to lightning but fails in case of face rotation or eye occlusion (e.g. by hair or eye-glasses). *Appearance-based methods* learn different types of eyes from a large dataset and are different to template matching. The learning process is on the basis of common photometric features of human eye from a collective set of eye images with different head poses. The paper develops the last one-appearance-based method.

2 Cascade Classifiers Using Haar-Like Masks

Such a system was developed by [Viola and Jones 2001] as a face detector. The detector combines three techniques: the use of a comprehensive set of Haar-like *masks* (also called ‘features’ by Viola and Jones) that are in analogy to base functions of the Haar transform, the application of a boosted algorithm to select a set of masks for classifier training, and forming a cascade of strong classifiers by merging week classifiers. Haar-like masks are defined by adjacent dark and light rectangular regions; see Fig. 1.

Selection process of the object is based on the value distributions in dark or light regions of a mask that models expected intensity distributions. For example, the mask in Fig. 2, left, relates to the idea that in a face there are darker regions of eyes compared to the bridge of the nose. similarly, the mask in Fig. 2, right, models that the central part of an eye (the iris) is darker than the sclera area.

Computing Mask Values. Mean values in rectangular mask regions are calculated by applying the integral image as proposed in [Viola and Jones 2001]; see Fig. 3. For a given $M \times N$ picture P, at first the *integral image*

$$I(x,y) = \sum_{0 \le i \le x \wedge 0 \le j \le y} P(i,j) \tag{1}$$

is calculated. The sum $P(R_1)$ of all P-values in rectangle region R_1 (see Fig. 3) is then given by $I(D) + I(A) - I(B) - I(C)$. Analogously we calculate sums $P(R_2)$ and $P(R_3)$ from corner values in the integral image I. Values of contributing regions are weighted by reals ω_i that create *regional mask values* in form of

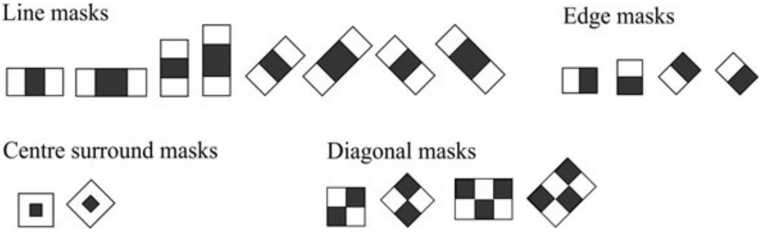


Fig. 1. Four different sets of masks for calculating Haar-like masks

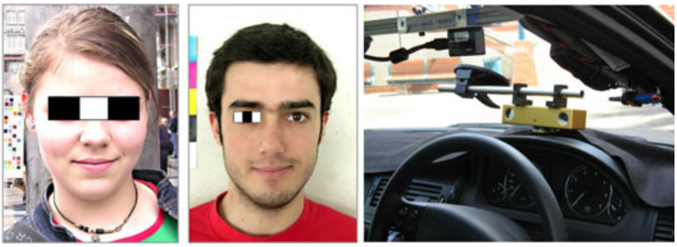


Fig. 2. *Left*: Application of two triple masks for collecting mean intensities in bright or dark regions. *Right*: Camera assembly in HAKA1 for driver distraction detection.

$v_i = \omega_i \cdot P(R_i)$, and then a *total mask value*; for the shown example this is $V_i = \omega_1 \cdot P(R_1) + \omega_2 \cdot P(R_2) + \omega_3 \cdot P(R_3)$. Signs of ω_i's are opposite for light and dark regions. In generalizing this approach, we also allow for arbitrary rotations. R_i is now defined by five parameters x, y, w, h, and φ, where x and y are coordinates of the lower-right corner, w and h are width and height, and φ is the rotation angle [Zhang and Zhang 2010]. For example, $P_\varphi(R_1) = I_\varphi(B) + I_\varphi(C) - I_\varphi(A) - I_\varphi(D)$ and for $\varphi = 45°$ we have

$$I_{45°}(x, y) = \sum_{|x-i| \le y-j \;\wedge\; 0 \le j \le y} P(i, j) \tag{2}$$

For any angle φ, the calculation of all $M \times N$ integral values I_φ takes time $\mathcal{O}(M \times N)$. This allows for real-time calculation of features on Haar-like masks.

Cascaded Classifiers via Boosted Learning. In a search window of 24×24 pixel there are more than 180,000 different rectangular masks of different shape, size, or rotation. However, only a small number of masks (usually less than 100) is sufficient to detect a desired object in an image (e.g. eye). In addition to defining regional mask weight w_i, using a boosting algorithm, the classifier can learn to sort out the prominent masks μ_i based on their overall wight W_i. Such wights determine the importance of each mask in an object detection process so we arrange all the masks in cascaded nodes as Fig. 4.

Each node (weak classifier) tries to determine whether the object (e.g. an eye) is inside the search window or not. The first classifier simply reject non-objects if

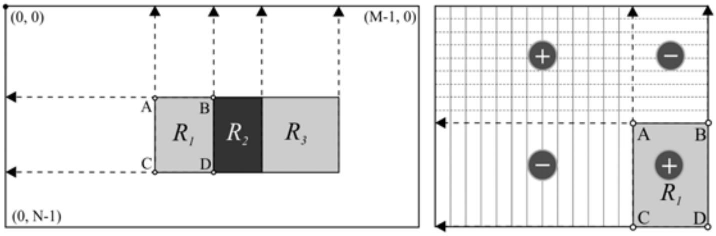


Fig. 3. Illustration for calculating a mask value using integral images. The coordinate origin is in the upper left corner.

the main masks (such as in Fig. 2) do not exist. If they exist then more detailed masks will be evaluated in next classifiers and the process continues. Actually each node represents a boosted classifier adjusted not to miss any object while it is rejecting non-objects if not matching the desired masks. Although each node is a weak classifier but all of them are considered a strong classifier and reaching the final node means that all non-objects have already been rejected and we have only one object (here: an eye). The function μ_i returns +1 if the mask value V_i is greater or equal to a trained threshold, and -1 if not:

$$\mu_i = \begin{cases} +1 & \text{if } V_i \geq T_i \\ -1 & \text{if } V_i < T_i \end{cases} \tag{3}$$

$\mu_i = +1$ means that the current weak classifier matches the object and we can proceed to the next classifier. Statistically about 75% of non-objects are rejected by the first two classifiers; the remaining 25% are for a more detailed analysis. This speeds up the process of object detection. In order to train the the algorithm we need a database of positive images (e.g. eyes)and on the first pass through the positive image database, we learn threshold T_1 for μ_1 such that it best classifies the input. Then boosting uses the resulting errors to calculate the overall weight W_1. Once the first node is trained then boosting continues for other nodes but with some other masks that are more sophisticated than previous ones [Freund et al. 1996].

Assume that each node (a weak classifier) is trained to correctly match and detect objects of interest with the true rate of $p = 99.9\%$ (true positive, TP). Since each stage alone is a weak classifier it is expected to be many false detections of non-objects, say $f = 50\%$ (false positive, FP)in each stage. This is still acceptable because, due to the serial nature of cascade classifiers, the overall detection ratios remains high (near 1) but it leads to a logarithmic decrease in the false positive rate (approaches to 0).

3 Scenarios and 3D Cascaded Classifiers

Most of eye detection algorithms such as [Wang et al. 2010] just look for the eyes in an already localized face. Therefore, eye detection simply fails if there is

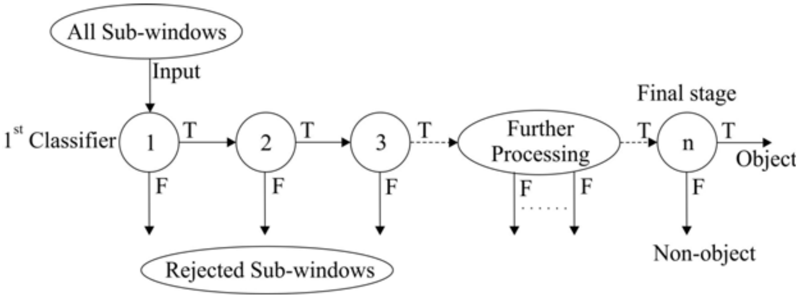


Fig. 4. Structure of cascaded classifiers for object detection

no full frontal view of a face, or some parts of face be occluded, or if parts of a face are outside of the camera viewing angle .

Our method follows a dynamic approaches, if the initial result of face detection is positive then we just look through the face region. Detection of an eye in a previously detected face region supports a double confirmation, and more confidence for the validity of eye detection. But if the face is not detected our 3D cascade looks for eye in the whole image. In our particular context we consider driver fatigue, drowsiness, distraction, or drunkenness when the driver misses to look forward on the road, or when the eyes are closed for some long uninterrupted period of time (say 1 sec. or more). As an example, when driving with a speed of 100 km/h, just one second eye closure means passing of 28 meters without paying attention. This can easily cause lane drift and a fatal crash. In our method we assume two status of *Looking Forward* and *Open Eyes* as important properties for judging driver's vigilance. for the face detection we follow the classifier in [Lienhart et al. 2003] for face detection and for the eye status detection we design our own classifiers. the proposed 3D designed classifier is able to detect and define 5 different scenarios while driving as below (see Fig. 5 from left to right):

Scenario 1: Obviously eyes are in the upper half of face region. By assessing 200 different faces from different races we derived that human eyes are geometrically located in segment A between 0.55 to 0.75 of the face's height. Applying this rough estimation in eye localization we already increased the search speed by factor 5 compared to a blind search, as we are only looking into 20% of the face's region. An eye pair is findable in segment A while the driver is looking forward.

Scenario 2: Some rare times happens that only one eye is detectable in segment A when the driver tilts his face. In that case we need to look for the second eye in segment B in the opposite half of the face region. Segment B is considered to be between 0.35 to 0.95 of the face's height; this covers more than ±30 degrees of face tilt. The size of the search window in segment B is 30% of the face region. In that case of a tilted face we search both sections A and B (in total, 50% of the face's region). In Scenarios 1 and 2, the driver is looking forward to the roadway. So if we detect two open eyes then we decide that the driver is in the *Aware* state.

Scenario 3: If a frontal face is not detectable and just one of the eyes is detected, then this can be due to more than 45° of face rotation. The driver is looking

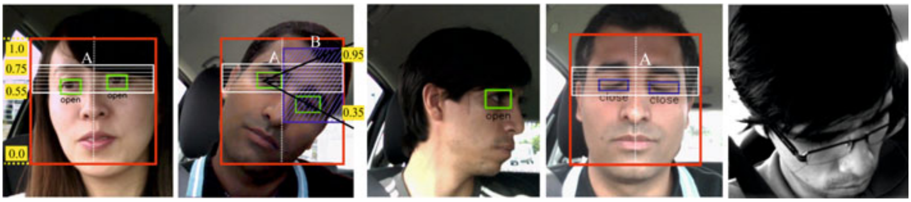


Fig. 5. Left to right: Scenarios 1 to 5 for driver's face and eye poses; see text for details

towards the right or left such that the second eye is occluded by the nose. The system immediately measures the period of time that the driver is looking to other sides instead of forward. This scenario also happens when the driver looks to side mirrors (but this takes normally less than second). Depending on the ego-vehicles speed, any occurrence of this scenario that takes more than 1 sec is considered as a sign of *Distraction* and the system will raise an alarm.

Scenario 4: Detection of closed eyes. Here we use an individual classifier for close eye detection. A closed-eye status happens frequently for normal eye blinking, and the eye *closure time* t_c is normally less than 0.3 sec. Any longer eye closures is a strong evidence of fatigue, drowsiness, or drunkenness. The system will raise an alarm for *Drowsiness* status if there is no open eye and at least one closed eye is detected.

Scenario 5: The worst case is when neither face, nor open eyes, nor closed eyes are detectable. This case occurs, for example, when the driver is looking over the shoulder, when the head falls in, or when the driver is performing secondary tasks. The system will raise an alarm for a detected *Risky Driving* status.

Considering all active detectors (face, open-eye, and close-eye detectors), we have cascaded classifiers in three dimensions that work in parallel. Implementing separate detectors for open and closed eye detection is important because at some times the open eye detector may fail to detect open eyes, but this does not necessarily mean that the eyes are closed. Missing eyes may be because of a specific head pose or bad lightening conditions. Having a separate closed-eye detector is a step toward high accuracy in driver distraction detection.

4 Training Image Database

The process of selecting positive and negative images is a very important step that affects the overall performance considerably. After several experiments it is determined that, although a larger number of positive and negative images can improve the detection performance in general, there is also an increase of the risk of mask mismatching during the training process. Thus, a careful consideration for number of positive and negative images and their content is essential. In addition, the multi-dimensionality of training parameters and the complexity of the feature space defines challenges. We propose optimized values of training parameters as well as unique features for our robust database.

In the initial negative image database, we removed all images that contained any objects similar to human eye (e.g. animal eyes). We prepared the training database by manually cropping closed or open eyes from positive images. Important questions needed to be answered: how to crop the eye regions and in what shapes (e.g. circular, isothetic rectangles, squares)? There is a general believe that circles or horizontal rectangles are best for fitting eye regions. However, we obtained the best experimental results by cropping eyes in square form. We fit the square enclosing full eye-width; for the vertical positioning we select balanced portions of skin area below and above the eye region. We cropped 12,000 eyes from selected positive images of our own database plus six other databases: FERET database sponsored by the DOD Counterdrug Technology Development Program Office [Phillips et al. 1998, Phillips et al. 2000], Radbound face database [Langner et al. 2010], Yale facial database B [Lee et al. 2005], BioID database [Jesorsky et al. 2001], PICS database [PICS], and the "Face of Tomorrow" [FTD]. The positive database includes more than 40 different poses and emotions for different faces, eye types, ages, and races:

– Gender and age: females and males between 6 to 94 years old,
– Emotion: neutral, happy, sad, anger, contempt, disgusted, surprised, feared,
– Looking angle: frontal (0°), ±22.5°, and profile (±45.0°), and
– Race: East-Asians, Caucasians, dark-skinned people, and Latino-Americans.

The generated multifaceted database is unique, statistically robust and competitive compared to other training databases.

We also selected 7,000 negative images (non-eye and non-face images) including a combination of common objects in indoor or outdoor scenes. Considering a search window of 24 × 24 pixel, we had about 7,680,000 sub-windows in our negative database. An increasing number of positive images in the training process caused a higher rate for true positive cases (TP) which is good, and also increased false positive cases (FP) which is bad. Similarly, when the number of negative training images increased, it lead to a decrease in both FP and TP. Therefore we needed to consider a good trade-off for the ratio of number of negative sub-windows to the number of positive images. For eye classifiers, we got the highest TP and lowest rate for false negative detection when we arranged the ratio of $N_p/N_n = 1.2$ (this may vary for face detection).

5 AdaBoost Learning Parameters and Experiments

We implemented the training algorithm in OpenCV 2.1. With respect to our database we gained a maximum performance by applying the following settings: Size of mask-window: 21×21 pixel. Total number of classifiers (nodes): 15 stages; any smaller number of stages brought a lot of false positive detection, and a larger number of stages reduced the rate of true positive detection. The minimum of acceptable hit rate for each stage: 99.80% and increasing; a rate too close to 100% may cause the training process to take for ever or early failure. The maximum acceptable false alarm for the 1st stage: 40.0% per stage; this error goes to zero exponentially when the number of iterations increases. Weight trimming threshold: 0.95; this is the similarity weight to pass or fail an object in each stage. Boosting algorithm: among four types of boosting (Discrete AdaBoost, Real AdaBoost, Logit AdaBoost, and Gentle AdaBoost), we got about 5% more TP detection rate with Gentle AdaBoost. [Lienhart et al. 2003] also proved that GAB will result into lower FP ratios for face detection.

We performed a performance evaluation test on 2,000 images from the second part of the FERET database plus on 2,000 other image sequences recorded by HAKA1, our research vehicle (see Fig. 2, right). None of the test images were

Table 1. Classifiers accuracy (in %) in terms of true positive and false positive rate

	Open-eye detection		Closed-eye detection	
Facial status	*TP*	*FP*	*TP*	*FP*
Frontal face	98.6	0.0	97.7	0.20
Tilted face (up to ±30°)	98.2	0.002	97.1	0.54
Rotated face (up to ±45°)	96.8	0.0	96.8	0.7

included before in the training process and all the images are recorded in daylight condition. Table 1 shows the final results of open and closed eye detection rate.

6 Conclusions

With the aim of driver distraction detection, we implemented a robust 3D detector based on Haar-like masks and AdaBoost machine learning that is able to inspect for face pose, open eyes and closed eyes at the same time. Despite the similar research that are only able to work on frontal faces, The developed classifier is also able to works for tilted and rotated faces in real-time driving applications. There are no comprehensive data about performance evaluation for eye detection. Comparing results in [Kasinski and Schmidt 2010], [Niu et al. 2006], [Wang et al. 2010] and in [Wilson and Fernandez 2006] with our results (see Table 1), the method appears to be superior in a majority of cases. The method still needs improvement for dark environments. High-dynamic range cameras or some kind of preprocessing might be sufficient to obtain satisfactory detection accuracy also at night or in low-light environments.

References

[Langner et al. 2010] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the Radbound faces database. Cognition Emotion 24, 1377–1388 (2010)

[Freund et al. 1996] Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Machine Learning, pp. 148–156 (1996)

[Jesorsky et al. 2001] Jesorsky, O., Kirchberg, K., Frischholz, R.: Robust face detection using the Hausdorff distance. J. Audio Video-based Person Authentication, 900–995 (2001)

[Kasinski and Schmidt 2010] Kasinski, A., Schmidt, A.: The architecture and performance of the face and eyes detection system based on the Haar cascade classifiers. J. Pattern Analysis Applications 3, 197–211 (2010)

[FTD] Face of tomorrow database (2010), http://www.faceoftomorrow.com/posters.asp

[Lee et al. 2005] Lee, K.C., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Pattern Analysis Machine Intelligence 27, 684–698 (2005)

[Lienhart et al. 2003] Lienhart, R., Kuranov, A., Pisarevsky, V.: Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: Michaelis, B., Krell, G. (eds.) DAGM 2003. LNCS, vol. 2781, pp. 297–304. Springer, Heidelberg (2003)

[Niu et al. 2006] Niu, Z., Shan, S., Yan, S., Chen, X., Gao, W.: 2D cascaded AdaBoost for eye localization. In: ICPR, vol. 2, pp. 1216–1219 (2006)

[Phillips et al. 1998] Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.: The FERET database and evaluation procedure for face recognition algorithms. J. Image Vision Computing 16, 295–306 (1998)

[Phillips et al. 2000] Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Analysis Machine Intelligence 22, 1090–1104 (2000)

[PICS] PICS image database: University of Stirling, Psychology Department (2011), http://pics.psych.stir.ac.uk/

[Viola and Jones 2001] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR, pp. 511–518 (2001)

[Wang et al. 2010] Wang, H., Zhou, L.B., Ying, Y.: A novel approach for real time eye state detection in fatigue awareness system. IEEE Robotics Automation Mechatronics, 528–532 (2010)

[Wardlaw 2011] Wardlaw, C.: 2012 Mercedes-Benz C-Class preview (2011), http://www.vehix.com:80/articles/auto-previewstrends/2012-mercedes-benz-c-class-preview

[Wilson and Fernandez 2006] Wilson, P.I., Fernandez, J.: Facial feature detection using Haar classifiers. J. Computing Science 21, 127–133 (2006)

[Zhang and Zhang 2010] Zhang, C., Zhang, Z.: A survey of recent advances in face detection. MSR-TR-2010-66, Microsoft Research (2010)

Non–destructive Detection of Hollow Heart in Potatoes Using Hyperspectral Imaging

Angel Dacal-Nieto[1], Arno Formella[1], Pilar Carrión[1],
Esteban Vazquez-Fernandez[2], and Manuel Fernández-Delgado[3]

[1] Computer Science Department, Universidade de Vigo,
Campus As Lagoas 32004 Ourense, Spain
angeldacal@uvigo.es
[2] GRADIANT, Galician R&D Center in Advanced Telecommunications, Spain
[3] Centro de Investigación en Tecnoloxías da Información (CITIUS),
Universidade de Santiago de Compostela, Spain

Abstract. We present a new method to detect the presence of the *hollow heart*, an internal disorder of the potato tubers, using hyperspectral imaging technology in the infrared region. A set of 468 hyperspectral cubes of images has been acquired from Agria variety potatoes, that have been cut later to check the presence of a hollow heart. We developed several experiments to recognize hollow heart potatoes using different Artificial Intelligence and Image Processing techniques. The results show that Support Vector Machines (SVM) achieve an accuracy of 89.1% of correct classification. This is an automatic and non-destructive approach, and it could be integrated into other machine vision developments.

Keywords: Hyperspectral, Infrared, Potato, SVM, Random Forest.

1 Introduction

Potatoes (*Solanum tuberosum*) are nowadays one of the most consumed products in the world: they are the world's fourth largest food crop. The annual production is 325 million tons and it moves an amount of global transactions of about 6 billion US dollars (2007 data). Thus, the world potato average consumption is 31 Kg per capita and year [1].

One of the internal characteristics of the potato tubers is the called *hollow heart*, a star–shaped cavity that grows into the potato. Some early studies point that there exist a relation between growing disorders and probability of the presence of a hollow heart [2]. Some contributions in the last years have tried to detect hollow hearts in potatoes using X–Ray examination [3] and acoustics [4,5], providing successful results (98%). However, [4] needs the potatoes to be isolated from noise and it can not detect tiny hollow hearts, meanwhile in [5] the potatoes are dropped to study the sound produced by the fall, which eventually bruises the samples. Moreover, both approaches are strongly dependent on the orientation of the potato. Despite these contributions, the main packaging companies in the North of Spain still use a human operator to deal with the problem,

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 180–187, 2011.

by removing bigger and amorphous tubers after destructively checking a small sample of the production, which causes subjectivity mistakes and possibly lower (but unknown) accuracy rates.

We propose a new automatic non–destructive method based on hyperspectral imaging, not dependent on the orientation, and with no potato isolation required. Hyperspectral imaging is a reliable approach to classical spectroscopy, because an object can be analysed in significantly less time, and always in a non-destructive way, despite a little loss of accuracy. This technology has become interesting in the field of food quality assessment [6], being used to predict the water content in potatoes [7], and to detect clods between a set of potato tubers [8]. Other contributions [9] use near–infrared (NIR) spectroscopy to predict specific gravity and dry matter in potatoes.

2 Image Acquisition System

The objective of hyperspectral imaging is to perform a spectroscopic analysis of the light reflected or transmitted by the object of interest. This is accomplished by coupling a spectrograph and a matrix camera, which obtains both spectral and spatial information. Our hyperspectral system has been designed for non-destructive food inspection in the NIR region. We coupled an infrared camera and a SWIR-NIR spectrograph, both sensitive from 900 nm to 1700 nm. Specifically, we used a Xenics Xeva 1.7-320 camera with an InGaAs 320 × 256 pixel sensor and USB connection. The spectrograph is a Specim Imspector N17E. The system has also three 50 W AC halogen lamps placed in the inspection plate to provide diffuse illumination to the potato surface. The diffuse light is obtained by the reflection in a plastic dome over the plate.

The spectrograph has a linear input (one pixel height), where the x-axis represents the same x-axis (spatial) of the object. The y-axis (spectral) is then *studied* to obtain how every pixel in the row varies along the spectral range.

With one spectral image, we are inspecting only one spatial line, so that we need to perform the inspection over the whole object. This is accomplished by joining a rotatory mirror scanner to the spectrograph. It is based on performing the mirror rotation, covering a 40° window over the object, taking care of synchronization between mirror stepping and image acquisition (Figure 1). Finally the images are transposed in order to obtain the hyperspectral cube (Figure 2).

To sum up, our system obtains 320 spectral images (320 × 240 pixels), that are transposed into a hyperspectral cube formed by 256 images with 320 × 320 pixels, corresponding to 256 consecutive wavelengths, equally spaced from 900 nm to 1700 nm.

3 Experiment

The objective of the experiment is to compare different algorithms for each Pattern Recognition stage in order to compose the combination of methods that maximizes the accuracy classifying hollow heart affected and healthy potatoes.

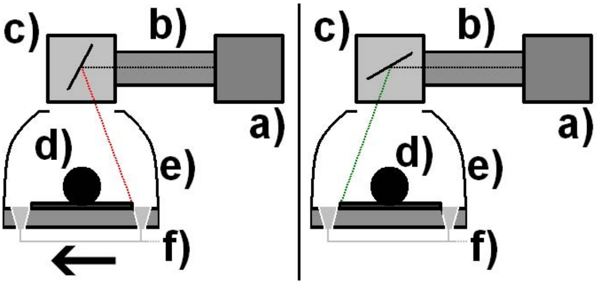


Fig. 1. Left: scanning initial position at 70°. Right: scanning final position at 110°. The arrow shows the direction of scanning. Hyperspectral system scheme: a) camera, b) spectrograph, c) mirror scanner, d) object, e) diffuse chamber, f) halogen lamp.

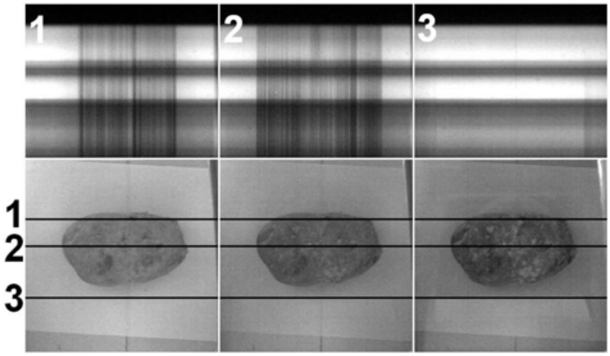


Fig. 2. Up: Three spectral images taken from different lines of the object. Down: 978 nm, 1173 nm, and 1608 nm spatial images.

The experiment uses 234 potato tubers (variety Agria) from Xinzo de Limia (Spain), that have been collected from some potato packing companies during 2009. The potatoes have been captured from two sides, using the system described in Section 2, and cut later to check the presence of hollow heart. They have been placed in a stable position, so that the biggest area is acquired.

3.1 Segmentation

Segmentation runs in several steps to obtain a mask to remove the background for the hyperspectral cube using the open source library OpenCV [10]. First, we binarize the image using Otsu's method [11], that calculates the optimum binarization threshold. Then, a Gaussian blurring clusters the noise in the image. Another binarization is needed for the next operation. A connected-component labelling is performed to remark contiguous areas in the image. At this point, we know that the blob with the largest area (excluding the background) is the potato. We select this blob and create the mask used to segment all the images in the hyperspectral cube. We call this segmentation method *full*.

Additionally, we have implemented three other segmentation methods. The *core* algorithm is intended to remove the external area of the potatoes, using a heavy erosion operation, so that we only take into account their central part. In the *border* algorithm, the aim is to remove the centre of the potato, so that the segmentation only makes visible a portion similar to a ring.

The last segmentation method (*scab*) has been developed in a parallel research [12], aimed to detect *common scab* (a skin disease in the potatoes) in an automatic and non-destructive way, using the same acquisition system. We use the result of the *scab* segmentation to obtain a hyperspectral cube free of common scab, which might be more accurate in the detection of hollow heart. The Figure 3 visualizes examples of the results given by these processes.

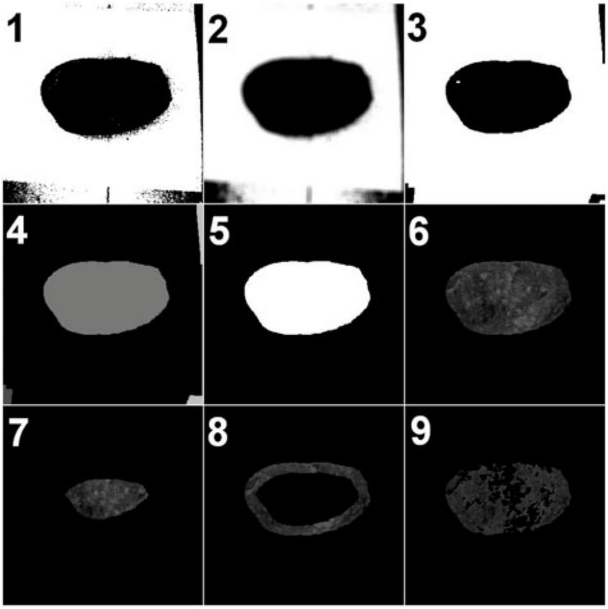


Fig. 3. 1: Binarization using Otsu's method. 2: smooth operation. 3: second binarization. 4: blob analysis. 5: mask used for hyperspectral cube segmentation. 6: *full* mask. 7: *core* mask. 8: *border* mask. 9: *scab* mask.

3.2 Feature Extraction

We calculated the average luminance value of the pixels belonging to the potato for each image in the hyperspectral cubes (i.e. for each of the 256 wavelengths). Depending on the segmentation method, we use the whole potato for this calculation (*full* mask), or different zones of the tuber (*core*, *border* and *scab* masks). Additionally, we included three morphological features in the feature list, namely the *area*, *perimeter* and *roundness* of the potato. Our objective is to test whether the potato size and roundness are relevant for the hollow heart detection. Hence, every hyperspectral cube is represented with 259 attributes (256 spectral and 3 morphological features). We used 468 samples (208 hollow heart affected and 260 healthy potatoes).

3.3 Feature Selection

This stage identifies which wavelengths are the optimal to detect hollow heart potatoes, in order to decrease the number of images to analyse. We used some algorithms implemented in Weka [13], using their default parameters: *Genetic* Search [14], *Scattered* Search [15], *Greedy* Stepwise [16], Linear Forward Selection (*LFS*) [17], and Correlation-based Feature Subset Selection (*CFS*) [18]. We also included the data set with all the features (*full*). We have discarded techniques such as Principal Component Analysis and Linear Discriminant Analysis, because they perform a linear combination of all the wavelengths, instead of selecting a subset, as the used feature selection methods do.

3.4 Classification

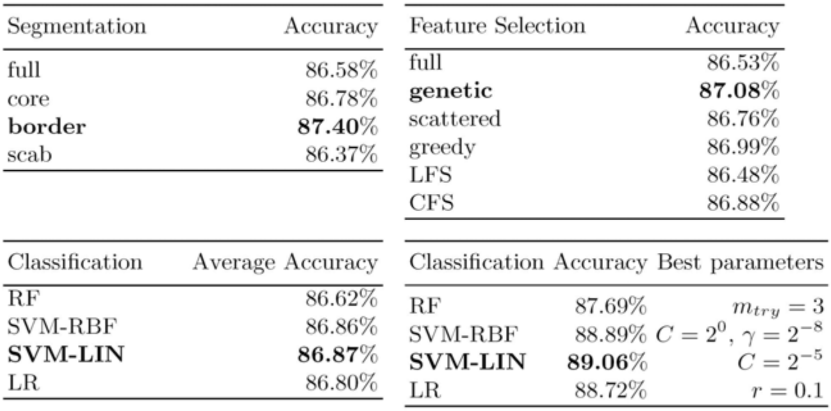
We present results of four classification algorithms: Random Forest (RF) [19], Support Vector Machines (SVM) [20] with Gaussian (SVM-RBF) and linear (SVM-LIN) kernels, and Logistic Regression (LR) [21]. Although LR is not among the most popular algorithms, it has been included in the experiment after good preliminary results with Weka [13].

Note that we have 4 segmentation methods and 6 feature selection methods (24 data sets) and 4 classification algorithms. In this stage we test each of these 96 options to solve our problem in order to evaluate which is the best solution. We randomly generated 10 permutations of the data sets. Each permutation was divided into three parts: training (50% of the samples), validation (25% of the samples, used for parameter tuning), and test (the remaining 25%). The samples were normalized (zero mean and standard deviation one) to avoid that attributes in greater numeric ranges influence excessively over those with smaller variation.

For each classifier, for each combination of tunable parameters and for each permutation, we trained the classifier using the 10 training sets. We tested its performance on the validation sets, selecting the parameter values with the best average accuracy over the 10 permutations. These parameters are: m_{try} (the number of features to use in random selection) for RF, using $m_{\text{try}} = p^0$, $m_{\text{try}} = \sqrt{p}$, $m_{\text{try}} = p/4$ and $m_{\text{try}} = p/2$, with p =number of features; the regularization parameter (C) and kernel spread (γ) for SVM-RBF, using $C = 2^n, n = -5 : 14$ and $\gamma = 2^n, n = -15 : 0$; SVM-LIN has just ($C$), using $C = 2^n, n = -5 : 14$, and LR has the ridge estimator (r), using $r = 10^k, k = -9 : 0$. Finally, for each permutation, we trained the classifier using the training sets tuned with the best parameters values, evaluating its accuracy on the 10 test sets.

4 Results and Discussion

The results are presented in Figure 4. Some average results using all the data sets are provided, in order to determine the best segmentation method (upper left panel in Figure 4), the best feature selection method (upper right panel), and the best classifier (lower left panel). The best data set uses the *border* segmentation method, the *genetic* feature selection method, and the SVM-LIN classification algorithm, achieving 89.06% of accuracy (lower right panel). The Table 1 shows the average confusion matrix achieved by the best data set–classifier pair using the test sets (117 samples).



Segmentation	Accuracy
full	86.58%
core	86.78%
border	**87.40%**
scab	86.37%

Feature Selection	Accuracy
full	86.53%
genetic	**87.08%**
scattered	86.76%
greedy	86.99%
LFS	86.48%
CFS	86.88%

Classification	Average Accuracy
RF	86.62%
SVM-RBF	86.86%
SVM-LIN	**86.87%**
LR	86.80%

Classification	Accuracy	Best parameters
RF	87.69%	$m_{try} = 3$
SVM-RBF	88.89%	$C = 2^0$, $\gamma = 2^{-8}$
SVM-LIN	**89.06%**	$C = 2^{-5}$
LR	88.72%	$r = 0.1$

Fig. 4. Upper left: average segmentation results using all the data sets. Upper right: average feature selection results using all the data sets. Lower left: average classification results using all the data sets. Lower right: results of the best data set (*border–genetic*).

Table 1. Average test confusion matrix achieved with the best combination of segmentation, feature selection and classification methods

Classified as / *Real*	Hollow heart	Healthy
Hollow heart	57.9	6.4
Healthy	6.4	46.3

It is interesting to note that the three morphological features were selected by all the feature selection algorithms in all the data sets, so that it seems they are very important information for the problem, which confirms [2] conclusions.

Finally, the Figure 5 presents the 10 wavelengths selected by the best feature selector (*genetic*), marked with black columns (wavelengths in 863, 905, 921, 1026, 1068, 1091, 1195, 1398, 1405, and 1434-1438 nm) over an example potato spectral chart. Although water absorption increases rapidly after 1450 nm [22], it is remarkable that all the selected wavelengths are below 1438 nm. This suggests that the water amount is not an important factor in the hollow heart detection.

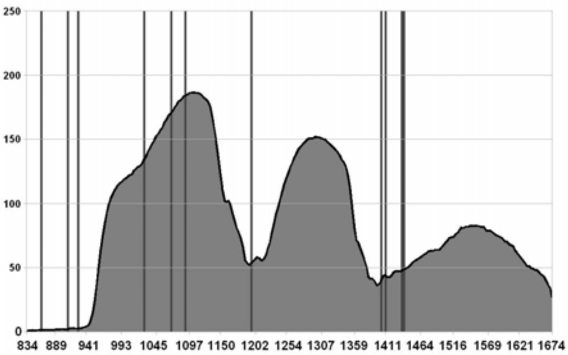


Fig. 5. Selected wavelengths on the best data set, marked with columns. The x-axis represents the bands. The y-axis represents the average grey level.

5 Conclusions

Infrared hyperspectral imaging has shown to be a good choice for hollow heart detection in potatoes of Agria variety. We developed an objective and non–destructive detection method using Pattern Recognition and Image Processing techniques, achieving accuracies of about 89.1%. The result can be interesting for the industry, because nowadays the process is still handled by human operators.

The *border* segmentation method seems slightly better than using the *full* potato. The results also indicate that removing the common scab from the hyperspectral cubes does not help the classification procedure and decreases the accuracy. The correlation between common scab and the presence of hollow heart will be studied in the future.

Regarding feature selection, size and roundness were detected to be essential features for the hollow heart detection, and should be taken into account. Besides, *genetic* has shown to be the most suitable feature selection algorithm.

In future work, it would be interesting to evaluate the system with other potato varieties, as well as researching the relationship between the optimal wavelengths and the biological causes of hollow heart.

Acknowledgements. This work was partly supported by the “Xunta de Galicia” (projects PGIDIT08TIC004CT and PGIDIT08MMA010402PR). We want to acknowledge the “Laboratorio Oficial de Metroloxía de Galicia (LOMG)” and the “Centro Tecnolóxico da Carne (CTC)” for their helpful collaboration.

References

1. Potato World, World-wide potato production statistics. International Year of the Potato (2008), http://www.potato2008.org/en/world/index.html
2. Rex, B.L., Mazza, G.: Cause, control and detection of hollow heart in potatoes: A review. Am. J. Potato Res. 66(3) (1989)

3. Finney, E.E., Norris, K.H.: X-Ray scans for detecting hollow heart in potatoes. Am. J. Potato Res. 55(2) (1978)
4. Jivanuwong, S.: Nondestructive detection of hollow heart in potatoes using ultrasonics. Master Thesis. Virginia Polytechnic Institute (1998)
5. Elbatawi, I.E.: An acoustic impact method to detect hollow heart of potato tubers. Biosyst. Eng. 100, 206–213 (2008)
6. Sun, D.: Hyperspectral Imaging for Food Quality Analysis and Control. Academic Press Elsevier, San Diego (2009)
7. Singh, B.: Visible and near-infrared spectroscopic analysis of potatoes. M.Sc. Thesis, McGill University, Montreal, PQ, Canada (2005)
8. Al-Mallahi, A., Kataoka, T., Okamoto, H., Shibata, Y.: Detection of potato tubers using an ultraviolet imaging-based machine vision system. Biosyst. Eng. 105, 257–265 (2009)
9. Kang, S., Lee, K., Son, J.: On-line internal quality evaluation system for the processing potatoes. In: Food Process. Autom. Conf. Proc., Providence, Rhode Island (2008)
10. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library. O'Reilly Media, Sebastopol (2008)
11. Otsu, N.: A threshold selection method for gray level histograms. IEEE Trans. Syst. Man Cybern. 9, 62–66 (1979)
12. Dacal-Nieto, A., Formella, A., Carrión, P., Vazquez-Fernandez, E., Fernández-Delgado, M.: Common scab detection on potatoes using an infrared hyperspectral imaging system. In: Proceedings of ICIAP 2011. LNCS, Springer, Heidelberg (2011)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
14. Goldberg, D.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading (1989)
15. García-López, F., García-Torres, M., Melián-Batista, B., Moreno-Pérez, J.A., Moreno-Vega, J.M.: Solving feature subset selection problem by a Parallel Scatter Search. Eur. J. Oper. Res. 169(2), 477–489 (2008)
16. Weihs, C.: Multivariate Exploratory Data Analysis and Graphics, A tutorial. J. Chemom. 7, 305–340 (1993)
17. Guetlein, M., Frank, E., Hall, M., Karwath, A.: Large Scale Attribute Selection Using Wrappers. In: Proc IEEE Symposium on Computational Intelligence and Data Mining, pp. 332–339 (2009)
18. Hall, M.: Correlation-based Feature Subset Selection for Machine Learning, Hamilton, New Zealand (1998)
19. Breiman, L.: Using Iterated Bagging to Debias Regressions. Mach. Learn. 45, 261–277 (2001)
20. Chang, C.C., Lin, C.J.: LIBSVM:a library for support vector machines (2008), http://www.csie.ntu.edu.tw/~cjlin/libsvm/
21. Le Cessie, S., Van Houwelingen, J.C.: Ridge Estimators in Logistic Regression. Appl. Stat. 41, 191–201 (1992)
22. Curcio, J.A., Petty, C.C.: The Near Infrared Absorption Spectrum of Liquid Water. J. Opt. Soc. Am. 41, 302–302 (1951)

Dice Recognition in Uncontrolled Illumination Conditions by Local Invariant Features

Gee-Sern Hsu*, Hsiao-Chia Peng, Chyi-Yeu Lin, and Pendry Alexandra

Department of Mechanical Engineering,
National Taiwan University of Science and Technology
jison@mail.ntust.edu.tw

Abstract. A system is proposed for the recognition of the number of the dots on dice in general table game settings. Different from previous dice recognition systems which use a single top-view camera and work only under controlled illumination, the proposed one uses multiple cameras and works for uncontrolled illumination. Under controlled illumination edges are the prominent features considered by most approaches. But strong specular reflection, often observed in uncontrolled illumination, paralyzes the approaches solely based on edges. The proposed system exploits the local invariant features robust to illumination variation and good for building homographies across multi-views. The homographies are used to enhance coplanar features and weaken noncoplanar features, giving a way to segment the top faces of the dice and make up the features ruined by possible specular reflection. To identify the dots on the segmented top faces, an MSER detector is applied for its consistency rendering local interest regions across large illumination variation. Experiments show that the proposed system can achieve a superb recognition rate in various uncontrolled illumination conditions.

Keywords: Object recognition, invariant feature, local descriptor.

1 Introduction

Dice is a popular table game in casinos, especially in Asia. As automatic or computer-controlled games are emerging and becoming popular, many are interested in the technologies able to assist or replace human bankers. A computer vision system is proposed in this paper for *dice recognition*, which refers to the automatic recognition of the numbers of dots on dice, in normal table game settings. Different from existing dice recognition systems, for example [4] and [5], which work under controlled illumination, the proposed system can work in uncontrolled illumination conditions. In controlled illumination edges are the prominent features considered. But specular reflection, often observed in uncontrolled illumination, paralyzes the approaches solely based on edges. Fig. 1 shows an image in the middle with strong specular reflection, on the left is its edge map

[*] Corresponding author.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 188–195, 2011.

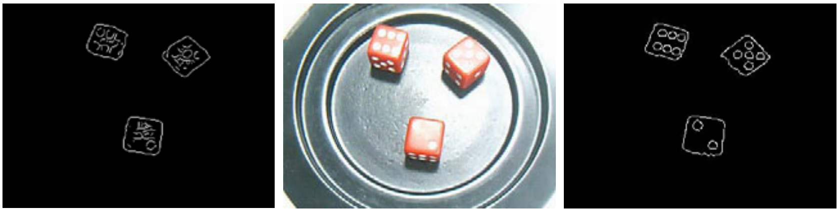


Fig. 1. Middle: specular reflection on the dice; Left: the edge map obtained by previous methods; Right: the edge map obtained by the proposed method

obtained by previous methods. Because it is not limited to controlled illumination, the proposed allows a much wider scope of applications, e.g., integration with table games or different designs of automatic dice games.

Existing dice recognition systems only consider the top view of dice. But a top-view camera is difficult to install on a game table as a specially designed camera support will be needed. To enable an easy integration with a game table, the proposed system considers tilted views to the dice captured by the cameras held on the peripheral supports around the table. Peripheral cameras are more friendly to install on a game table than top-view ones. However top views only capture the top faces of the dice, tilted views reveal the top and side surfaces. The latter is harder to handle as a method is required to segment the top faces and remove the side surfaces.

The proposed system consists of two major modules: dice segmentation and dots identification. To segment dice, it exploits the local invariant features robust to illumination variation and good for building homographies across multi-views. The homographies are used to enhance coplanar features, segment the top faces of the dice and make up the features ruined by possible specular reflection. To identify the dots on the segmented top faces, an MSER (Maximally Stable Extreme Region) [8] detector is applied for its consistency rendering local interest regions across large illumination variation. Although one can consider classifiers for the segmentation and identification, such as that proposed by Viola and Jones [12], they are not considered here as a large amount of training samples are required. The proposed only need a few samples as references.

The rest of this paper is organized as follows: the dice segmentation is presented in Section 2. The dot identification is elaborated in Section 3. Section 4 presents an experimental study of the proposed methods, followed by a conclusion in Section 5.

2 Dice Segmentation Using Local Invariant Features

Because dice can pose in arbitrary locations and orientations on a dice roller base and their sizes vary slightly according to the distance to the camera, local invariant features are explored in capturing these variations. Many local invariant feature detectors were proposed and applied in a broad range of applications. Reviews on these detectors can be found in [10], and [9], [3]. The invariant

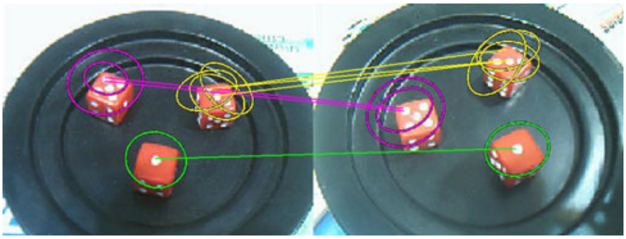


Fig. 2. Correspondences across two different views on the local invariant features detected by a multi-scale Harris-Hessian detector. Many of the detected correspondences are removed for better visual inspection.

feature detectors can be generally categorized into three types [11]. One detects corner-like features, e.g., Harris-affine, Harris-Laplace, and multi-scale Harris detectors.One detects blob-like features, e.g., Hessian-affine, Hessian-Laplace, multi-scale Hessian and Difference of Gaussians (DoG) [7]. Different from the former two types, region detectors extract homogeneous local areas, e.g., the MSER detector [8], which is used in this work for identifying the dots on dice, and will be addressed in details in Sec. 3.

Due to the limitation of Harris and Hessian detectors in handling multiple scales, both are modified with multiple scales and made scale-invariant in [1]. To determine the most appropriate scale for a local feature, Harris-Laplace and Hessian-Laplace both search for the characteristic scale with a Laplace operator added on top of the multi-scales. Harris-affine and Hessian-affine obtain the affine invariant corners or blobs by an iterative estimation of elliptical affine regions proposed by Lindeberg et al. [6]. The shape of the feature region is adapted to ensure that the same region is covered when extracted from a different viewpoint.

The performance of the aforementioned 8 invariant feature detectors in rendering the most accurate homographies between different viewpoints is evaluated by a comparison to the ground truth obtained using manually selected correspondences. All of the invariant regions (or interest regions) are represented in the form of SIFT descriptor [7] as it is experimentally proven as one of the most effective descriptors among others [10]. The match of the invariant features across views is measured by the Euclidean distance between the feature descriptors, and a threshold on this distance measure is determined to select correspondences. Because a dot on a die in a given view can appear quite similar to a different dot in another view, the scale factor in the local feature detectors is first chosen as that comes with the maximum number of correct correspondences. RANSAC [2] is then applied to filter out outliers and determine the most appropriate homographies across different views with matched correspondences. Our experiments reveal that the multi-scale Harris-Hessian detector gives the best performance. Fig. 2 shows an example of the correspondences across two viewpoints obtained using this detector. The settings and other details of the performance evaluation are reported in Section 4.

Given N different viewpoints of dice images, $N(N-1)/2$ homographies would be obtained using the invariant feature correspondences. In most cases $2 \leq N \leq 4$ suffices. Each homography and its inverse define the transformation between a pair of different viewpoints, and such a transformation only works for the top faces of the dice as these surfaces are *coplanar*. This property motivates the stacking of coplanar surfaces to segment the top faces of the dice even when specular reflection appears in certain viewpoints. One can choose a dice image of any viewpoint as a reference image and transform the rest $N-1$ images of different viewpoints to the reference one using the corresponding homographies.

Stacking of the reference image and $N-1$ transformed images does not just enhance the coplanar features but also weaken the non-coplanar features, as those on the lateral sides of the dice would be overlapped with features from different planes. As the specular reflection can be considered a view-dependent feature, different from the coplanar features observed in other majority of views, it can be removed by imposing a threshold on a similarity measure. An example with $N = 3$ is shown in Fig. 1, which in the middle shows a view of the dice with strong specular reflection, and on the right is the edge map of the image by stacking the homography-transformed images from the rest two views.

3 Dot Identification and Dice Recognition

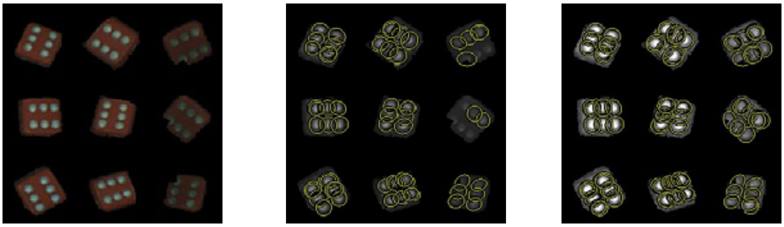
Given a segmented top face of a die, an MSER detector [8] is exploited to extract the dots from the segmented area because of its stability in rendering persistent or slowly varying edges around the dots as illumination varies. The extraction of MSER considers the set of all possible thresholds able to binarize an intensity image $I(\mathbf{x})$ into a binary image $E_{t_M}(\mathbf{x})$,

$$E_{t_M}(\mathbf{x}) = \begin{cases} 1 & if I(\mathbf{x}) \leq t_M \\ 0 & otherwise. \end{cases} \tag{1}$$

where t_M is the threshold. An MSER is a connected region in $E_{t_M}(\mathbf{x})$, with little change in its size for a range of thresholds, extracted with a watershed like segmentation algorithm. The homogeneous intensity regions extracted are stable over a wide range of thresholds. The number of thresholds that maintain the connected region similar in size is known as the *margin* of the region.

The dots on dice are blob-like objects and MSER usually anchors on the boundaries of such objects, and thus the dots can be better located by MSER compared to other interest region detectors. Fig. 3 shows the MSER regions detected on dice. With some preprocessing, as histogram equalization, MSER can achieve highly accurate identification rate. Fig. 3 shows a case with the segmented top faces, and the regions detected by MSER before and after preprocessing. Note that the MSER can detect incomplete or partial interest regions which can be due to imperfect segmentation.

The dots identified by the MSER are clustered by k-means (k happens to be the number of dice) subject to the constraints that the number of dots in a cluster must be less than 7 and the distance between the farthest dots must



(a) Segmentation of top faces

(b) Regions detected before preprocessing

(c) Regions detected after preprocessing

Fig. 3. The performance of MSER in the identification of the dots

be less than the diagonal of the dice. The spatial distribution of the dots in each cluster must be verified against the 6 known patterns. For example, 6-dot must contain two parallel rows of dots and 3 dots each row. 5-dot must have two crossing rows of dots, 3 dots each row and crossing each other at the same central dot. Specific patterns are configured for 4-, 3-, and 2-dot cases. Depending on the number of dots in a given cluster, the distribution pattern for that number is examined first, and if found incompatible, two possibilities would be verified. One is a non-dot spot falsely considered as a dot and the other is a valid dot failed to be identified as a dot. A large number of casts and experiments, with details given in Section 4, reveal that such a combination of size-constrained clustering and spatial pattern confirmation yields a superb recognition rate.

4 Experiments

The experimental setup follows a common dice table game ”sci-bo” with three dice, and three cameras of different viewpoints are installed on the sides of a game table. 12 different illumination conditions are configured to study the performance of the proposed system, 3 of them chosen as the *training set* and the rest 9 as the *test set*, as shown in Fig. 4. The intensity on the dice from the training set is 67, 108, and 138 in average, in 8-bit gray scale, with deviation 8, 10, and 11, respectively. The intensity on the test set is between 45 to 158 in average with deviation from 7 to 12. 120 random cast sessions and 30 manual placement sessions are carried out under each illumination condition. The manual placement attempts to create special layouts of the dice, such as three dice in a row and others.

4.1 Homography Based on Local Invariant Features

The training set is for the evaluation of the 8 invariant feature detectors, mentioned in Section 2, in creating homographies with least error across different illumination conditions. The error E_{F_i} is measured by the difference between the correspondences from the invariant-feature-based homography $\mathbf{H}_{F_i}$ and the ground-truth $\mathbf{H}_G$ obtained using manually selected correspondences, i.e.,

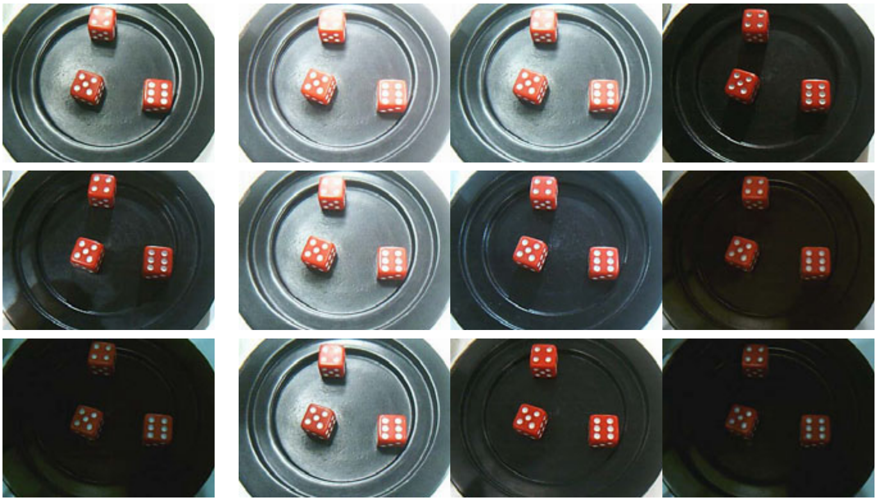


Fig. 4. First column from the left is the training set with 3 illumination conditions; the rest is the test set with 9 illumination conditions

$$E_{F_i}^{(a,b)} = \frac{||(\mathbf{H}_{F_i}^{(a,b)} - \mathbf{H}_G^{(a,b)})\mathbf{x}_{F_i}^b||}{N_{F_i}} \tag{2}$$

where $\mathbf{H}_{F_i}^{(a,b)}$ is the homography that transforms the invariant features $\mathbf{x}_{F_i}^b$ detected by the invariant feature detector F_i in the image I_b to the corresponding ones in I_a; $\mathbf{H}_G$ is the ground-truth homography obtained by manual selected correspondences between I_a are I_b, N_{F_i} is the number of features detected by F_i, and a, b denote two different viewpoints.

Additionally, it is also desired that the correspondences from the feature-based homographies can be consistent across different scales, as some features change with scales. To investigate what features are better than others in rendering desired homographies across illumination and scale, the original images in 320×240 pixels are scaled down to smaller sizes, and the error is computed in each size and averaged over the three illumination conditions in the training set. Fig. 5 shows

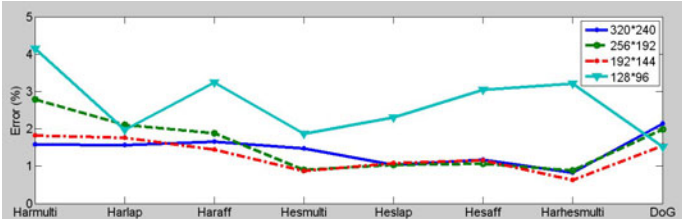


Fig. 5. Normalized error of feature-based homography across scales and three illumination conditions

this comparison, the smallest scale with 128 × 96 reveals relatively high errors, indicating that some details between the dice are lost in such a small scale and thus the accuracy in the homography estimation is degraded. Among the eight invariant feature detectors we tested, the multi-scale Harris-Hessian detector gives the lowest error at 0.87%, and it is about 1.7 pixels in a 192 × 144 image.

4.2 Dice Identification

The performance evaluation on the 9 test sets reveals the following observations and results:

- As long as the correspondences from the feature-based homography are consistent over at least two scales, the average match error can be kept below or near 1%, and the top faces of dice can be perfectly segmented in all tested conditions.
- Two identification rates are measured in each test illumination condition, one is the identification of the dots and the other is the identification of the dot number on each die. The former is shown by the bar on the left and the latter by the bar on the right at each indexed illumination condition in Fig. 6. Because the MSER dot detector has been adjusted to zero miss rate on the price of additional false positives on the training set, the imperfections in the dot identification in Fig. 6 are all caused by false positives. For example, in the brightest illumination condition, indexed ”1”, 1.8%(=1 − 98.2%) of the dots identified are false positives. All false positives are found caused by specular reflection or insufficient lightings. As the intensity of the illumination increases, specular reflection becomes stronger, causing more false positives to appear.
- The combination of size-constrained clustering and spatial pattern confirmation can effectively remove the false positives and yield superb dice recognition rates in all tested conditions, as shown by the right bar at each indexed illumination in Fig. 6.

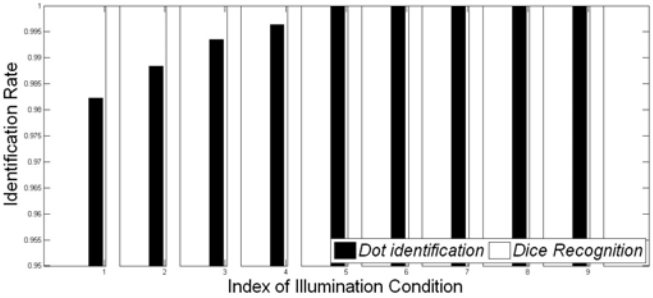


Fig. 6. Identification rates in 9 illumination conditions, indexed from 1 to 9; at each index the left bar shows the rate of dot identification, and the right bar shows the rate of dice number identification

5 Conclusion

A solution with invariant features across multiple views is proposed for dice recognition under uncontrolled illumination. An extensive comparison on the performance of various invariant feature detectors in rendering correct homographies under various test conditions and parameters shows that the multi-scale Harris Hessian is the best, and better than the commonly selected SIFT features. The homographies built on the multi-scale Harris Hessian features are exploited to enhance the coplanar features and weaken the non-coplanar features on the dice. This leads to an extraction of the coplanar features and the segmentation of the top faces of the dice even when the features, observed from some viewpoint, are ruined by specular reflection. An MSER detector is applied for the identification of dots on the top faces, followed by a pattern-specific confirmation of the spatial distribution of dots. Experiments reveal that, although false positives of dots are observed in few cases, as under strong or insufficient illumination, the numbers of the dots on the dice can still be recognized accurately by the proposed solution.

References

1. Dufournaud, Y., Schmid, C., Horaud, R.: Matching images with different resolutions. In: CVPR, pp. 1612–1618 (2000)
2. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24(6), 381–395 (1981)
3. Hsu, G.S.J.: Stereo Correspondence with Local Descriptors for Object Recognition. In: Advances in Theory and Applications of Stereo Vision, ch. 7, pp. 129–150. InTech (2011)
4. Huang, K.Y.: An auto-recognizing system for dice games using a modified unsupervised grey clustering algorithm. Sensors 8(2), 1212–1221 (2008)
5. Lai, Y.N., Hsu, S.T., Wang, C.Y., Tsai, M.T.: Method for recognizing dice dots. U.S. Patent No. 2009/0263008 A1 (October 2009)
6. Lindeberg, T., Gårding, J.: Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. Image Vision Comput. 15(6), 415–434 (1997)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
8. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: BMVC (2002)
9. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. International Journal of Computer Vision 65(1-2), 43–72 (2005)
10. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 27(10), 1615–1630 (2005)
11. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3), 177–280 (2007)
12. Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: CVPR, vol. (1), pp. 511–518 (2001)

Specularity Detection Using Time-of-Flight Cameras

Faisal Mufti[1] and Robert Mahony[2]

[1] Center for Advanced Studies in Engineering
faisal.mufti@ieee.org
[2] Australian National University
robert.mahony@anu.edu.au

Abstract. Time-of-flight (TOF) cameras are primarily used for range estimation by illuminating the scene through a TOF infrared source. However, additional background sources of illumination of the scene are also captured in the measurement process. This paper exploits conventional Lambertian and Phong's illumination models, developed for 2D CCD image cameras, to propose a radiometric model for a generic TOF camera. The model is used as the basis for a novel specularity detection algorithm. The proposed model is experimentally verified using real data.

Keywords: Time-of-flight, Radiometric Modelling, Specularity Detection, Reflectance Modelling.

1 Introduction

Objects and materials in real world appear differently to an observer depending on the nature of the light source that they are illuminated by and the manner in which the light is reflected to the observer. Computer vision [1] and computer graphics [2] researchers have extensively treated reflectance modelling for image analysis, rendering and scene geometry. Specular highlights can be used to provide information about the surface [12] and the illumination geometry [14] in a natural scene. However, saturation effects, due to specularity in intensity images, often create problems for image processing algorithms in real environments [13]. In addition, many computer vision algorithms [7,9] are dependent on surface illumination of an object and changing illumination conditions, such as highly saturated highlights interfere and adversely effect the camera image. It is therefore, important to detect specular highlights in image processing applications and algorithms. Since photometric understanding (using 2D CCD camera technology) of illumination modelling is focused on intensity, specularity detection methods [8,1] are normally based on chromaticity of the region.

3D time-of-flight (TOF) cameras provide information in addition to intensity that can be incorporated in a reflectance model. A TOF camera works on the principle of measuring time of flight of a modulated infrared light signal as phase offset after reflection and provides amplitude, phase and intensity data over a full image array at video frame rate [5].

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 196–203, 2011.

This paper presents a novel algorithm for specularity detection using TOF cameras. In the proposed radiometric framework, the background light sources and the dependencies between amplitude, intensity and phase/range measurements of a TOF camera are exploited. The model is utilized for specularity detection using real TOF camera data.

2 Reflectance Model

Time-of-flight (TOF) sensors estimate distance to a target using the time of flight of a modulated infrared (IR) wave between the target and the camera. The sensor illuminates/irradiates the scene with a modulated signal of amplitude A (*exitance*) and receives back a signal (*radiosity*) after reflection from the scene with background signal offset I_o that includes non-modulated DC offset generated by TOF camera as well as ambient light reflected from the scene. The amplitude, intensity offset I and phase of a modulated signal can be extracted by demodulating the incoming signal $A_i = A\cos(\omega t_i + \varphi) + I$; $(t_i = i \cdot \frac{\pi}{2\omega}, i = 0, \ldots 3)$ [5]. With known phase φ, modulation frequency f_{mod} and precise knowledge of speed of light c, it is possible to measure the un-ambiguous distance r from the camera [11].

The measurement parameters of amplitude A, intensity I, and range r are not independent but depend on the reflectance characteristics of the scene [11]. In this discussion a near-field IR point source for the camera's active LED array, an ambient illumination and a far-field source for background illumination is considered. The primary source of illumination in TOF cameras is an IR source that produces a modulated IR signal offset and a non-modulated DC signal. Based on Phong's illumination model [14], [3, p. 729], the following discussion incorporates diffuse $(.)_d$, ambient $(.)_a$, and specular $(.)_s$ components of illumination.

2.1 Modulated IR Source

Let P be a Lambertian surface in space with n_p denoting the normal to each point $p \in P$ on the surface as shown in Figure. 1. Following the laws of radiometry [15] the amplitude of total radiance $A_d(p)$ (called *radiosity*) leaving point p due to illumination by the modulated signal $A(s)$ is proportional to the *diffuse reflectance* or *albedo* $\rho_d(p)$ scaled by the cosine of arrival angle θ_p [10, p.68]. In the present analysis, the LED point sources of the camera are part of the compact IR array of the TOF camera, and can be approximated by a single virtual modulated point source [4, p. 78] with the centre of illumination aligned with the optical axis of the camera [6]. In this case, the integration for illuminating sources can be written as a function of the exitance of a single point source at S as [4, p. 77] [11]

$$A_d(p) := \frac{1}{\pi}\rho_d(p)\frac{A(s)\cos\theta_p\cos\theta_s}{r^2}. \tag{1}$$

The irradiance of an image point x is obtained [10, p. 48] as

$$A_d(x) = \Upsilon A_d(p), \tag{2}$$

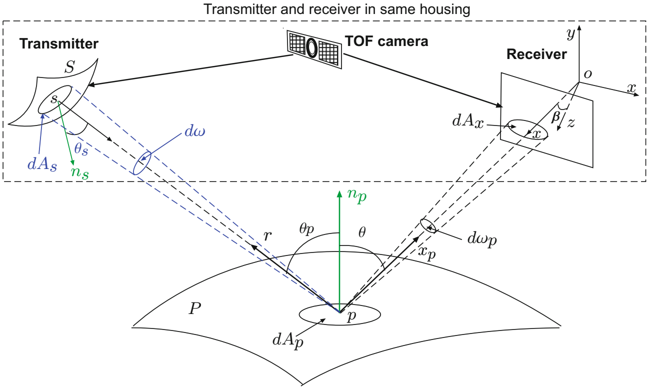


Fig. 1. Geometry of reflectance model for time-of-camera. Note that although the LED source and receiver of a physical TOF camera are co-located, it is difficult to provide a visualisation of this geometry. Here the source is shown separately to make is easier to see notation. However, in practice the directional vectors r and x_p are equal. Note that time variation (discussed in Section 2) of $A(s)$ does not need to be modelled as only the relative magnitude of $A(s)$ is of interest.

where $\Upsilon := \Upsilon(x)$[1] is the lens collection [15] representing the vignetting due to aperture size and irradiance fall-off with cosine-fourth law.

2.2 Non-modulated IR Source

The TOF camera IR source produces a DC signal from the same IR source LEDs. This signal will have the same reflectance model as has been derived for the modulated IR source (see (1)). The received signal $I_{c_d}(x)$ is given by [11]

$$I_{c_d}(x) = \Upsilon I_{c_d}(p). \tag{3}$$

The effect of this signal is an added offset to the modulated signal that provides better illumination of the scene.

2.3 Far-Field Background Illumination

For a point source $q \in Q$ that is far away compared to the area of the target surface, the exitance $I_{b_d}(q)$, does not depend on the distance from the source or the direction in which the light is emitted. Such a point source can be treated as constant [4, p. 76]. The radiosity perceived by a TOF image plane as a result of this IR source is given by [11]

$$\begin{aligned} I_{b_d}(x) &= \frac{\Upsilon}{\pi}\rho_d(p) I_b(q)\cos\theta_q \\ &= \Upsilon I_{b_d}(p). \end{aligned} \tag{4}$$

where θ_q is the angle between normal to the surface point p.

[1] := Defination of a symbol

2.4 Ambient Background Illumination

Consider an ambient background illumination of the scene i.e an illumination that is constant for the environment [4, p. 79] and produces a diffuse uniform lighting over the object [3, p. 273]. Let I_a be the intensity (called *exitance*) of the ambient illumination, then the received intensity $I_a(p)$ from a point p is expressed in an image plane as [11]

$$\begin{aligned} I_a(x) &= \frac{\Upsilon}{\pi}\rho_a(p)I_a \\ &= \Upsilon I_{_a}(p), \end{aligned} \tag{5}$$

where ρ_a is the *ambient reflection coefficient* which is often estimated empirically instead of relating it to the properties of a real material [3, p. 723].

2.5 Specular Illumination

Specular reflection is observed from a shiny surface when light is reflected in a single direction where the angle of incidence θ_p and angle of reflection θ are equal around the normal to the surface. The fall off effect of specular reflectance from shiny surfaces is modelled by $\cos^n \alpha$, where n is the *specular reflection exponent* [3] and α is the angle between direction of reflection and the view point. For imperfect shiny surfaces, specular reflection is spread over an angle α around the direct reflection. The received illumination components for intensity and amplitude observed in the image plane due to specularity are given by

$$I_{(.)_s}(x) := \frac{\Upsilon \rho_s(p)}{\pi}\frac{I_s(s)\cos^n \alpha \cos\theta_s}{r^2}; A_{_s}(x) := \frac{\Upsilon \rho_s(p)}{\pi}\frac{A_s(s)\cos^n \alpha \cos\theta_s}{r^2} \tag{6}$$

where the *specular reflection coefficient* $\rho_s(p)$, effects the brightness of specularity. Typical values of n vary from 0 to several hundred depending upon the surface material. A value of 1 gives a broad fall off of specular reflectance and a high value results in sharp fall-off of the specular reflectance.

3 Specularity Detection

From the principles of TOF camera (see Section 2) signals one knows that intensity component of TOF carries information for both, amplitude of the modulated signal and the background offset I_o. The radiometric intensity measured by a TOF camera is then

$$I := A + I_o. \tag{7}$$

The background offset I_o is composed of DC offset I_c, due to the DC component of the illumination by the TOF camera LED array and background illumination that are modelled by an ambient illumination I_a and a background illumination I_b due to an infrared far field source present in the environment such as the Sun or other light source. Indexing the point p in the scene by the TOF receiving pixel x, one has

$$I_o(x) = I_a(x) + I_{b_d}(x) + I_{b_s}(x) + I_{c_d}(x) + I_{c_s}(x), \tag{8}$$

Using the total intensity of the TOF camera (7) and background offset (8) and dividing it by the total (diffuse plus specular) received amplitude, one has

$$\frac{I(x)}{A(x)} = 1 + \frac{I_a(x)}{A(x)} + \frac{I_b(x)}{A(x)} + \frac{I_c(x)}{A(x)}, \tag{9}$$

where $I_b(x) = I_{b_d}(x) + I_{b_s}(x)$, $I_c(x) = I_{c_d}(x) + I_{c_s}(x)$ and $A(x) = A_d(x) + A_s(x)$. Using the reflectance models derived earlier (2) (3), and the specular model (6), it is now possible to re-arrange (9) as

$$\frac{I(x)}{A(x)} = 1 + \kappa_c(x) + \kappa_a \frac{r^2(x)\rho_a(p)}{\cos\theta_s[\rho_d(p)\cos\theta_p + \rho_s(p)\cos^n\alpha]}, \tag{10}$$

where $\theta_s := \theta_s(x)$ is a known function of a pixel and κ_a is defined as the ratio of background ambient light I_a to modulated TOF IR source $A(s)$. Observe that κ_a does not depend upon scene or camera geometry and hence is a constant parameter over the full image array. Also for an indoor environment (such as the one with no direct sunlight effect) the terms involving I_b in (9) is ignored in order to simplify the model and only ambient illumination (due to indoor lighting) component I_a is considered. The parameter $\kappa_c = \kappa_c(x)$, is defined as the ratio of TOF non-modulated IR source $I_c(s)$ to TOF modulated IR source $A(s)$. Since the two sources of illumination originating from the TOF camera IR LED source have the same ray geometry, they are in direct proportion where $\kappa_c(x)$ is a camera based pixel $x \in \mathbb{R}^2$ parameter independent of the scene for an entire image [11].

Thus for each pixel x, one can re-write (10) as

$$\kappa_a(x)\frac{\rho_a(p)}{[\rho_d(p)\cos\theta_p + \rho_s(p)\cos^n\alpha]} := \left(\frac{I(x)}{A(x)} - \kappa_c(x) - 1\right)\frac{\cos\theta_s}{r^2(x)}. \tag{11}$$

Define the specular measurement criterion $\check{\kappa}_a(x)$ based on measurements taken from the camera at a given time as

$$\boxed{\check{\kappa}_a(x)\frac{\rho_a(p)}{[\rho_d(p)\cos\theta_p + \rho_s(p)\cos^n\alpha]} := \left(\frac{\check{I}(x)}{\check{A}(x)} - \hat{\kappa}_c(x) - 1\right)\frac{\cos\theta_s}{\check{r}^2},} \tag{12}$$

where $\hat{\kappa}_c(x) \in \mathbb{R}^2$ is an estimate of camera based pixel parameter for an entire image. Since $\kappa_c(x)$ is scene independent, it can be measured offline as $\hat{\kappa}_c(x)$ in a set of calibration experiments.

For a TOF camera, the light direction of the source and the receiver are collinear as a result specularity is only observed in the direction of IR light from the camera. For maximum specularity $\alpha = 0$, the angle $\cos\theta_p = 1$ and the left hand side of (12) is scaled by a constant term C_ρ as

$$\check{\kappa}_a(x)C_\rho := \left(\frac{\check{I}(x)}{\check{A}(x)} - \hat{\kappa}_c(x) - 1\right)\frac{\cos\theta_s}{\check{r}^2}, \tag{13}$$

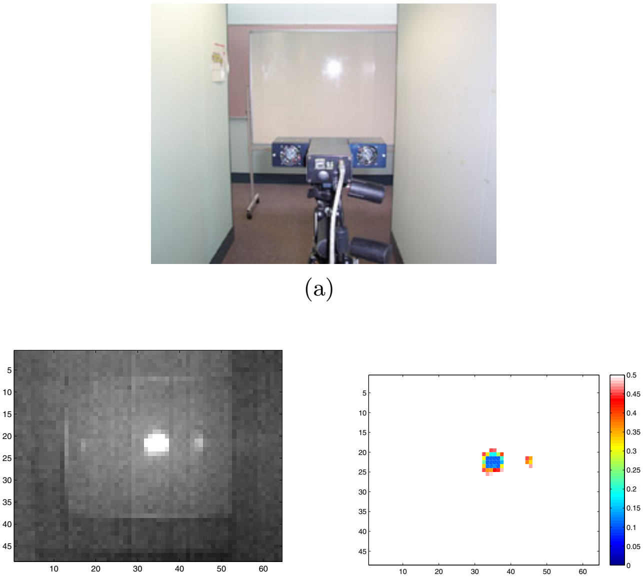
where C_ρ is given by

$$C_\rho := \frac{\rho_a(p)}{\rho_d(p) + \rho_s(p)}. \tag{14}$$

Note that maximum specularity occurs when the angles $(\cos\theta_p, \cos^n \alpha)$ have maximum values resulting in a higher denominator term with respect to numerator irrespective of the scaled constant terms of reflectivity coefficients C_ρ (with decreasing denominator for non-perfect specularity regions). The specularity criterion can be easily formulated without explicit angle estimation. For a specular region, the TOF camera receives sufficient signal and the range data is reliable. As a result the specular radiometric criterion (13) has only one dominating parameter κ_a representing the ratio of ambient offset to TOF IR amplitude. Consequently, specularity can be detected from a $\breve{\kappa}_a(x)$ plot where

$$\text{specularity} = \min \ |\breve{\kappa}_a(x)| \ \ \forall(x) \tag{15}$$

due to high IR amplitude signal and low background offset of intensity. A specular lobe around this point would be indicative of the surface material encoded by n.

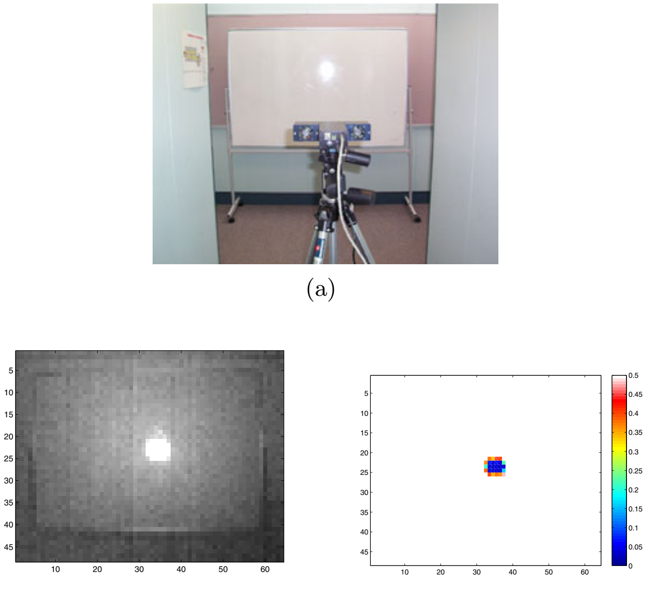


(a)

(b)

(c)

Fig. 2. (a) Picture taken from a normal CCD camera of the experimental setup showing TOF camera and a white board (b) Grayscale intensity image as observed in the TOF camera with specular lobe visible due to IR reflection from the board to camera. (c) Segmentation of specular and non-specular regions based on $\breve{\kappa}_a(x)$ of a frame.



(a)

(b)

(c)

Fig. 3. (a) Picture taken from a normal CCD camera of the experimental setup (b) Intensity image as observed in TOF camera. (c) Segmentation of specular and non-specular regions of a complete frame.

4 Experiments

An indoor environment was chosen for experiments as shown in Figure 2(a) and 3(a) using a white board placed in front of the camera. The board provided sufficient specular reflectance due to its surface material.

In the first case specular reflectance was picked up by the camera. The minimum point of $\breve{\kappa}_a(x)$ space represented the point of maximum specularity with the fall-off forming a lobe of specularity due to the surface material. Since the camera was placed in corridor, a few side reflections from the wall on the board caused a secondary specular lobe as observed in Figure 2(a) and Figure 2(b). These were picked up by the algorithm along with the main specular lobe as illustrated in Figure 2(c). In another experimental setup, as shown in Figure 3(a), the algorithm has picked the specular reflection (see Figure 3(c)) that was observed in Figure 3(b).

5 Conclusion

Unlike conventional cameras where only a single parameter is measured as intensity, TOF camera measures three independent parameters of amplitude, intensity and phase. These measurements along with the illumination conditions of the environment facilitated in deriving a radiometric model for specular highlights in TOF cameras. The proposed framework proved robust and effective for specular highlights detection in imaging algorithms.

References

1. Angelopoulou, E.: Specular highlight detection based on the fresnel reflection coefficient. In: Proc. IEEE 11th International Conference on Computer Vision ICCV 2007, pp. 1–8 (2007)
2. Cook, R.L., Torrance, K.E.: A reflectance model for computer graphics. ACM. Trans. on Graphics 1(1), 7–24 (1982)
3. Foley, J.D., Da, A., Feiner, S.K., Hughes, J.F.: Computer Graphics: Principles and Practices. Addisin-Wesley Publishing Company, Inc., Reading (1997)
4. Forsyth, D.A., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall, Englewood Cliffs (2003)
5. Kahlmann, T., Remondino, F., Guillaume, S.: Range imaging technology: new developments and applications for people identification and tracking. In: Proc. SPIE-IS&T Electronic Imaging, San Jose, CA, USA, vol. 6491 (January 2007)
6. Kuhnert, K.D., Stommel, M.: Fusion of stereo-camera and PMD-camera data for real-time suited precise 3D environment reconstruction. In: Proc. IEEE/RSJ Int. Conf. Intell. Robot. Systs. (2006)
7. Lee, Y.B., You, B.J., Lee, S.W.: A real-time color-based object tracking robust to irregular illumination variations. In: Proc. ICRA Robotics and Automation IEEE Int. Conf., vol. 2, pp. 1659–1664 (2001)
8. Lin, S., Lee, S.W.: Detection of specularity using stereo in color and polarization. In: Proc. 13th International Conference on Pattern Recognition, vol. 1, pp. 263–267 (1996)
9. Lin, S., Lee, S.W.: Using chromaticity distributions and eigenspace analysis for pose-, illumination-, and specularity-invariant recognition of 3d objects. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 426–431 (1997)
10. Ma, Y., Soatto, S., Košecká, J., Sastry, S.S.: An Inviation to 3-D Vision: From Images to Geomtric Models. ch. 2. Springer, Heidelberg (2003)
11. Mufti, F., Mahony, R.: Radiometric range image filtering for time-of-flight cameras. In: Proc. Int. Conf. on Computer Vision Theory and Applications (VISAPP 2010), Angers, France, vol. 1, pp. 143–152 (May 2010)
12. Nehab, D., Weyrich, T., Rusinkiewicz, S.: Dense 3d reconstruction from specularity consistency. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR 2008, pp. 1–8 (2008)
13. Park, J.B., Kak, A.C.: A truncated least squares approach to the detection of specular highlights in color images. In: Proc. IEEE International Conference on Robotics and Automation ICRA 2003, vol. 1, pp. 1397–1403 (2003)
14. Phong, B.T.: Illumination for computer generated pictures. Comm. ACM 18, 311–317 (1975)
15. Sillion, F.X., Puech, C.: Radiosity and Global Illumination. Morgan Kaufmann, San Francisco (1994)

Symmetry Computation in Repetitive Images Using Minimum-Variance Partitions

Manuel Agustí-Melchor, Angel Rodas-Jordá, and José M. Valiente-González

Grupo de Visión por Computador, DISCA, Universitat Politècnica de València
Camino de Vera, s/n 46022 Valencia, Spain
{magusti,arodas,jvalient}@disca.upv.es

Abstract. The symmetry computation has recently been recognized as a topic of interest in many different fields of computer vision and image analysis, which still remains as an open problem. In this work we propose an unified method to compute image symmetries based on finding the minimum-variance partitions of the image that best describe its repetitive nature. We then use a statistical measurement of these partitions as symmetry score. The principal idea is that the same measurement can be used to score symmetries (rotation, reflection, and glide reflection). Finally, a feature vector composed from these symmetry values is used to classify the whole image according to a symmetry group. An increase in the success rate, compared to other reference methods, indicates the improved discriminative capabilities of the proposed symmetry features. Our experimental results improve the state of the art in wallpaper classification methods.

Keywords: Symmetry features, plane symmetry groups, symmetry analysis.

1 Introduction

Images of repetitive patterns are very common in industrial sectors, such as ceramics, textile or graphic arts. They also appear in specific applications, such as architecture designs, medical imaging or geographic analysis. These images are usually composed by patterns or motifs that are repeated in some parts of the image or, in many cases, completely fill the image. In the last case, the images are commonly referred to as *regular mosaics*, *wallpaper images*, or simply *wallpapers*. Some examples of wallpapers obtained from textile collections are shown in Fig. 1.

The study and definition of feature sets that define the structure and contents of such repetitive images brings the possibility of building Content-Based Image Retrieval systems (CBIR), specifically designed for applications such as identifying buildings in photographs, recovering similar designs from textile databases, or dating ancient mosaics.

A wallpaper pattern is a regular tiling made by repetition of a parallelogram shaped subimage or motif, called *Unit Lattice* (UL) or *Unit Tile*. A *symmetry* of this UL can be described through the geometrical transformation that transforms it on itself (isometry). The standard isometries are: displacements (translational symmetry), rotations (n-fold symmetry), reflections (specular symmetry), and glide reflections (specular plus lateral displacement). Depending on the UL image content only certain isometries hold. For

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 204–211, 2011.

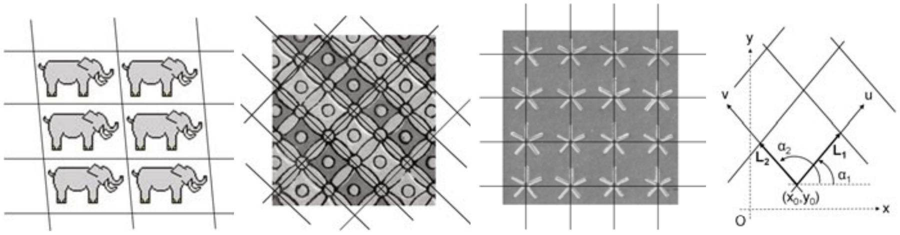


Fig. 1. Wallpaper images obtained from textile collections. The repetitive pattern (lattice) is marked out with a grid that can be located anywhere. (Right) Grid parameters.

example, the pattern in Fig. 1 (left) is only translational. In contrast, the other patterns of Fig. 1 have 180° rotations and reflections. When several isometries are applicable to the pattern, they form a *symmetry group*. The well-known Symmetry Groups Theory (Horne 2000) established that, due to geometric constraints, only a limited number of symmetry groups can be defined. Specifically, in the 2D case there are 17 *Plane Symmetry Groups* (PSG). Figure 2 shows the details of each 17 PSG as well as their standard notation. For example, the patterns in Fig. 1 belong, respectively, to symmetry groups P1, PMM and PM.

The interest in the algorithmic treatment of symmetries has been recognized by a recent tutorial (Liu 2010), which includes an extended discussion and comparison of the state of the art of symmetry detection algorithms. The work of Liu et al. (2004) uses the Mean of Square Differences between original and transformed image to compute symmetries and a *rule-based classifier* (RBC) to classify the images. In a recent work (Agustí et al. 2011) we proposed an alternative Mean of Absolute Differences method and a *Prototype Based Classifier* (PBC) for the same purpose.

In this work we propose an unified method to compute image symmetries based on finding the minimum-variance partitions of the image that best describe its repetitive nature. We then use a statistical measure of these partitions as symmetry score. The main idea is that the same measure can be used to score all symmetries (rotation, reflection, glide reflection).

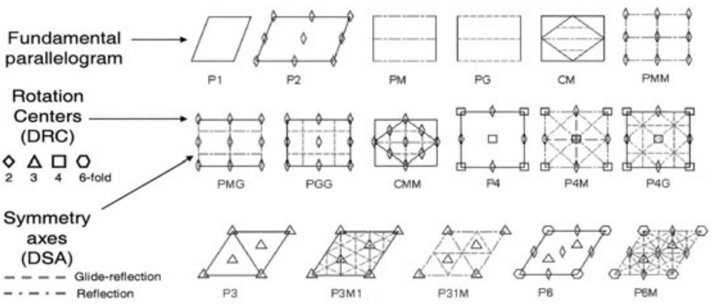


Fig. 2. Representation of the 17 wallpaper groups, their standard notation and their internal symmetries. The UL is referred as the fundamental parallelogram.

2 Calculation of Symmetries through the Image Partition

As indicated before, a wallpaper pattern is generated by the repetition of a parallelogram shaped sub-image (UL) in two directions $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$, which are defined by four parameters $(L_1, \alpha_1, L_2, \alpha_2)$. The geometry of this lattice can be seen as a grid imposed on the pattern (see Fig. 1 (right)). It should be noted that this geometry is translation invariant, which means that it is independent of the starting point used to draw the grid. We can find the repetitiveness of a wallpaper image by dividing or partitioning the image using this regular lattice. The lattice geometry that makes equal all subimages will denote the perfect translational symmetry. On the contrary, a wrong lattice geometry will produce very different lattice subimages. In this work, we introduce a symmetry measure based on the variance of the image partition.

A gray level image $I(x, y)$ can be seen as a set of n points defined by grand mean $\bar{g}$ and total variance S^2. This total variance can be partitioned by decomposing the image into r disjoined groups $P = P_1, P_2, ..., P_r$ of n_i points each $(n_1 + n_2 + ... + n_r = n)$, with mean $\bar{g_i}$ and variance S_i^2. According to the *Law of the Total Variance*:

$$S^2 = \frac{1}{n}\sum_{i=1}^{r} n_i \cdot S_i^2 + \frac{1}{n}\sum_{i=1}^{r} n_i \cdot (\bar{g_i} - \bar{g})^2 \tag{1}$$

If we divide booth terms by the total variance S^2:

$$1 = \frac{\frac{1}{n}\sum_{i=1}^{r} n_i \cdot S_i^2}{S^2} + \frac{\frac{1}{n}\sum_{i=1}^{r} n_i \cdot (\bar{g_i} - \bar{g})^2}{S^2} = FVE + FVU \tag{2}$$

The first term is the *Fraction of Variance Explained* (FVE) statistic, which is a measure of how well the performed partition P predicts the image variability. The complementary term is the *Fraction of Variance Unexplained* (FVU). In presence of translational symmetry, the image partition P that makes similar every subset P_i will also make one of the terms tends to 0 and the other to 1. Moreover, it depends on how the subsets P_i are formed, with two extreme cases:

Continuous point selection: Every point in each UL parallelogram belongs to same subset P_i. As the UL is regularly repeated, every subset P_i has the same mean and variance, and the mean of all variances is similar to the total variance. In this way, the statistic FVE tends to 1 and the FVU tends to 0.

Scattered point selection: Only one point of each UL parallelogram belong to same P_i. The points of P_i are selected at grid steps. In this way the same point of every UL belongs to one subset P_i so, if these points are regularly repeated, their variance will be 0 and the mean of all variances (FVE) tends to 0 and the value FVU to 1.

The first method requires that each subset P_i has the same number of points. Partial parallelograms in image sides have fewer points, so the mean values $\bar{g_i}$ are very different to the rest, biassing the FVE. Therefore, these partial parallelograms must be discharged and the total variance S^2 be re-computed in each case. We prefer the second method, because all image points are considered and each subset P_i may have different number of points. Besides, it is easier to compute the statistic FVU, so we prefer to maximize the factor $FVU = 1 - FVE$, instead of minimising FVE.

However these image partitions only take into account the translational nature of the image. To consider other possible internal symmetries we propose a partition method that incorporates the transformation function T involved in each symmetry.

The proposed partition method performs a two-dimensional division of the whole image into $L_1 x L_2$ subsets, $P^T = \left\{ \bigcup P_{kl}^T \right\}$, $k = 1, \ldots, L_1$ $l = 1, \ldots, L_2$. The points at the same relative position (k, l) inside each UL, after applying the function T, belong to same subset P_{kl}^T. This can be done using non-orthogonal grid coordinate system defined by the lattice direction vectors $\boldsymbol{L_1} = (L_1 \cdot cos\alpha_1, L_1 \cdot sin\alpha_1)$ and $\boldsymbol{L_2} = (L_2 \cdot cos\alpha_2, L_2 \cdot sin\alpha_2)$. The cartesian coordinates of any image point $p = (x_i, y_j)$ can be transformed into grid coordinates (u_i, v_j) through:

$$\begin{pmatrix} u_i \\ v_j \end{pmatrix} = \begin{pmatrix} L_1 cos\alpha_1 & L_2 cos\alpha_2 \\ L_1 sin\alpha_1 & L_2 sin\alpha_2 \end{pmatrix}^{-1} \cdot T \cdot \begin{pmatrix} x_i - x_0 \\ y_j - y_0 \end{pmatrix} \tag{3}$$

where (x_0, y_0) is the grid origin and T is the lattice transformation function. The integer part $Int(u_i, v_j)$ represents the lattice coordinates (grid intersection points), and the fractionary part $Frac(u_i, v_j)$ represents the internal position of each UL parallelogram. If an internal symmetry is held, the contents of this parallelogram is the same after applying the transformation T. In short, a partition of the image P^T can be obtained by accumulating the gray values of each subset P_{kl}^T, through an image scan, as follows:

$$\begin{matrix} I(x_i, y_j) \in P_{kl}^T \\ \forall i \, \forall j \end{matrix} \Leftrightarrow \begin{pmatrix} k \\ l \end{pmatrix} = Frac \left[\begin{pmatrix} L_1 cos\alpha_1 & L_2 cos\alpha_2 \\ L_1 sin\alpha_1 & L_2 sin\alpha_2 \end{pmatrix}^{-1} \cdot T \cdot \begin{pmatrix} x_i - x_0 \\ y_j - y_0 \end{pmatrix} \right] \cdot \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} \tag{4}$$

Table 1 shows the transformation functions T for each of the involved symmetries.

Table 1. Transformation functions T for each of symmetries. (homogeneous coordinates)

<table>
<tr><th>Rotation of angle α</th><th>Reflection about axis L with angle β</th></tr>
<tr><td>$R_\alpha = \begin{pmatrix} cos\alpha & sin\alpha & 0 \\ -sin\alpha & cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}$</td><td>$Re_{L_\beta} = \begin{pmatrix} cos\beta & sin\beta & 0 \\ -sin\beta & cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} cos\beta & -sin\beta & 0 \\ sin\beta & cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix}$</td></tr>
<tr><th colspan="2">Glide reflection about axis L with angle β and displacement d</th></tr>
<tr><td colspan="2">$GRe_{L_{\beta,d}} = \begin{pmatrix} cos\beta & sin\beta & 0 \\ -sin\beta & cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & d \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} cos\beta & -sin\beta & 0 \\ sin\beta & cos\beta & 0 \\ 0 & 0 & 1 \end{pmatrix}$</td></tr>
</table>

Once the subsets P_{kl}^T have been accumulated, and n_{kl} and $\bar{g}_{kl}^T$ values are obtained, the FUV statistic can be computed as follows:

$$FVU(P^T) = \frac{\frac{1}{L_1 \cdot L_2} \sum_{k=1}^{L_1} \sum_{l=1}^{L_2} n_{kl} \cdot (\bar{g}_{kl}^T - \bar{g})2}{S2} \tag{5}$$

Under ideal conditions (e.g. synthetic repetitive images) this measure will be 1, denoting perfect symmetry. In real cases, FVU will be high if the symmetry is present and low otherwise, so we can use it as a symmetry score. Note that the transformation involved in (4) is equivalent to making an image transformation, but gray level interpolation of resulting image points is not needed. In this way, every image points participates, because they are, independent of the transformation considered.

To obtain the translational symmetry we have to find an image partition (grid geometry) that maximises the FVU value. As stated before, this grid geometry is translation invariant, so we can select $(x_0, y_0) = (0, 0)$ as the grid origin. We then perform a search for a maximimum by varying the grid angles (α_1, α_2) and sides (L_1, L_2) and the computing a map of FUV values. The maximum of this map represents the *Translational Symmetry* (TS) of the image. Note that this value will always be high because repeatability is required in this type of image. A low value of TS indicates that the content of the image is not repeated, or is excessively distorted. The four parameters $(L_1, L_2, \alpha_1, \alpha_2)$ at the maximum position indicate the lattice geometry for the image. This is a costly brute-force procedure that can be speed up by optimisation techniques.

As indicated in Fig. 2, the internal symmetries to be computed are: 2-fold, 3-fold, 4-fold and 6-fold rotational symmetries, and reflection and glide-reflection symmetries around sides and diagonals of the UL. As the lattice geometry of the image has already been obtained previously, we now propose a *Extended Partition* (EP), formed by the union of the original translational partition P^I and the image partitions P^T obtained after applying the appropriate transformation T to the lattice geometry. Table 2 shows the proposed partition and the search parameters for each case.

Table 2. Symmetry features and their corresponding partition and transformation function

Symmetry	Name	Partition P^T	T	Parameters	Search space
Translation	TS	P^I	I	$(x_0, y_0) = (0, 0)$	$(L_1, \alpha_1, L_2, \alpha_2)$
2-fold	RS_2	$P^I \cup P^{R_{180}}$	R_{180}	$\alpha = 180°$	$(x_0, y_0) \in$ UL
3-fold	RS_3	$P^I \cup P^{R_{120}}$	R_{120}	$\alpha = 120°$	$(x_0, y_0) \in$ UL
4-fold	RS_4	$P^I \cup P^{R_{90}}$	R_{90}	$\alpha = 90°$	$(x_0, y_0) \in$ UL
6-fold	RS_6	$P^I \cup P^{R_{60}}$	R_{60}	$\alpha = 60°$	$(x_0, y_0) \in$ UL
Reflection side 1	ReS_{L_1}	$P^I \cup P^{Re_{L_1}}$	Re_{L1}	$\beta = \alpha_1$	$(x_0, y_0) \perp L_1$
Reflection side 2	ReS_{L_2}	$P^I \cup P^{Re_{L_2}}$	Re_{L2}	$\beta = \alpha_2$	$(x_0, y_0) \perp L_2$
Reflection diagonal 1	ReS_{D_1}	$P^I \cup P^{Re_{D_1}}$	Re_{D1}	$\beta = \alpha_{D1}$	$(x_0, y_0) \perp D_1$
Reflection diagonal 2	ReS_{D_2}	$P^I \cup P^{Re_{D_2}}$	Re_{D2}	$\beta = \alpha_{D2}$	$(x_0, y_0) \perp D_2$
Glide reflect. side 1	$GReS_{L_1}$	$P^I \cup P^{GRe_{L_1}}$	GRe_{L1}	$\beta = \alpha_1 \; d = L_1/2$	$(x_0, y_0) \perp L_1$
Glide reflect. side 2	$GReS_{L_2}$	$P^I \cup P^{GRe_{L_2}}$	GRe_{L2}	$\beta = \alpha_2 \; d = L_2/2$	$(x_0, y_0) \perp L_2$
Glide reflect. diag. 1	$GReS_{D_1}$	$P^I \cup P^{GRe_{D_1}}$	GRe_{D1}	$\beta = \alpha_{D1} d = D_1/2$	$(x_0, y_0) \perp D_1$
Glide reflect. diag. 2	$GReS_{D_2}$	$P^I \cup P^{GRe_{D_2}}$	GRe_{D2}	$\beta = \alpha_{D2} \; d = D_2/2$	$(x_0, y_0) \perp D_2$

The rotation depends on rotation center, so we make a search moving the grid origin point (x_0, y_0) in the scope of an UL, maintaining the original P^I. The maximum obtained indicates the rotational symmetry score as well as the best position of the rotation center. In all cases, a minimum value is also obtained, indicating a reference value for symmetry 'absence'. These maximum and minimum values will be later used for feature normalization.

The reflection about an axis depends on the axis angle β and position. The angles are known, because they are drawn from the lattice geometry previously obtained. The position is unknown so we again make a search by moving the axis L parallely to angle β, which implies moving the grid origin point (x_0, y_0) in a perpendicular direction to axis L_β.The maximum obtained indicates the reflection symmetry score as well as the best position of the reflection axis. Similarly for the Glide Reflection Symmetries.

As in reference works (Liu et al. 2004 and Agustí et al. 2011), we put everything together into a normalised *Symmetry Feature Vector* of twelve symmetry scores S_i as follows:

$$\begin{aligned} \boldsymbol{SFV} = (&RS_2^n, RS_3^n, RS_4^n, RS_6^n, ReS_{L_1}^n, ReS_{L_2}^n, ReS_{D_1}^n, ReS_{D_2}^n, \\ &GReS_{L_1}^n, GReS_{L_2}^n, GReS_{D_1}^n, GReS_{D_2}^n) \\ S_i^n = &\frac{S_i^{max} - S_i^{min}}{TS - S_i^{min}} \cdot 100 \end{aligned} \tag{6}$$

Table 3 shows the symmetry feature vectors obtained for the three images of Fig. 1. It includes a classical MAD-based (Mean of Absolute Differences) feature vector $\boldsymbol{SFV_{MAD}}$, reported in previous work (Agustí et al. 2011), and the proposed partition-based feature vector $\boldsymbol{SFV}$, both normalized as in (6). Values in bold indicate the symmetries that should be high, according to its symmetry group. You can see how the proposed symmetry features work better than the classic features, due to wider range of values between the presence and absence of symmetry in each sample. In addition, the classic features fail in the third case, while the new ones perfectly describe the symmetries present in that image.

Table 3. Symmetry feature vectors of images in Fig. 1 and the corresponding PSG

MAD symmetry features $\boldsymbol{SFV_{MAD}}$ –												PSG
70.93	43.32	51.45	48.58	69.48	58.32	49.07	45.68	53.72	36.88	48.47	45.58	P1
89.18	17.26	74.91	17.76	**89.00**	**83.31**	13.89	17.00	44.76	45.93	13.70	17.50	PMM
94.69	99.32	96.47	100.0	95.57	**99.23**	97.75	99.03	93.11	96.31	94.25	95.34	PM
Proposed symmetry features $\boldsymbol{SFV}$ –												PSG
69.18	3.78	11.40	4.16	37.13	25.00	6.54	16.92	36.79	26.85	6.54	16.92	P1
98.15	4.05	70.34	4.69	**97.93**	**98.27**	6.01	6.86	62.84	51.89	6.01	6.86	PMM
62.24	9.59	45.35	7.10	60.63	**96.30**	45.07	48.23	42.07	33.92	45.07	48.23	PM

3 Experiments and Results

To establish a comparison between the two sets of symmetry features, classical and proposed, we made an experiment of classifying an image testbed. The performance measurement was the percent of success in the classification (*accuracy*). As a standard image database is not publicly available, we selected several image collections from known websites. We picked image datasets from Wallpaper (2007), Wikipedia (2010), Quadibloc (2010), and SPSU (2009), resulting in a test bed of 218 images. All images were hand-labelled to make the ground truth.

Several classifiers were selected: a Bayes classifier (NaiveBayes), a decision tree (J48), a neural network (Perceptron) and a statistical Nearest Neighbour (NN). Also, two other classifiers specifically adapted for this type of application where chosen: Liu's Rule-Based classifier (RBC) and Agustí's Prototype-Based classifier (PBC). In these cases some threshold values were needed. In the first case to binarise the feature vector and, in the second case, to distinguish symmetry/no_symmetry in the prototypes. A threshold of 80% was experimentally obtained for the MAD features and a pair 80–100 was selected for the prototypes, as indicated in Agusti et al. (2011). Then, the Weka tool was used to make the experiments, with 10-fold cross-validation.

The results are summarized in Table 4. It can be seen that, using the PBC classifier, the classic feature set gets an absolute maximum of 72.94% of success, which grows up to 80.74% with the proposed symmetry features – an improvement of 8%. With respect to the RBC method, an increase from 67.43% to 80,28% is achieved – an improvement of nearly 13 points. Similar behaviour were obtained with the remaining classifiers. As a general result, it can be concluded that the proposed symmetry features have higher capabilities to express the symmetries present in an image that other classic methods. In relation to the classifier, the PBC and RBC are, obviously, the best choices.

Table 4. Classification results for several classifier types using classic and proposed feature sets

– Classifier –	– Classic –	– Proposed –	– Classifier –	– Classic –	– Proposed –
NaiveBayes	55.50%	64.22%	NN	61.93%	63.30%
J48	49.54%	71.10%	RBC	67.43%	80.28%
PERCEPTRON	60.09%	65.14%	PBC	72.94%	80.73%

A final experiment was conducted to explored the behaviour of the proposed feature set in classifying images from the different wallpaper collections. The results are showed in Table 5. These results showed that the proposed features behave very well with images collections composed of very geometric and low-noise images, such as the Wallpaper and Quadibloc collections, and even with collections of intermediate complexity, such as Wikipedia. The SPSU collections behaves poorly, because many of SPSU images are noisy or strongly distorted.

Table 5. RBC and PBC classification results for each image collection (number of samples)

Classifier	Wallpaper (17)	Wikipedia (68)	Quadibloc (47)	SPSU (86)	FULL-SET (218)
RBC	100.0%	75.00%	97.87%	70.93%	80.28%
PBC	100.0%	75.00%	95.75%	73.26%	80.73%

It often happens that the image has wrong aspect ratio, probably due to the acquisition process (e.g. squares become rectangles). In these cases the translational symmetry remains high, but other symmetries decrease or disappear so the original symmetry group (PSG) changes. In other cases, the presence/absence of a certain symmetry is due to small details in the image, so much so that the difference between the original and the transformed image is only a few pixels. In these cases, the computed scores are in the noise level, so that they are not distinguishable from the correct value. But even with these drawbacks, the symmetry features proposed have proved to be more discriminative than other proposals in the literature.

4 Conclusions

This paper had presented a novel framework for computing symmetry features in repetitive images. The indicated symmetries are: n-fold rotation, reflection and glide reflection symmetries. To achieve this, the classical approaches are based on computing the differences between the original and transformed images (MSD or MAD features). In this work, we propose an image partition based on scattered point selection at lattice intervals, which can be achieved with just an image scan. If the symmetry holds the formed sub-sets have minimal intra-group variance, or equivalently maximal inter-group variance. This idea is picked up by the statistic FVU, or *Fraction of Variance Unexplained*, which depends on the image content and the lattice geometry. The key point is that the same statistic can be computed using different image partitions, adapted to the type of symmetry. Finally, a *Symmetry Feature Vector* is composed, joining and normalising twelve symmetry scores.

The performance of the proposed symmetry feature set is evaluated through image classification. The results show the higher discriminative capabilities of the proposed feature set, obtaining an improvement of near 8–13% in success rates with respect to the MAD feature set. The goodness of the specific method is related to its uniform treatment of all symmetries. The badness is due to its parametric configuration, which implies the use of minimisation algorithms. As future work, we are looking for extending the test beds, and propose using this results in recovery tasks (CBIR systems) by computing a list of similarity for every group that can be sorted from highest to lower values and so, for example, detect images that are near to several groups.

References

Agustí, M., Rodas, Á., Valiente, J.M.: Computational Symmetry via prototype distances for symmetry groups classification. In: Conf. on Computer Vision Th. and Applications, pp. 85–93 (2011)

Edwards, S.: Tiling Plane & Fancy (2009), http://www2.spsu.edu/math/tile/index.htm

Horne, C.: Geometric Symmetry in Patterns and Tilings. Woodhead Publishing, Abington Hall, England (2000)

Joyce, D.E.: Wallpaper Groups Plane Symmetry Groups, http://www.clarku.edu/~djoyce/ (last visited January 2011)

Liu, Y., Collins, R.T., Tsin, Y.: A computational model for periodic pattern perception based on frieze and wallpaper groups. IEEE Trans. on PAMI 26(3), 354–371 (2004)

Liu, Y., Hel-Or, H., Kaplan, C.S., Van Gool, L.: Computational Symmetry in Computer Vision and Graphics. Foundations and Trends in Computer Graphics and Vision 5(1-2), 1–195 (2010)

Savard, J.G.: Basic tilings: The 17 wallpaper groups, http://www.quadibloc.com/math/tilint.htm (last visited January 2011)

Wikipedia: Wallpaper group, http://www.wikipedia.org (last visited January 2011)

Tensor Method for Constructing 3D Moment Invariants

Tomáš Suk and Jan Flusser

Institute of Information Theory and Automation of the ASCR
{suk,flusser}@utia.cas.cz

Abstract. A generalization from 2D to 3D of the tensor method for derivation of both affine invariants and rotation, translation and scaling (TRS) invariants is described. The method for generation of the 3D TRS invariants of higher orders is automated and experimentally tested.

Keywords: Recognition, tensor, moment, rotation, invariant, 3D.

1 Introduction

Pattern recognition of objects in two-dimensional (2D) images has been an important part of image analysis for many years. The images are often geometrically distorted; the distortion of a flat scene can be modeled as a combination of translation, rotation and scaling (TRS) in the case of a scanning device parallel to the scene and as a projective transformation in the opposite case.

An efficient approach to the recognition of deformed objects is using certain features that do not vary in the transformation; we call them *invariants*. Thus, TRS invariants can be applied to the recognition of objects distorted by TRS, and projective invariants to the recognition of objects distorted by the projective transform. Since it is difficult to derive global projective invariants describing an entire object, the projective transform is often approximated by an affine transformation. If the distance of the scanned scene and the camera is large, this approximation is accurate enough.

Recently, the scanning devices of 3D objects (computer tomography (CT), magnetic resonance imaging (MRI), rangefinders, etc.) become more and more affordable, which arises the need of recognition of 3D objects and thus the need of having 3D invariants. One of the most popular family of 3D invariants is based on image moments. While moment invariants in 2D have been studied extensively for decades (see [3] for a survey and [7] for the latest results), the theory of 3D moment invariants has not been fully explored. The first attempts to derive 3D rotation moment invariants are relatively old (see e.g. [6], [4]), but the generalization to higher moment orders has not been reported as it is rather complicated.

The topic of this paper is generalization of the tensor method for deriving rotation moment invariants in 3D. The main contribution is not only finding a closed-form solution for 3D rotation and an affine invariant but principally an algorithm, which generates all invariants up to a given order.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 212–219, 2011.

2 Moments and Tensors

Geometric moments of image f (2D or 3D) are defined as

$$\begin{aligned} m_{pq} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q f(x,y)\,\mathrm{d}x\,\mathrm{d}y, \\ m_{pqr} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p y^q z^r f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z. \end{aligned} \tag{1}$$

The sum of the indices is called the *order* of the moment. To provide the translation invariance, we often use the *central geometric moments*

$$\mu_{pqr} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-x_c)^p (y-y_c)^q (z-z_c)^r f(x,y,z)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z, \tag{2}$$

where $x_c = m_{100}/m_{000}$, $y_c = m_{010}/m_{000}$ and $z_c = m_{001}/m_{000}$ are centroid coordinates of the image $f(x,y,z)$.

We can define a *moment tensor* [1] for using tensor calculus for derivation of moment invariants

$$M^{i_1 i_2 \cdots i_k} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^{i_1} x^{i_2} \cdots x^{i_k} f(x^1,x^2,x^3)\,\mathrm{d}x^1\,\mathrm{d}x^2\,\mathrm{d}x^3, \tag{3}$$

where $x^1 = x$, $x^2 = y$ and $x^3 = z$. If p indices equal 1, q indices equal 2 and r indices equal 3, then $M^{i_1 i_2 \cdots i_k} = m_{pqr}$. The definition in 2D is analogous. The behavior of the moment tensor under an affine transform (in Einstein notation[1]) is

$$\begin{aligned} M^{i_1 i_2 \cdots i_r} &= |J| p^{i_1}_{\alpha_1} p^{i_2}_{\alpha_2} \cdots p^{i_r}_{\alpha_r} \hat{M}^{\alpha_1 \alpha_2 \cdots \alpha_r} \\ \hat{M}^{i_1 i_2 \cdots i_r} &= |J|^{-1} q^{i_1}_{\alpha_1} q^{i_2}_{\alpha_2} \cdots q^{i_r}_{\alpha_r} M^{\alpha_1 \alpha_2 \cdots \alpha_r}, \end{aligned} \tag{4}$$

where p^i_α is the matrix of the direct affine transform and q^i_α is the matrix of the inverse affine transform (without translation). This means that the moment tensor is a relative contravariant tensor with the weight -1.

2.1 Affine Invariants in 2D and in 3D

The tensor method for affine invariants in 2D is described e.g. in [5]: we arrange our measurements into a tensor, multiply the tensors so the number of contravariant indices equals the number of covariant indices and compute the total contraction of the product. Since the moment tensor is purely contravariant, we need to multiply them by some covariant tensors. In such a case, we can use so-called unit polyvectors.

[1] A. Einstein introduced this notation to simplify expressions with n-dimensional coordinates, see e.g. [9] for explanation.

The *unit polyvector* is an antisymmetric tensor over all indices and the component with indices $1, 2, \ldots, n$ equals 1. It can be both covariant, i.e. $\epsilon_{12\cdots n} = 1$ and contravariant, i.e. $\epsilon^{12\cdots n} = 1$. The term *antisymmetric* means that the tensor component changes its sign and preserves its magnitude when interchanging two arbitrary indices. In 2D, it means that (in matrix notation except for the unit polyvector is not multiplied like a matrix)

$$\varepsilon_{i_1 i_2} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{5}$$

Then a proper tensor product can be used for derivation of a relative affine invariant, e.g.

$$M^{ij} M^{kl} \epsilon_{ik} \epsilon_{jl} = 2(m_{20} m_{02} - m_{11}^2).$$

After modification, we obtain an absolute affine invariant

$$I_1^{2D} = (\mu_{20}\mu_{02} - \mu_{11}^2)/\mu_{00}^4.$$

The exponent $\omega = 4$ of μ_{00} equals the number of factors in the tensor product, i.e. the moment tensors and the unit polyvectors. Other example

$$M^{ijk} M^{lmn} M^{opq} M^{rst} \epsilon_{il} \epsilon_{jm} \epsilon_{ko} \epsilon_{lr} \epsilon_{ps} \epsilon_{qt} :$$
$$I_2^{2D} = (-\mu_{30}^2\mu_{03}^2 + 6\mu_{30}\mu_{21}\mu_{12}\mu_{03} - 4\mu_{30}\mu_{12}^3 - 4\mu_{21}^3\mu_{03} + 3\mu_{21}^2\mu_{12}^2)/\mu_{00}^{10}.$$

A question how to generate all relevant tensor products arises. We can employ an idea, that the tensor products can be described by graphs and then generation of all tensor products means generation of all graphs with the corresponding number of nodes and edges; this is a so-called graph method [3].

In 3D, the unit polyvector looks like

$$\varepsilon_{i_1 i_2 1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \ \varepsilon_{i_1 i_2 2} = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \ \varepsilon_{i_1 i_2 3} = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{6}$$

There are two differences between 2D and 3D; the unit polyvector has three indices, i.e. we need fewer factors in the tensor product, and each index can have three values 1, 2 and 3. In the sense of Einstein notation, we sum over these three values. Some examples

$$M^{ij} M^{kl} M^{mn} \epsilon_{ikm} \epsilon_{jln} :$$
$$I_1^{3D} = (\mu_{200}\mu_{020}\mu_{002} + 2\mu_{110}\mu_{101}\mu_{011} - \mu_{200}\mu_{011}^2 - \mu_{020}\mu_{101}^2 - \mu_{002}\mu_{110}^2)/\mu_{000}^5.$$

and

$$M^{ijk} M^{lmn} M^{opq} M^{rst} \epsilon_{ilo} \epsilon_{jmr} \epsilon_{kps} \epsilon_{nqt} :$$

$$\begin{aligned}
I_2^{3D} = (&\mu_{300}\mu_{003}\mu_{120}\mu_{021} + \mu_{300}\mu_{030}\mu_{102}\mu_{012} + \mu_{030}\mu_{003}\mu_{210}\mu_{201} - \mu_{300}\mu_{120}\mu_{012}^2 \\
&-\mu_{300}\mu_{102}\mu_{021}^2 - \mu_{030}\mu_{210}\mu_{102}^2 - \mu_{030}\mu_{201}^2\mu_{012} - \mu_{003}\mu_{210}^2\mu_{021} - \mu_{003}\mu_{201}\mu_{120}^2 \\
&-\mu_{300}\mu_{030}\mu_{003}\mu_{111} + \mu_{300}\mu_{021}\mu_{012}\mu_{111} + \mu_{030}\mu_{201}\mu_{102}\mu_{111} + \mu_{003}\mu_{210}\mu_{120}\mu_{111} \\
&+\mu_{210}^2\mu_{012}^2 + \mu_{201}^2\mu_{021}^2 + \mu_{120}^2\mu_{102}^2 - \mu_{210}\mu_{120}\mu_{102}\mu_{012} - \mu_{210}\mu_{201}\mu_{021}\mu_{012} \\
&-\mu_{201}\mu_{120}\mu_{102}\mu_{021} - 2\mu_{210}\mu_{012}\mu_{111}^2 - 2\mu_{201}\mu_{021}\mu_{111}^2 - 2\mu_{120}\mu_{102}\mu_{111}^2 \\
&+3\mu_{210}\mu_{102}\mu_{021}\mu_{111} + 3\mu_{201}\mu_{120}\mu_{012}\mu_{111} + \mu_{111}^4)/\mu_{000}^8.
\end{aligned}$$

The idea of the graphs is more complicated in 3D, we cannot use the ordinary graphs, we need so-called three-uniform hypergraphs, where each hyperedge connects three nodes.

2.2 Rotation Invariants in 2D and in 3D

When the transformation in question is not a full affine transform, but is a mere TRS, then a slightly different approach is suitable to yield simpler invariants. So-called *Cartesian tensors* are suitable for derivation of the rotational invariants. The ordinary tensor behaves in the affine transformation according to the rule

$$\hat{T}^{\beta_1,\beta_2,\ldots,\beta_{k_2}}_{\alpha_1,\alpha_2,\ldots,\alpha_{k_1}} = q^{\beta_1}_{j_1} q^{\beta_2}_{j_2} \cdots q^{\beta_{k_2}}_{j_{k_2}} p^{i_1}_{\alpha_1} p^{i_2}_{\alpha_2} \cdots p^{i_{k_1}}_{\alpha_{k_1}} T^{j_1,j_2,\ldots,j_{k_2}}_{i_1,i_2,\ldots,i_{k_1}}, \tag{7}$$

while the Cartesian tensor behaves in the rotation according to the rule

$$\hat{T}_{\alpha_1,\alpha_2,\ldots,\alpha_k} = r_{\alpha_1 i_1} r_{\alpha_2 i_2} \cdots r_{\alpha_k i_k} T_{i_1,i_2,\ldots,i_k}, \tag{8}$$

where r_{ij} is an arbitrary orthonormal matrix. The distinction of the covariant and contravariant tensors has no meaning in the case of the Cartesian tensors and we can perform the total contraction of the moment product without the unit polyvectors.

The simplest example is the total contraction of the second-order moment tensor M_{ii}. In 2D

$$M_{11} + M_{22} : \Phi_1^{2D} = (\mu_{20} + \mu_{02})/\mu_{00}^2,$$

in 3D

$$M_{11} + M_{22} + M_{33} : \Phi_1^{3D} = (\mu_{200} + \mu_{020} + \mu_{002})/\mu_{000}^{5/3}.$$

Another example is $M_{ij}M_{ij}$. In 2D

$$\Phi_2^{2D} = (\mu_{20}^2 + \mu_{02}^2 + 2\mu_{11}^2)/\mu_{00}^4,$$

in 3D

$$\Phi_2^{3D} = (\mu_{200}^2 + \mu_{020}^2 + \mu_{002}^2 + 2\mu_{110}^2 + 2\mu_{101}^2 + 2\mu_{011}^2)/\mu_{000}^{10/3}.$$

The last example of the second order is $M_{ij}M_{jk}M_{ki}$. In 2D it is

$$\Phi_3^{2D} = (\mu_{20}^3 + 3\mu_{20}\mu_{11}^2 + 3\mu_{11}^2\mu_{02} + \mu_{02}^3)/\mu_{00}^6,$$

while in 3D it becomes

$$\begin{aligned}\Phi_3^{3D} = (&\mu_{200}^3 + 3\mu_{200}\mu_{110}^2 + 3\mu_{200}\mu_{101}^2 + 3\mu_{110}^2\mu_{020} + 3\mu_{101}^2\mu_{002} + \mu_{020}^3 \\ &+3\mu_{020}\mu_{011}^2 + 3\mu_{011}^2\mu_{002} + \mu_{002}^3 + 6\mu_{110}\mu_{101}\mu_{011})/\mu_{000}^5.\end{aligned}$$

The scaling normalization is slightly more difficult; the exponent is

$$\omega = \frac{w}{n} + s, \tag{9}$$

where w is the sum of the indices of all moments in one term, s is the number of the moments in one term and n is the dimension.

2.3 Automated Generation of the Rotation Invariants in 3D

We recall the idea of using graphs for constructing invariants. In the case of a 3D rotation, the ordinary graphs, where each edge connects two nodes, are sufficient. These graphs can include self-loops, see e.g. Fig. 1a.

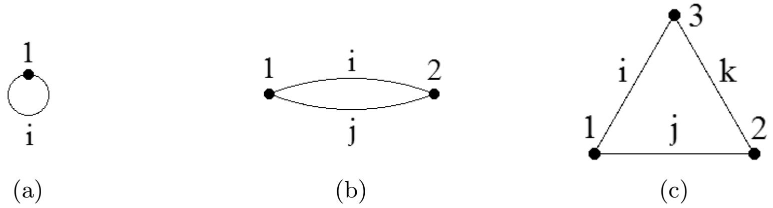


Fig. 1. The generating graphs of (a) both Φ_1^{2D} and Φ_1^{3D}, (b) both Φ_2^{2D} and Φ_2^{3D} and (c) both Φ_3^{2D} and Φ_3^{3D}

An important part of this process is an elimination of the linearly dependent invariants, i.e. zeros, identical invariants, products and linear combinations. The invariants remaining after this elimination are called *irreducible*; those which were eliminated are called *reducible* invariants. It should be noted that irreducibility does not mean independence. There may be polynomial dependencies among irreducible invariants which are not discovered in the elimination algorithms. As will be shown, there must be a large number of them but their identification is an extremely complex problem even in 2D.

The method of finding all irreducible invariants we propose here is a generalization of our method for 2D case [3]. We first generate all possible invariants (graphs) and then, by exhaustive search, eliminate the reducible ones. Everything is carried out on symbolic level, independent of any particular image data. As a result of this algorithm, we obtain closed-form expressions for all irreducible invariants along with automatically generated data for their calculation. Taking into account the computing complexity on the one hand and the capability of our computers on the other, the maximum number of graph edges which are feasible to construct is 8. Table 1 summarizes their numbers.

The first row of the table shows the orders of the invariants, the second row contains the cumulative number of the irreducible invariants up to the given order which were actually constructed by the proposed algorithm (note that there is no guarantee that all existing irreducible invariants were found because of the limitation to maximum of 8 graph edges), and the third row contains the theoretical maximum number of the independent invariants. The number of the independent invariants was estimated as a difference between the number of moments and the number of degrees of freedom of the transformation, i.e. $\binom{t+3}{3} - 7$ up to the order t in our case. As previously mentioned, currently we are not able to systematically find these independent invariants (i.e. to identify polynomial dependencies among irreducible invariants). This extremely difficult task will be a subject of future research.

Table 1. The numbers of the 3D irreducible and independent rotation invariants

Order	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
irred.	3	42	242	583	840	1011	1098	1142	1164	1174	1180	1182	1184	1184	1185
indep.	3	13	28	49	77	113	158	213	279	357	448	553	673	809	962

We generated explicit forms of all 1185 invariants, they are available on our website [2]. However, the corresponding pdf file contains more than 10 000 pages.

3 Numerical Experiment

The following simple experiment verifies that the constructed irreducible invariants are actually invariant to rotation, i.e. they were derived correctly. Moreover, it demonstrates that two similar but different objects have different values of (at least some) invariants.

The experiment was carried out on real data. We used two ancient Greek amphoras scanned using a laser rangefinder from various sides. Consequently all measurements were combined to obtain a 3D binary image of the amphora. Since the rangefinder cannot get inside the amphora, it is considered filled up and closed on the top. To compress the data, the surface of the amphora was divided into small triangles (42 400 and 23 738 triangles, respectively). The amphoras were then represented only by their triangulated surfaces. The test data are shown in Fig. 2. The photographs are for illustration only, no graylevel/color information was used in moment calculation.

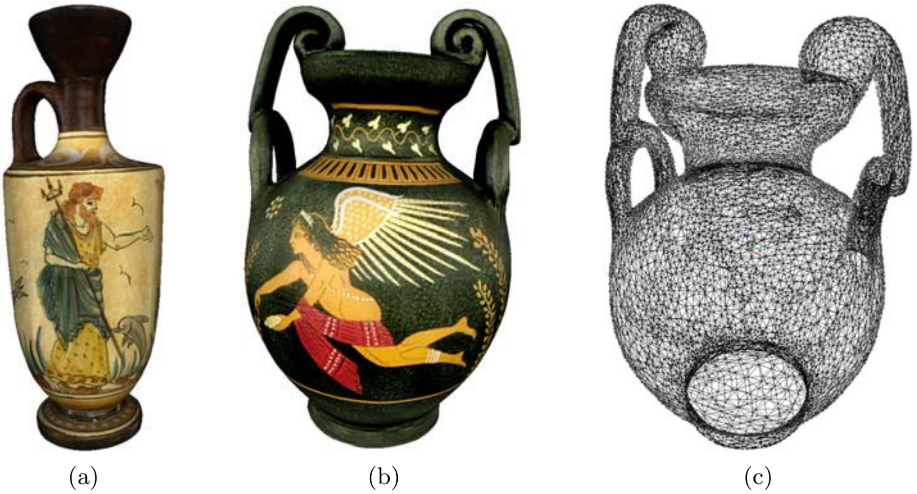


Fig. 2. The amphoras: (a) photo of A1, (b) photo of A2, (c) wire model of the triangulation of A2

For each amphora we generated 10 random rotations and translations of its triangular representation[2] and calculated the values of the first 242 invariants up to the 4th order. Here we have two possibilities as to what moments to use for computing invariants – we can employ either traditional 3D volume moments or *surface moments* [10]. Surface moments are calculated by double integration over the object surface only. We used both approaches.

The maximum relative standard deviation was $3.2 \cdot 10^{-13}$ in the case of invariants computed from the volume moments and $4.5 \cdot 10^{-13}$ for the invariants from the surface moments, which illustrates a perfect invariance in all cases. On the other hand, the maximum relative deviations between two different amphoras A1 and A2 were 1.01 for volume invariants and 1.55 for surface invariants, which proves discriminability – different objects have distinct values of the invariants.

3.1 Computing Moments of Triangulated Objects

Now we explain how the 3D moments (both volume and surface) were actually calculated in this experiment. It would be of course possible to calculate them from definition but since we already have the triangular representation, the moments can be calculated in a more efficient way [8].

We complete each triangle to a tetrahedron such that the new vertex coincides with the coordinate origin. The triangles must have the same orientation with respect to the object, e.g. counterclockwise when seeing from outside to inside of the object. The object volume is then divided into these disjoint tetrahedrons and the volume moment is given as a sum of moments of all tetrahedrons. Using the vertices of the triangulation only, the volume moment is calculated as

$$m_{pqr} = \frac{p!q!r!}{(p+q+r+n)!} \sum_{(k_{ij})\in\mathcal{K}} \frac{\prod_{j=1}^{3}((\sum_{i=1}^{3} k_{ij})!)}{\prod_{i,j=1}^{3}(k_{ij}!)} \sum_{\ell=1}^{N} A_\ell \prod_{i,j=1}^{3} \left(a_{ij}^{(\ell)}\right)^{k_{ij}}, \qquad (10)$$

where N is the number of the triangles, $\mathcal{K}$ is a set of such 3×3 matrices k_{ij} with non-negative integer values that $\sum_{j=1}^{3} k_{1j} = p$, $\sum_{j=1}^{3} k_{2j} = q$ and $\sum_{j=1}^{3} k_{3j} = r$, $a_{ij}^{(\ell)}$ is a matrix of the vertex coordinates of the ℓ-th triangle, i is the number of the coordinate and j is the number of the vertex. $A_\ell = \det\left(a_{ij}^{(\ell)}\right)$, i.e. it is a 6-multiple of the oriented tetrahedron volume and the dimension $n = 3$.

In the case of the surface moments the formula is basically the same except for $A_\ell = \left\|\left(a_{i2}^{(\ell)} - a_{i1}^{(\ell)}\right) \times \left(a_{i3}^{(\ell)} - a_{i1}^{(\ell)}\right)\right\|$ is twice the oriented area of the triangle. The surface moments have different scaling normalization, the dimension $n = 2$ in (9) and in (10).

[2] A more correct way would be to rotate the amphora physically in the capturing device and scan it again in each position. However, this would be extremely costly and the results would be comparable.

4 Conclusion

We have proposed and implemented a tensor method for generation of 3D rotation moment invariants of arbitrary orders. We tested this method on invariants up to the order 16. We constructed 1185 irreducible invariants, a vast majority of them being published for the first time. Our method includes elimination of linearly dependent invariants, but for now does not contain identification of polynomial dependencies among the invariants.

Acknowledgments. Thanks to the grant No. P103/11/1552 of the Czech Science Foundation for financial support. Thanks to Clepsydra (ΚΛΕΨΥΔΡΑ), The Digitization Center of Cultural Heritage in Xanthi, Greece (`http://clepsydra.ipet.gr`) and especially to prof. Christodoulos Chamzas for the models of amphoras.

References

1. Cyganski, D., Orr, J.A.: Object recognition and orientation determination by tensor methods. In: Huang, T.S. (ed.) Advances in Computer Vision and Image Processing, pp. 101–144. JAI Press, Greenwich (1988)
2. Department of Image Processing: 3D rotation moment invariants, `http://zoi.utia.cas.cz/3DRotationInvariants`
3. Flusser, J., Suk, T., Zitová, B.: Moments and Moment Invariants in Pattern Recognition. Wiley, Chichester (2009)
4. Lo, C.H., Don, H.S.: 3-D moment forms: Their construction and application to object identification and positioning. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(10), 1053–1064 (1989)
5. Reiss, T.H.: Recognizing Planar Objects Using Invariant Image Features. LNCS, vol. 676. Springer, Heidelberg (1993)
6. Sadjadi, F.A., Hall, E.L.: Three dimensional moment invariants. IEEE Transactions on Pattern Analysis and Machine Intelligence 2(2), 127–136 (1980)
7. Suk, T., Flusser, J.: Affine moment invariants generated by graph method. Pattern Recognition 44(9), 2047–2056 (2011)
8. Tuzikov, A.V., Sheynin, S.A., Vasiliev, P.V.: Efficient computation of body moments. In: Skarbek, W. (ed.) CAIP 2001. LNCS, vol. 2124, pp. 201–208. Springer, Heidelberg (2001)
9. Wikipedia: Einstein notation, `http://en.wikipedia.org/wiki/Einsteinnotation`
10. Xu, D., Li, H.: 3-D surface moment invariants. In: Proceedings of the 18th International Conference on Pattern Recognition ICPR 2006, pp. 173–176. IEEE Computer Society, Los Alamitos (2006)

Multi-camera 3D Scanning with a Non-rigid and Space-Time Depth Super-Resolution Capability

Karima Ouji[1], Mohsen Ardabilian[1], Liming Chen[1], and Faouzi Ghorbel[2]

[1] LIRIS, Lyon Research Center for Images and Intelligent Information Systems,
Ecole Centrale de Lyon. 36, av. Guy de Collongue, 69134 Ecully, France
{karima.ouji,mohsen.ardabilian,liming.chen}@ec-lyon.fr

[2] GRIFT, Groupe de Recherche en Images et Formes de Tunisie, Ecole Nationale
des Sciences de l'Informatique, Tunisie
faouzi.ghorbel@ensi.rnu.tn

Abstract. 3D imaging sensors for the acquisition of three dimensional faces have created, in recent years, a considerable degree of interest for a number of applications. Structured light camera/projector systems are often used to overcome the relatively uniform appearance of skin. In this paper, we propose a 3D acquisition solution with a 3D space-time non-rigid super-resolution capability, using three calibrated cameras coupled with a non calibrated projector device, which is particularly suited to 3D face scanning, i.e. rapid, easily movable and robust to ambient lighting conditions. The proposed solution is a hybrid stereovision and phase-shifting approach, using two shifted patterns and a texture image, which not only takes advantage of the assets of stereovision and structured light but also overcomes their weaknesses. The super-resolution process is performed to deal with 3D artifacts and to complete the 3D scanned view in the presence of small non-rigid deformations as facial expressions. The experimental results demonstrate the effectiveness of the proposed approach.

Keywords: Stereovision, Phase-shifting, Space-time, Multi-camera, Super-resolution, Non-rigid matching, 3D frames.

1 Introduction

Real-time 3D imaging sensors for the acquisition of three dimensional faces have created, in recent years, a considerable degree of interest for a wide range of applications, including biometry, facial animation and aesthetic surgery. Structured light camera/projector systems are often used to overcome the relatively uniform appearance of skin. These systems require explicit user cooperation and controlled lighting conditions [1,2]. Depth information is recovered by decoding patterns of a projected structured light which include gray codes, sinusoidal fringes, etc. Current solutions mostly utilize more than three phase-shifted sinusoidal patterns to recover the depth information, thus impacting the acquisition delay; they further require projector-camera calibration whose accuracy is crucial for phase to depth estimation step; and finally, they also need an unwrapping

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 220–228, 2011.

stage which is sensitive to ambient light, especially when the number of patterns decreases [3]. An alternative to projector-camera systems consists of recovering depth information by stereovision using a multi-camera system as proposed in [2,4]. A stereo matching step finds correspondence between stereo images and the 3D information is obtained by optical triangulation [2,5]. However, the model computed in this way generally is quite sparse. To upsample and denoise depth images, researchers looked into super-resolution techniques. Kil et al. [9] applied super-resolution for laser triangulation scanners by regular resampling from aligned scan points with associated gaussian location uncertainty. Super-resolution was especially proposed for time-of-flight cameras which have very low data quality and a very high random noise by solving an energy minimization problem [10].

In this paper, we propose a 3D acquisition solution with a 3D space-time and non-rigid super-resolution capability, using three calibrated cameras coupled with a non calibrated projector device, which is particularly suited to 3D face scanning, i.e. rapid, easily movable and robust to ambient lighting conditions. The proposed solution is a hybrid stereovision and phase-shifting approach which not only takes advantage of the assets of stereovision and structured light but also overcomes their weaknesses. According to our method, first an automatic primitives sampling is performed from stereo-matching to provide a 3D facial sparse model with a fringe-based resolution and a subpixel precision. Second, an intra-fringe phase estimation densify the 3D sparse model using the two sinusoidal fringe images and a texture image, independently from the left, middle and right cameras. The left, middle and right 3D dense models are merged to produce the final 3D model which constitutes a spatial super-resolution.

Also, we propose to carry out a temporal super-resolution process which considers the facial deformable aspect. The temporal super-resolution corrects the 3D information and completes the 3D scanned view. In contrast to conventional methods, our method is less affected by the ambient light thanks to the use of stereo in the first stage of the approach, replacing the phase unwrapping stage. Also, it does not require a camera-projector off-line calibration which constitutes a tedious and expensive task. Moreover, our approach is applied only to the region of interest which decreases the whole processing time. Section (2) details the primitives sampling to generate the 3D sparse model. In Section(3), we highlight the spatial super-resolution from the three calibrated cameras. Section(4) explains how the 3D non-rigid temporal super-resolution is carried out. Section (5) discusses the experimental results and section (6) concludes the paper.

2 Primitives Sampling for 3D Sparse Model Generation

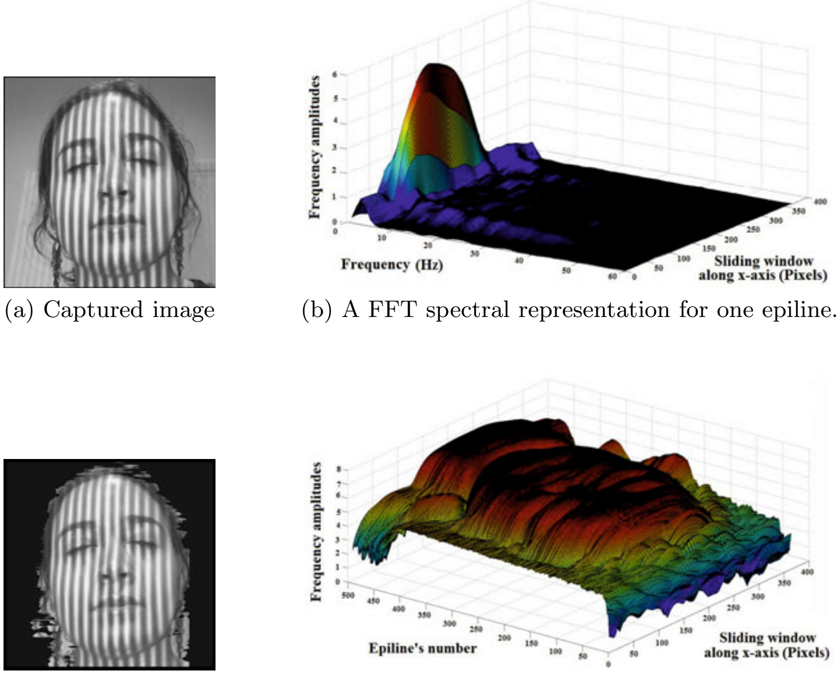
First, an offline strong stereo calibration computes the intrinsic and extrinsic parameters of the cameras, estimates the tangential and radial distortion parameters, and provides the epipolar geometry as proposed in [8]. In online process, two π-shifted sinusoid patterns and a third white pattern are projected onto the face. Three sets of left, middle and right images are captured, undistorted and rectified. The proposed model is defined by the system of equations (1). It constitutes a variant of the mathematic model proposed in [3].

$$\begin{aligned} I_p(s,t) &= I_b(s,t) + I_a(s,t) \cdot \sin(\phi(s,t)), \\ I_n(s,t) &= I_b(s,t) + I_a(s,t) \cdot \sin(\phi(s,t)+\pi), \\ I_t(s,t) &= I_b(s,t) + I_a(s,t). \end{aligned} \tag{1}$$

At time t , $I_p(s,t)$, $I_n(s,t)$, $I_t(s,t)$ constitute the intensity term of the pixel s on respectively the positive image, the negative one and the texture one. $I_b(s,t)$ represents the texture information and the lighting effect. $\phi(s,t)$ is the local phase defined at each pixel s. Solving (1), $I_b(s,t)$ is computed as the average intensity of $I_p(s,t)$ and $I_n(s,t)$. $I_a(s,t)$ is then computed from the third equation of the system (1) and $\phi(s,t)$ is estimated by equation (2).

$$\phi(s,t) = \arcsin\left[\frac{I_p(s,t)-I_n(s,t)}{2\cdot I_t(s,t)-I_p(s,t)-I_n(s,t)}\right]. \tag{2}$$

Also, we suggest an automatic region-of interest localization to decrease the whole processing time. The idea is to benefit from the contrast variation and carry out a spectral analysis to localize the low frequencies on captured images.



(a) Captured image

(b) A FFT spectral representation for one epiline.

(c) Segmented image

(d) 2D facial region segmented by a FFT spectral analysis.

Fig. 1. Pattern-based face localization

First, we compute FFT on a sliding window for each epiline which provides for each pixel a 2D curve of FFT frequency amplitudes. A 3D spectral distribution is obtained which highlights the facial region for the current epiline as shown in figure 1.b. We propose to keep only pixels belonging to this highlighted region. Thus, for each pixel in the epiline, we consider a weighted sum of only the low-frequency amplitudes and we apply an adequate thresholding to obtain the region-of-interest as illustrated by figure 1.d.

Finally, the sparse 3D model is generated through a stereovision scenario. It is formed by the primitives situated on the fringe change-over which is the intersection of the sinusoidal component of the positive image and the second π-shifted sinusoidal component of the negative one [5]. Therefore, the primitives localization has a sub-pixel precision. Corresponding multi-camera primitives necessarily have the same Y-coordinate in the rectified images. Thus, stereo matching problem is resolved in each epiline separately using Dynamic Programming. The 3D sparse point cloud is then recovered by computing the intersection of optical rays coming from the pair of matched features. When projecting vertical fringes, the video projector can be considered as a vertical adjacent sources of light. Such a consideration provides for each epiline a light source point O_{Prj} situated on the corresponding epipolar plane. The sparse 3D model is a serie of adjacent 3D vertical curves obtained by the fringes intersection of the positive and the negative images. Each curve describes the profile of a projected vertical fringe distorted on the 3D facial surface. We propose to estimate the 3D plane containing each distorted 3D curve separately. As a result, the light source vertical axis of the projector is defined as the intersection of all the computed 3D planes. This estimation can be performed either as an offline or online process unlike conventional phase-shifting approaches where the projector is calibrated on offline and cannot change its position when scanning the object.

3 3D Multi-camera Spatial Super-Resolution

Here, the idea is to find the 3D coordinates for each pixel situated between two successive fringes in either left, middle or right camera images to participate separately on the 3D model elaboration. Therefore, we obtain a 3D point cloud from each camera set of images. The spatial super-resolution consists of merging the left, middle and right 3D point clouds. The 3D coordinates of each pixel are computed using phase-shifting analysis. Conventional phase-shifting techniques estimates the local phase in $[0..2\pi]$ for each pixel on the captured image. Local phases are defined as wrapped phases. Absolute phases are obtained by phase unwrapping. In the proposed approach, the sparse model lets us retrieve 3D intra-fringe information from wrapped phases directly. In fact, each point P_i in the sparse model constitutes a reference point for all pixels situated between P_i and its next neighbor P_{i+1} on the same epiline of the sparse model. For a pixel P_k situated between $P_i(X_i, Y_i, Z_i)$ and $P_{i+1}(X_{i+1}, Y_{i+1}, Z_{i+1})$, we compute its local phase value ϕ_k using equation (2). The phase value of P_i is $\phi_i = 0$ and the phase value of P_{i+1} is $\phi_{i+1} = \pi$.

The phase ϕ_k which belongs to $[0..\pi]$ has monotonous variation if $[P_iP_{i+1}]$ constitutes a straight line on the 3D model. When $[P_iP_{i+1}]$ represents a curve on the 3D model, the function ϕ_k describes the depth variation inside $[P_iP_{i+1}]$. Therefore, the 3D coordinates $(X(\phi_k), Y(\phi_k), Z(\phi_k))$ of P_k corresponding to the pixel point G_k are computed by a geometric reconstruction as shown in figure 2.

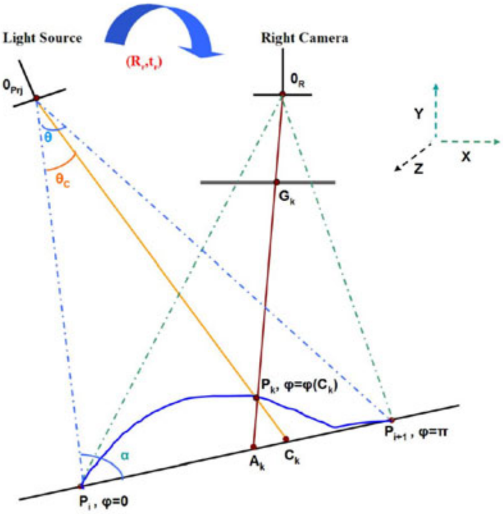


Fig. 2. Intra-fringe 3D information retrieval scheme

The 3D intra-fringe coordinates computation is carried out for each epiline i separately. An epipolar plane is defined for each epiline and contains the optical centers O_L, O_M and O_R of respectively left, middle and right cameras and all 3D points situated on the current epiline i. Each 3D point P_k is characterized by its own phase value $\phi(P_k)$. The light ray coming from the light source into the 3D point P_k intersects the segment $[P_iP_{i+1}]$ in a 3D point C_k having the same phase value $\phi(C_k) = \phi(P_k)$ as P_k. To localize C_k, we need to find the distance P_iC_k. This distance is computed by applying the sine law in the triangle $O_{Prj}P_iC_k$ as described in equation (3).

$$\frac{P_iC_k}{sin(\theta_C)} = \frac{O_{Prj}P_i}{sin(\pi - (\theta_C + \alpha))}. \tag{3}$$

The distance $O_{Prj}P_i$ and the angle α between $(O_{Prj}P_i)$ and (P_iP_{i+1}) are known. Also, the angle θ between $(O_{Prj}P_i)$ and $(O_{Prj}P_{i+1})$ is known. Thus, the angle θ_C is defined by equation (4). After localizing C_k, the 3D point P_k is identified as the intersection between (O_RG_k) and $(O_{Prj}C_k)$.

$$\theta_C = \frac{\pi}{\theta}.\phi(C_k). \tag{4}$$

Conventional super-resolution techniques carry out a registration step between low-resolution data, a fusion step and a deblurring step. Here, the phase-shifting analysis provides a registrated left, middle and right point clouds since their 3D coordinates are computed based on the same 3D sparse point cloud. Also, left, middle and right point clouds present homogeneous 3D data and need only to be merged to retrieve the high-resolution 3D point cloud.

4 3D Non-rigid Temporal Super-Resolution

We propose to perform a 3D temporal super-resolution to correct the 3D information provided by the spatial super-resolution and to deal with 3D artifacts caused by either an expression variation, an occlusion or even a facial surface reflectance. First, our temporal super-resolution approach performs a non-rigid registration for each couple of successive 3D point sets M_{t-1} and M_t at each moment t. The 3D non-rigid registration problem is formulated as a maximum-likelihood estimation problem since the deformation between two successive 3D faces is non rigid in general.

We employ the CPD (Coherent Point Drift) algorithm proposed in [11] to registrate the 3D point set M_{t-1} with the 3D point set M_t. The CPD algorithm considers the alignment of two point sets M_{src} and M_{dst} as a probability density estimation problem and fits the GMM (Gaussian Mixture Model) centroids representing M_{src} to the data points of M_{dst} by maximizing the likelihood as described in [11]. N_{src} constitutes the number of points of M_{src} and $M_{src} = \{s_n | n = 1, ..., N_{src}\}$. N_{dst} constitutes the number of points of M_{dst} and $M_{dst} = \{d_n | n = 1, ..., N_{dst}\}$. To create the GMM for M_{src}, a multi-variate Gaussian is centered on each point in M_{src}. All gaussians share the same isotropic covariance matrix $\sigma^2 I$, I being a $3X3$ identity matrix and σ^2 the variance in all directions [11]. Hence the whole point set M_{src} can be considered as a Gaussian Mixture Model with the density $p(d)$ as defined by equation (5).

$$p(d) = \sum_{m=1}^{N_{dst}} \frac{1}{N_{dst}} p(d|m), \qquad d|m \propto N(s_m, \sigma^2 I). \tag{5}$$

Once registered, the 3D point sets M_{t-1} and M_t and also their correponding 2D texture images are used as a low resolution data to create a high resolution 3D point set and its corresponding texture. We apply the 2D super-resolution technique as proposed in [12] which solves an optimization problem of the form:

$$minimize \qquad E_{data}(H) + E_{regular}(H). \tag{6}$$

The first term $E_{data}(H)$ measures agreement of the reconstruction H with the aligned low resolution data. $E_{regular}(H)$ is a regularization or prior energy term that guides the optimizer towards plausible reconstruction H. The 3D model M_t

cannot be represented by only one 2D disparity image since the points situated on the fringe change-over have sub-pixel precision. Also, the left, middle and right pixels participate separately in the 3D model since the 3D coordinates of each pixel is retrieved using only its phase information as described in section (3). Thus, we propose to create for each camera three 2D maps defined by the X, Y and Z coordinates of the 3D points. The optimization algorithm and the deblurring are applied for each camera separately to compute high-resolution images of X, Y, Z and texture from the low-resolution images. We obtain for each camera a high-resolution 3D point cloud using high-resolution data of X, Y and Z. The final high-resolution 3D point cloud is retrieved by merging the left, middle, and right obtained 3D models which are already registrated since all of them contain the 3D sparse point cloud.

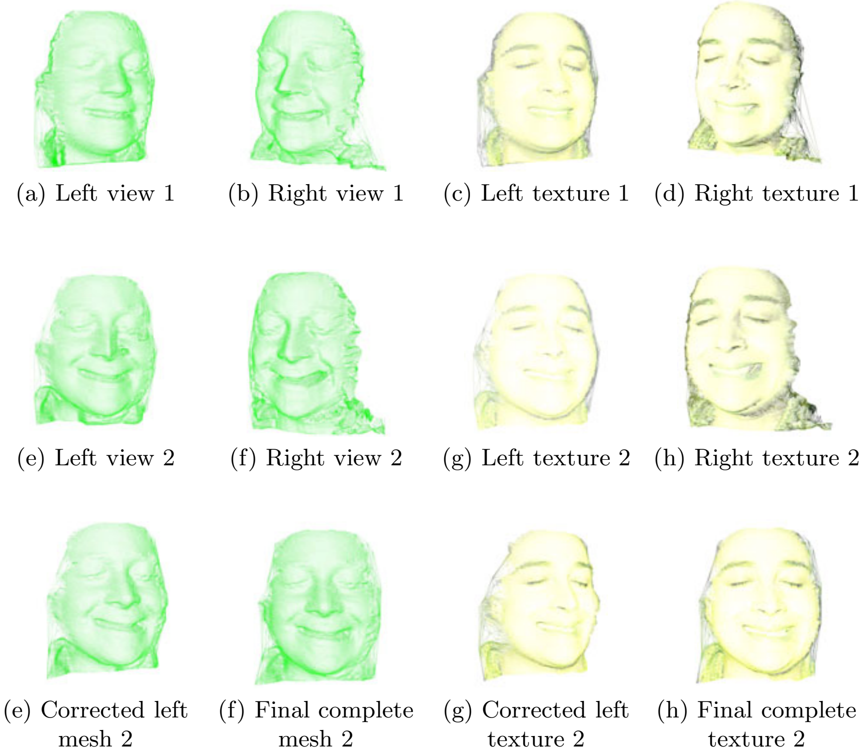
5 Experimental Results

The stereo system hardware is formed by three network cameras with 30 fps and a 480x640 pixel resolution and a LCD video projector. The precision of the reconstruction is estimated using a laser 3D face model scanned by a MINOLTA VI-300 non-contact 3D digitizer. We perform a point-to-surface variant of the 3D rigid matching algorithm ICP (Iterative Closest Point) between a 3D face model provided by our approach and a laser 3D model of the same face. The mean deviation obtained between them is 0.3146mm. Figure 3 presents the primitives extracted and the reconstruction steps to create one facial 3D view with neutral expression from only two cameras. The localization of the face is carried out as described in [5].

At time t, the left, middle and right cameras provide two 3D facial views which can present some artifacts as shown in figure 4 especially for the left 3D view of the second 3D frame shown in 4.e. To deal with these errors, 3D information from the first and second 3D frames are merged despite their non-rigid deformation thanks to the super-resolution approach proposed in section (4). As shown in figure 5, the non-rigid matching algorithm CPD matchs efficiently the preceeding 3D frame with the current 3D left view with a mean deviation of 0.0493$mm/pixel$. Also, the non-rigid matching localizes and clears the artifacts which represent a high spatial deviation with the preceeding 3D frame.

(a) Left view (b) Right view (c) Sparse model (d) Dense mesh (e) Texture

Fig. 3. Reconstruction steps to create one facial 3D view from two cameras


(a) Left view 1 (b) Right view 1 (c) Left texture 1 (d) Right texture 1

(e) Left view 2 (f) Right view 2 (g) Left texture 2 (h) Right texture 2

(e) Corrected left mesh 2 (f) Final complete mesh 2 (g) Corrected left texture 2 (h) Final complete texture 2

Fig. 4. 3D space-time results

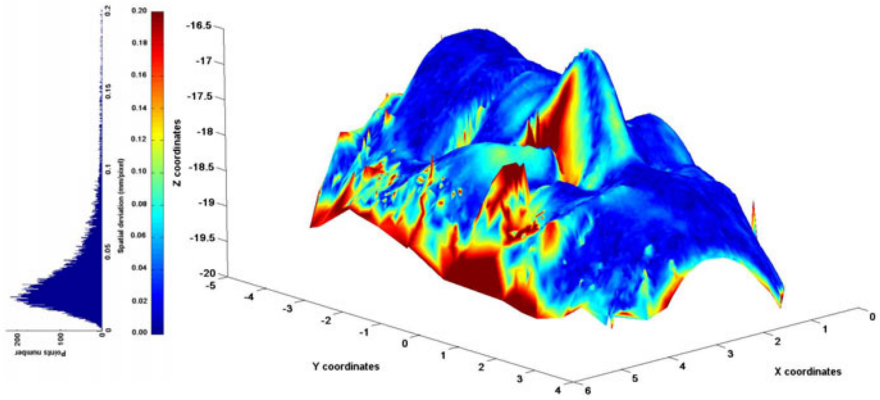


Fig. 5. Non-rigid matching result in presence of artifacts

6 Conclusion and Future Work

This paper proposes a multi-camera 3D face acquisition solution with a 3D space-time super-resolution capability. The proposed super-resolution scheme is particularly suited to 3D speaking face with an expression variation between successive 3D frames. A scanned 3D face model can present some artifacts caused by either an expression variation, an occlusion or even a facial surface reflectance. Super-resolution aims to enhance the quality of the face and to complete the 3D scanned view. The temporal super-resolution fails to correct the 3D face when two successive 3D models present severe artifacts which can propagate through the following 3D frames. As a future work, we suggest to consider more 3D frames through the time axis to enhance the 3D video quality.

Acknowledgments. This research is supported in part by the ANR project FAR3D under the grant ANR-07-SESU-003.

References

1. Zhang, S., Huang, P.S.: High-resolution, real-time three-dimensional shape measurement. Optical Engineering 45, 123601 (2006)
2. Zhang, L., Curless, B., Seitz, S.M.: Rapid shape acquisition using color structured light and multipass dynamic programming. In: 3DPVT Conference (2002)
3. Zhang, S., Yau, S.: Absolute phase-assisted three-dimensional data registration for a dual-camera structured light system. Applied Optics. 47, 3134–3142 (2008)
4. Cox, I., Hingorani, S., Rao, S.: A maximum likelihood stereo algorithm. Computer Vision and Image Understanding. 63, 542–567 (1996)
5. Ouji, K., Ardabilian, M., Chen, L., Ghorbel, F.: Pattern Analysis for an Automatic and Low-Cost 3D Face Acquisition Technique. In: ACIVS Conference, pp. 666–673 (2009)
6. Lu, Z., Tai, Y., Ben-Ezra, M., Brown, M.S.: A Framework for Ultra High Resolution 3D Imaging. In: CVPR Conference (2010)
7. Klaudiny, M., Hilton, A., Edge, J.: High-detail 3D capture of facial performance. In: 3DPVT Conference (2010)
8. Zhang, Z.: Flexible Camera Calibration by Viewing a Plane from Unknown Orientations. In: ICCV Conference (1999)
9. Kil, Y., Mederos, Y., Amenta, N.: Laser scanner super-resolution. In: Eurographics Symposium on Point-Based Graphics (2006)
10. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: LidarBoost: Depth Superresolution for ToF 3D Shape Scanning. In: CVPR Conference (2009)
11. Myronenko, A., Song, X., Carreira-Perpinan, M.A.: Non-rigid point set registration: Coherent Point Drift. In: NIPS Conference (2007)
12. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Fast and robust multi-frame super-resolution. IEEE Trans. Image Processing (2004)

A New Algorithm for 3D Shape Recognition by Means of the 2D Point Distance Histogram

Dariusz Frejlichowski

West Pomeranian University of Technology, Szczecin,
Faculty of Computer Science and Information Technology,
Zolnierska 52, 71-210, Szczecin, Poland
dfrejlichowski@wi.zut.edu.pl

Abstract. A new algorithm for the recognition of three-dimensional objects is proposed in this paper. The algorithm is based on the rendering of several 2D projections of a 3D model, from various positions of the camera. Similarly to the proposition given in [1], the vertices of the dodecahedron enclosing the processed model contain the cameras for this purpose. The obtained projections are stored in bitmaps and the Point Distance Histogram for the description of the planar shapes extracted from them is applied. The obtained histograms represent a 3D model. The experiments performed have confirmed the high efficiency of the proposed algorithm. It outperformed five other algorithms for the representation of 3D shapes.

Keywords: 3D model recognition, 3D shape description, Point Distance Histogram.

1 Introduction and Motivation

Appropriate description of three-dimensional objects is a difficult and challenging task. An algorithm for this purpose has to be robust to many problems and deformations of a model. The most important issue is the invariance of the method to the transformations of an object that may occur. Usually, the affine transformations, e.g. translation, rotation, scaling, shear are taken into account. However, in some cases certain more difficult problems have to be taken into consideration. An example is the partial similarity between objects; the occlusion is the another one. The result of noise and small perturbations in vertices' locations is also very challenging. Moreover, some other specific deformations can be characteristic for particular applications. These also have to be considered.

When dealing with 3D shape representation, recognition or retrieval, some additional properties of an algorithm can be desirable [2]. The first one is the compact representation, which would significantly speed up the process of retrieval and indexing. This characteristic is especially important in dealing with large databases of 3D shapes. However, one has to bear in mind the fact that a representation, which is too compact, may sometimes decrease the efficiency of the approach. It is a common problem in retrieval regardless of the type

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 229–236, 2011.

of processed data. The second desirable property of 3D model representation algorithms, which is often recalled, is the efficiency in discrimination between particular classes.

The 3D shape description algorithms can be assigned to four groups [3]:

- **geometrical** (e.g. Extended Gaussian Image [4], Complex Extended Gaussian Image [5], Light Field Descriptor [1], 3D moments [2], Shape Histograms [6], 3D SIFT Deescriptor [7], Shape Distributions [8]);
- **structural** (e.g. Multiresolutional Reeb Graph [9], Weighted Structural Histogram [10], Skeleton based descriptor [11]);
- **symmetrical** (e.g. Reflective Symmetry Descriptor [12], Planar-Reflective Symmetry Transform [13]);
- **local** (e.g. canonical geometric scale-space analysis [14], Multi-Scale Hierarchical 3D Shape Representation [15]).

The experimental comparison of several 3D model representation algorithms, performed in [3] indicated that good results can be obtained by means of the Light Field Descriptor [1]. This approach is based on rendering several two-dimensional projections of an object, from 20 points of view with cameras placed in vertices of dodecahedron. The obtained projections are stored in bitmaps and further represented by means of a 2D shape descriptor. Originally, the Fourier Descriptors were applied for this purpose. However, better results were achieved when the polar-Fourier transform was applied instead [16]. In this paper a new approach is proposed. It is based on the application of the Point Distance Histogram for the representation of the projected planar shapes. The experiments on 3D shape recognition confirmed that this method is more efficient than previously applied algorithms.

The remaining part of the paper is organised as follows. Section 2 presents the proposed method. Section 3 briefly describes approaches applied in the experiment for comparison with the proposed algorithm. Section 4 is devoted to the presentation of the condition and results of the experiment. Finally, the last section concludes the paper.

2 Description of the 3D Shape Representation Algorithm Based on the 2D Point Distance Histogram

The first part of the proposed approach is based on rendering the projection of a represented 3D model, which results in the creation of two-dimensional shapes. Similarly to the Light Field Descriptor [1] and Polar-Fourier 3D Shape Descriptor [16] this operation is performed for 20 positions of the cameras, in the vertices of dodecahedron enclosing the model. Before the projections are obtained, the object's middle point L is calculated:

$$L = (L_x, L_y, L_z) = (\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i, \frac{1}{n}\sum_{i=1}^{n} z_i), \tag{1}$$

where:
(x_i, y_i, z_i) — denotes the co-ordinates of a vertex of an object,
n — is the number of vertices for particular 3D shape.

All vertices are translated in order to place the centroid in the origin of the co-ordinate system. For a vertex R this procedure can be formulated as follows $(i = 1, 2, \ldots, n)$:

$$R_i = (x_i, y_i, z_i) = (x_i - L_x, y_i - L_y, z_i - L_z). \tag{2}$$

Later, the co-ordinates are normalised according to the maximal distance from the centre of gravity L:

$$M = \max_i \{\|R_i - L\|\}, \tag{3}$$

where: $i = 1, 2, \ldots, n$.

And:

$$R_i = (\frac{x_i}{M}, \frac{y_i}{M}, \frac{z_i}{M}). \tag{4}$$

Now, the above-mentioned projections are obtained, giving in result planar shapes that can be described using the algorithm for the description of two-dimensional objects. In this paper the Point Distance Histogram [19] was applied for the contour of a shape. For each of the 20 projected two-dimensional shapes the same procedure is performed and the final descriptions for those shapes represents a 3D model. It starts with the calculation of the centroid O of the planar shape:

$$O = (O_p, O_q) = (\frac{1}{s}\sum_{i=1}^{s} p_i, \frac{1}{s}\sum_{i=1}^{s} q_i). \tag{5}$$

where:
s — is the number of points in a contour of a planar shape,
p_i, q_i — Cartesian coordinates of the i-th point of the projected shape.

The obtained polar coordinates are put into two vectors Θ^i for angles (in degrees) and P^i for radii:

$$\rho_i = \sqrt{(p_i - O_p)^2 + (q_i - O_q)^2}, \qquad \theta_i = atan\left(\frac{q_i - O_q}{p_i - p_x}\right). \tag{6}$$

The resultant values in θ_i are converted into nearest integers:

$$\theta_i = \begin{cases} \lfloor \theta_i \rfloor, & if\ \theta_i - \lfloor \theta_i \rfloor < 0.5 \\ \lceil \theta_i \rceil, & if\ \theta_i - \lfloor \theta_i \rfloor \geq 0.5 \end{cases}. \tag{7}$$

The next step is the rearrangement of the elements in Θ^i and P^i according to the increasing values in Θ^i. This way we achieve the vectors Θ^j, P^j. For equal elements in Θ^j only the one with the highest corresponding value P^j is selected. That gives a vector with at most 360 elements, one for each integer angle. For further work only the vector of radii is taken — P^k, where $k = 1, 2, ..., m$ and m

is the number of elements in P^k ($m \leq 360$). Now, the normalization of elements in vector P^k is performed:

$$G = \max_k \{\rho_k\}, \qquad \rho_k = \frac{\rho_k}{G}, \tag{8}$$

The elements in P^k are assigned to r bins in histogram (ρ_k to l_k):

$$l_k = \begin{cases} r, & if\ \rho_k = 1 \\ \lfloor r\rho_k \rfloor, & if\ \rho_k \neq 1 \end{cases} . \tag{9}$$

The obtained histograms representing two objects — one for a template and one for a test object — can be matched together by means of any dissimilarity measure. In the paper the Euclidean distance was applied. The template object representing a base class with the smallest dissimilarity measure according to the test object was indicating the recognized class.

3 Brief Description of the Algorithms Selected for the Experimental Comparison with the Proposed Approach

This section tackles the presentation of the description algorithms for 3D shapes used in the experiment along with the method described in the previous section.

The first one — the Extended Gaussian Image (EGI, [4]) is one of the oldest algorithms used in the description of 3D models. In this method, the points on the Gaussian sphere are associated with each point on an object's surface with the same surface orientation.

The second approach, which has been compared with the proposed algorithm, was the Shape Distribution (SD, [8]). The construction of the description starts with the selection of a function representing a model. The authors of the method have proposed five functions, based on the various ways of measuring the distances between particular points (e.g. centroid, varying number of random points) on a surface. In this paper the D2 function was applied, which measures the distance between two random points on a surface. It was the most effective one from the functions proposed by authors. For this function N samples are calculated and later a histogram is constructed, containing information about how many of those samples fall into B bins. From the histogram a piecewise linear function is derived, with V equally spaced vertices, $V \leq B$. The authors suggested the following values of the mentioned parameters: $N = 1024^2$ samples, $B = 1024$ bins, and $V = 64$ vertices.

The Shape Histograms (SH, [6]) was the third descriptor used. It applies the partitioning of the space, where a 3D object lies. For the surface of an object the histogram is built using the obtained cells. The authors have proposed three methods for the decomposition of the space — a shell model, a sector model, and a spider-web model.

The Light Field Descriptor (LFD, [1]) was the another method selected for the experiment. In fact its main idea (the rendering of 2D projections of a 3D model) was the basis for the algorithm described in this paper. The shape descriptions for planar objects placed in the obtained projected bitmaps are compared in order to indicate the similar models. Similarly to the algorithm presented in the previous section the construction of LFD starts with the translation of the vertices according to the origin of the Cartesian co-ordinates system and normalization of the co-ordinates according to the maximal one. Later, rendered planar projections are obtained for twenty various angles. They are stored in bitmaps and afterwards represented using the Fourier Descriptors. The pictorial description of the LFD algorithm is provided in Fig. 1.

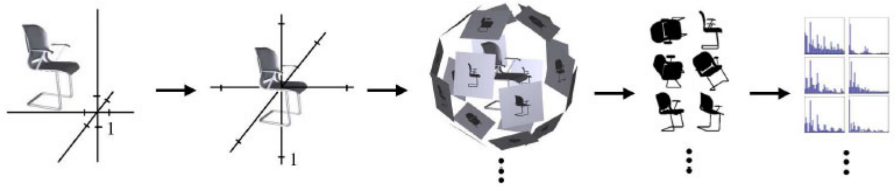


Fig. 1. The pictorial representation of the main steps in the process of determining the Light Field Descriptor [17]

The experimental results of the LFD descriptor have confirmed its high efficiency in the problem of 3D shape recognition [3]. However, as it turned out, the method can by easily improved, if replaced FD with some other 2D shape descriptor. In [16] the polar-Fourier transform of the planar boundary was applied for this purpose. This approach was also selected for comparison with the proposed algorithm, since it is very similar and effective.

4 Methodology and Results of the Experiment

The methodology of the experiment evaluating the proposed in the paper approach was similar to the experimental comparison of the four well-known and popular 3D model representation methods, described in [3]. Later, those methods were compared in the same way with the new algorithm in [16]. The Princeton Shape Benchmark [18] was used during tests — 312 objects belonging to 13 different classes (some examples are presented in Fig. 2). The idea of the experiment was simple. The recognition was considered successful if the Euclidean distance between a test object and a template was the smallest for the same class. Obviously, both were represented using particular description algorithm. The precise results of the recognition obtained for investigated approaches are provided in Table 1. As one can notice, the proposed algorithm outperforms the other explored 3D model representation techniques. Its average recognition rate (RR) is higher than 76%. The Polar-Fourier 3D Descriptor was slightly worse — by 2 per cent. Other explored algorithms obtained the accuracy lower than 70%.

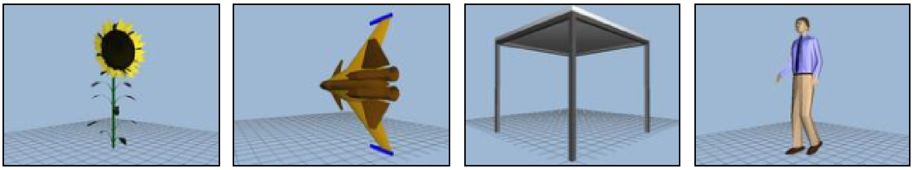


Fig. 2. Examples of the 3D models used during the experiment, from the Princeton Shape Benchmark database [18]

Table 1. Experimental results — percentage of the successful identification (recognition rate) for investigated 3D shape description algorithms

Class no.	EGI	SH	SD	LFD	P-F 3D	Proposed approach
1.	57.75	29.58	78.87	78.87	83.10	88.73
2.	65.71	48.57	57.14	85.71	85.71	88.57
3.	52.63	21.05	84.21	57.89	63.16	68.42
4.	53.13	56.25	34.38	56.25	62.50	68.75
5.	80.00	20.00	30.00	10.00	40.00	50.00
6.	50.00	44.44	72.22	88.89	83.33	88.89
7.	66.67	33.33	50.00	50.00	66.67	66.67
8.	66.67	0.00	0.00	33.33	66.67	33.33
9.	65.12	67.44	27.91	74.42	76.74	74.42
10.	70.00	10.00	60.00	60.00	70.00	50.00
11.	60.61	9.09	54.55	66.67	75.76	72.73
12.	50.00	12.50	12.50	25.00	37.50	50.00
13.	100.00	16.67	50.00	16.67	33.33	33.33
Overall	**60.26**	**36.86**	**56.09**	**68.91**	**74.68**	**76.28**

5 Conclusions

In the paper a new algorithm for representation, identification, recognition and retrieval of three-dimensional models was proposed. It rests on the idea applied in the Light Field Descriptor — the rendering of several projections of a 3D object (for cameras placed in the vertices of dodecahedron enclosing the model). However, here, the obtained 2D shapes are represented by means of the Point Distance Histogram, the shape descriptor that combines the polar transform and the histogram.

The proposed method was experimentally evaluated in the problem of 3D shape recognition, using the data from the Princeton Shape Benchmark [18]. The obtained accuracy has indicated that the proposed method works better in the problem than the other explored approaches. It achieved above 76% recognition rate, while the second best method — Polar-Fourier 3D Descriptor was worse by 2 per cent. The rest of the investigated algorithms gave the recognition rate below 70%. The Light Field Descriptor obtained 69% accuracy, the Extended Gaussian

Image — 60%, the Shape Distributions — 56%, and the worst approach, Shape Histograms — 37%.

References

1. Chen, D.-Y., Ouhyoung, M., Tian, X.-P., Shen, Y.-T.: On visual similarity based 3D model retrieval. In: Computer Graphics Forum, pp. 223–232 (2003)
2. Novotni, M., Klein, R.: Shape Retrieval Using 3D Zernike Descriptors. Computer-Aided Design 36(11), 1047–1062 (2004)
3. Frejlichowski, D.: 3D Shape Description Algorithms Applied to the Problem of Model Retrieval. Central European Journal of Engineering 1(1), 117–121 (2011)
4. Horn, B.: Extended Gaussian Images. Proc. of the IEEE A.I. Memo, no. 740 72(12), 1671–1686 (1984)
5. Kang, S., Ikeuchi, K.: Determining 3-D Object Pose Using the Complex Extended Guassian Image. In: Proc. of the CVPR, pp. 580–585 (1991)
6. Ankerst, M., Kastenmuller, G., Kriegel, H., Seidl, T.: 3D Shape Histograms for Similarity Search and Classification in Spatial Databases. In: Proc. of the 6th Int. Symp. on Spatial Databases, pp. 207–226 (1999)
7. Flitton, G., Breckon, T.P., Megherbi, N.: Object Recognition using 3D SIFT in Complex CT Volumes. In: Proc. of the British Machine Vision Conference 2010, pp. 11.1–11.12 (2010)
8. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Matching 3D Models with Shape Distributions. In: Proc. of Int. Conf. SMI 2008, pp. 154–166 (2001)
9. Hilaga, M., Shinagawa, Y., Kohmura, T., Kunii, T.L.: Topology Matching for Fully Automatic Similarity Estimation of 3D Shapes. In: Proc. of the 28th Conference on Computer Graphics and Interactive Techniques, pp. 203–212 (2001)
10. Lu, T., Gao, R., Wang, T., Yang, Y.: 3D Similarity Search Using a Weighted Structural Histogram Representation. In: Qiu, G., Lam, K.M., Kiya, H., Xue, X.-Y., Kuo, C.-C.J., Lew, M.S. (eds.) PCM 2010. LNCS, vol. 6297, pp. 348–356. Springer, Heidelberg (2010)
11. Sundar, H., Silver, D., Gagvani, N., Dickinson, S.: Skeleton Based Shape Matching and Retrieval. In: Proc. of the IEEE Shape Modeling International, May 12–15, pp. 130–139 (2003)
12. Kazhdan, M., Chazelle, B., Dobkin, D., Funkhouser, T., Rusinkiewicz, S.: A Reflective Symmetry Descriptor for 3D Models. Algorithmica 38, 201–225 (2003)
13. Podolak, J., Shilane, P., Golovinskiy, A., Rusinkiewicz, S., Funkhouser, T.: A Planar-Reflective Symmetry Transform for 3D Shapes. ACM Trans. on Graphics (Proc. SIGGRAPH) 25(3), 549–559 (2006)
14. Novatnack, J., Nishino, K.: Scale-Dependent/Invariant Local 3D Shape Descriptors for Fully Automatic Registration of Multiple Sets of Range Images. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 440–453. Springer, Heidelberg (2008)
15. Iyer, N., Jayanti, S., Lou, K., Kalyanaraman, Y., Ramani, K.: A Multi-scale Hierarchical 3D Shape Representation for Similar Shape Retrieval. In: Proceedings of TMCE 2004, vol. 2, pp. 1117–1118 (2004)
16. Frejlichowski, D.: A Three-Dimensional Shape Description Algorithm Based on Polar-Fourier Transform for 3D Model Retrieval. In: Heyden, A., Kahl, F. (eds.) SCIA 2011. LNCS, vol. 6688, pp. 457–466. Springer, Heidelberg (2011)

17. Shen, Y.-T., Chen, D.-Y., Tian, X.-P., Ouhyoung, M.: 3D Model Search Engine Based on Lightfield Descriptors. In: EUROGRAPHICS Interactive Demos, Granada, Spain (2003)
18. Shilane, P., Min, P., Kazhdan, M.M., Funkhouser, T.A.: The Princeton Shape Benchmark. In: Proc. of the SMI 2004, Genova, Italy, pp. 145–156 (2004)
19. Frejlichowski, D.: An Experimental Comparison of Seven Shape Descriptors in the General Shape Analysis Problem. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6111, pp. 294–305. Springer, Heidelberg (2010)

Wide Range Face Pose Estimation by Modelling the 3D Arrangement of Robustly Detectable Sub-parts

Thiemo Wiedemeyer[1], Martin Stommel[2], and Otthein Herzog[3]

TZI Center for Computing and Communication Technologies,
University Bremen, Am Fallturm 1, 28359 Bremen, Germany
raider@informatik.uni-bremen.de, {mstommel,herzog}@tzi.de

Abstract. A highly accurate solution for the estimation of face poses over a wide range of 180 degree is presented. The result has been achieved by modeling the 3D arrangement of 15 facial features and its mapping to the image plane for different poses. A voting scheme is used to compute the mapping for a given image in a bottom-up procedure. The voting is based on a robust classification of the appearance of the sub-parts. However, equal importance must be ascribed to the extension of the annotation scheme of the Feret data base, also including the correction of existing misannotations.

Keywords: face pose estimation, facial feature detection, Feret data base.

1 Introduction

Face recognition is an important step in the analysis, archiving and retrieval of TV or movie productions. Since most video productions present stories and news about people, the automatic detection of faces together with the estimation of the pose and possibly the identification of the person can give important clues to the content of a video. Pose estimation is useful in two respects. First, it might serve as a clue to the scene layout by indicating the relationship between different people. Secondly, it is usually considered as a preprocessing step for the identification of an individual. A careful pose estimation can therefore alleviate the problem that differences in pose cause stronger numerical differences in appearance than changes between two individuals [1,16].

Current methods for pose estimation achieve the best results for close to frontal camera views [6,10]. Many methods also treat pose estimation as a classification problem and hence require a previous face detection. Model-based methods often require a previous detailed registration of face parts.

In this paper, we present a method that performs a simultaneous detection, registration and pose estimation with equally good results over yaw angles in a broad range of 180 degree. Although the proposed method is based on our previous work [17], the voting mechanism is completely different. Additional to

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 237–244, 2011.

new optimisations and heuristics, the new method uses a 3D constellation of face parts, whereas the previous method only had a 2D model.

There are also significant differences in the experimental setup. Because of felt deficiencies in the Feret data base, we reannotated[1] a higher number of 15 parts for about 11 000 samples of all poses. This solves the problem of missing annotations for side views, as well as misannotations for certain angles.

2 Related Work

Face detection and pose estimation is mostly considered as a problem where the local appearance of a face must be combined with the general geometrical structure. Therefore, many approaches exist that model the local appearance based on sets of spatially separated feature descriptors [13,14,17] or wavelets [2,4]. The local appearance is usually learned from a training set [6,13]. Different techniques exist to superimpose the geometrical layout on the parts, e.g. voting [17], graphs [2], or active appearance models [3]. Given a proper registration of facial landmarks, it is even possible to adapt highly detailed geometrical models [5]. Recent developments show a certain revival of embedding techniques [8], where the face appearance across pose is modeled holistically [16]. However, the application of standard methods like e.g. Isomap requires a dense sampling of the manifold for the interpolations in tangent space to be valid. Often additional information proves useful in face recognition, such as dynamics [15], symmetry [10], or large data bases [9,11].

3 Data Set

Our experiments are conducted on the Feret [1], Graz'02 [7] data bases. The Feret data base contains about 11 000 face images taken under controlled illumination in front of homogeneous backgrounds. The data base contains samples from 12 poses with yaw angles from the range of +90° to −90°. The original annotation contains the positions of 4 facial features for a limited range of poses. For our compositional approach to be applicable, we manually annotated 15 different facial features for all samples. The rare pose −75° is not used. Figure 1 shows the annotated features for an example. To extend our approach to uncontrolled scenarios, we train and test our method also on the background set of the Graz'02 data base. This set does not contain any faces, so it allows for the measurement of false positives. Since the Feret data base only considers yaw angles, the Basel face model [12] is used to create synthetic images of additional poses.

4 Detection of Facial Features

Our method proceeds in two steps. At first, facial features are detected in an image. Afterwards, a voting is performed to estimate the pose.

[1] The annotation (without the Feret images) is available on request. Please contact one of the authors.

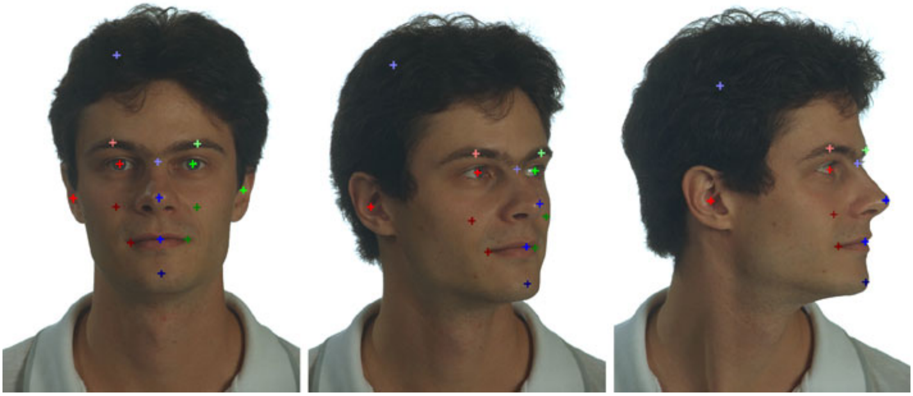


Fig. 1. Example for the extended annotation scheme

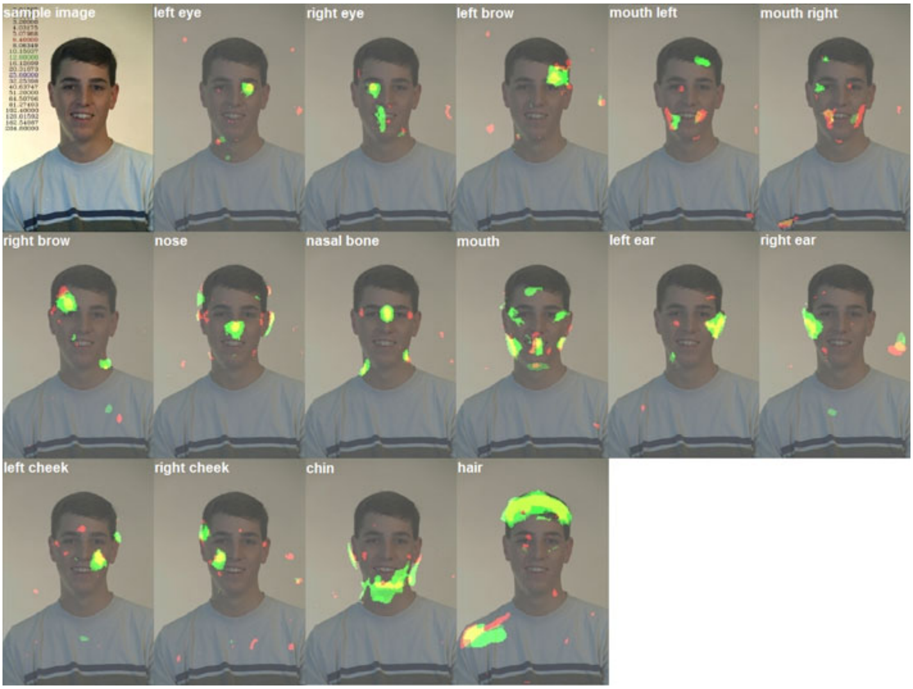


Fig. 2. Sample image with detected parts highlighted

The feature detection combines a dense SIFT feature extraction with a SVM classification at each pixel coordinate. Except from using the original SIFT features in this paper, the method is similar to our previous work [17].

Both the training as well as the (disjoint) test set contain descriptors from the annotated coordinates of 100 Feret images per pose, as well as 1520 unique,

Table 1. Precision and recall [%] for the detection of facial features

Class	Precision	Recall	Class	Precision	Recall
background	91.52	93.85	left mouth corner	97.46	96.97
left eye	96.19	96.46	right mouth corner	95.76	95.04
right eye	96.75	95.13	left cheek	95.06	96.31
left brow	94.50	94.93	right cheek	94.28	93.83
right brow	93.86	92.69	left ear	96.81	92.91
nose	94.74	96.77	right ear	95.62	93.53
nasal bone	98.39	96.79	chin	96.10	95.18
mouth	93.56	94.59	hair	96.51	94.21
average	94.87	94.87			

randomly selected background features from the Graz data base. As Tab. 1 shows, the method achieves a high precision and recall of 95% averaged over all classes. As Fig. 2 shows, the detections form rather loosely bounded areas in the image plane.

5 3D Model and Pose Estimation

The pose estimation is based on a voting mechanism that matches a 3D model to the detections of facial features.

The 3D model consists of the coordinates of the annotated facial features relative to a reference point. The model is computed by triangulation of the annotated feature points from different angles. The centre between nasal bone and chin is chosen as the reference point because it is visible in most samples and achieves the lowest mean absolute error between model and annotation. Translation and scale invariance is achieved by normalising the shape of the feature constellations to the vector from the nasal bone to the chin. Symmetry is imposed on the model by mirroring annotated feature coordinates. The resulting mean absolute error of the coordinates of facial features when transforming the model according to the annotated pose is 3.2 pixels. In comparison to the reasonably coarse quantisation of the pose in 11 intervals, this indicates a high accuracy of the annotated feature coordinates.

The voting is based on the feature detections $d_1, d_2, \ldots, d_n$, where each detection $d = (d_x, d_y, d_l)$ consists of a coordinate x, y and a label l of the feature class (e.g. left eye). For a certain feature label d_l, a pose ϕ represented by the yaw, pitch and roll angles, and the scale s, the R-Table $R(d, \phi, s) = (r_x, r_y)$ returns the relative screen position r_x, r_y of the reference point of the model according to the coordinate transformation given by the pose and scale. The resulting absolute screen position $a_x = r_x + d_x, a_y = r_y + d_y$ is the sum of both coordinates. To account for noise, a kernel function $k(s_x, s_y, s, R, d)$ is centered at a_x, a_y to smooth the transformed detection over adjacent image coordinates s_x, s_y and scales s. By summation in an accumulator $A(s_x, s_y, s, phi) = \sum_d k$, the set of detections give votings for a particular position, pose and scale of a face. It can be detected by finding the maximum $= \arg\max_{s_x, s_y, s, phi} A$ in the accumulator.

The (uniform) kernel function introduces a dilatation to allow for certain spatial displacements. The radius has been estimated experimentally. The kernel function also interpolates votings along a line between the next smaller and higher scales. The scales correspond to those of the SIFT descriptors.

6 Experimental Results for the Pose Estimation

Table 2 shows the test results for the estimation of the yaw angle for the Feret test set (cf. Sec. 4). The Graz data base is not used here. In contrast to the strong increase of the error rate that is often observed for non-frontal poses (e.g. [6]), the mean absolute error is almost equally high over all poses in our experiments. To a certain degree, this might also be a result of using a sufficient number of non-frontal samples.

Table 2. Per each pose the mean absolute error (MAE) in degree and the recall for an area around the expected pose is listed

Pose	MAE	±0.0°	±7.5°	±15.0°	±22.5°
−90.0	11.62	32%	68%	83%	90%
−67.5	29.55	0%	8%	24%	45%
−45.0	12.67	21%	49%	71%	94%
−22.5	9.82	18%	62%	93%	96%
−15.0	7.50	23%	86%	96%	96%
0.0	2.78	67%	96%	100%	100%
15.0	7.13	22%	85%	98%	100%
22.5	8.40	31%	69%	93%	99%
45.0	12.52	14%	59%	71%	91%
67.5	31.35	1%	6%	30%	39%
90.0	4.65	68%	86%	94%	96%
With ±67.5°	12.42	25%	62%	78%	86%
Overall	8.57	33%	73%	89%	96%

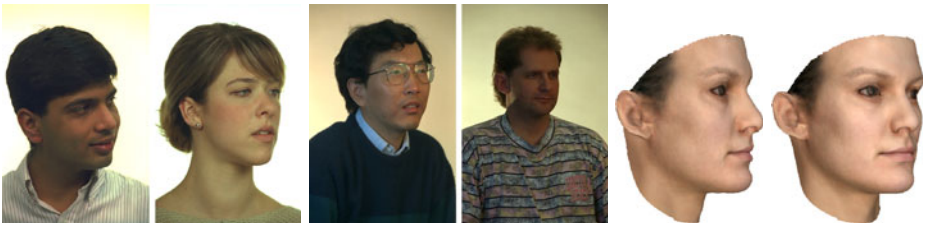


Fig. 3. Hypothesised labelling error: The left four images from the Feret data base are labelled as −67.5°. For comparison two synthetic images based on the Basel face model are shown for the poses −67.5° and −45.0°. Given that in the −67.5° case only one eye is visible, the synthetic images suggest a pose of −45.0° rather than −67.5°.

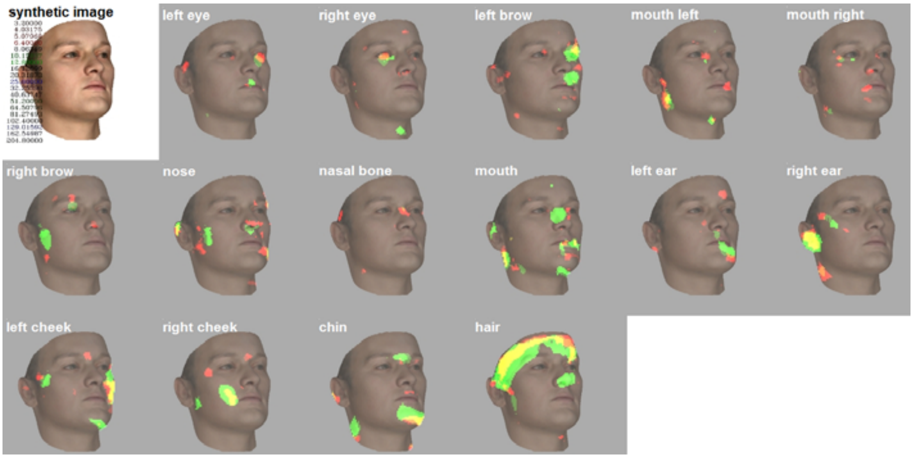


Fig. 4. Detected facial features for a synthetic image

Table 3. Confusion matrix for the pose estimation

Output	Annotated Pose [°]										
Pose	−90.0	−67.5	−45.0	−22.5	−15.0	0.0	15.0	22.5	45.0	67.5	90.0
−90.0	22	0	0	0	0	0	0	0	0	0	0
−82.5	39	6	0	0	0	0	0	0	0	0	0
−75.0	19	1	0	0	0	0	0	0	0	0	0
−67.5	9	0	0	0	0	0	0	0	0	0	0
−60.0	3	7	5	0	0	0	0	0	0	0	0
−52.5	3	10	12	0	0	0	0	0	0	0	0
−45.0	0	21	23	0	1	0	0	0	0	0	0
−37.5	3	23	31	1	0	0	0	0	0	0	0
−30.0	0	11	12	4	1	0	0	0	0	0	0
−22.5	1	13	14	18	19	0	0	0	0	0	0
−15.0	0	5	2	45	26	1	0	0	0	0	0
−7.5	0	0	0	24	42	19	0	0	0	0	0
0.0	0	0	1	5	9	68	12	2	0	1	0
7.5	0	2	0	3	0	9	43	11	1	0	0
15.0	1	1	0	0	1	3	28	23	7	13	0
22.5	0	0	0	0	1	0	10	29	11	19	0
30.0	0	0	0	0	0	0	5	14	13	12	0
37.5	0	0	0	0	0	0	2	10	16	16	1
45.0	0	0	0	0	0	0	0	9	14	9	0
52.5	0	0	0	0	0	0	0	2	33	21	0
60.0	0	0	0	0	0	0	0	0	5	4	4
67.5	0	0	0	0	0	0	0	0	0	1	5
75.0	0	0	0	0	0	0	0	0	0	1	9
82.5	0	0	0	0	0	0	0	0	0	3	26
90.0	0	0	0	0	0	0	0	0	0	0	55

However, for yaw angles of ±67.5°, a drop in recall is visible. Tab. 3 shows a frequent confusion with the neighboring pose of ±45.0°. A manual comparison of the images with annotated ±67.5° yaw angle from the Feret-Database to synthetically generated images using the Basel face model shows that most of the faces annotated as ±67.5° are more likely shown from ±45.0° (cf. Fig. 3). A comparison of the images and ground truth data of other poses from the Feret-Database shows, that the accuracy of the annotation decreases the stronger faces are rotated and outliers occur more frequently. Table 2 therefore gives additional results without testing the angle ±67.5°.

To test the estimation of all three angles, 10 faces of the Basel face model are synthesised each for 7 yaw angles from $-90° - -0°$ and three roll and pitch angles from the interval ±15°. The application of the feature detection trained on the Feret images (cf. Fig. 4) yields a higher proportion of misdetections. With mean absolute errors of $12.9°, 6.9°$, and 6.9° for yaw, pitch and roll, the results are therefore slightly inferior to the results on real images.

7 Conclusion

A face detection and pose estimation system is presented that achieves high recognition rates over a wide range of poses. The method is evaluated using the Feret and Graz data bases, as well as the Basel face model.

The excellent precision and recall of 95% in the detection of a set of 15 facial features over 11 poses shows that facial features can be successfully classified even in the presence of strong visual variations. The training of such numerically heterogeneous clusters requires annotated data, however.

The high recognition rates on feature level allow for a compositional modelling and voting of the face pose. The pose estimation yields a mean absolute error of 8.6° which is close to the accuracy of the annotation. Decreases of the recognition rate for certain angles could be traced back to inaccuracies in the annotation of some Feret samples.

References

1. Phillips, P.J., Rauss, P.J., Der, S.Z.: FERET (Face Recognition Technology) Recognition Algorithm Development and Test Results, Army Research Lab technical report 995 (October 1996)
2. Kruger, N., Potzsch, M., von der Malsburg, C.: Determination of face position and pose with a learned representation based on labelled graphs. In: British Machine Vision Conference on Image and Vision Computing, vol. 15(8), pp. 665–673 (1997)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
4. Elagin, E., Steffens, J., Neven, H.: Automatic pose estimation system for human faces based on bunch graph matching technology. In: Proc. IEEE International Conference on Automatic Face and Gesture Recognition, April 14–16, pp. 136–141 (1998)

5. Blanz, V., Vetter, T.: A Morphable Model for The Synthesis of 3.D Faces. In: Computer Graphics, SIGGRAPH Proceedings, Los Angeles, CA, pp. 187–194 (August 1999)
6. Gourier, N., Hall, D., Crowley, J.L.: Estimating Face orientation from Robust Detection of Salient Facial Structures. In: FG Net Workshop on Visual Observation of Deictic Gestures, POINTING (2004)
7. Opelt, A., Fussenegger, M., Pinz, A., Auer, P.: Weak Hypotheses and Boosting for Generic Object Detection and Recognition. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3022, pp. 71–84. Springer, Heidelberg (2004)
8. Fu, Y., Huang, T.S.: In: Graph Embedded Analysis for Head Pose Estimation. In: Seventh IEEE International Conference on Automatic Face and Gesture Recognition (FG 2006), pp. 3–8 (2006)
9. Grujic, N., Ilic, S., Lepetit, V., Fua, P.: 3D Facial Pose Estimation by Image Retrieval. In: 8th IEEE Int'l Conference on Automatic Face and Gesture Recognition (2008)
10. Pathangay, V., Das, S., Greiner, T.: Symmetry-based Face Pose Estimation from a Single Uncalibrated View. In: Proc. International Conference on Face and Gesture Recognition (FG 2008), Amsterdam, Netherlands (September 2008)
11. Aghajanian, J., Prince, S.J.D.: Face Pose Estimation in Uncontrolled Environments. In: British Machine Vision Conference, BMVC (2009)
12. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D Face Model for Pose and Illumination Invariant Face Recognition. In: IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) for Security, Safety and Monitoring in Smart Environments. IEEE, Los Alamitos (2009)
13. Kozakaya, T., Shibata, T., Yuasa, M., Yamaguchi, O.: Facial feature localization using weighted vector concentration approach. In: Pantic, M., Sebe, N., Cohn, J.F., Huang, T.S. (eds.) Image and Vision Computing, Special Issue: Best of Automatic Face and Gesture Recognition 2008, vol. 28(5), pp. 772–780 (2010)
14. Mayer, C., Wimmer, M., Radig, B.: Adjusted pixel features for robust facial component classification. In: Pantic, M., Sebe, N., Cohn, J.F., Huang, T.S. (eds.) Image and Vision Computing, Special Issue: Best of Automatic Face and Gesture Recognition 2008, vol. 28(5), pp. 762–771 (2010)
15. Morency, L.-P., Whitehill, J., Movellan, J.: Monocular head pose estimation using generalized adaptive view-based appearance model. In: Pantic, M., Sebe, N., Cohn, J.F., Huang, T.S. (eds.) Image and Vision Computing, Special Issue: Best of Automatic Face and Gesture Recognition 2008, vol. 28(5), pp. 754–761 (2010)
16. Sarfraz, M.S., Hellwich, O.: Probabilistic learning for fully automatic face recognition across pose. In: Pantic, M., Sebe, N., Cohn, J.F., Huang, T.S. (eds.) Image and Vision Computing, Special Issue: Best of Automatic Face and Gesture Recognition 2008, vol. 28(5), pp. 744–753 (2010)
17. Stommel, M., Herzog, O.: Learning of Face Components in Coherent and Disturbed Constellations. In: Int'l Conf. on Image and Vision Computing New Zealand (IVCNZ), Queenstown, New Zealand, November 8-9 (2010)

Single Image Restoration of Outdoor Scenes

Codruta Orniana Ancuti, Cosmin Ancuti, and Philippe Bekaert

Hasselt University - tUL -IBBT,
Expertise Center for Digital Media, Wetenschapspark 2, Diepenbeek, 3590, Belgium
firstname.secondname@uhasselt.be

Abstract. We present a novel strategy to restore outdoor images degraded by the atmospheric phenomena such as haze or fog. Since both the depth map of the scene and the airlight constant are unknown, this problem is mathematically ill-posed. Firstly, we present a straightforward approach that is able to estimate accurately the airlight constant by searching the regions with the highest intensity. Afterwards, based on a graphical Markov random field (MRF) model, we introduce a robust optimization framework that is able to transport the local minima over large neighborhoods while smoothing the transmission map but also preserving the important depth discontinuities of the estimated depth. The method has been tested extensively for real outdoor images degraded by haze or fog. The comparative results with the existing state-of-the-art techniques demonstrate the advantage of our approach.

1 Introduction

The outdoor applications such as video surveillance and intelligent vehicles are in general more challenging due to the additional issues introduced by the weather conditions. Atmospheric phenomena such as haze or fog may alter substantially the scene visibility of outdoor images and videos. A similar issue is typical as well for underwater and aerial images. Since this effect depends on the depth map of the considered scene, restoration of such spoilt images represents a difficult task.

When examining an outdoor scene from an elevated position, features gradually appear to become lighter and fading as they are closer towards the horizon. Only a percentage of the reflected light reaches the observer as a result of the absorption in the atmosphere. Furthermore, this light gets mixed with the *airlight* [1] color vector, and due to the scattering effects the scene color is shifted.

Early techniques have used additional information such as images [2], approximate depth map of the scene [3] and hardware [4]. Obviously, these techniques are in general impractical since in most of the cases this extra information is not available to the common users.

Recently, however, several solutions [5,6,7,8,9,10,11] to restore hazy images by processing only the degraded input image have been introduced. Because this problem is highly underconstrained, different assumptions have been made

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 245–252, 2011.

in order to estimate as accurate as possible both the transmission and latent image.

In this paper we present an alternative solution to restore such outdoor degraded images. Since both the depth map of the scene but also the airlight constant are unknown, this problem is mathematically ill-posed. Our strategy is first to provide an accurate estimate of the airlight constant. To achieve this goal, we describe a straightforward approach that searches the regions with the highest intensity. Afterwards, based on a graphical Markov random field (MRF) model, we introduce a robust framework that is able to transport the local minima over large neighborhoods constrained by smoothing, but also to preserve the important depth discontinuities in the transmission map. To speed-up our technique, this step is implemented effectively by a fast belief propagation scheme.

Graphical models (MRF) have been used as well in several previous approaches [6,5,9]. The method of Tan [6] aims to locally increase the contrast of the recovered regions, however the results may present artifacts due to the patch-based final composition. More related approaches to ours are the recent works of Fattal [5] and Kratz and Nishino [9]. Different than these approaches that are considerably more complex, a direct smoothing constraint of the transmission map estimate is imposed in our optimization process. This step represents a key insight of our approach since the airlight contribution in general varies smoothly in such outdoor scenes but also as can be seen in figure 1, smoothing the transmission map plays a crucial role in recovering the finest details.

The method has been tested extensively for real images of the outdoor scene degraded by haze. The comparative results with the existing state-of-the-art techniques demonstrate the utility of our approach.

2 Our Restoration Approach

The captured image of a hazy scene $\mathcal{I}_h$ is represented by a linear combination of *direct attenuation* $\mathcal{D}$ and *airlight* $\mathcal{A}$ contributions:

$$\mathcal{I}_h = \mathcal{D} + \mathcal{A} = \mathcal{I} * t\,(x) + A_\infty * [1 - t\,(x)] \tag{1}$$

where $\mathcal{I}_h$ is the image spolit by haze, $\mathcal{I}$ is the scene radiance or haze-free image, A_∞ is the constant airlight color vector and t is the transmission along the cone of vision. This ill-posed problem requires to recover the unknowns $\mathcal{I}$, A_∞ and $t(x)$ from only a single input image $\mathcal{I}_h$. Practically, our main goal is to estimate accurate values of the transmission and the airlight constant in order to recover the degraded image.

First, based on the assumption that the airlight gain increases proportionally with the optical depth, we develop a straightforward technique to estimate the airlight constant A_∞ that proves robustness even for the most challenging cases (e.g. non-sky images). Regarding the estimation of transmission map, we have inspired our approach by the recent dark channel [7] approximation. However, in order to preserve the discontinuities the technique of He et al. [7] requires an expensive refinement post processing step (alpha matting) that in many cases

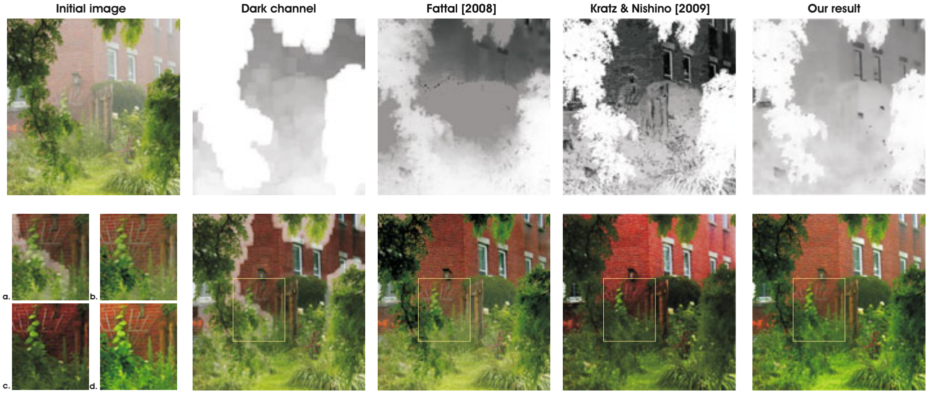


Fig. 1. Comparative results obtained employing estimated transmission maps yielded by different approaches. From left to right on the first line: input image; transmission maps of dark channel estimated as in [7]; transmission of Fattal [5]; transmission of Kratz and Nishino [9] ; our estimated transmission. From left to right on the second line: close-up regions of (a) dark channel, (b) Fattal [5] , (c) Kratz and Nishino [9], (d) ours; restoration result obtained by employing the dark channel; restoration result of Fattal [5]; restoration result of Kratz and Nishino [9]; our restoration result. Please notice that by simply employing the dark channel, the result will present artifacts along depth discontinuities.

does not guarantee the convergence to the local minima. To overcome these limitations we introduce an effective framework based on graphical Markov Random Field (MRF) model that is able to transport the local minima over large neighborhoods while smoothing the initial depth estimate, and also to preserve the important depth discontinuities of the transmission map.

2.1 Airlight Color (A_∞) Estimation

Since there is an important correlation between the optical depth and the airlight [12,4] the airlight gain $\mathcal{A}$ is assumed to increase proportionally with the optical depth. By analyzing the optical model previously described (equation 1), it results that two surfaces characterized by different reflectance properties but located on the same distance from the observer, have similar airlight gains $\mathcal{A}$. Consequently , since the transmission $t(x)$ is considered to vary smoothly except for depth discontinuities, the values of $\mathcal{A}$ in a small region around a given scene point will vary in a similar way. As a result, we assume that the constant A_∞ can be estimated with good accuracy from those parts of the scene with the highest airlight gain, commonly represented by the brightest image regions.

These properties of hazy images, have been exploited as well by the previous approaches in order to estimate the airlight constant A_∞. For example, the method of Tan [6] searches the regions with the highest intensity, assuming that the captured scene includes the sky and there are no saturated pixels. However, since this constraint is not always satisfied, especially when the sky is not present in the image, in the method of He et al. [7] the airlight is found in the top 0.1% of the brightest pixels of the dark channel.

Similarly, inspired as well by the technique of Narasimhan and Nayar [2], we aim to estimate the airlight constant A_∞ in the most degraded (hazy) regions. However, for the non-sky images (e.g. images that do not contain sky regions) but also for images characterized by additional light sources, we observed that the pixels with the highest intensity may not correspond to the airlight color. Therefore, since we suppose that we deal with hazy/foggy images, we consider that the median value of the top 50% brightest pixels of the hazy blurred image (we apply a simple Gaussian blur to remove some of the finest texture transitions) contains the airlight color. However, this value may not have always the highest desired intensity value. Practically, to obtain the final airlight constant value A_∞, our strategy searchs for the pixels with the highest intensity that have similar hue value with the median value of the considered region.

2.2 Transmission Estimation

The restoration of such images aims to increase the local contrast that decreases with the airlight contribution and the scene depth. Practically, the visible contrast is the consequence of the luminance difference yielded by the difference in the amount of the reflected light from two surfaces in their vicinities. This demonstrates the known fact that the contrast depends by the local variations but also explains the disparity that allows an observer to perceive separately objects from the background. Analyzing the optical model equation, since A_∞ is constant (the value of A_∞ is estimated as was explained in the previous subsection) it implies that local contrast can be enhanced only if the transmission map is relatively smooth except for the depth transitions.

To compute a first estimate of the transmission map we use the recent strategy of He et al. [7] that computes a rough version of the depth map based on the dark channel observation. The dark channel ($\mathcal{I}_{d-c}$) is expressed straightforwardly as:

$$\mathcal{I}_{d-c}(x) = \min_{y\in\Omega(x)}[\min_{c\in r,g,b}(\mathcal{I}_c(y))] = \min_{y\in\Omega(x)}[\mathcal{I}_{c-min}] \tag{2}$$

where $\mathcal{I}_c$ represents a color channel of the hazy image $\mathcal{I}_h$ and $\Omega(x)$ represents a patch centered at location x. As discuss afterwards (equation 6), the value of $\mathcal{I}_{c-min}$ is used to initialize our optimization framework.

Based on the previous formulation, the transmission estimate $t_i\,(x)$ is computed by the following expression:

$$t_i\,(x) = 1 - \omega \min_{y\in\Omega(x)}[\min_{c\in r,g,b}(\frac{\mathcal{I}_c(y)}{A_\infty})] \tag{3}$$

where $\omega = 0.95$ and A_∞ is the airlight color constant.

Since the estimated transmission , obtained by this simple operation, is not able to properly preserve the existing depth transition, further refinement is

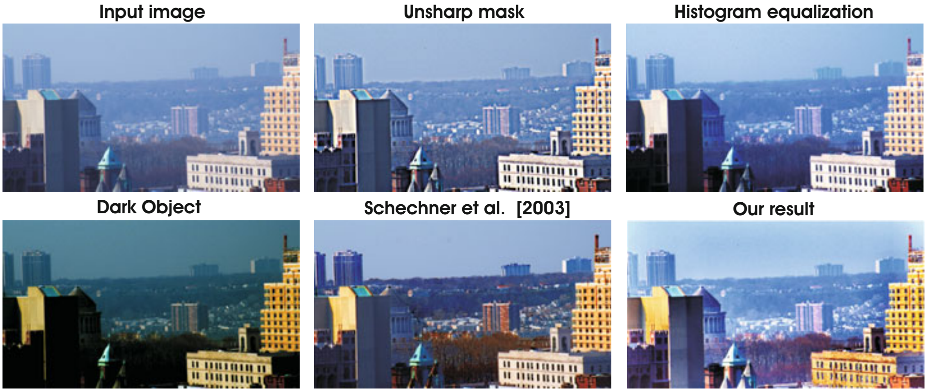


Fig. 2. Standard contrast enhancement filters such as histogram equalization and unsharp mask but also the earlier dark object technique of Chavez [14] are relatively limited to restore such images. Moreover, compared with the polarized-based technique of Schechner et al. [15] that employs two images, our technique is able to recover more effectively the color and finest details.

required. The technique of He et al. [7] also smooths their initial estimate of transmission map but applying an expensive post processing alpha matting strategy that shown to be highly dependent by tweaking the parameters, and as a result, it may yield poor results. The technique of He et al. [7] is limited to properly preserve edges, which is caused mainly by the employed erosion filter during the stage of computing the dark channel.

After we initiate the value of transmission based on dark channel, the next step is to smooth this estimate over large neighborhoods while maintaining the abrupt transmissions. In order to estimate the final accurate transmission (depth) map we design a graphical Markov random field (MRF) model that is optimized by an effective belief propagation-based strategy. This approach is inspired by the work of [13] that demonstrate the utility of belief propagation for several low level vision tasks such segmentation and inpainting.

The MRF framework is defined as a sum of the data costs and discontinuity costs. Considering that $\mathcal{P}$ is the set of pixels in the input image, and $\mathcal{L}$ is a set of labels that corresponds to estimated transmission. f represents a labeling that matches a label $f_p \in \mathcal{L}$ to each pixel from the image $p \in \mathcal{P}$. This definition satisfies the assumption that values need to change smoothly everywhere, except the boundaries between depth discontinuities. As a result, we define the energy function as following:

$$E(f) = \sum_{(p,q)\in\mathcal{N}} V(f_p, f_q) + \sum_{p\in\mathcal{P}} D_p(f_p) \tag{4}$$

where $V(f_p, f_q)$ represents the discontinuity cost and discloses the cost for assigning the labels f_p and f_q to neighboring pixels. $D_p(f_p)$ is referred as the data costs function and represents the cost of assigning the label f_p to pixel p. Finally, $\mathcal{N}$ represents the edges in the four-connected image grid graph. The labeling estimation problem that stands for the minimum energy can be expressed as a maximum a posteriori (MAP) problem.

Since direct computation of marginal probabilities in a MRF framework would take exponential times (intractable) the optimization of MRF cost functions is performed in general by approximate solutions such as belief propagation and graph cuts. Belief propagation represents an efficient way to solve inference problems formulated as maximizing marginal using the idea of passing local massages around the nodes through edges. However, standard belief propagation is relatively slow.

Therefore, we searched for more effective ways to compute the messages. As a result, similarly as in [13] we employ a common linear model to compute the discontinuity cost that increases proportionally linear with the difference between the labels f_p and f_q up to a specified level:

$$V(f_p, f_q) = min(s||f_p - f_q||, d) \tag{5}$$

where s represents the rate of the cost increase (default value is s=1) while d is a constant that indicates when the cost stops to increase (default value is d=20). This strategy reduces the computation time of a single image from $O(k^2)$ to $O(k)$.

Different than most of existing dehazing strategies, in our approach, based on previous observations regarding the transmission map, we constrained directly the smoothness of the depth map. As a result, the data costs function $D_p(f_p)$ has been defined assuming the brightness constancy over initial estimate obtained by dark channel (equation 3):

$$D_p(f_p) = min(||\mathcal{I}_{c-min}(p) - f_p||, \tau) \tag{6}$$

where $\mathcal{I}_{c-min}$ is the min value of the local minimal on the R, G and B channels and τ represents a truncation value (default tau=120). The factor τ assures the robustness of the algorithm to brightness violation and to the occlusions. The labels in this implementation correspond to the transmission values. We defined the cost of assigning a particular transmission for a pixel that is based on the difference between the transmission and the observed value. Theoretically, for this problem we have to consider 256 values (0-255 intensity levels). However, in practice, observed as well by Kratz and Nishino [9], approximating the range to only 100 values is generally adequate.

The effective belief propagation employed in our strategy, reduces substantially the processing time needed to optimize the cost function of equation 4. Each message is a vector of dimension given by the number of possible labels. Our optimization strategy demonstrated to speed up the standard belief propagation algorithm technique from $O(Nk^2T)$ to $O(Nk)$, where N represents the number of image pixels, k represents the number of possible labels and T represents the number of iterations.

3 Results and Discussion

Figure 2 shows comparative results with several approaches. Since the haze degradation effects depend on the distance, standard enhancement filters such as histogram equalization and unsharp mask are relatively limited to restore such images. We also considered the well-known dark object technique [14]. Moreover, our single image restoration technique yields more pleasing results compared even with the method Schechner et al. [15] that is a polarization-based approach that employs two images - the worst and the best polarization states among the existing image versions.

Some more comparative results against the recent single image dehazing techniques are shown in figure 3. Compared with the method of Tan [6] but also with Tarel and Hautiere [8], we can isolate better the transitions between different

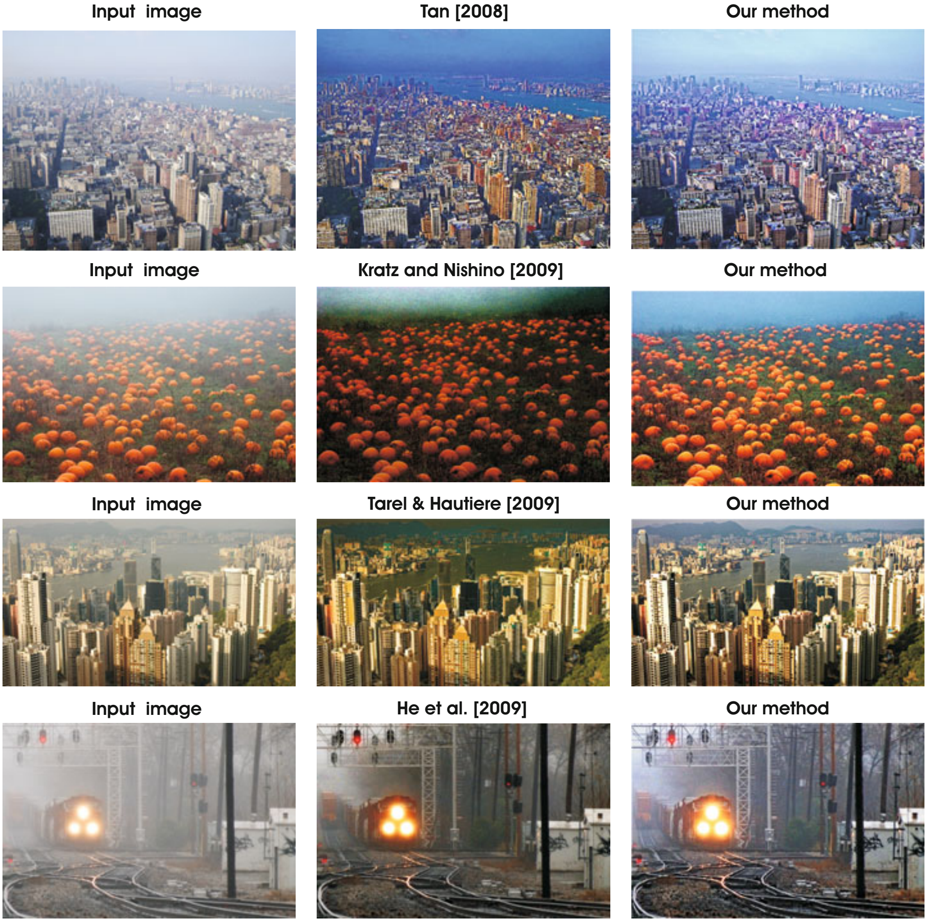


Fig. 3. Comparative results against the recent single image dehazing techniques

depth objects. As well the method of Kratz and Nishino [9] is limited to restore the distant parts of the scene introducing some unpleasing artifacts in those regions (please observe the skyline part of the pumpkins field).

In general our technique is able to yield more accurate results that existing techniques being able to recover the original appearance and details as well of the most distant parts of the scene. However, the main limitation of our technique is given by the initial dark channel estimate of transmission that may be unreliable when the airlight has a similar level with the considered scene object.

In future work we would like to extend our framework as well for more complex scene where the homogeneity of the haze does not hold but also to the problem of enhancing outdoor videos.

References

1. Koschmieder, H.: Theorie der horizontalen sichtweite. In: Beitrage zur Physik der freien Atmosphare (1924)
2. Narasimhan, S., Nayar, S.: Contrast Restoration of Weather Degraded Images. IEEE Trans. on Pattern Analysis and Machine Intell. (2003)
3. Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., Lischinski, D.: Deep photo- Model-based photograph enhancement and viewing. ACM Transactions on Graphics (2008)
4. Treibitz, T., Schechner, Y.Y.: Polarization: Beneficial for visibility enhancement? In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
5. Fattal, R.: Single image dehazing. SIGGRAPH, ACM Transactions on Graphics (2008)
6. Tan, R.T.: Visibility in bad weather from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
7. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
8. Tarel, J.P., Hautiere, N.: Fast visibility restoration from a single color or gray level image. In: IEEE International Conference on Computer Vision (2009)
9. Kratz, L., Nishino, K.: Factorizing scene albedo and depth from a single foggy image. In: IEEE International Conference on Computer Vision (2009)
10. Ancuti, C.O., Ancuti, C., Bekaert, P.: Effective single image dehazing by fusion. In: IEEE International Conference on Image Processing, ICIP (2010)
11. Ancuti, C.O., Ancuti, C., Hermans, C., Bekaert, P.: A fast semi-inverse approach to detect and remove the haze from a single image. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part II. LNCS, vol. 6493, pp. 501–514. Springer, Heidelberg (2011)
12. Henry, R.C., Mahadev, S., Urquijo, S., Chitwood, D.: Color perception through atmospheric haze. Opt. Soc. Amer. A 17, 831–835 (2000)
13. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. International Journal of Computer Vision (2006)
14. Chavez, P.: An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data. Remote Sensing of Environment (1988)
15. Schechner, Y.Y., Narasimhan, S.G., Nayar, S.K.: Polarization-based vision through haze. App. Opt., 511–525 (2003)

Exploiting Image Collections for Recovering Photometric Properties

Mauricio Diaz and Peter Sturm

Laboratoire Jean Kuntzmann & INRIA Grenoble Rhône-Alpes, Montbonnot, France
{Mauricio.Diaz,Peter.Sturm}@inrialpes.fr

Abstract. We address the problem of jointly estimating the scene illumination, the radiometric camera calibration and the reflectance properties of an object using a set of images from a community photo collection. The highly ill-posed nature of this problem is circumvented by using appropriate representations of illumination, an empirical model for the nonlinear function that relates image irradiance with intensity values and additional assumptions on the surface reflectance properties. Using a 3D model recovered from an unstructured set of images, we estimate the coefficients that represent the illumination for each image using a frequency framework. For each image, we also compute the corresponding camera response function. Additionally, we calculate a simple model for the reflectance properties of the 3D model. A robust non-linear optimization is proposed exploiting the high sparsity present in the problem.

Keywords: illumination conditions, reflectance estimation, radiometric calibration, photo collections.

1 Introduction

Capturing the photometric properties of a scene is a complex process and requires exhaustive work. A scene captured in a digital image is completely described, in a photometrical sense, if we are able to represent the objects by an appearance model, if we estimate the illumination conditions during the acquisition time and finally if we know the radiometric calibration of the camera. Once the parameters for the models that describe these processes are estimated, they may be used, for example, in augmented reality, relighting or realistic rendering applications. Several issues arise in the formulation and computation of models for object appearance, lighting and camera radiometric response using only images. The main obstacle is that intensity values registered by the sensor are the result of the interaction between surface geometry, object reflectance, scene illumination and camera properties. In order to estimate one or several of these, it is thus important to take into account or be robust to the respective other factors.

In this work we aim at estimating the photometric description for a particular scene using an unstructured and heterogeneous set of images. The use of photo collections is motivated by the goal of exploiting the richness of appearance variations present in these repositories. Recent works have shown the potential

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 253–260, 2011.

of using large databases in computer vision applications. For example, nowadays it is possible to create an acceptable 3D geometrical model, using images taken under completely casual conditions. But estimating the photometric properties from these datasets is still one of the most difficult conundrums for the experts.

Our problem can be formulated as follows: for a 3D geometry obtained from a set of M unstructured images and that is composed by J surface elements, and given the camera's geometric calibration, we wish to determine the camera response function (CRF) for every image, along with the image illumination conditions and the surface reflectance properties for every surface element. We start from scratch, recovering 3D structure of the scene and the camera pose from an image collection. The points belonging to the recovered structure are called surface elements. Depending on the representation, these points are vertices of a mesh or 3D coordinates of a cloud of points. The 3D reconstruction is done using publicly available tools [17,5]. To represent the image intensities we model the scene radiance as the result of a linear combination of the radiometric camera model, the illumination for each acquisition time and the material reflectance. We estimate the parameters by grouping the unknowns in two subsets: the camera–illumination variables (CRF coefficients and spherical harmonic coefficients) and point variables (albedos). This notation allows us to exploit the high sparsity present in the problem using a robust estimation algorithm.

2 Related Work

Radiometric Calibration. A common strategy used to estimate the CRF consists on posing specific calibration objects (*e.g.* color charts) into the scene at the acquisition time [2,9]. Other methods require multiples images taken under variable exposure times [3,13]. The main drawback of these algorithms is that physical access to the scene during the acquisition must be guaranteed. On the other hand, researchers have proposed approaches exploiting image characteristics that reflect the non-linearity produced by the camera response function. For example, [10] and [11] use low level representations such as edges in regions with constant color to extrapolate the CRF. [14] uses geometric invariants looking for the same goal. Noise present in a simple digital image is used in [18] to infer the CRF. A common point in most of the methods above mentioned is the use of a simple, but realistic model for the CRF. In [7], Grossberg and Nayar proposed an empirical model based on the principal component analysis of real world CRF's. The non-linear radiometric response of the camera is composed by a few coefficients multiplying a precomputed basis. A different camera model is also introduced in [1]. In this work, authors model the CRF as the product of a white-balance transform matrix and a polynomial of fifth degree. In this approach the space of possible CRFs is dramatically large and there is no guarantee that the estimated CRFs correspond to the real ones. In the work [4], authors estimate CRFs using a photo collection, but without inferring information about the illumination or the surface reflectance.

Illumination and reflectance estimation. On the side of illumination estimation, we found also different approaches. The most common method, known also as inverse lighting in the computer graphics community, uses a reflective sphere inserted in the scene to recover natural illumination [19]. Non invasive methods have been possible from the formulation of the "signal processing framework" introduced in [16]. This work has allowed to simplify the integration over the hemisphere of the BRDF and the light source as the multiplication of some coefficients in the appropriate space. For the case of Lambertian surfaces, illumination estimation becomes a simple operation, at least in analytical terms. Other methods estimate at the same time different unknowns. For example, Luong *et al.* [12] use, like our approach, a 3D model to perform radiometric calibration and illumination estimation. They used a linear model for the CRF's and required that several images be taken with the same camera, under controlled conditions. Their model for the illumination consists of a point light source. However, most of the time a linear model for the CRF's is not accurate enough. Haber *et al.* [8] have developed an approach to recover reflectance properties and illumination using a wavelet framework. In this work, authors assume that images extracted from photo collections can be photometrically corrected by mapping with a traditional "gamma correction" curve.

3 Image Formation and Estimation Problem

Image irradiance coming from a Lambertian surface under distant illumination, is a magnitude dependent on the surface orientation and the incoming incident illumination. In this work we assume that illumination sources are distant and can be modeled via an environment map. Also, we assume that studied surfaces are characterized by a Lambertian reflectance with spatially varying albedo. Cast shadowing and interreflections are ignored (we explain later how to alleviate this restriction when applying our algorithm to real world cases). Under these considerations, image irradiance E for a surface element j with albedo ρ_j and normal $\mathbf{n}_j$ under an illumination L is computed by:

$$E(\rho_j, \mathbf{n}_j, L) = \rho_j \int_{\Omega} L(\theta_i, \phi_j) \cos\theta_j d\Omega \ \ , \tag{1}$$

where θ_i and ϕ_j are the inclination and azimuth angles respectively of the light directions, represented in a local coordinate system around the surface normal $\mathbf{n}_j$ and Ω denotes the hemisphere of all possible incoming light directions.

Camera Response Function. The mapping between image irradiance and intensity values is determined by the CRF. A simple but efficient model for the CRF is proposed in [7]. Authors found that CRF's belonging to real world cameras lie in a small part of a theoretical function space that can be spanned using a small basis. This result allows to express the CRF's in terms of N coefficients. For an image i, irradiance E is related to image intensities B by a linear combination of an average CRF h_0 and N principal components h_n:

$$B = f_i(E) = h_0(E) + \sum_{n=1}^{N} w_{in} h_n(E) \ \ . \tag{2}$$

The basis CRF's are thus represented as polynomials of degree D: $h_n(E) = \sum_{d=0}^{D} c_{nd}E^d$. Note that according to [7], the h_n are expressed relative to normalized brightness and irradiance, such that $c_{n0} = 0$ and $\sum c_{nd} = 1$ for all $n = 1 \cdots N$. The degree of the polynomials was chosen for an adequate representation of the curves forming the basis, which was obtained with $D = 9$. These polynomials are known; unknown are the coefficients w_{in} of their linear combination (2).

Global Illumination Model. Representing illumination is a key factor to obtaining 3D models from real world images. An approach that has recently gained importance is to analyze illumination conditions in a frequency framework. We represent the lighting falling on a surface element by a hemisphere centered in the normal position, using spherical harmonics. In this context, Ramamoorthi [15] has shown that for convex Lambertian surfaces, the image irradiance is well represented by a linear combination of coefficients and an orthonormal basis set. This basis is composed by the spherical harmonics $Y_{lm}(\mathbf{n_j})$ rotated around the plane defined locally by the surface normal. The indices obey $l \geq 0$ and $-l \leq m \leq l$ (there are $2l + 1$ basis functions for a given order l). Equation (1) is expressed in terms of this basis as a combination of L coefficients as follows:

$$E(\rho_j, \mathbf{n}_j, E_{lm}) = \rho_j \sum_{l=0}^{L} \sum_{m=-l}^{l} E_{lm} Y_{lm}(\mathbf{n_j}) \ . \tag{3}$$

Given a geometric reconstruction of a 3D surface, the normal $\mathbf{n}_j$ becomes a known value and the inverse lighting estimation problem is reduced to the computation of the coefficients E_{lm} that best fit the basis of spherical harmonics rotated at the point normal. Thus, image irradiance per channel is eventually parametrized as a function of the material albedo and the illumination spherical coefficients: $E(\rho_j, \mathbf{n}_j, E_{lm})$.

Estimation Problem. Having defined a linear representation for the CRF's and for the illumination, the intensity value for a surface element j is calculated using the normal at the point $\mathbf{n}_j$ and equations (2) and (3). Since we are dealing with a set of images taken with different illumination conditions but keeping static surface properties, the image irradiance emitted by a surface element depends on the lighting and the material reflectance properties. Additionally, each camera has a different CRF. Therefore, if we denote B_{ij} as the intensity value for a particular color channel, describing the surface element j and the image i, the normalized intensity is estimated by:

$$\tilde{B}_{ij} = f_i(E(p_j, \mathbf{n}_j, E^i_{lm})) = f_i(\rho_j \sum_{l=0}^{L} \sum_{m=-l}^{l} E^i_{lm} Y_{lm}(\mathbf{n_j})) \ . \tag{4}$$

To simplify notations we express equation (4) as a vector multiplication, where the vector $\mathbf{E_i}$ is the set of 9 spherical coefficients that describe illumination in camera i (E^i_{lm} with $L = 2$ and $-l \leq m \leq l$) and $\mathbf{Y_j}$ is the spherical harmonics basis expressed in terms of the coordinate plane around the normal in point j. Combining this vector multiplication with the equation (2), intensity is:

$$\tilde{B}_{ij,\mathrm{ch}} = h_0(\rho_{j,\mathrm{ch}} \mathbf{E}_{i,\mathrm{ch}}^{\mathrm{T}} \mathbf{Y}_j) + \sum_{n=1}^{N} w_{in,\mathrm{ch}} h_n(\rho_{j,\mathrm{ch}} \mathbf{E}_{i,\mathrm{ch}}^{\mathrm{T}} \mathbf{Y}_j) \ , \tag{5}$$

where ch is a suffix indicating the color channel to evaluate (red, green, blue).

Let us denote the vector $\mathbf{a}_i = \left[\mathbf{E}_{iR}^{\mathrm{T}}\ \mathbf{E}_{iG}^{\mathrm{T}}\ \mathbf{E}_{iB}^{\mathrm{T}}\ \mathbf{w}_{iR}^{\mathrm{T}}\ \mathbf{w}_{iG}^{\mathrm{T}}\ \mathbf{w}_{iB}^{\mathrm{T}}\right]^{\mathrm{T}}$ describing illumination and CRF per channel. The vector $\mathbf{E}_{i,\mathrm{ch}}$ has dimension O while the vector $\mathbf{w}_{i,\mathrm{ch}}$ has N components. Then, the dimension of $\mathbf{a}_i$ is $3{\times}O{+}3{\times}N$. The vector $\mathbf{b}_j$ of dimension 3 represents the surface material albedo: $\mathbf{b}_j = \left[\rho_{jR}\ \rho_{jG}\ \rho_{jB}\right]$. We define our estimation function $B(\mathbf{a}_i, \mathbf{b}_j)$ as a function from $\Re^{3\times(O+N+1)} \rightarrow \Re^3$. To estimate the unknowns $\mathbf{a}_i$, $\mathbf{b}_j$, we minimize the difference between the observed and predicted intensity values. An optimal solution to calculate the unknowns requires a full non-linear optimization of the cost function, defined as the squared difference between the measured intensity and its correspondent estimation. Given a set of J surface elements projected in M images, the optimization problem is formulated as follows:

$$\min_{\mathbf{a}_i,\mathbf{b}_j} \sum_{i=1}^{M} \sum_{j=1}^{J} \left(B_{ij} - v_{ij} \hat{B}(\mathbf{a}_i, \mathbf{b}_j)\right)^2 \ . \tag{6}$$

The scalars v_{ij} are booleans, a value of 1 indicating that surface element j is visible in image i, otherwise the value being 0. Note that the unknowns can not be estimated without ambiguity: albedos ρ_j and lighting coefficients $\mathbf{E}_i$ can only be estimated up to one global scale factor. Additionally, we impose a constraint on the monotonicity of the estimated CRF's (plausible CRF's are monotonic).

In our experiments, we initialized the optimization algorithm using a vector $\mathbf{b}_j$ containing the mean values of all observed intensities of surface element j. In the case of the vector $\mathbf{a}_i$, the spherical harmonic coefficients ($\mathbf{E}_i$) are initialized with ones while the CRF coefficients ($\mathbf{w}_i$) correspond to a vector of zeros (the initial CRF's are $f_i = h_0$). To avoid the problems related with outliers (*i.e.* intensity samples not included in the 3D model, cast-shadowing, imperfections on the camera pose estimation, surface materials with specular reflection properties, interreflexions, *cf.* section 3), the least squares minimization presented in equation (6) is transformed to a robust estimation problem using the Iterative Reweighted Least Squares (IRLS) algorithm [20].

4 Results

We evaluate the performance of our algorithm in real world conditions using two databases. Both collections target architectural structures in outdoor environments. Images were taken during different periods of the day with natural illumination and different cameras. For the first database (DB1) we had access to the scene and the acquisition equipments. This database is composed by 120

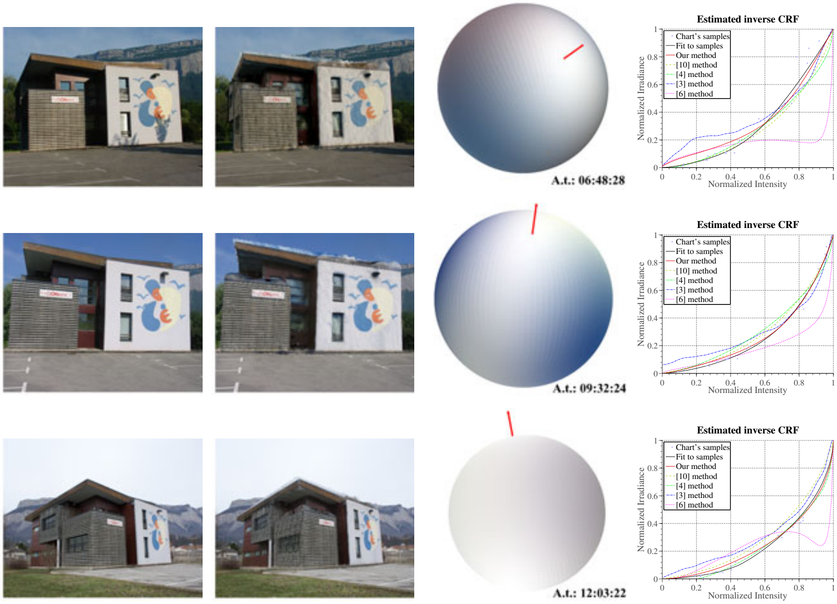


Fig. 1. The 1st column shows some image samples of DB1 used in the reconstruction. On the 2nd column we present corresponding rendered images using the estimated CRFs, illumination conditions and albedos. 3rd and 4th column show representations of the estimated parameters (*cf.* text).

images used to reconstruct a mesh with 112,504 vertices. We got 10 extra images containing a color checker board inside the scene, also we took multiple exposure images for these extra samples, just seconds after the image used for 3D reconstruction was taken. The second database (DB2) was collected from an internet repository and 928 images were used to reconstruct a mesh with 80,444 vertices. In our implementation, we modeled the CRF with 3 coefficients ($N = 3$) and we used 9 spherical harmonics coefficients to model the illumination ($O = 9$). The number of parameters to estimate is $3 \times (J + M \times (O + N))$.

CRF Estimation. To validate our results, we compared the estimated CRF with the ground truth, obtained by placing a color chart in the scene depicted when usin DB1. Four column of figure 1 shows our results with the CRF computed using the *HDRShop* software [3] and the technique described in [6]. These algorithms present poor estimations due to the difficulty of having perfectly aligned images when shooting outdoor scenes (shadows, reflections may change rapidly). CRF estimation using the single image method described in [10] is included. We also show the CRF obtained with the algorithm presented in [4]. When using DB2, we compare our estimated CRF with results of algorithms that do not require physical access to the scene [10,4].

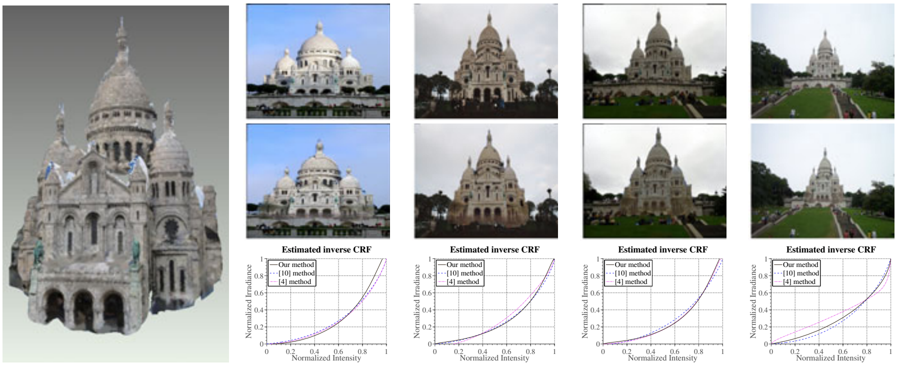


Fig. 2. At the left, the 3D model rendered with the average of the pixel intensities. At the right, four sample images of DB2 and the projection of 3D model with estimated parameter over the original background. Estimated CRF is shown in the third row.

Illumination Estimation. The performance of our technique when estimating the illumination is evaluated by rendering a synthesized image using the calculated lighting. We evaluated the root mean square difference (RMS) between the rendered and original images. These values are calculated for intensities scaled between zero and one. Using DB1 the median RMS difference is 1.2% of the full pixel intensity while using DB2 is around 1.8%. We performed a cross-validation test, using one subset of the database and rendering the synthesized images with the illumination and CRF calculated in a different subset (see figure 1, columns 1-2). For this case the median RMS difference was 7.7% for DB1 and around 23% for DB2. When using DB2, RMS error increases, since the original images contain sometimes pedestrians or objects not taken into account in the 3D model. Third column of figure 1 represents the computed spherical harmonics projected on a sphere viewed from the same point of view as the original images. An arrow indicates the maximum point, the direction where the illumination is strongest. It was mentioned in section 3 that albedos and illumination coefficients can only be estimated up to a global scale factor. This is the case for all three color channels. Hence, in order to display RGB illumination models and surface colors, we first have to estimate the ratios of these scales, between color channels. These scales are calculated by selecting a portion of the sky and projecting its pixels on the sphere. We found the right scale by fitting the spherical harmonics coefficients to the color of some manually selected pixels projected on the surface of the sphere. Images where the presence of a directional light source can be deduced form shadows show a correct estimation of the illumination direction. For cloudy skies, illumination is more uniform (*cf.* third image) and the maximum is less pronounced.

5 Discussion and Conclusion

We have presented a method to estimate jointly photometric properties for a scene. The computed CRFs show good performance, similar to state–of–the–art methods, with the added value that our method provides illumination and reflectance information. Illumination estimation presents a suitable environmental lighting to render new views for the captured scene. One limitation remains on the use of Lambertian reflectance. Although this constraint is contoured using a robust optimization, if we wish to calculate accurately the surface reflectance properties, a more complete model must be used. In that case, the number of parameters to compute may increase dramatically because the illumination and the reflectance interact over all directions of the upper hemisphere centered at the normal of a surface point. Other frameworks may be explored.

References

1. Chakrabarti, A., Scharstein, D., Zickler, T.: An empirical camera model for internet color vision. In: BMVC (2009)
2. Chang, Y.C., Reid, J.: RGB calibration for color image analysis in machine vision. IEEE Trans. on Image Processing 5(10), 1414–1422 (1996)
3. Debevec, P., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: SIGGRAPH (August 1997)
4. Diaz, M., Sturm, P.: Radiometric calibration using photo collections. In: ICCP (2011)
5. Furukawa, Y., Ponce, J.: Patch-based multi-view stereo software. Web (2009), http://grail.cs.washington.edu/software/pmvs
6. Grossberg, M.D., Nayar, S.K.: What is the space of camera response functions? In: CVPR, vol. 2, p. 602 (2003)
7. Grossberg, M.D., Nayar, S.K.: Modeling the space of camera response functions. IEEE Trans. on PAMI 26(10), 1272–1282 (2004)
8. Haber, T., Fuchs, C., Bekaer, P., Seidel, H.P., Goesele, M., Lensch, H.: Relighting objects from image collections. In: CVPR, pp. 627–634 (June 2009)
9. Ilie, A., Welch, G.: Ensuring color consistency across multiple cameras. In: ICCV, pp. 1268–1275 (2005)
10. Lin, S., Gu, J., Yamazaki, S., Shum, H.Y.: Radiometric calibration from a single image. In: CVPR 2004, vol 2, pp. II–938 – II–945 (2004)
11. Lin, S., Zhang, L.: Determining the radiometric response function from a single grayscale image. In: CVPR 2005, vol 2, pp. 66 – 73 (2005)
12. Luong, Q., Fua, P., Leclerc, Y.: The radiometry of multiple images. IEEE Trans. on PAMI 24(1), 19–33 (2002)
13. Mitsunaga, T., Nayar, S.K.: Radiometric self calibration. In: CVPR (July 1999)
14. Ng, T.T., Chang, S.F., Tsui, M.P.: Using geometry invariants for camera response function estimation. In: CVPR, pp. 1 – 8 (2007)
15. Ramamoorthi, R.: Modeling illumination variation with spherical harmonics. Face Processing: Advanced Modeling and Methods (2002)
16. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: SIGGRAPH, pp. 117–128 (2001)
17. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. In: IJCV (January 2008)
18. Takamatsu, J., Matsushita, Y., Ikeuchi, K.: Estimating camera response functions using probabilistic intensity similarity. In: CVPR, pp. 1 – 8 (2008)
19. Yu, Y., Debevec, P., Malik, J., Hawkins, T.: Inverse global illumination: recovering reflectance models of real scenes from photographs. In: SIGGRAPH (1999)
20. Zhang, Z.: Parameter estimation techniques: A tutorial with application to conic fitting. In: Image and Vision Computing (January 1997)

Human Visual System for Complexity Reduction of Image and Video Restoration

Vittoria Bruni[1], Daniela De Canditiis[2], and Domenico Vitulano[2]

[1] Dept. of SBAI, Faculty of Engineering, University of Rome ”La Sapienza”
bruni@dmmm.uniroma1.it
[2] Istituto per le Applicazioni del Calcolo ”M. Picone” (CNR) - Rome
d.decanditiis-d.vitulano@iac.cnr.it

Abstract. This paper focuses on the use of Human Visual System (HVS) rules for reducing the complexity of image and video restoration algorithms. Specifically, a fast HVS based block classification is proposed for distinguishing image blocks where restoration is necessary from the ones where it is useless. Some experimental results on standard test images and video sequences show the capability of the proposed method in reducing the computing time of de-noising algorithms, preserving the visual quality of the restored sequences.

Keywords: Human Visual System, Block classification, Complexity Reduction, Image and Video Restoration.

1 Introduction

A wide literature has been dedicated to image and video restoration with particular attention to both quality and computational effort. In particular, the latter is fundamental for real time applications and for codecs transportability on common devices. Even though the more recent literature has taken a great advantage of using Human Visual System (HVS) mechanisms for improving coding performance [1,2], for guiding image enhancement [3,4] or for detecting image anomalies [5], to the best of authors knowledge, the benefit from using HVS rules in restoration schemes for computational purposes has not yet been investigated.

This paper aims at presenting a fast HVS-based blocks classification to be embedded into any de-noising algorithm in order to reduce its computing time. It is related to the Structural SIMilarity index (SSIM) [6], that is used for evaluating the visual difference between two images. The proposed classification actually aims at distinguishing between: *i)* blocks where both noise and motion are perceived, *ii)* blocks where noise is perceptible while motion is not and *iii)* blocks where human eye is insensitive to both noise and motion. The goal is to adapt the restoration process to each block, according to its visible content. In particular, de-noising is inhibited if noise is imperceptible, while motion vector is not estimated if motion is not perceived. The computational gain depends on the processed frames and the selected restoration scheme. The larger the number of blocks where operations are inhibited and the more negligible the

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 261–268, 2011.

additional computational effort of classification, the higher the computational saving. Therefore, the latter is convenient whenever motion estimation and/or de-noising are time consuming. First results in case of moderate noise show that the computational cost of restoration can be reduced of about 50% on average, without compromising the visual quality of the restored images.

The paper is organized as follows. Blocks classification is presented in Section 2 while some evaluations about its computing time are contained in Section 3. Finally, some experimental results on standard test images and video sequences along with concluding remarks are the topic of Section 4.

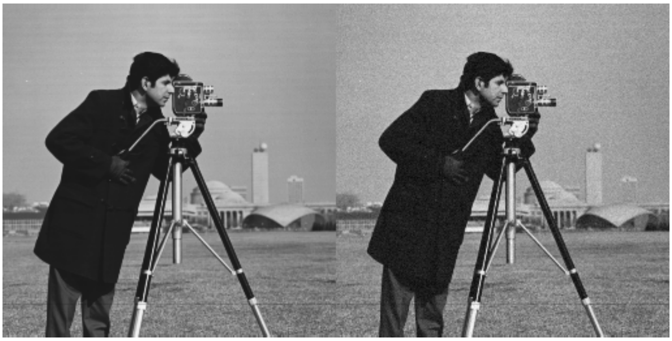


Fig. 1. $256 \times 256 \times 8$ bits *Cameraman image*: original (*left*) and noisy (*right*) corrupted by Gaussian noise $N(0, \sigma^2)$ with $\sigma^2 = 225$

2 Block Classification

Let us start with a typical example of image corrupted by additive Gaussian noise, as shown in Fig. 1. Although noise spreads over the whole image, it is clearly visible on flat regions (e.g. on the sky) and in correspondence to the object contours (e.g. on the man shoulder), while it is not perceived on the grass since it is masked by grass texture. It is obvious that masking occurs whenever noise amplitude does not exceed image components. For that reason, additive and signal independent noise with moderate variance will be considered in the sequel. On the other hand, motion of a scene is perceived if some objects of the scene move, i.e. if their contours (edges) change their location from one frame to another one. Bearing in mind these observations, frame (or image) blocks can be classified as follows: *i)* ***flat block*** — only noise is perceived, then only de-noising is required; *ii)* ***textured block*** — both noise and motion are not perceptible, then any operation is necessary; *iii)* ***edge block*** — both noise and motion are perceived, then both de-noising and motion estimation are required. Representative examples are shown in Fig. 2.

The literature offers a large variety of image blocks classification methods [7,8,9] that are able to distinguish between flat, textured and edge blocks. However, they often involve expensive data transformations as well as sophisticated statistical models that are too computationally demanding for our task. On the other hand, some recent neurological studies [10,11] have shown that human observation process is mainly guided by first and second order statistical moments. It turns out that a visual perception-based classification procedure can be successfully derived from a proper combination of image local mean and variance.

The Structural SIMilarity index [6] (SSIM) is a perception based reference measure that compares the original image with its corrupted copy using corresponding local means and variances. In particular, if I and J are corresponding blocks of two images to be compared, then their visual similarity index is

$$SSIM(I,J) = \underbrace{\frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1}}_{luminance\ adaptation} \underbrace{\frac{2\sigma_I\sigma_J + C_2}{\sigma_I^2 + \sigma_J^2 + C_2}}_{contrast\ masking} \underbrace{\frac{\sigma_{IJ} + C_3}{\sigma_I\sigma_J + C_3}}_{spatial\ correlation}, \qquad (1)$$

where μ_I, μ_J, σ_I and σ_J respectively are the sample means and standard deviations of I and J, σ_{IJ} is the sample covariance between I and J and C_1,C_2 and C_3 are numerical stabilizing constants such that $0 \leq SSIM(I,J) \leq 1$. The first two terms of SSIM give the difference in terms of luminance mean and variance, while the third one measures the structural difference in terms of blocks covariance. The larger SSIM value, the more similar the visual appearance of the two blocks. For example, noise masking effect in textured regions gives a large $SSIM$ value (0.9583) in the bottom-right part (white rectangle) of the images in Fig. 3. On the contrary, $SSIM$ is small (0.2829) in correspondence to the sky (flat region in the black rectangle), since noise is more visible.

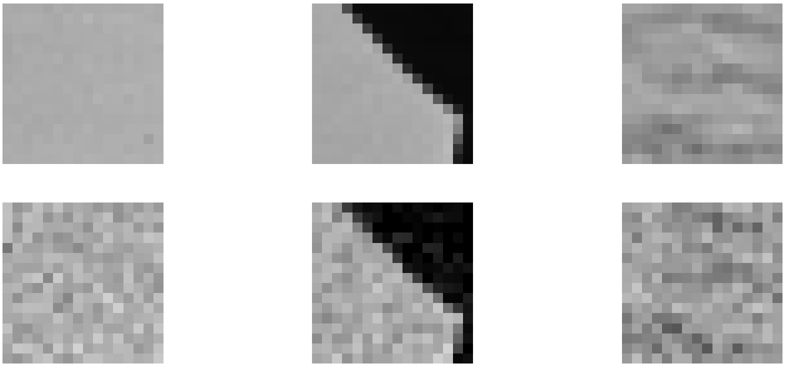


Fig. 2. Flat, edge and textured blocks: original (*top*), noisy (*bottom*)

Two tests are then required for block classification: the first one checks the flatness of the block, i.e. if its noisy version is visually similar to pure noise; the second one evaluates the visual homogeneity of the block — if block luminance content is not stationary, with high probability it contains edges.

First test. It evaluates the visual similarity between the noisy block B and a noisy flat block $B_\nu = \mu_B + \nu$, having the same noise ν and average μ_B of B. In this case, the first term of $SSIM(B, B_\nu)$ is equal to 1 and the empirical covariance $\sigma_{B\,B_\nu}$ is σ^2, then $SSIM$ comes from a comparison between the variance σ_B^2 of B and noise variance σ^2. This is an intuitive way of checking block flatness — see for example [8], where the energy of the noisy block turns out to be its sample variance. However, troublesome choices of appropriate energy thresholds can be avoided using the χ^2 statistical test: it tests whether the variance of B is σ^2, under the assumption that B is normally distributed. Therefore,

$$\text{if} \quad \sum_{i=1}^{N}(B(i)-\mu_B)^2/\sigma^2 \quad < \quad Z_{\chi^2_{N-1,1-\alpha}}, \quad \text{then } \textit{the block is flat}, \tag{2}$$

where N is the block size, $Z = F^{-1}_{\chi^2_{N-1}}(1-\alpha)$, with F the cumulative distribution function of a chi-square with $N-1$ degrees of freedom, while α is the level of the test. α gives the confidence in properly classifying B with probability $1-\alpha$, when the block is really flat. If the block is not flat (e.g. textures) the test can fail and its effectiveness is directly evaluated on the restoration results and the required computing time.

Second test. The second test aims at establishing if a not flat block is visually homogeneous. Therefore, a partition $\{b_i\}_{i=1,\ldots,k}$ of B into k distinct sub-blocks and a straightforward generalization of SSIM to more than two blocks are required. If the k sub-blocks are visually similar, then B contains a texture. Otherwise, at least one sub-block is visually different from the others and B contains edges with high probability. Specifically, if b_is are visually similar, then there exists a positive constant T: $0 \leq T < 1$ such that

$$\frac{2\sum_{i<j}\mu_{b_i}\mu_{b_j}}{(k-1)(\mu_{b_1}^2+\cdots+\mu_{b_k}^2)} \; \frac{2\sum_{i<j}\sigma_{b_i}\sigma_{b_j}}{(k-1)(\sigma_{b_1}^2+\cdots+\sigma_{b_k}^2)} \quad > \quad T. \tag{3}$$

The closer T to 1, the more similar b_is. Eq. (3) involves just the first two terms of SSIM. Spatial correlation is expensive and not really useful in this case, since similar sub-blocks are not required to have exact spatial correspondences. Moreover, for normally distributed luminance values, as it often happens for textures, the comparison of the first two sample moments is a good distribution similarity index. The choice of a threshold T close to 1 should prevent us against misclassification of edge blocks as textured blocks. The opposite would be less serious for the final quality of the restored image.

3 Computational Cost

In this section, **a**, **m**, **d** and **c** respectively denote additions, multiplications, divisions and comparisons. N is the total number of pixels in the block B, k is the number of sub-blocks $\{b_i\}$, $n = N/k$ is the number of pixels in each b_i, μ_i is the luminance mean of b_i and $\sigma_B^2 = \frac{1}{N-1}\left(\sum_{i=1}^{k}(\sum_{j=1}^{n} b_{i,j}^2) - N\mu_B^2\right)$ is the

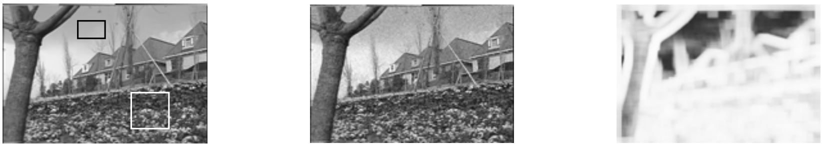


Fig. 3. 40th frame of Flower sequence: black and white rectangles respectively include flat and textured regions *(left)*; noisy frame with additive Gaussian noise with variance $\sigma^2 = 225$ (*middle*); point-wise SSIM index between original and noisy frame (*right*).

variance of B. The flatness test in eq. (2) requires $2N-1\mathbf{a}$, $k+2\mathbf{d}$, $N+3\mathbf{m}$ and $1\mathbf{c}$. In fact, by rewriting eq. (2) as follows

$$\frac{(N-1)\sigma_B^2}{\sigma^2} = \frac{N-1}{\sigma^2}\left(\sum_{i=1}^{k}(\sum_{j=1}^{n} b_{i,j}^2) - N\left(\frac{\sum_{i=1}^{k}\mu_i}{k}\right)^2\right) < Z_{\chi^2_{N-1,1-\alpha}},$$

$k(2n-2)\mathbf{a}$, $k\mathbf{d}$ and $kn\mathbf{m}$ are necessary for the computation of μ_i and $\sum_{j=1}^{n} b_{i,j}^2$, while $2k-1\mathbf{a}$, $2\mathbf{d}$, $3\mathbf{m}$ and $1\mathbf{c}$ are for the left hand side of previous inequality.

The visual homogeneity test in eq. (3) requires $(k^2+2k-4)\mathbf{a}$, $(k^2+5)\mathbf{m}$, $k+1\mathbf{d}$, $1\mathbf{c}$ and k **sqr**, where **sqr** is the square root. In fact, only $k\mathbf{d}$, $k\mathbf{m}$, $k\mathbf{a}$ and k **sqr** are necessary for the evaluation of σ_i, σ_i^2 and μ_i^2, since $\sum_{j=1}^{n} b_{i,j}^2$ and μ_i have already been computed in the previous step. Finally, if 16 operations are assigned to the square root computation, according to Bakhshali algorithm [12], the whole classification requires

$$3 + \frac{2k^2+20k+8}{N} \quad \text{operations per pixel } (opp). \tag{4}$$

Let us now consider the gain G in terms of computational cost whenever the proposed classification is embedded in an image or video de-noising scheme. The computational effort of the restoration technique alone is $(C_{den} + C_{ME})N_{tot}$, where C_{den} and C_{ME} respectively are the computing time for de-noising and motion estimation in a single block, while N_{tot} is the number of blocks contained in the video sequence to be processed. The proposed classification splits N_{tot} into $N_{flat} + N_{textured} + N_{edge}$, respectively the number of flat, textures and edge blocks of the whole video sequence. Hence, the complexity of the HVS based classification embedded in a restoration framework is $C_{den}N_{flat} + (C_{den} + C_{ME})N_{edge} + C_{classification}N_{tot}$, where $C_{classification}$ is defined in eq. (4). The computational gain is then $G = \frac{(C_{den}+C_{ME})N_{tot}}{C_{den}N_{flat}+(C_{den}+C_{ME})N_{edge}+C_{classification}N_{tot}}$.

It turns out that if $C_{classification}$ is negligible with respect to $(C_{den} + C_{ME})$ and the number of edge blocks is not predominant in the video sequence, the gain in computing time is not negligible at all.

4 Experimental Results and Concluding Remarks

The proposed HVS-based blocks classification has been tested on a large data set of images. In all tests 16×16 blocks ($N = 256$) and 8×8 sub-blocks ($k = 4, n = 64$) have been adopted while $\alpha = .1$ and $T = .95$ have been set in eqs. (2) and (3). For the same set of parameters, the proposed classification requires 3.4688 operations per pixel. Some results are shown in Fig. 4, where images having a different amount of flat, textured and edge information have been considered. As also *Cameraman* SSIM matrix shows, image blocks are properly classified according to both block content and amount of noise. In particular, the lower regions in *FlowerGarden* sequence are correctly classified as textures; the same happens for the grass in *Cameraman* image and for the sea in *Coastguard*. On the contrary, in *Tennis* sequence the flat table is separated from the wall, whose texture hides moderate noise. Finally, the cameraman contour and tripod are correctly recognized as edge regions in *Cameraman* image as well as the arm and the ball in *Tennis* sequence. It is worth observing that as the noise level increases, the number of flat blocks increases while the number of edge blocks diminishes since the masking effect is less evident — see Table 1. In fact, noise becomes predominant with respect to image content so that it becomes more visible than the actual image information.

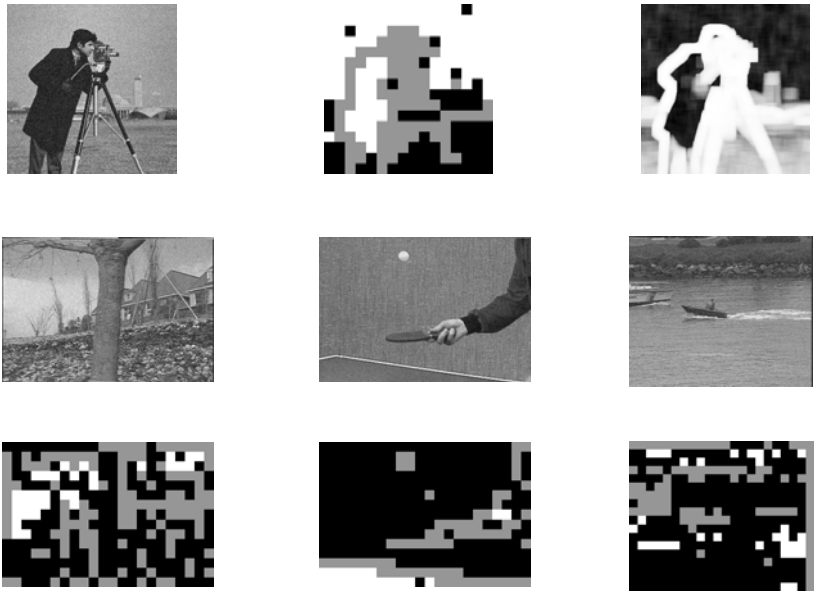


Fig. 4. Top) *Cameraman* ($\sigma = 15$) image, its classification map and SSIM matrix. **Central)** Noisy *FlowerGarden* ($\sigma = 15$), *Tennis* ($\sigma = 10$) and *Coastguard* ($\sigma = 15$) frames. **Bottom)** Their HVS-based classification maps — flat, textured and edge blocks respectively are white, black and light gray.

Table 1. Number of blocks in each class for different videos and noise variances

No. of blocks	Noisy Video sequence								
	Flower ($\sigma = 5$)	Flower ($\sigma = 15$)	Flower ($\sigma = 30$)	Coast ($\sigma = 10$)	Coast ($\sigma = 15$)	Tennis ($\sigma = 10$)	Tennis ($\sigma = 20$)	Foreman ($\sigma = 5$)	Foreman ($\sigma = 15$)
N_{flat}	374	1191	2147	1199	5556	2420	12862	5428	19597
N_{edge}	14209	12422	9994	58012	49301	15626	12746	78235	53933
$N_{textured}$	5217	6187	7659	59193	63547	31124	23562	34741	44874
N_{tot}	19800	19800	19800	118404	118404	49170	49170	118404	118404

Table 2. SNR, SSIM and gain in computing time for three video sequences using FA+BMA [13] with and without HVS-based classification

Video sequence ($\sigma = 15$)	Noisy		Without cl.		With cl.		comp. gain
	SNR	SSIM	SNR	SSIM	SNR	SSIM	
Foreman (150 fr.)	21.07	0.6503	23.38	0.7250	24.34	0.7881	2.16
Coastguard (300 fr.)	19.37	0.7160	21.00	0.7689	21.22	0.7809	2.36
Flower (61 fr.)	20.06	0.8064	21.01	0.8431	21.23	0.8875	1.50

Table 2 shows the computational gain that has been reached by embedding the proposed block classification into a simple frame averaging (FA) video de-noiser combined with block matching algorithm (BMA) for motion estimation [13]. In case of moderate noise variance, the more textured the video sequence (large $N_{textured}$), the higher the computational gain. Moreover, the presence of a small number of edge blocks strongly reduces the computational effort of the restored framework whenever motion estimation is computationally demanding in terms of number of operations per pixel, as it is the case of BMA. It is worth stressing that the inclusion of HVS based classification in the restoration algorithm does not compromise the visual quality of the restored images. On the contrary, it sometimes allows to slightly improve it, as SNR (Signal to Noise Ratio) and SSIM values in the last two columns of Table 2 show. In fact, destructive denoising operations in correspondence to textured regions are avoided, reducing over-smoothing whenever noise is not visually annoying as well as misalignments due to wrong motion estimation — see, for example, the right side of the tree trunk in Fig. 5. It is also worth stressing that blocking artifacts can occur in case

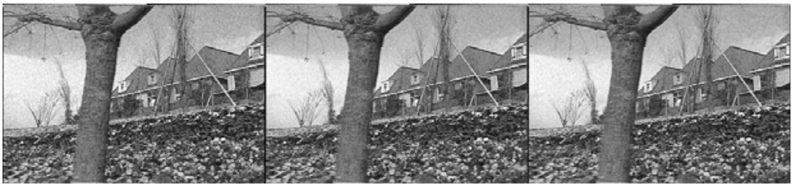


Fig. 5. From left to right) 20^{th} noisy frame of FlowerGarden ($\sigma = 15$) and its denoised copy using FA without and with HVS-based block classification

of high levels of noise. These artifacts can be reduced working with overlapping blocks in the denoising procedure.

Future research will be oriented to refining the proposed block classification by introducing more complex rules of human vision as well as to investigate about the possibility of a further reduction of its computational effort.

References

1. Hontsch, I., Karam, L.J.: Adaptive image coding with perceptual distortion control. IEEE Trans. on Image Processing 11(3), 213–222 (2002)
2. Watson, A.B., Yang, G.Y., Solomon, J.A., Villasenor, J.: Visibility of wavelet quantization noise. IEEE Trans. on Image Processing 6(8), 1164–1174 (1997)
3. Agaian, S., Silver, B., Panetta, K.: Transform coefficient histogram based image enhancement algorithms using contrast entropy. IEEE Trans. on Image Processing 16(3), 741–758 (2007)
4. Panetta, K., Wharton, E.J., Agaian, S.S.: Human visual system-based image enhancement and logarithmic contrast measure. IEEE Trans. on Syst., Man and Cyber. 38(1), 174–188 (2008)
5. Bruni, V., Vitulano, D.: A generalized model for scratch detection. IEEE Trans. on Image Processing 13(1), 44–50 (2004)
6. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. on Image Processing 13, 600–612 (2004)
7. Tong, H.H.Y., Venetsanopoulos, A.N.: A perceptual model for jpeg applications based on block classification, texture masking and luminance masking. In: Proc. of International Conference on Image Processing (ICIP 1998), vol. 3 (1998)
8. Wong, T.S., Bouman, C.A., Pollak, I., Fan, Z.: A document image model and estimation algorithm for optimized JPEG decompression. IEEE Trans. on Image Processing 18(11), 2518–2535 (2009)
9. Zhang, X., Lin, W., Xue, P.: Just-noticeable difference estimation with pixels in images. Journal of Visual Communication and Image Representation 19, 30–41 (2008)
10. Monte, V., Frazor, R.A., Bonin, V., Geisler, W.S., Corandin, V.: Independence of luminance and contrast in natural scenes and in the early visual system. Nature Neuroscience 8(12) (2005)
11. Frazor, R.A., Geisler, W.S.: Local luminance and contrast in natural in natural images. Vision Research 46, 1585–1598 (2006)
12. Channabasappa, M.N.: On the square root formula in the Bakhshali manuscript. Indian J. History Sci. 11(2), 112–124 (1976)
13. Shi, Y.Q., Sun, H.: Image and video compression for multimedia engineering: fundamentals, algorithms and standards. CRC Press, Boca Raton (2000)

Optimal Image Restoration Using HVS-Based Rate-Distortion Curves

Vittoria Bruni[1], Elisa Rossi[2], and Domenico Vitulano[2]

[1] University of Rome 'Sapienza', Dept. of SBAI, Faculty of Engineering, Via A.
Scarpa 16, 00161 Rome, Italy
bruni@dmmm.uniroma1.it

[2] Istituto per le Applicazione del Calcolo "M. Picone", C.N.R., Via dei Taurini 19,
00185 Rome, Italy
{rossi,vitulano}@iac.rm.cnr.it

Abstract. This paper proves that the Jensen-Shannon Divergence (JSD) is a good information theory measure of the visibility cost of a degraded region in a pictorial scene. Hence, it can be combined with Michelson contrast for building a visual rate-distortion curve. The latter allows to optimize parameters of restoration algorithms. Some results on both synthetic and real data show the potential of the proposed approach.

Keywords: Visual Rate-Distortion, Jensen-Shannon Divergence, Human Perception, Visual Contrast, Image Restoration, Occam Razor.

1 Introduction

In the last years, an increasing research effort has been devoted to models and techniques based on Human Perception (HP). Within image processing field, it seems that HP allows to easily optimize systems designed for acquisition, compression, restoration etc. [1]. The optimization of parameters in a given framework is not a novel topic in image processing. For example, the Occam razor (*"the simplest explanation is most likely the correct one"*) offers a simple and straightforward solution in different optimization problems searching for optimal parameters as trade-off between rate and distortion. In its original version [2], it was successfully applied to estimate the amount of noise in corrupted images in order to properly set the parameters of the successive de-noising procedure. Unfortunately, it is not always possible to directly measure the strength of any kind of distortion by means of simple and significant parameters. Notwithstanding, HP seems to offer the possibility of extending Occam Razor since it measures the distortion perceived by human observers. In fact, in a visibility context, the "strength" of image degradation i.e., the distortion in the rate-distortion curve, can be easily evaluated using various contrast measures (Weber, Michelson etc. [1]). On the contrary, a measure that is able to really quantify visual information in terms of compressed size (bits per pixel — *bpp*) still misses. As a matter of fact, there have been various attempts in the literature for describing a visual scene using information theory

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 269–276, 2011.

language. Some examples concerning the combination of contrast and entropy-based measures for compression purposes can be found in [3,4,5], where optimization of bit allocation, quantization masking and bit saving in video transmission via a fovea-based model are dealt with. More general approaches 'for capturing' real world visive structures as well as for visual quality assessment have also been proposed in [6,7,8,9,10,11,12]. A greater attention has been recently devoted to the very tiny class of natural images [10,13]. These latter are believed to contain a particular kind of information that drove the evolution of Human Visual System (HVS) [14]. Hence, understanding of natural images statistics and features should be equivalent to better characterize HVS. However, a deep and useful knowledge about how to measure visual information in terms of bits per pixel seems to lack yet [10]. In particular, there is still not a complete agreement on the best value of the just noticeable detection threshold [1,15] that is able to agree with the complex mechanisms guiding human eye in the observation process.

The goal of this paper is to provide the visual cost of a localized distortion in terms of the number of *bits per sample*. This can be achieved by exploiting the formal link between the visual information tied to a distortion in a natural image and its information theory content, in terms of Jensen-Shannon divergence. This visual rate measure, combined with the visual contrast, is then useful for implementing an Occam razor based strategy for parameters optimization in restoration algorithms. It will be proved that the Jensen-Shannon divergence can be written in terms of the intersection between the supports of the distributions of the two involved kinds of information: the original and the distorted one. The knowledge of the amount of common information can be then exploited during the restoration process for tuning some of its variables or for automatically define the stopping criterion, in case of iterative algorithms. In this way, it is not necessary to use empirical thresholds for the visual contrast of the restored image. Experimental results show the efficacy of the proposed approach on both synthetic and real examples.

2 Visual Rate-Distortion Curve

In order to build a visual rate-distortion curve, it is necessary to first define a suitable rate measure that is able to quantify the visual information in terms of bits per pixel, and then to define a simple measure for the allowed loss of information (distortion). With regard to the rate, a relative measure is necessary. In fact, human eye works as a differential operator: it is able to detect degradation just because it is different from the surrounding information. This is the reason why the Jensen-Shannon divergence is more appropriate than classical entropy. As distortion measure, the significance of the visual contrast of the degraded area with respect to the one of the surrounding information will be considered.

Visual Rate. Let I be an image depicting a real-world scene and $I \sim p$, where p is its probability density function (pdf). If I is subjected to a given distortion $\mathcal{T}(x)$ in a subset Ω of its domain, we can denote with $O = \mathcal{T}[I(\Omega)]$ the corresponding distorted image and with q its pdf, i.e. $O \sim q$. We can also denote with B, the eventual not distorted region of I, then $B \sim p$. It is well known that, if

$h(X)$ is the differential entropy of the random variable X, for a linear distortion it holds $h(X+\beta) = h(X)$ and $h(\alpha X) = h(X) + \log(|\alpha|)$. Then, a luminance shift β of I leaves its entropy $h(I)$ unchanged, while a rescaling α increases or decreases it according to α absolute value. Unfortunately, a luminance shift, i.e. $\mathcal{T}(x) = x + \beta$ cannot be negligible from HVS point of view. It turns out that classical entropy is not able to describe human perception and then it is not a good measure for the visual cost of image degradation. On the contrary, the Jensen-Shannon divergence [16,17] is more proper for this task. It measures the distance between two pdfs p and q, i.e.

$$D_{JS}(p,q) = \frac{1}{2}(D_{KL}(p||m) + D_{KL}(q||m)) = -\frac{h(p)+h(q)}{2} + h(m), \qquad (1)$$

where $D_{KL}(p||q) = \int_{-\infty}^{-\infty} p(x)\log(\frac{p(x)}{q(x)})dx$ is Kullback-Leibler divergence [18] of two random variables X and Y with distributions p and q, $m = \frac{p+q}{2}$,

$$h(m) = 1 - \frac{1}{2}\left(\int_{S_p \cup S_q} (p(x)+q(x))\ \log(p(x)+q(x))dx\right), \qquad (2)$$

while S_p and S_q respectively are p and q supports. D_{JS} is then a relative measure that agrees with human perception. In particular, it can be proved that the D_{JS} between original and distorted images depends on the intersection between their pdfs supports.

Proposition. If $B \sim p$, $O = \mathcal{T}[I(\Omega)] \sim q$, where $\mathcal{T}(x)$ is the distortion operator, $\overline{S} = S_p \cap S_q$ with S_p and S_q respectively the supports of p and q, then

$$D_{JS}(p,q) \approx \begin{cases} 1 & \overline{S} = \emptyset \\ 1 - \frac{\overline{p}+\overline{q}}{2}|\overline{S}| & \overline{S} \neq \emptyset, \end{cases} \qquad (3)$$

where $|\overline{S}|$ is the length of $\overline{S}$, while $\overline{p}$ and $\overline{q}$ are proper p and q values in $\overline{S}$.

The proof is in Appendix. D_{JS} is the additional "visual" cost that is required for interpreting degradation within the original scene. In fact, it explicitly depends on the intersection between competing objects i.e., on how much information the original and degraded/restored image share in the degraded area. The more the common information, the smaller the D_{JS} value, the more visually similar the compared images and the closer their compressed size.

Visual Contrast Ratio. A degraded area cannot be perceived in a given context if its content is similar to the surrounding information. This means that, by denoting with C_R the visual contrast ratio between the degraded area and its surrounding information i.e.,

$$C_R = \frac{C_q}{C_p}, \qquad (4)$$

where $C_q = \frac{|S_q|}{2\,\overline{S}_q}$ and $C_p = \frac{|S_p|}{2\,\overline{S}_p}$ are the Michelson contrasts [1] of the degraded and not degraded region respectively, while $\overline{S}_p$ and $\overline{S}_q$ indicate the mid-points of p and q, it holds

$$|C_R - 1| \leq \epsilon, \qquad (5)$$

where ϵ is a suitable threshold. ϵ is related to the global visual quality of the degraded/restored image and it gives the greatest amount of distortion that is allowed for a given degradation to be not visible. It can then represent the allowed visual loss to use in the definition of the visual rate distortion curve according to the Occam razor. For a given distortion and restoration algorithm, the C_R versus D_{JS} curve can be built by increasing the allowed loss ϵ. Similarly to [2], the point where the second derivative of the curve attains its maximum represents the optimal point of the curve, i.e. the point that gives the amount of distortion that is allowed to the analysed image — the one that is not perceived in the image. For example, using a linear image adjustment algorithm [19] for enhancing the visual quality of poorly contrasted images, the $C_R - D_{JS}$ curve in Fig. 1 is achieved, where the loss ϵ corresponds to the dilation of the histogram of the output image (if $\epsilon = 0$, the support of the output histogram corresponds to the one of the original clean image, otherwise it is smaller). As it can be observed, the maximum of the second derivative gives the value of C_R for a restored image that is visually similar to the original one — the visual similarity has been measured by means of the Structural Similarity Index (SSIM) [8]. On the other hand, if

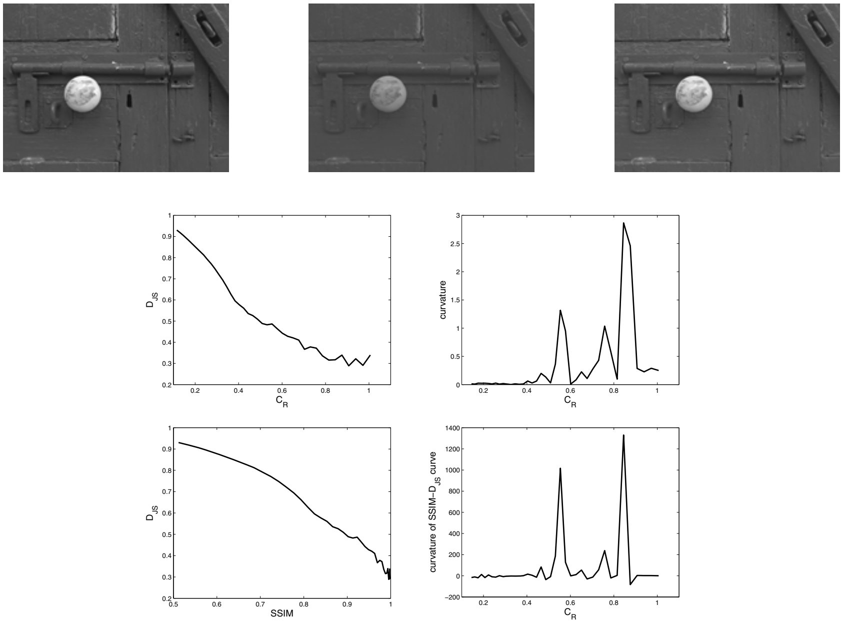


Fig. 1. *Top:* Original, low contrasted (SSIM = 0.8729) and enhanced image (SSIM = 0.9875) (original and distorted images are from TID2008 database [22]). *Middle:* D_{JS} versus C_R curve that has been built using a conventional linear image adjustment algorithm (*left*) and its $2nd$ derivative (*right*). The maximum of the $2nd$ derivative corresponds to the enhanced image. *Bottom:* D_{JS} versus $SSIM$ curve (*left*) and its $2nd$ derivative (*right*). The maximum of the $2nd$ derivative corresponds to the one of the $C_R - D_{JS}$ curve.

SSIM is used instead of C_R, the same restored image is selected. It turns out that two simple measures (C_R and D_{JS}) that only depend on the pdfs of involved images can be used for getting information about the optimal parameters of a restoration framework instead of sophisticated but computationally expensive visual distortion measures.

It is worth noticing that C_R and D_{JS} depend in a different way on the pdfs of the involved images: C_R depends on their supports and midpoints (respectively $S_p, S_q, \overline{S}_p, \overline{S}_q$), while D_{JS} only depends on the information within their intersection $\overline{S}$. It turns out that $\overline{S}$ is the additional information that can be derived from the visual rate distortion curve to be used in the assessment of restoration methods.

The computation of both C_R and D_{JS} requires the knowledge of the original and degraded pdf. It is not always possible in real applications, especially in the presence of global distortions like noise or poor contrast. On the contrary, this result is useful in the restoration of localized degradation, i.e. whenever the latter involves just a small part of the whole scene. In this case, the information surrounding the degradation preserves the features of the original pdf and then C_R and D_{JS} can be still directly evaluated on the degraded image, as it is shown in the experimental results, along with the maximum point of the second derivative of the corresponding rate-distortion curve. Taking into account the relation in eq. (3), this point gives information about the amount of information the two pdfs are required to share in order to be not seen as different objects of the scene. This represents an "a priori" information about the final restoration result that can be used to automatically set restoration parameters or to fix the stopping criterion in iterative methods.

3 Experimental Results and Conclusions

Fig. 2 shows two semi-transparent blotches on archived photographs. They are common local defects whose main characteristic is the semi-transparency: they do not completely hide original image content in the degraded area. In this case, affine restoration models are used in order to preserve the original image content. In the sequel we will consider two restoration algorithms: the additive multiplicative model in [20] and the HVS-based restoration method in [21]. They suppose the same degradation model but, while the former uses global affine parameters, the latter makes use of point-wise affine parameters that need an iterative refinement strategy to be optimally tuned. We will show that if the latter is used for constructing the rate-distortion curve, as described in the previous section, the optimal point of the curve provides the iteration where the iterative algorithm reaches a high visual quality results as well as an a priori information that could be eventually used in the estimation of the global affine parameters in the first algorithm. For the isolated blotch in Fig. 2.left the optimal point of the rate-distortion curve corresponds to $D_{JS} \approx 0.1540$ and $C_R = 1.0668$, while the initial values were $D_{JS} \approx 0.2886$ and $C_R = 1.2538$. On the contrary, for

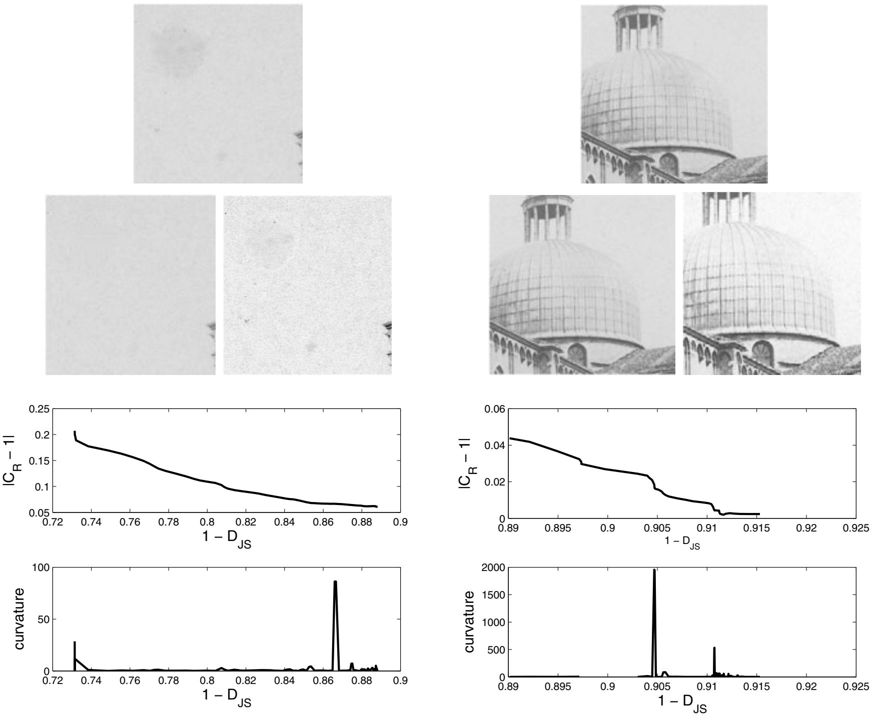


Fig. 2. Top: Isolated (*left*) and Church (*right*) images. **Second row:** Recovered images using the restoration algorithms in [21] (*left*) and [20] (*right*). **Third row:** Rate-distortion curves obtained using the algorithm in [21]. **Bottom:** The corresponding curvature whose optimal value realizes the maximum.

the blotch in Church image (Fig. 2.right), it corresponds to $D_{JS} \approx 0.0954$ and $C_R = 1.0577$, while the initial values were $D_{JS} \approx 0.1112$ and $C_R = 0.9904$. The inversion of D_{JS} allows us to derive the amount of the common information $|\overline{S}|$ between the restored and clean information that is enough for masking the presence of degradation. In particular, $|\overline{S}| = 29$ for the image in Fig. 2.left and $|\overline{S}| = 108$ for the image in Fig. 2.right, while their initial values respectively were $|\overline{S}| = 23$ and $|\overline{S}| = 99$. The estimated value for $|\overline{S}|$ is then used for setting the affine parameters in the additive/multiplicative restoration model in [20]. A visual inspection by ten observers having experience with image restoration confirms that the degradation is really invisible in the restored images.

These preliminary results show that the Jensen-Shannon divergence combined with visual contrast measures is able to provide the cost in bits of human eye tolerance to a given distortion in a given context. Moreover, it offers an alternative way to adaptively select the just noticeable detection thresholds for the image under study. Future research will be oriented to deeper investigations about the potential use of a perception-based Jensen-Shannon divergence for image quality assessment.

4 Appendix

Proof of Proposition. If p and q are two p.d.f with supports S_p and S_q and $\overline{S} = S_p \cap S_q$, the integral in eq. (2) can be split as follows:

$$\int_{S_p - \overline{S}} p(x)\log(p(x))dx + \int_{\overline{S}} (p(x)+q(x))\log(p(x)+q(x))dx +$$

$$+ \int_{S_q - \overline{S}} q(x)\log(q(x))dx.$$

Summing and subtracting $\int_{\overline{S}}(p(x)\log(p(x)) + q(x)\log(q(x)))dx$, we have:

$$-h(p) - h(q) + \int_{\overline{S}} p(x)[\log(p(x)+q(x)) - \log(p(x))]dx +$$

$$+ \int_{\overline{S}} q(x)[\log(p(x)+q(x)) - \log(q(x))]dx.$$

Since $0 \leq p(x) \leq 1$ and $0 \leq q(x) \leq 1 \quad \forall x$, the first order Taylor expansion of the log function around $p(x)$ (in the first integral) and around $q(x)$ (in the second integral) combined with eq. (1) gives

$$D_{JS}(p,q) \simeq 1 - \frac{1}{2}\int_{\overline{S}} (q(x)+p(x))dx. \tag{6}$$

Let the couples of points (P_1, P_2) and (Q_1, Q_2) respectively be the extremes of S_p and S_q. S_p and S_q can be such that: *i)* $\overline{S} = S_p \cap S_q = \emptyset$, i.e. $Q_2 < P_1$ or $Q_1 > P_2$; *ii)* $\overline{S} \neq \emptyset$ with $\overline{S} \subset S_p$ and $\overline{S} \subset S_q$, i.e. $Q_1 < P_1 \leq Q_2 < P_2$ or $P_1 < Q_1 \leq P_2 < Q_2$; *iii)* $\overline{S} = S_p$ or $\overline{S} = S_q$, i.e. $Q_1 \leq P_1 < P_2 \leq Q_2$ or $P_1 < Q_1 < Q_2 < P_2$. If $\overline{S}$ is empty, then $D_{JS} = 1$ (*case i)*). Otherwise, if $P_1 \leq Q_2$ (*case ii)*), then the mean value theorem in eq. (2) gives

$$D_{JS} = 1 - (\overline{q} + \overline{p})\frac{|\overline{S}|}{2}, \tag{7}$$

where $\overline{q} = q(x_1)$, $\overline{p} = p(x_2)$ and $x_1, x_2 : P_1 \leq x_1, x_2 \leq Q_2$. Using the same arguments for case *iii)* there exists $x_1 : P_1 \leq x_1 \leq P_2$ such that, setting $\overline{p} = p(x_1)$, eq. (6) can be rewritten as

$$D_{JS} = \frac{1}{2} - \frac{\overline{p}|\overline{S}|}{2}. \qquad \bullet \tag{8}$$

Remark. $D_{JS}(p,q) = 1 - \frac{1}{2}\left(\int_{\overline{S}} p(x)\log(1 + \frac{q(x)}{p(x)})dx + \int_{\overline{S}} q(x)\log(1 + \frac{p(x)}{q(x)})dx\right)$ $\approx 1 - \frac{1}{2}\int_{\overline{S}}(p(x)+q(x))[1 - V(x)]dx$, where $V = \frac{1}{4\ln(2)}\frac{(q-p)^2}{pq}$ is the second order error term of the Taylor expansion of *log* functions around 1. $V \geq 0$, $V = 0$ iff $p = q$ and it grows as the distance between p and q increases. It turns out that the approximation in eq. (6) always gives a smaller value than the real one.

References

1. Winkler, S.: Digital Video Quality, Vision Models and Metrics. Wiley, Chichester (2005)
2. Natarajan, B.K.: Filtering Random Noise from Deterministic Signals via Data Compression. IEEE Trans. on Signal Processing 43(11), 2595–2605 (1995)
3. Andre', T., Antonini, M., Barlaud, M., Gray, R.M.: Entropy-Base Distortion Measure and Bit Allocation for Wavelet Image Compression. IEEE Trans. on Image Processing 16(12), 3058–3064 (2007)
4. Hontsch, I., Karam, L.: Adaptive Image Coding with Perceptual Distortion Control. IEEE Trans. on Image Processing 11(3), 213–222 (2002)
5. Lee, H., Lee, S.: Visual Entropy Gain for Wavelet Image Coding. IEEE Sig. Proc. Let. 13(19) (2006)
6. Daughman, J.G.: Entropy Reduction and Decorrelation in Visual Coding by Oriented Neural Receptive Fields. IEEE Trans. on Biomedical Engineering 36(1) (1989)
7. Rivera, M., Ocegueda, O., Marroquin, J.L.: Entropy-Controlled Quadratic Markov Measure Field Models for Efficient Image Segmentation. IEEE Trans. on Image Processing 16(12), 3047–3057 (2007)
8. Wang, W., Wang, Y., Huang, Q., Gao, W.: Measuring Visual Saliency by Site Entropy Rate. In: Proc. of CVPR 2010, pp. 2368–2375 (2010)
9. Hou, Z., Yau, W.Y.: Visible Entropy: A Measure for Image Visibility. In: Proc. of ICPR 2010, pp. 4448–4451 (2010)
10. Sheikh, H.R., Bovik, A.C.: Image Information and Visual Quality. IEEE Trans. on Image Proc. 15(2) (2006)
11. Agaian, S., Silver, B., Panetta, K.: Transform Coefficient Histogram Based Image Enhancement Algorithms Using Contrast Entropy. IEEE Trans. on Image Processing 16(3) (2007)
12. Wang, Z., Li, Q.: Information Content Weighting for Perceptual Image Quality Assessment. To Appear on IEEE Trans. on Image Proc. (2011)
13. Sristava, A., Lee, A.B., Simoncelli, E.P., Zhu, S.-C.: On Advances in Statistical Modeling of Natural Images. J. Math. Imag. Vis. 18, 17–33 (2003)
14. Simoncelli, E.P., Olshausen, B.A.: Natural Image Statistics and Neural Representation. Ann. Rev. NeuroSc. 24, 1193–1216 (2001)
15. Pappas, T.N., Safranek, R.J.: Perceptual criteria for image quality evaluation. In: Handbook of Image and Video Processing. Academic Press, London (2000)
16. Lin, J.: Divergence Measures based on the Shannon Entropy. IEEE Trans. on Inf. Th. 37(1) (1991)
17. Endres, D.M., Schinelin, J.E.: A New Metric for Probability Distributions. IEEE Trans. on Information Theory 49(7) (2003)
18. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley & S, Chichester (1991)
19. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Prentice Hall, Englewood Cliffs (2002)
20. Stanco, F., Tenze, L., Ramponi, G.: Virtual Restoration of Vintage Photographic Prints Affected by Foxing and Water Blotches. J. of Elect. Imaging 14(4) (2005)
21. Bruni, V., Crawford, A.J., Kokaram, A., Vitulano, D.: Semi-transparent Blotches Removal from Sepia Images Exploiting Visibility Laws. In: Signal, Image and Video Processing. Springer, Heidelberg (2011), Online First doi:10.1007/s11760-011-0220-1
22. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. Advances of Modern Radioelectronics 10, 30–45 (2009), http://www.ponomarenko.info/tid2008.htm

A Parallel Implementation of the Thresholding Problem by Using Tissue-Like P Systems

Francisco Peña-Cantillana[1], Daniel Díaz-Pernil[2],
Ainhoa Berciano[2,3], and Miguel Angel Gutiérrez-Naranjo[1]

[1] Research Group on Natural Computing - Dept. of Computer Science and AI
University of Seville, Spain
[2] CATAM Research Group - Dept. of Applied Mathematics I
University of Seville, Spain
[3] Departament of Didactic of Mathematics and Experimental Sciences
University of the Basque Country, Spain
frapencan@gmail.com, {sbdani,magutier}@us.es, ainhoa.berciano@ehu.es

Abstract. In this paper we present a parallel algorithm to solve the thresholding problem by using Membrane Computing techniques. This bio-inspired algorithm has been implemented in a novel device architecture called CUDA$^{\text{TM}}$, (Compute Unified Device Architecture). We present some examples, compare the obtained time and present some research lines for the future.

1 Introduction

In Computer Vision [13], segmentation is the process of splitting a digital image into sets of pixels in order to make it simpler and easier to analyze. One of its main uses is the localization of objects and boundaries. Technically, the process consists of assigning a label to each pixel, in such way the pixels with the same label form a meaningful region. *Thresholding* is one of the simplest and most widely used image segmentation techniques.

In this paper, we present a bio-inspired algorithm to solve the thresholding problem. The treatment of digital images has several features which make it suitable for techniques inspired by nature. One of them is that the treatment of the image can be parallelized and locally solved. Another interesting feature is that the basic necessary information can be easily encoded by bio-inspired representations. Throughout this paper, we use techniques taken from Membrane Computing. This is a theoretical model of computation inspired by the structure and functioning of cells as living organisms able to process and generate information. The computational devices in Membrane Computing are called *P systems* [12,15]. Roughly speaking, a P system consists of a membrane structure, in whose compartments one places multisets of objects which evolve according to given rules which are usually applied in a synchronous non-deterministic maximally parallel manner. In particular, we consider antiport rules, which were introduced as communication rules for P systems in [11]. Such rules are inspired

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 277–284, 2011.

on the communication among cells. In the antiport rules, objects residing at both sides of the membrane cross it simultaneously in opposite directions.

In the literature, one can find several attempts for bridging problems from Digital Imagery with Natural Computing as the work by K.G. Subramanian *et al.* [1] or the work by Chao and Nakayama where Natural Computing and Algebraic Topology are linked by using Neural Networks [2] (extended Kohonen mapping). Recently, new approaches have been presented in the framework of Membrane Computing [3,4,5].

The algorithm has been implemented by using a novel device architecture called CUDA™, (Compute Unified Device Architecture) [9,10,14]. CUDA™ is a general purpose parallel computing architecture that allows the parallel NVIDIA Graphics Processors Units (GPUs) to solve many complex computational problems in a more efficient way than on a CPU.

The paper is organised as follows: Firstly, we briefly recall some basics on the theoretical Membrane Computing framework and, in Section 2.1, the family of P systems which solves the thresholding problem is presented. Next, we present our parallel implementation, with several examples and comparisons. The paper finishes with some final remarks.

2 Methods

There are different models of P systems in the Membrane Computing framework. In this paper we will consider the so-called *tissue-like P systems* [8]. They have two biological inspirations: intercellular communication and cooperation between neurons. The common mathematical model of these two mechanisms is a network of processors dealing with symbols and communicating these symbols along channels specified in advance.

Formally, a *tissue-like P system* with input of degree $q \geq 1$ is a tuple

$$\Pi = (\Gamma, \Sigma, \mathcal{E}, w_1, \ldots, w_q, \mathcal{R}, i_\Pi, o_\Pi),$$

where

1. Γ is a finite *alphabet*, whose symbols will be called *objects*, $\Sigma(\subset \Gamma)$ is the input alphabet, $\mathcal{E} \subseteq \Gamma$ is the alphabet of the objects in the environment,
2. $w_1, \ldots, w_q$ are strings over Γ representing the multisets of objects associated with the cells at the initial configuration,
3. $\mathcal{R}$ is a finite set of communication rules of the following form: $(i, u/v, j)$ for $i, j \in \{0, 1, 2, \ldots, q\}, i \neq j$, $u, v \in \Gamma^*$,
4. $i_\Pi \in \{1, 2, \ldots, q\}$ is the input cell and $o_\Pi \in \{0, 1, 2, \ldots, q\}$ is the output cell.

A tissue-like P system of degree $q \geq 1$ can be seen as a set of q cells (each one consisting of an elementary membrane) labelled by $1, 2, \ldots, q$. We will use 0 to refer to the label of the environment, i_Π denotes the input region and o_Π denotes the output region (which can be the region inside a cell or the environment).

The strings $w_1, \ldots, w_q$ describe the multisets of objects placed in the q cells of the system at the initial configuration. We interpret that $\mathcal{E} \subseteq \Gamma$ is the set of objects placed in the environment, each one of them available in an arbitrary large amount of copies.

The communication rule $(i, u/v, j)$ can be applied over two cells labelled by i and j such that u is contained in cell i and v is contained in cell j. The application of this rule means that the objects of the multisets represented by u and v are interchanged between the two cells. Note that if either $i = 0$ or $j = 0$ then the objects are interchanged between a cell and the environment.

Rules are used as usual in the framework of membrane computing, that is, in a maximally parallel way (a universal clock is considered). In one step, each object in a membrane can only be used for one rule (non-deterministically chosen when there are several possibilities), but any object which can participate in a rule of any form must do it, i.e., in each step we apply a maximal set of rules.

In what follows we assume the reader is already familiar with the basic notions and the terminology underlying P systems [12].

2.1 Thresholding of Digital Images

Thresholding is a method of image segmentation whose basic aim is to obtain a binary image from a color one. The idea is to split the set of pixels into two sets (black and white) depending on its bright and a fixed value, the *threshold*. If the bright of the pixel is greater than the threshold, then the pixel is labeled as *object*. Otherwise, it is labeled as *background*. After the labeling, a new binary image is created by coloring each pixel white or black.

The basic thresholding method can be generalized in a natural way to *quantization*. Instead of using $\{0, 1\}$ as labels to obtain a binary image, we can consider a larger set of labels, $\{1, \ldots, k\}$ and get a final image with k levels. Another natural generalization is to replace the color information by another scale on the features of the pixel (bright, intensity, gray scale, etc.). In this section, we present a family of tissue-like P systems which solves the thresholding problem by considering a 4-neighborhood between pixels.

Let $\mathcal{C}$ be the alphabet of colors. The key idea is to divide $\mathcal{C}$ into classes and to assign a representative of each class. After this choosing, the color of each pixel is changed by the representatives of their class. In this paper, we choose the first color of each class as representatives of them.

2.2 A Family of Tissue-Like P Systems

Given a digital image I with n^2 pixels and $n \in \mathbb{N}$, we define a tissue-like P system whose input is given by the pixels of the image encoded by the objects a_{ij}, where $1 \leq i, j \leq n$ and $a \in \mathcal{C}$. For each image of size n^2 and k the number of intervals in which the set of colors $\mathcal{C}$ is divided, ($k, m, n \in \mathbb{N}$, with $n = k \times m$), we consider the following tissue-like P system with input of degree 1, $\mathbf{\Pi}(k, n) = (\Gamma, \Sigma, \mathcal{E}, w_1, \mathcal{R}, i_\Pi, o_\Pi)$, where

- $\Gamma = \Sigma = \mathcal{E} = \{a_{ij} \,:\, a \in \mathcal{C},\ 1 \leq i, j \leq n\}$,
- $w_1 = \emptyset$,

- R is the following set of communication rules:
  $(1, b_{ij}/a_{ij}, 0)$, for $1 \leq i, j \leq n$; $l = 0, 1, 2, \ldots, n - m$, $a = m \cdot l$ and $b \in \mathcal{C}$, $a < b \leq a + (m - 1)$.
  These rules are used to divide the set of colors in k intervals of length m.
- $i_{\Pi} = o_{\Pi} = 1$.

Next, we will give some outlines how to prove that the *thresholding problem* can be solved in one step using this family of tissue-like P systems $\mathbf{\Pi}(k, n)$: the alphabet of colors is split into k intervals and the first element of each interval is chosen as a representative. The objects representing pixels are placed in the membrane labeled by 1 and the representatives are placed in the environment. The rules are applied in parallel and simultaneously each pixel is traded against the corresponding representative.

The question is how we can decide the number of intervals dividing $\mathcal{C}$. We could divide the alphabet in two intervals to obtain a *binary image*. If we divide a color image in 255 intervals then we will obtain a *grey scale image*. Other possibility is, given a digital image I, to obtain the medium color of I (μ_I).

3 Results

GPUs constitute nowadays a solid alternative for high performance computing, and the advent of CUDA™ allows programmers a friendly model to accelerate a broad range of applications. The way GPUs exploit parallelism differ from multi-core CPUs, which raises new challenges to take advantage of its tremendous computing power. In this paper, we present a parallel software tool based on our membrane solution for image quantization and binarization. It has been developed by using Microsoft Visual Studio 2008 Professional Edition (C++) with the plugging Parallel Nsight (CUDA™) under Microsoft Windows 7 Professional with 32 bits. CUDA™ C, an extension of C for implementations of executable kernels in parallel with graphical cards NVIDIA has been used to implement the P systems. It has been necessary the *nvcc compiler* of CUDA™ Toolkit and some libraries from openCV to the treatment of input and output images.

The experiments have been performed on a computer with a CPU Intel Pentium 4 650, with support for HT technology which allows to work like two CPUs of 32 bits to 3412 MHz. The computer has 2 MB of L2 cache memory and 1 GB DDR SDRAM of main memory with 64 bits bus wide to 200 MHz.

The graphical card (GPU) is an NVIDIA Geforce 8600 GT composed by 4 *Stream Processors* with a total of 32 cores to 1300 MHz and executes 512 threads per block as maximum. It has a 512 MB DDR2 main memory, but 499 MB could be used by processing in a 128 bits bus to 700 MHz. So, the transfer rate obtained is by 22.4 Gbps. For constant memory used 64 KB and for shared memory 16 KB (It is not a good data for a good CUDA™ graphical card). Its Compute Capability is 1.1 (from 1.0 to 2.1), then we can obtain a lot of improvements in the efficiency of the algorithms.

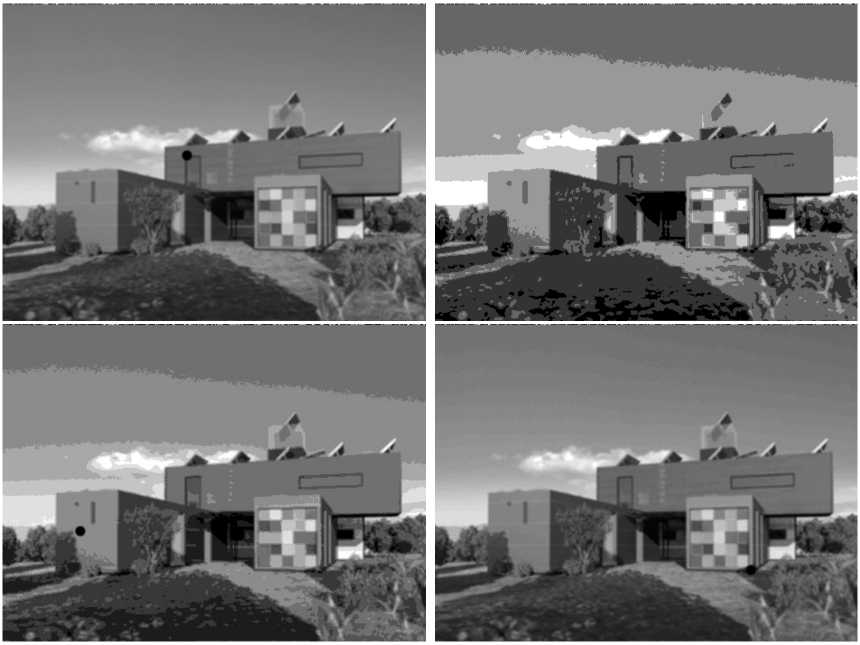


Fig. 1. Quantization with classes of same size

3.1 Examples

Although our algorithm can work with color images, our software is implemented to work with grey scale. In this section, we show the results obtained by our tool to do a quantization of images. We consider an image of size 468 × 351 (see left up image of Fig. 1). Notice that, when we work with different thresholds we obtain different results, as the rest of images of Fig. 1 show.

The simplest thresholding method is binarization where the image is divided into two set of pixels. We use our tool to do a binarization of images with grey scale, but we need to choice a threshold to divide the image into two sets of different sizes. In Fig. 2 we show three original images (see up) with sizes 3456 × 2592, 1536 × 2048 and 960 × 1280, respectively. As we can see, only the third image has been binarized in an appropriate way. In fact, we can perfectly distinguish the climber in the resulting image in our software. But, when we see the first two images, it is clear that they are not good binarizations. In the first one, when the binarization is done, we keep one of the fishes, but the figure of the bird is partially lost with the background of the image and, in the second one, when the binarization is done, we see, for example, the problems with the legs of the dog. The processing time for each image of Fig. 2 are 363.161 ms, 307.558 ms and 243.461 ms, respectively.

Then, the question is to select an appropriate threshold to the binarization. So, we take the Hamadani algorithm [6] for the chosen thresholds and implement this in our tool (in a sequential way). In this case we consider a linear combination of

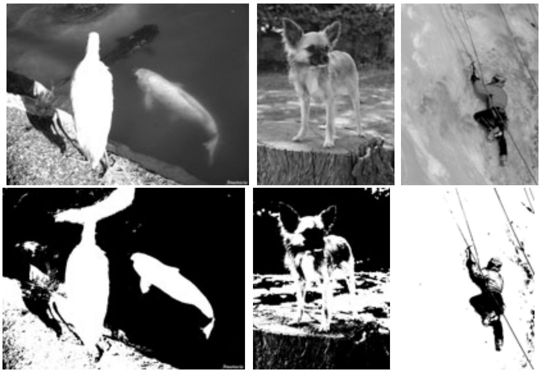


Fig. 2. Binarization of three images using as threshold k=80, 100 and 115, respectively

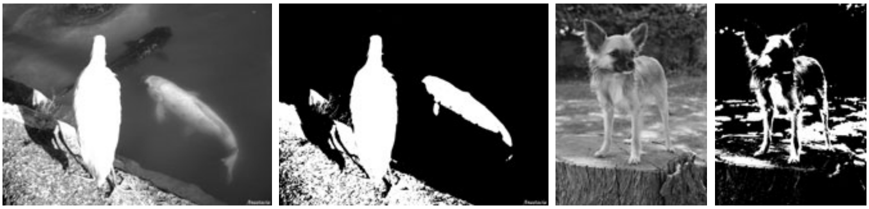


Fig. 3. Binarization using the Hamadani Algorithm. Thresholds are 172 and 155 to each image, respectively.

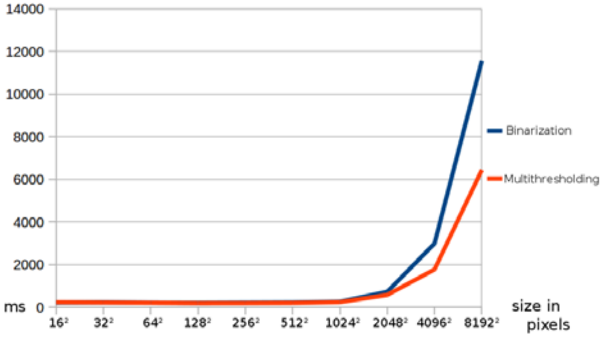


Fig. 4. Comparative

the mean, μ, and the standard deviation, σ, of the values of pixels of the images: $k = k_1 \cdot \mu + k_2 \cdot \sigma$. So, the question is the chosen of the parameters k_1 and k_2. For example, we have taken $k_1 = k_2 = 1$ in the two images with problems in the binarization, and we can see in Fig. 3 how the binarizations are better with this chosen. We need 619.889 ms and 346.034 ms to each binarization, respectively.

Finally, we have some proofs with our software to check the needed running time to process images of different sizes. We show the results in the Fig. 4. Obviously, we need more time to the binarization because we have added the (sequential) implementation of the Hamadani algorithm. Moreover, we should advise, our tool needs much more time to very weighted images because our graphical card is not professional, for example, the shared memory is very small.

4 Conclusions

The bio-inspired computing techniques have features as the encapsulation of the information, a simple representation of the knowledge and parallelism, which are appropriate with dealing with digital images. Nonetheless, the use of the intrinsic parallelism of these paradigms can hardly be implemented in current one-processor computers, so the potential advantages of the theoretical design are lost.

In this paper we show that the drawback of using one-processor computers for implementing Membrane Computing designs can be avoided by using the parallel architecture CUDA™. This new technology provides the hardware needed for a real parallel implementation of Membrane Computing algorithms.

By following this research line, several questions are open: In this paper, the Hamadani algorithm has been implemented in sequential mode. In order to exploit all the possibilities of the parallel architecture, a new parallel implementation should be designed.

¿From Digital Imagery, new parallel algorithms (for example, the Otsu algorithm [7] for multilevel thresholding), can be adapted to the new technology; from the Membrane Computing side, new design or different P system models can be explored; from the hardware point of view, the advances in the new technology CUDA™ with the new boards Tesla and Fermi open new possibilities for going on with the research.

Acknowledgements. DDP and MAGN acknowledge the support of the projects TIN2008-04487-E and TIN-2009-13192 of the Ministerio de Ciencia e Innovación of Spain and the support of the Project of Excellence with *Investigador de Reconocida Valía* of the Junta de Andalucía, grant P08-TIC-04200. HC acknowledges the support of the project MTM2009-12716 of the Ministerio de Educación y Ciencia of Spain and the project PO6-TIC-02268 of Excellence of Junta de Andalucía, the "CATAM" PAICYT research group FQM-296 and EHU09/04 project of University of the Basque Country.

References

1. Ceterchi, R., Gramatovici, R., Jonoska, N., Subramanian, K.G.: Tissue-like P systems with active membranes for picture generation. Fundamenta Informaticae 56(4), 311–328 (2003)

2. Chao, J., Nakayama, J.: Cubical singular simplex model for 3D objects and fast computation of homology groups. In: 13th International Conference on Pattern Recognition (ICPR 1996), vol. IV, pp. 190–194. IEEE Computer Society, Los Alamitos (1996)
3. Christinal, H.A., Díaz-Pernil, D., Gutiérrez-Naranjo, M.A., Pérez-Jiménez, M.J.: Thresholding of 2D images with cell-like P systems. Romanian Journal of Information Science and Technology (ROMJIST) 13(2), 131–140 (2010)
4. Christinal, H.A., Díaz-Pernil, D., Real, P.: Segmentation in 2D and 3D image using tissue-like P system. In: Bayro-Corrochano, E., Eklundh, J.-O. (eds.) CIARP 2009. LNCS, vol. 5856, pp. 169–176. Springer, Heidelberg (2009)
5. Díaz-Pernil, D., Gutiérrez-Naranjo, M.A., Molina-Abril, H., Real, P.: A bio-inspired software for segmenting digital images. In: Nagar, A.K., Thamburaj, R., Li, K., Tang, Z., Li, R. (eds.) Proceedings of the 2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications BIC-TA, vol. 2, pp. 1377–1381. IEEE Computer Society, Los Alamitos (2010)
6. Hamadani, N.: Automatic target cueing in IR imagery. Master's thesis, Air Force Institute of Technology, WAFP (December 1981)
7. Liao, P.S., Chen, T.S., Chung, P.C.: A fast algorithm for multilevel thresholding. Journal of Information Scence and Engineering 17(5), 713–727 (2001)
8. Martín-Vide, C., Păun, G., Pazos, J., Rodríguez-Patón, A.: Tissue P systems. Theoretical Computer Science 296(2), 295–326 (2003)
9. Nickolls, J., Buck, I., Garland, M., Skadron, K.: Scalable parallel programming with cuda. Queue 6, 40–53 (2008)
10. Owens, J.D., Houston, M., Luebke, D., Green, S., Stone, J.E., Phillips, J.C.: GPU Computing. Proceedings of the IEEE 96(5), 879–899 (2008)
11. Păun, A., Păun, G.: The power of communication: P systems with symport/antiport. New Generation Computing 20(3), 295–306 (2002)
12. Păun, G., Rozenberg, G., Salomaa, A. (eds.): The Oxford Handbook of Membrane Computing. Oxford University Press, Oxford (2010)
13. Shapiro, L.G., Stockman, G.C.: Computer Vision. Prentice Hall PTR, Upper Saddle River (2001)
14. NVIDIA Corporation. NVIDIA CUDA™ Programming Guide, http://www.nvidia.com/object/cuda_home_new.html
15. P system web page, http://ppage.psystems.eu

P Systems in Stereo Matching

Georgy Gimel'farb, Radu Nicolescu, and Sharvin Ragavan

Department of Computer Science, University of Auckland, Auckland, New Zealand
{g.gimelfarb,r.nicolescu}@auckland.ac.nz, srag010@aucklanduni.ac.nz

Abstract. Designing parallel versions of sequential algorithms has attracted renewed attention, due to recent hardware advances, including various general-purpose multi-core, multiple core and many-core processors, as well as special-purpose FPGA implementations. P systems consist of networks of autonomous cells, such that each cell transforms its input signals in accord with symbol-rewriting rules and feeds the output results into its immediate neighbours. Inherent intra- and inter-cell parallelism make the P systems a prospective theoretical testbed for designing parallel algorithms. This paper discusses capabilities of P systems to implement the symmetric dynamic programming algorithm for stereo matching, with due account to binocular or monocular visibility of 3D surface points.

Keywords: Parallel systems, membrane computing, stereo matching, symmetric dynamic programming stereo (SDPS).

1 Introduction

Essentially, a P system is a network of data processing cells, inspired from the structure and interaction of living cells [1,4,8,9,10,11,12,13]. Each cell transforms input and local symbols in accord with rewriting rules and sends some of the resulting symbols out, to its close neighbours. Rules of the same cell can be applied in parallel (if possible) and all cells work in parallel, in the synchronous mode. The underlying network is a *digraph* or a more specialized version, such as a directed acyclic graph (*dag*) or a *tree* (which is the most studied case). Advanced scenarios consider cases when cells or arcs can dynamically appear, disappear or move.

A couple of previous papers suggested that P systems offer a theoretically efficient testbed for the design of parallel versions of various sequential image analysis tasks, such as *segmentation* [2,3]. This paper discusses a P system based implementation of the *symmetric dynamic programming stereo* (SDPS) [7]. To the best of our knowledge, this is the first paper that introduces a P system model for stereo matching and it reveals and separates the inherently parallel and sequential processing stages of the SDPS algorithm.

SDPS searches for the maximal scanline-to-scanline similarity, in explicit or implicit Cyclopean space of (x, y)-coordinates and d-disparities [7]. The most obvious and "trivial" avenue for parallel implementation of the SDPS is common

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 285–292, 2011.

for all so-called 1D stereo matching algorithms: similarity between each pair of corresponding scanlines in a rectified (co-aligned) stereo pair of images can be maximised independently of other pairs. SDPS offers an additional parallelisation avenue, exploited in this paper, by computing similarity scores in parallel, for all the disparities associated with each current x-coordinate. Together, both properties suggest a massively parallel 3D membrane computing framework combining the 2D parallel forward and 1D sequential backward processing. The main part of this framework consists of a 3D (x, y, d) uniform array of similar cells, connected each to the near neighbours, along the x-coordinate and d-disparity axes. The cells are linked to image memory for simultaneous initialisation by embedding image data into each cell. Fixed data transmission links between the neighbouring cells facilitate the forward propagation of 2D processing, in parallel along the 3D array, and the backward trace, required to reconstruct the goal 3D surface, consisting of independently found 2D cross-sections, or profiles.

2 General P Model

The basic definition of *simple P modules* in [4] covers many common P systems, such as *cell-like* (based on trees), *hyperdag* (based on dags), and *neural* P systems (based on directed graphs). This definition is further generalised, by introducing new features, which appear useful for modelling the SDPS.

Definition 1. A simple P module with duplex channels *is a system $\Pi = (O, K, \delta)$, where O is a finite non-empty alphabet of objects; K is a finite set of cells, and δ is an irreflexive binary relation on K, representing a set of structural arcs between cells (essentially a digraph), with duplex communication capabilities.*

Each cell, $\sigma_i \in K$, has the initial configuration $\sigma_{i0} = (Q_i, s_{i0}, w_{i0}, R_i)$, and the current configuration $\sigma_i = (Q_i, s_i, w_i, R_i)$, where: Q_i is a finite set of states; $s_{i0} \in Q_i$ is the initial state; $s_i \in Q_i$ is the current state; $w_{i0} \in O^*$ is the initial multiset of objects; $w_i \in O^*$ is the current multiset of objects; and R_i is a finite ordered set of multiset rewriting rules of the form: $s\ x \rightarrow_\alpha s'\ x'\ (u)_{\beta_\gamma}$, where $s, s' \in Q$, $x, x' \in O^*$, $u \in O^*$, $\alpha \in \{\texttt{min}, \texttt{max}\}$, $\beta \in \{\uparrow, \downarrow, \updownarrow\}$, and $\gamma \in \texttt{repl} \cup K$. If $u = \lambda$ (the empty multiset of objects), this rule can be abbreviated as $s\ x \rightarrow_\alpha s'\ x'$.

A cell evolves by applying one or more rules, which can change its content and state and can send objects to its neighbours. For cell $\sigma_i = (Q_i, s_i, w_i, R_i)$, a rule $s\ x \rightarrow_\alpha s'\ x'\ (u)_{\beta_\gamma} \in R_i$ is applicable if $s = s_i$ and $x \subseteq w_i$. The application of a rule takes two sub-steps, after which the cell's current state s is replaced by target state s', the current content x is replaced by x', and multiset u is sent as specified by the transfer operator β_γ (as further described below). The rules are applied in the weak priority order [11], i.e. the higher priority applicable rules are applied before the lower priority ones, and a lower priority applicable rule is applied only if it indicates the same target state as the previously applied rules.

The rewriting operators $\alpha = \texttt{min}$ and $\alpha = \texttt{max}$ indicate that an applicable rewriting rule of R_i is applied once or as many times as possible, respectively. Multisets u represent messages, sent to digraph neighbours, up and/or down *structural arcs*, according to the indicated transfer operators β_γ. (*i*) If $\beta_\gamma = \downarrow_{\texttt{repl}}$, then, for each application of this rule, a copy of multiset u is sent to each cell in $\delta(i)$ (if any). (*ii*) If $\beta_\gamma = \downarrow_j$, then, for each application of this rule, a copy of multiset u is sent to cell $\sigma_j \in K$, provided that $j \in \delta(i)$. (*iii*) If $\beta_\gamma = \uparrow_j$, then, for each application of this rule, a copy of multiset u is sent to cell $\sigma_j \in K$, provided that $j \in \delta^{-1}(i)$. (*iv*) In all other cases, message u is silently lost. Other (not used in this paper) operators are described in [5].

The above definition allows each cell to have its own state and rule sets. However, we prefer scenarios when all the cells share the same state and rule set; differing only by their initial content and their neighbourhood relations. Intuitively, such cells are created identical, by a virtual "cell factory" and initialised to the same quiescent state, typically designated as s_0. A state is called *quiescent* if no rules can be applied for empty cells in this state (i.e. an empty quiescent cell does not evolve until some object appeared in this cell, e.g. was sent from this cells' neighbours). Then, the cells are allocated different positions in the structural digraph and possibly initialised with different contents.

Extensions. We extend the basic simple P module concept with the following three features, which are useful in complex scenarios. (*i*) Arcs can have *labels* to be used in transfer operators, instead of cell labels. For example, if k is the label of an outgoing arc $(\sigma_i, \sigma_j) \in \delta$, the occurrence of $(u)_{\downarrow_k}$ in the right-hand side of a rule indicates that message u is to be sent from cell σ_i to cell σ_j. Although not globally unique, arc labels are locally unique, i.e. unique in each local neighbourhood. (*ii*) A rule can send several messages (not just one), each one to a different neighbour, e.g. the occurrence of $(u)_{\downarrow_k}(u')_{\downarrow_{k'}}$ in the right-hand side of a rule indicates that messages u, u' are to be sent via arcs labelled k, k', respectively. (*iii*) A pair of neighbouring nodes can be connected by several labelled arcs (not just one), i.e. the supporting structural digraph becomes a *multigraph*, with labelled arcs. These extensions support scaling up the problem size, without increasing the alphabet size or the number of the rules, as several cells can reuse the same labels for their outgoing arcs.

3 SDPS: P System Design

Given an epipolar stereo pair of rectified images, let $\mathbf{C} = \mathbb{XYD}$ be a discrete space of 3D points $\mathbf{p} = (x, y, d)$ of optical (visible) surfaces reduced to the left image plane $\mathbb{XY}$ with integer planar (x, y)-coordinates $x \in \mathbb{X} = \{0, 1, \ldots, n-1\}$; $y \in \mathbb{Y} = \{0, 1, \ldots, m-1\}$ and specified by the integer disparities, $d \in \mathbb{D} = \{d_{\min}, d_{\min}+1, \ldots, d_{\max}\}$, of corresponding points (x, y) and $(x-d, y)$ depicting $\mathbf{p}$ in the left and right image, respectively. Each 2D profile relating to a conjugate pair of scanlines with the same y-coordinate is given by a sequence of points $\mathbf{d}_y = \{(x, y, d) : x \in \mathbb{X};\ d \in \mathbb{D}\}$ such that for each two successive points (x', d') and (x, d) either $x = x' + 1$ and $d \in \{d', d' + 1\}$ or $x = x'$ and $d = d' - 1$.

Let $\mathbf{g}_{1:y} = \{g_1(x, y) : x = 0, 1, \ldots, n-1\}$ and $\mathbf{g}_{2:y} = \{g_2(x, y) : x = 0, 1, \ldots, n-1\}$; $y = 0, 1, \ldots, m-1$, denote grey values along the conjugate scanlines (wlg, we assume that these two lines have the same length). Let $s \in$

$\{B, M_1, M_2\}$ indicate visibility of a 3D point, i.e. the binocular, B, or only monocular observation by the left (M_1) or right (M_2) sensor. Under an assumed single continuous surface, along each profile, the visibility rules constrain the following relations between two successive points, (x', y, d', s') and (x, y, d, s): (i) $s = B$ or M_1, $s' = B$ or M_2, $x' = x - 1$, $d' = d$; (ii) $s = B$ or M_1, $s' = M_1$, $x' = x - 1$, $d' = d - 1$; (iii) $s = M_2$, $s' = B$ or M_2, $x' = x$, $d' = d + 1$.

A simplified version of the SDPS, which does not adapt for possible local contrast and offset deviations of the corresponding signals, relates the point-wise signal similarity to the absolute difference $\delta(x, y, d) = |g_1(x, y) - g_2(x - d, y)|$ between the corresponding signals for the binocularly visible point (x, y, d, B) or a regularising score $w_{\text{occl}} > 0$ for the monocularly visible ones (see [7] for more detail). Due to unequal numbers of points in profile variants to be compared by their total signal similarity, the point-wise similarities $\varphi_y(x, d, s|x', d', s')$ are integrated between the successive points with due account of their actual shifts in Cyclopean space: $\varphi_y(x, d, B|x', d', s') = \delta(x, y, d)$ if $s' \in \{B, M_2\}$ or $0.5\delta(x, y, d)$ if $s' = M_1$; $\varphi_y(x, d, M_1|x', d', s') = w_{\text{occl}}$ if $s' \in \{B, M_2\}$ or $0.5w_{\text{occl}}$ if $s' = M1$, and $\varphi_y(x, d, M_2|x', d', s') = 0.5w_{\text{occl}}$ for $s' \in \{B, M_2\}$. The SDPS maximises the similarity score between the conjugate scanlines:

$$\mathbf{d}_y^* = \arg\max_{\mathbf{d}_y} \varPhi_y(\mathbf{d}_y) = \sum_{i=1,2,\ldots} \varphi_y(x_i, d_i, s_i|x_{i-1}, d_{i-1}, s_{i-1}) \tag{1}$$

The forward pass computes potentially optimal similarity scores $F_y(x, d, s)$ and backward transitions $T_y(x, d, s)$, for each $y \in \mathbb{Y}$ by sequential pass along $x \in \mathbb{X}$ and from $d_{\max}$ to $d_{\min}$, for each $d \in \mathbb{D}$ (it can be partly parallel in $\mathbb{XD}$ plane):

$$\begin{aligned} F_y(x, d, s) &= \max_{(x',d',s') \in \varOmega(x,d,s)} \{F_y(x', d'.s') + \varphi_y(x, d, s|x', d', s')\} \\ T_y(x, d, s) &= \arg \max_{(x',d',s') \in \varOmega(x,d,s)} \{F_y(x', d', s') + \varphi_y(x, d, s|x', d', s')\} \end{aligned} \tag{2}$$

where the sets of back-transitions $\varOmega(x, d, s)$ follow from the visibility conditions. The backward pass computes the optimal profiles $\mathbf{d}_y$ in parallel across $y \in \mathbb{Y}$ and sequentially along $x \in \mathbb{X}$:

$$\begin{aligned} (x^* = n - 1, d^*, s^*) &= \arg\max_{(d,s) \in \mathbb{D} \times \mathbb{S}} \{F_y(n - 1, d, s)\} \\ (x'^*, d'^*, s'^*) &= T_y(x^*, d^*, s^*) \text{ while } x'^* > 0 \end{aligned} \tag{3}$$

where $n - 1$ is the x-coordinate of the rightmost point of each profile variant.

Design issues. Without loss of generality, our P system was designed for a pair of conjugate scanlines, which arguably is the most challenging parallelisation task. The solution can be further extended to a stereo pair of images in a straightforward manner, using "trivial" parallel processing of each pair of conjugate scanlines. We also take $d_{\min} = 0$, so the disparities range over the integer interval $[0, d_{\max}]$, and we encode integer numbers by repeating symbols, starting with one occurrence for zero (as in traditional λ-calculus), e.g., the base symbol a gives the following encodings: $0 \to a$, $1 \to a^2$, $2 \to a^3$, etc. We discuss our solution using the following example: $n = 6$, the occlusion weight $w_{\text{occl}} = 18$, and the left and right scanlines with of pixel values: $g_1 = \{15, 10, 30, 50, 15, 10\}$; $g_2 = \{10, 30, 50, 50, 15, 10\}$. The above SDPS algorithm finds the optimal similarity score $= 36$ and the profile: $d = \{0, 1, 1, 1, 0, 0\}$, for $x = \{0, 1, 2, 3, 4, 5\}$.

Cells. We construct a P system, Π, consisting of the following subcomponents. (i) L: a list of cells, $\sigma_i^l, i \in [0, n-1]$, which represent the left scan line, initialized with the left pixel values, encoded in base x. In our example, these cells are initialized, in order, with the following multisets: x^{16}, x^{11}, x^{31}, x^{51}, x^{16}, x^{11}. (ii) R: a list of cells, $\sigma_i^r, i \in [0, n-1]$, which represent the right scan line, initialized with the right pixel values, encoded in base y. In our example, these cells are initialized, in order, with the following multisets: y^{11}, y^{31}, y^{51}, y^{51}, y^{16}, y^{11}. (iii) D: a list of cells, $\sigma_i^d, i \in [0, n-1]$, which represent the disparity line, initially empty. At the end of our P program, these cells will contain the optimal disparity values, in base d. In our example, these cells will contain, in order, the following multisets: d^1, d^2, d^2, d^2, d^1, d^1. (iv) W: an array of cells, $\sigma_{ij}^w, i \in [0, n-1], j \in [0, d_{\max}]$, which represent the main workspace. Each cell σ_{ij} encloses *three virtual subcells*, respectively holding *monocular-left* values (M_1), *binocular* values (B), and *monocular-right* values (M_2). The initialisation of these cells is described later on. (v) M: a list of cells, $\sigma_j^m, j \in [0, d_{\max}]$, which represent a secondary workspace, which identifies the optimum (minimal) score, among $d_{\max} + 1$ possible scores. Figure 1 shows all these cells, for our example.

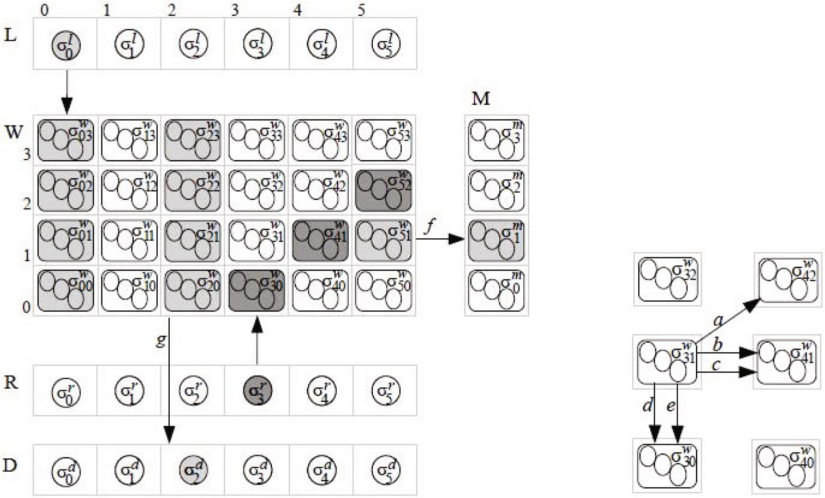


Fig. 1. Cells of Π (left) and typical arcs between W-cells (right)

Arcs. The cells are linked by arcs in a dag δ, incrementally constructed by the following enumeration. (1) Each L-cell is parent of its corresponding W-column, in the S direction: $(\sigma_i^l, \sigma_{ij}^w) \in \delta$, $i \in [0, n-1]$, $j \in [0, d_{\max}]$. (2) Each R-cell is parent of a corresponding W-diagonal, running S/W to N/E: $(\sigma_i^r, \sigma_{ij_i}^w) \in \delta$,

$i \in [0, n-1]$, $j_i \in [i, \max(n, i+d_{\max})]$. (3) Each W-cell before the last column and below the top row is parent, via an a-labelled arc, of the W-cell in the N/E direction: $(\sigma^w_{ij}, \sigma^w_{i+1j+1}) \in \delta$, $i \in [0, n-2]$, $j \in [0, d_{\max}-1]$. (4) Each W-cell before the last column is *twice* parent, via b- and c-labelled arcs, of the W-cell following it, in the E direction: $(\sigma^w_{ij}, \sigma^w_{i+1j}) \in \delta$, $i \in [0, n-2]$, $j \in [0, d_{\max}]$. (5) Each W-cell above the bottom row is *twice* parent, via d- and e-labelled arcs, of the W-cell below it, in the S direction: $(\sigma^w_{ij}, \sigma^w_{ij-1}) \in \delta$, $i \in [0, n-1]$, $j \in [1, d_{\max}]$. (6) Each rightmost column W-cell is parent, via an f-labelled arc, of its corresponding M-cell, in the E direction: $(\sigma^w_{nj}, \sigma^m_j) \in \delta$, $j \in [0, d_{\max}]$. (7) Each W-cell is parent, via a g-labelled arc, of its corresponding D-cell, in the S direction: $(\sigma^w_{ij}, \sigma^d_i) \in \delta$, $i \in [0, n-1]$, $j \in [0, d_{\max}]$.

Even for our simple example, a full picture of all these arcs is impossible, because of its sheer complexity. However, we can suggest a representative fragment of these arcs. By highlighting cells that are dag neighbours, Figure 1 (left) suggests the following arcs: (i) L-cell σ^l_0 is the parent of W-cells σ^w_{03}, σ^w_{02}, σ^w_{01}, σ^w_{00}; (ii) R-cell σ^r_3 is the parent of W-cells σ^w_{30}, σ^w_{41}, σ^w_{52}; (iii) W-cell σ^w_{51} is the parent of M-cell σ^m_1, via f-labelled arc; (iv) W-cells σ^w_{23}, σ^w_{22}, σ^w_{21}, σ^w_{20} are parents of D-cell σ^d_2, via g-labelled arcs. Also, Figure 1 (right) shows the following arcs between W-cells: (i) σ_{31} is parent of σ_{42}, via an a-labelled arc; (ii) σ_{31} is twice parent of σ_{41}, via b- and c-labelled arcs; (iii) σ_{31} is twice parent of σ_{30}, via d- and e-labelled arcs.

Workspace initialisation. Some of the W-cells are initialised as follows. (i) Each cell above the first S/W to N/E diagonal, σ^w_{ij}, $i \in [0, n-1]$, $j \in [i+1, d_{\max}]$, contains one copy of symbols w_l, w_b and w_r, interpreted as the infinite values of its monocular-left (M_1), binocular (B) and monocular-right (M_2) subcells, respectively. (ii) Each top-row cell, $\sigma^w_{id_{\max}}$, $i \in [0, n-1]$, contains one copy of symbols w_l and w_r (interpreted the same way as above). (iii) The top-left cell, $\sigma^w_{0d_{\max}}$, contains one copy of symbol k, which triggers the whole computation.

Evolution—bird's eye view. At a very high level, the system Π evolves in two major phases, closely related to the above description of the SDPS.

First, the *forward pass* phase proceeds on slope 2 diagonals, starting from the top-left cell (initially marked with k); all cells on the same diagonal work in parallel. In our sample scenario, the computational wave moves, in order, over the diagonals $\{\sigma^w_{03}\}$, $\{\sigma^w_{02}\}$, $\{\sigma^w_{01}, \sigma^w_{13}\}$, $\{\sigma^w_{00}, \sigma^w_{12}\}$, $\{\sigma^w_{11}, \sigma^w_{23}\}$, $\{\sigma^w_{10}, \sigma^w_{22}\}$, $\{\sigma^w_{21}, \sigma^w_{33}\}$, $\ldots$, $\{\sigma^w_{40}, \sigma^w_{52}\}$, $\{\sigma^w_{51}\}$ and $\{\sigma^w_{50}\}$.

During the *forward pass*, in the general case, each cell σ^w_{ij} determines: (i) in its binocular subcell, a preliminary score, which is the absolute value of the differences between the pixel values received from σ^l_i (left scanline) and σ^r_i (right scanline); (ii) in its binocular and monocular-left subcells, the sum between its previous score (cf. step i) and the minimum value of those received via: (1) arc a: σ^w_{i-1j-1}'s monocular-left score; (2) arc b: σ^w_{i-1j}'s binocular score; (3) arc c: σ^w_{i-1j}'s monocular-right score; (iii) in its monocular-right subcell, the sum between its previous score (cf. step i) and the minimum score of those received via: (1) arc d: σ^w_{ij+1}'s binocular score; (2) arc e: σ^w_{ij+1}'s monocular-right score.

Additionally, each W-cell keeps pointers to the origin of the *minimum* received scores, i.e., the label of the corresponding via arc (a, b, c, d, e), which will be used in the *backward pass* phase. When the wave reaches the rightmost column of W, the final scores are sent to the corresponding M-cells. These evaluate the minimum score and triggers the *backward pass*, which essentially follows back the pointers stored in the *forward pass*, during the evaluation of the minimum score. W-cells traversed during the *backward pass* send their disparity positions to the corresponding D-cells, which in the end return the problem's solution.

Rules. All cells start in the same initial quiescent state s_0 and they share the same ruleset, applied in weak priority order [11]. Later, cells begin to differentiate, by entering different states (according to their contents), which implicitly partitions the initial ruleset. We outline the full ruleset by dividing it into logical fragments; where each fragment is introduced by one or more expressions of type $s \overset{X:k}{\Longrightarrow} s'$, indicating a set of rules which transform the state of X cells, from s into s', in k P steps.

Here we only introduce the first two rule fragments; a complete version of these rules appears in our technical report [6], together with a trace table for our example.

$\mathbf{s_0} \overset{L,R:1}{\Longrightarrow} \mathbf{s_{99}}$, $\mathbf{s_0} \overset{W,M,D:1}{\Longrightarrow} \mathbf{s_1}$. L-cells and R-cells transfer their pixel values to their corresponding W children cells and transit to state s_{99}; all other cells, i.e. W, M and D, transit to state s_1; everything in one P step.

1. $s_0\ x \rightarrow_{\texttt{max}} s_{99}\ x\ (l_b)_{\downarrow}$ 2. $s_0\ y \rightarrow_{\texttt{max}} s_{99}\ y\ (r_b)_{\downarrow}$ 3. $s_0 \rightarrow_{\texttt{min}} s_1$

$\mathbf{s_1} \overset{W:4}{\Longrightarrow} \mathbf{s_4}$, $\mathbf{s_1} \overset{M,D:1}{\Longrightarrow} \mathbf{s_{27}}$. Each W-cell computes the absolute value of the differences between the received left and right pixel values and transits to state s_4, in four P steps. Each other cell, i.e. M and D, transits to state s_{27}, in one P step.

1. $s_1\ w_b \rightarrow_{\texttt{min}} s_2\ w_b$
2. $s_1\ l_b\ r_b \rightarrow_{\texttt{min}} s_3\ z_b$
3. $s_1\ l_b\ r_b \rightarrow_{\texttt{max}} s_3$
4. $s_1\ l_b \rightarrow_{\texttt{max}} s_3\ z_b$
5. $s_1\ r_b \rightarrow_{\texttt{max}} s_3\ z_b$
6. $s_1\ l_b \rightarrow_{\texttt{max}} s_2$
7. $s_1\ r_b \rightarrow_{\texttt{max}} s_2$
8. $s_1 \rightarrow_{\texttt{min}} s_{27}$
9. $s_2 \rightarrow_{\texttt{min}} s_4$
10. $s_3 \rightarrow_{\texttt{min}} s_4$

The asymptotic runtime complexity of our P model, which is arguably optimal, is summarized by Theorem 1. The proof follows from the structure and their execution of the above rules in the P system. In contrast, the best implementation on existing parallel hardware is limited, by hardware resources, to $O(nd/q)$, where q is the number of available processors.

Theorem 1. *The P system described in Section 3 completes in $O(n+d)$ P steps.*

4 Conclusion

In this paper, we presented a massively parallel P model for implementing a critical part of the SDPS algorithm. Our solution is based on simple P modules, with extensions which allow our model to adapt to complex cases such as SDPS. Our model processes in parallel all potentially optimal similarity scores that trace candidate decisions, for all the disparities associated with each current x-coordinate. We plan to further extend the solution, from pairs of scanlines to stereo pairs of full images. To avoid complex and lengthy arguments, in this paper we did not discuss how all the cells have been created and initialised. We also plan to investigate an advanced model, which starts as one single "ur-cell" and then grows, until it reaches the proper size and shape, required by the SDPS solution. Since we only managed to implement a sequential version of simulator while designing this P model, we intend to implement an improved simulator to leverage the parallel features of this design on parallel hardwares.

References

1. The P systems webpage, http://ppage.psystems.eu
2. Carnero, J., Díaz-Pernil, D., Molina-Abril, H., Real, P.: Image segmentation inspired by cellular models using hardware programming. In: González-Díaz, R., Real-Jurado, P. (eds.) 3rd International Workshop on Computational Topology in Image Context, pp. 143–150 (2010)
3. Christinal, H.A., Díaz-Pernil, D., Real, P.: P systems and computational algebraic topology. Mathematical and Computer Modelling 52(11-12), 1982–1996 (2010)
4. Dinneen, M.J., Kim, Y.B., Nicolescu, R.: A faster P solution for the Byzantine agreement problem. In: Gheorghe, M., Hinze, T., Păun, G. (eds.) CMC 2010. LNCS, vol. 6501, pp. 175–197. Springer, Heidelberg (2010)
5. Dinneen, M.J., Kim, Y.B., Nicolescu, R.: P systems and the Byzantine agreement. Journal of Logic and Algebraic Programming 79(6), 334–349 (2010)
6. Gimel'farb, G., Nicolescu, R., Ragavan, S.: P systems in stereo matching. Report CDMTCS-401, Centre for Discrete Mathematics and Theoretical Computer Science, The University of Auckland, Auckland, New Zealand (April 2011), http://www.cs.auckland.ac.nz/CDMTCS//researchreports/401NG.pdf
7. Gimel'farb, G.L.: Probabilistic regularisation and symmetry in binocular dynamic programming stereo. Pattern Recognition Letters 23(4), 431–442 (2002)
8. Nicolescu, R., Dinneen, M.J., Kim, Y.B.: Towards structured modelling with hyperdag P systems. International Journal of Computers, Communications and Control 2, 209–222 (2010)
9. Păun, G.: Computing with membranes. Journal of Computer and System Sciences 61(1), 108–143 (2000)
10. Păun, G.: Membrane Computing: An Introduction. Springer-Verlag New York, Inc., Secaucus (2002)
11. Păun, G.: Introduction to membrane computing. In: Ciobanu, G., Pérez-Jiménez, M.J., Păun, G. (eds.) Applications of Membrane Computing. Natural Computing Series, pp. 1–42. Springer, Heidelberg (2006)
12. Păun, G., Pérez-Jiménez, M.J.: Solving problems in a distributed way in membrane computing: dP systems. International Journal of Computers, Communications and Control 5(2), 238–252 (2010)
13. Păun, G., Rozenberg, G., Salomaa, A.: The Oxford Handbook of Membrane Computing. Oxford University Press, Inc., New York (2010)

Functional Brain Mapping by Methods of Evolutionary Natural Selection

Mohammed Sadeq Al-Rawi[1] and João Paulo Silva Cunha[2]

[1] IEETA-Instituto de Engenharia Electrónica e Telemática de Aveiro,
University of Aveiro, 3810-193 Aveiro, Portugal
al-rawi@ua.pt
[2] IEETA- Instituto de Engenharia Electrónica e Telemática de Aveiro, Dept. of Electronics,
Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal
Brain Imaging Network (ANIFC), Coimbra, Portugal
jcunha@ua.pt

Abstract. We used genetic algorithms to detect active voxels in the human brain imaged using functional magnetic resonance images. The method that we called EVOX deploys multivoxel pattern analysis to find the fitness of most active voxels. The fitness function is a classifier that works in a leave-one-run-out cross-validation. In each generation, the fitness value is calculated as the average performance over all cross-validation folds. Experimental results using functional magnetic resonance images collected while humans (subjects) were responding to attention visual stimuli showed certain situations that EVOX has could be useful compared to univariate ANOVA (analysis of variance) and searchlight methods. EVOX is an effective multivoxel evolutionary tool that can be used to tell where in the brain patterns responding to stimuli are.

Keywords: genetic algorithms, fMRI, human brain mapping, neuroimaging, multivoxel pattern analysis.

1 Introduction

Neuroscientists worldwide have made extensive efforts to understand the functional and neuroanatomical design of one of nature's topmost achievements, the human brain. A correlate to brain neuronal responses could be measured via various methods, e.g., EEG, MEG, PET, fMRI, TMS, DOT, etc. Functional magnetic resonance imaging (fMRI) provides a neat noninvasive technique that can be used to measure the blood oxygenation level dependent (BOLD) in the brain. It is widely believed/accepted that BOLD signals correlates to neuronal activity.

In fMRI studies, however, there is a desire to know what parts of the brain responded preferentially to some stimuli. The most commonly used method to detect such peek activations is based on univariate analysis, for example using analysis of variance (ANOVA). Searchlight (Kriegeskorte and Goebel, 2006), use multivoxel pattern analysis (MVPA) (Detre et al., 2006; Pereira et al, 2009) which is based on defining a local spherical mask, centered in turn on every possible voxel inside the brain, and perform a multivoxel test on each of those spherical regions. Searchlight

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 293–299, 2011.

has the disadvantage that its performance is restricted by adjacent voxels in the spherical mask. Both ANOVA and searchlight give a score value for each voxel and one can select the active/informative voxels by using a threshold. While ANOVA is a univariate (univoxel) approach, searchlight is multivoxel but it is restricted to search a neighborhood of voxels. In this work, we plan to expand the range of what voxels to be searched for activations using evolutionary genetic algorithms.

Based on the notion of evolutionary natural selection (Holland, 1992), we present a method we named EVOX (evolutionary based voxel search) for the purpose of dimensionality reduction, feature selection, and functional brain mapping. Because EVOX evolves stochastically through time and may take all the brain voxels as possible inputs, it might enable us to find solutions that are out of the scope of both ANOVA and searchlight.

2 Methods and Data

Genetic algorithms are evolutionary algorithms that form an effective solution for optimization problems. They can also can be considered as probabilistic search algorithms (Holland, 1992). They operate on a set of individuals called population, where each individual is an encoding of the problem (in our case brain voxels) and is called a chromosome, and each individual's fitness is calculated using an objective function. In genetic algorithms terminology, each iteration of the search is called a generation. From each generation, the fittest individuals are selected and pooled out to form a base for a new population with better characteristics. To sum, genetic algorithms are characterized by attributes such as fitness function, encoding of the input data, crossover, mutation, population size, migration, etc.

In this work, EVOX starts iteration with a random population of chromosomes where each chromosome is a mask that selects a set of voxels from the whole-brain volume or regions-of-interest, and then a score for each chromosome is estimated using a fitness function. Our fitness function in this work is the average accuracy of a classifier that implements a leave-one-epoch/run out cross-validation by reserving one run from the time series to be used for testing and training using the remaining runs.

2.1 Parameters Used in EVOX

FMRI data are acquired for each subject in a session and each session usually has several chunks of brain volumes that are called runs. The stimulus presentation usually varies across runs. Below we list the parameters used in EVOX:

Encoding: Each chromosome is represented as a binary string, ‘1’ for voxel inclusion and ‘0’ for voxel exclusion. A chromosome that contains all ‘1’s would select all the brain voxels. Thus, the chromosome is size is determined by the number of voxels in the region to be searched. It would be more feasible that each chromosome in the population represents a region-of-interest in the brain, or even several regions. In such case, the chromosome contains the voxels of these regions, if all the values in the chromosome are ‘1’s means all the voxels are included, but through evolution, more voxels will be in or out, but usually less than the whole-region voxels.

Fitness function: The purpose is finding the fitness value for each chromosome in the population. In this work, the fitness function is a classifier that executes in a leave-one-run-out cross-validation folds and the fitness value is calculated as the average performance over all folds. Gaussian Naïve Bayes (GNB) is fast but Support Vector Machines (SVM) and Logistic Regression (LOGREG) are slow and more accurate, see (Detre et al., 2006; Pereira et al., 2009) for MVPA and these classifiers. To detect the maximum response to a single category ω, we can use the following fitness function;

$$fitness = \frac{\text{mean classification accuracy of } \omega}{\text{mean classification accuracy of all classes except } \omega}. \quad (1)$$

See Fig. 1 that illustrates the basic idea of finding the fitness in EVOX.

EVOX levels (*n*): First, the population of chromosomes is randomly initialized to binary values between 0 and 1, then after #*m* generations, EVOX is stopped and we get the fittest chromosome. To run the next level, we increase the crossover fraction and pass the resultant fittest chromosome of the previous level and replace it with one individual of the randomly initialized population and start EVOX again for up to #*m* generations. This approach helps speeding up EVOX since each fittest chromosome is smaller than its parents are. GNB is used at the first few levels, then, when the fittest chromosome size, which will be feed in to the next level, is small enough, an SVM and/or a LOGREG classifier can be used. The value of *n* which mimics the number of levels is chosen by the user.

Elite count (1): here, one individual with the highest (classifier performance) fitness value is chosen to survive to the next generation.

Crossover fraction set :{0.5}: This specifies the fraction that forms crossover children from the population, other than elite children. We may start with low a value (eg, 0.1) to enable high mutation, a then increase it with EVOX levels.

Mutation fraction: the remaining individuals from the population other than elite and crossover children are mutation children.

Population size (16): We set the population size with a low number because intensive computations are needed for determining an individual's fitness, however, a higher population size might give better results.

EVOX stopping criteria: There are many ways to stop EVOX, stall time, stall iterations, number of EVOX levels, number of generations, fitness value, and the number of selected voxels. In fact, we can run EVOX and continue running it until the minimum number of maximally responsive voxels is fixed which might be of interest to many applications.

EVOX with specific problem constraints: Lots of constraints could be enforced on the voxel selectivity problem, for example, it is possible to select one universal set of voxels for all epochs, or even subjects when EVOX searches for voxels among subjects. One can also change the encoding to select a set of adjacent voxels rather than probable sparse voxels.

2.2 Data

In the dataset we have adopted for this work which is publicly available from the fMRIDC (accession no. 2-2000-1113D), five right-handed subjects with normal vision participated in the experiment and performed delayed matching (delay-match-to-sample) task with photographs and with line drawings of houses, faces, and chairs. In the delayed matching task, a couple of choice stimuli followed each single-sample stimulus after a delay and subjects pressed a button with the right or left thumb to indicate which stimulus matched the sample. Tasks were presented in 21-s blocks using the same category of stimuli. All blocks with meaningful stimuli were separated by control blocks which contained nonsense images. The order of blocks with meaningful stimuli was counterbalanced across runs. Each run contained six blocks with meaningful stimuli, two for each stimuli category, and each run contained either blocks with either matching photographs tasks or matching line drawings tasks. The dataset was acquired using echo-planar imaging (EPI), and the volumes were registered with an iterative method, spatially smoothed in plane with a Gaussian filter (FWHM was 3.75 mm along the x- and y-axis).

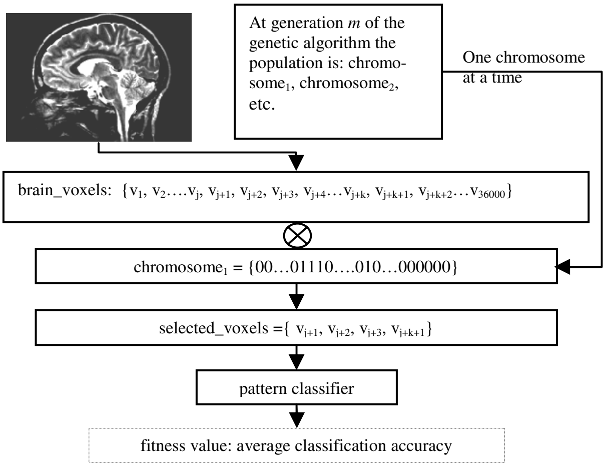

At generation m of the genetic algorithm the population is: chromosome$_1$, chromosome$_2$, etc.
One chromosome at a time
brain_voxels: $\{v_1, v_2 \ldots v_j, v_{j+1}, v_{j+2}, v_{j+3}, v_{j+4} \ldots v_{j+k}, v_{j+k+1}, v_{j+k+2} \ldots v_{36000}\}$
$\otimes$
chromosome$_1$ = {00…01110….010…000000}
selected_voxels ={ $v_{j+1}, v_{j+2}, v_{j+3}, v_{j+k+1}$}
pattern classifier
fitness value: average classification accuracy

Fig. 1. An example of how a chromosome selects five voxels from a brain that contains 36000 voxels as performed in the method proposed in this work. The set brain_voxels is masked with the chromosome to yield the five selected voxels. Genetic algorithms population contains several chromosomes and the fitness of each chromosome is calculated using a pattern classifier. After many generations, the ultimate fittest chromosome will be the output and it will resemble the best representative voxels, or the most active voxels in due to the performed task.

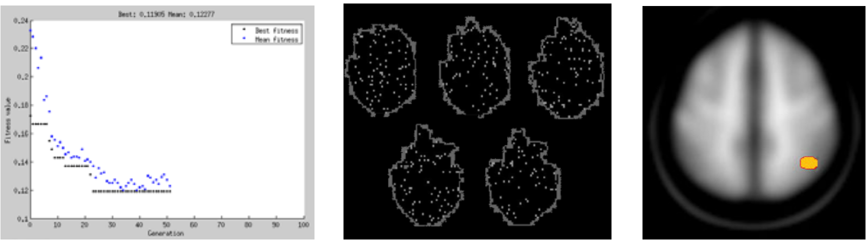
FMRI data of each subject has 64×64×18×1092 voxels, where 1092 is the number of brain volumes. In cross-validation experiments, 1001 volumes are used in training and 92 volumes are used in testing for each of the 12 folds. Since the used fMRI dataset contained several conditions, we have suppressed all conditions except those that contained grayscale photographs and line drawings of the same stimuli category. We also used an intracranial mask to filter out unwanted voxels from each volume. Photographs and line drawings do not necessarily belong to the same object, but they do belong to the same stimuli category, e.g. face stimulus.

3 Results

We implemented several EVOX experiments to detect peek brain activity using subjects who responded to line drawings and grayscale photographs of the same stimuli category. For comparison purpose, we performed the analysis using whole-brain data, ANOVA feature selection, and GLM. We found that patterns of brain activity when subjects responded to line drawings differ from their responses to photographs. Our leave-one-run-out classifier gave high accuracy for most of the subjects that have been used in this study. Although genetic algorithms produced brain maps that need further investigation, the results provide evidence that the mechanisms involved in processing shape information greatly depend on visual cues. Fig. 2-a shows the convergence of genetic algorithms when searching for the most discriminative voxels of houses versus faces, and Fig. 2-b shows the location of these discriminative voxels.

In another experiment, we used two categories from subject no. 2, face photographs, and face line drawings. Running EVOX for many levels reached 100% classification accuracy. The number of active (max responsive) voxels at each of the 12 runs is {52, 66, 65, 76, 66, 56, 56, 67, 48, 68, 68, 49}. These voxels defines the region of maximum response in the brain to face stimulus. These results are depicted in Fig. 2-c. For comparison purpose, we run other tests using ANOVA on raw fMRI data (No active voxels were detected using ANOVA on z-scored fMRI data), p_thresh=1e-120. We also tried using searchlight of radius 2 voxels on z-scored fMRI, a GNB classifier is used as the multivoxel statistical test for each sphere.

In another experiment, we added a clustering constraint to the fitness function of EVOX and we were able to detect peek activation at the superior parietal lobule, as illustrated in Fig. 2-b, when subjects responded to face line drawings and to face photographs in a delayed-match-to-sample. Due to using genetic algorithms, EVOX is slow, for example, using 200 iterations of searching the ventral temporal cortex for some activation (with one-level EVOX that implements Gaussian Naïve Bayes classifier) will take ten minutes. The execution time may go higher if larger brain regions are included in the search.



(a) (b) (c)

Fig. 2. a) EVOX convergence using fMRI data of a subject who responded to house delayed matching and face delayed matching stimuli. A logistic regression classifier was used in this experiment to calculate the fitness function. b) Detected active voxels responding to house delayed matching and face delayed matching after EVOX converged and stopped. c) Peek activation detected using EVOX (genetic algorithms) at the right superior parietal lobule for subjects responded to face photographs and face line drawings.

Aside from the experiments shown above, we performed another experiment using an fMRI dataset that contains responses to face, house, shoe, bottle, scissor, chair, cat, scrambled images (Haxby et al., 2001). We used genetic algorithms to find the voxels that are maximally responsive to face category. The resultant face response region is depicted in Fig. 3 and is compared to the region representing face versus other objects that was provided by Haxby et al (2001).

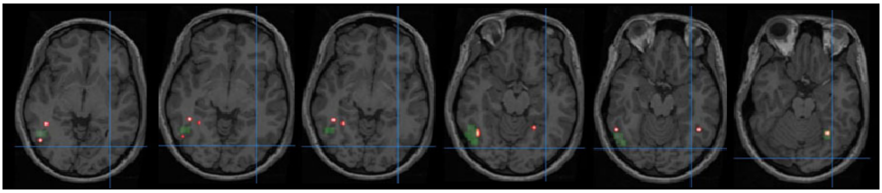


Fig. 3. Face-responsive area detected with EVOX is close to face area produced by general linear model. The red dots are those detected with EVOX and the green area is the maximally responsive face area detected by univariate analysis (voxels that had maximal responses averaged across all runs). EVOX only searched the ventral temporal cortex no clustering option was used. FMRI data were shifted to compensate for the hemodynamic lag, detrended, and z-scored.

4 Conclusions

The proposed method can be used in situations where across conditions covariance of functional data is very low, as in the case discussed in this work where we were interested in studying response to face photograph versus face line drawing. EVOX, which is the method we proposed based on genetic algorithms, dynamically searches the whole-brain or a predefined region of the brain for active voxels or clusters. Functional brain maps obtained using the EVOX shows distinctive activations compared to ANOVA and searchlight. Although EVOX is a bit slow, the concept of EVOX levels as well using a low number of 1's in chromosomes would enhance its speed. The maximally responsive regions detected using EVOX were close to those detected using univariate analysis but do not coincide with them exactly.

Acknowledgment. This work was supported by the "Fundação para a Ciência e a Tecnologia (FCT)", PTDC/PSI/67381/2006, GRID/GRI/81833/2006, C2007-UA/IEETA/SP/02. We would like to thank Alumit Ishai, James Haxby and the rest of their group for putting their fMRI dataset publically available at fMRIDC, and Princeton MVPA project and those who are working on its development.

References

Detre, G.J., Polyn, S.M., Takerkart, S., Natu, V.S., Benharrosh, M.S., Singer, B.D., Cohen, J.D., Haxby, J.V., Norman, K.A.: The Multi-Voxel Pattern Analysis (MVPA). In: 12th Meeting of the Organization of Human Brain Mapping, Florence, Italy (2006)

Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fMRI: A tutorial overview. Neuroimage 45, S199-S209 (2009)

Kriegeskorte, N., Goebel, R., Bandettini, P.: Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868 (2006)

Holland, J.H.: Genetic Algorithms. Sci. Am. 267, 66–72 (1992)

Ishai, A., Ungerleider, L.G., Martin, A., Schouten, H.L., Haxby, J.V.: Distributed representation of objects in the human ventral visual pathway. Proc. Natl. Acad. Sci. U. S. A. 96, 9379–9384 (1999)

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430 (2001)

Interactive Classification of Remote Sensing Images by Using Optimum-Path Forest and Genetic Programming

Jefersson Alex dos Santos[1], André Tavares da Silva[2], Ricardo da Silva Torres[1], Alexandre Xavier Falcão[1], Léo P. Magalhães[2], and Rubens A.C. Lamparelli[3]

[1] Institute of Computing
[2] School of Electrical and Computer Engineering
[3] Center for Research in Agriculture
University of Campinas, Campinas, SP, Brazil
{jsantos,rtorres,afalcao}@ic.unicamp.br, atavares@dca.fee.unicamp.br,
leopini@fee.unicamp.br, rubens@cpa.unicamp.br

Abstract. The use of remote sensing images as a source of information in agribusiness applications is very common. In those applications, it is fundamental to know how the space occupation is. However, identification and recognition of crop regions in remote sensing images are not trivial tasks yet. Although there are automatic methods proposed to that, users very often prefer to identify regions manually. That happens because these methods are usually developed to solve specific problems, or, when they are of general purpose, they do not yield satisfying results. This work presents a new interactive approach based on relevance feedback to recognize regions of remote sensing. Relevance feedback is a technique used in content-based image retrieval (CBIR) tasks. Its objective is to aggregate user preferences to the search process. The proposed solution combines the Optimum-Path Forest (OPF) classifier with composite descriptors obtained by a Genetic Programming (GP) framework. The new approach has presented good results with respect to the identification of pasture and coffee crops, overcoming the results obtained by a recently proposed method and the traditional Maximimun Likelihood algorithm.

Keywords: Remote Sensing Image Classification, Genetic Programming, Optimum-Path Forest, Relevance Feedback.

1 Introduction

Agriculture productivity is strongly dependent on monitoring and planning activities. Production estimation and land use are the basis for government policies to finance agricultural activities. Thus, there is a huge demand for information systems that allow to store, analyze, and handle geographic data. This is the purpose of Geographic Information Systems (GISs). Most of existing GIS-based applications rely on the use of Remote Sensing Images (RSIs) to crop monitoring.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 300–307, 2011.

Because RSIs are raster data, to obtain vectorial information is necessary to extract regions of interest. In addition to the typical problems in pattern recognition research, identifying crop areas in RSIs faces hard problems associated, for instance, with terrain distortions. Besides that, RSIs, contrary to common images, do not encode just human visible information, but also other spectral bands (for example, infrared). For this reason, the recognition task normally needs classification strategies which exploit RSI properties related to both spectral and texture patterns.

The process of recognizing regions is called classification and can be done both automatically or manually. Sometimes users prefer to identify regions manually because the results of automatic approaches are unsatisfactory. The most successful RSI classification methods are normally created to a specific target or data [14]. General purpose methods, however, are very sensitive to noise. Furthermore, the spectral response and the texture patterns observed for a given crop can be different. A crop can be planted in different ways and this factor, allied to the different phases of plants, tends to create distinction between regions where the same culture is being cultivated. Therefore, in practical situations, the results of automatic methods need to be revised.

This work aims to present a new interactive approach for classifying regions in RSIs. The proposed solution relies on the use of an interactive strategy, called *relevance feedback* (RF), based on which the classification system can learn what regions are of interest, given what is indicated by users. The proposal is a new hybrid method, named $GOPF$, which uses a GP framework to create composite image descriptors [5] and the optimum-path forest (OPF) classifier [11] to determine regions of interest. OPF is a classification method which represents each class of objects by one or multiple optimum-path trees rooted at key samples, called prototypes.

2 Related Work

In [9], Lu & Weng present an overview of the problem of classification of remote sensing images including the steps which comprise the process (extraction of features, segmentation, classification and accuracy assessment) and the research challenges faced. For each step, most of the existing techniques until 2005 are presented, grouped by the approach adopted (such as techniques that exploits classification by pixels or regions).

The classification algorithm based on pixels, MaxVer (*Maximum Likelihood Classification*) [12], is still one of the most popular. On the other hand, the growth of classification approaches based on regions, for instance, is analyzed in [3]. The article proves that the growth in the number of new approaches published accompanies the increase of the accessibility to high-resolution images and, hence, the development of alternative techniques to the classification based on pixels.

Apart from the classification methods presented in [14,12,3], several others have been proposed recently [8,13,1,10,2]. The novelty of these approaches relies on: resolution and number of bands of ISRs; type of extracted features; used learning technique, and level of discrimination among the classes of the image (some studies include all the vegetation types in the same class, for example). Li et. al [8] proposed a method based on a regions adjacency graph for segmentation of images with high resolution. The regions are segmented according to the combination of shape, color and texture features. The RSIs classification methods proposed by Munoz-Mari et al. [10] and Basi & Melgani [2] are based on Support vector machines (SVMs). The first [10] proposes a supervised classification method called SVDD. The experiments were made by differentiating various classes of vegetation (corn, grass, pasture, trees, etc.). As far as [2] is concerned, they proposed an RSIs classification system based on SVM in which Genetic Algorithms (GA) are used to find the best set of parameters of SVM. The extracted features are based on pixels. Low-resolution aerial images were used. Besides [2], both RSIs classification methods proposed by Tseng et. al [13] and Bandyopadhyay et. al [1] use Genetic algorithms (GA). The former uses GA to find the configuration parameters of a neural network while the latter uses GA for clustering the pixels of the RSIs. Although both use the pixel information, in [13], the vegetation classes are distinguished and images of middle and high resolution are used.

The main advantages of the proposed method against the aforementioned approaches are the use of a runtime technique for combination of features (genetic programming) and the refinement of the OPF-based classification system by taking into account the user interaction. Moreover, most of the mentioned works do not address the problem of specific crops recognition. They group different classes into larger sets or even into a single one. We have recently proposed an interactive method for classification of remote sensing images based on Genetic Programming, GP_{SR} [6]. That method allows users to interact with the classification system, indicating regions of interest (and those which are not). This feedback information is employed by a genetic programming approach for learning user preferences and combining image region descriptors that encode spectral and texture properties. One remarkable advantage of the proposed method when compared with GP_{SR} is that it does not need thresholds for selecting seeds to be used in the segmentation process. At the end of each relevance feedback interaction, the classifier itself defines the relevance level for all of the subimages. GP_{SR} is used as baseline in our experiments.

3 The *GOPF* Approach

The *GOPF* approach is a framework for recognition of regions of interest in remote sensing images combining *OPF* [11] classifier and *GP* [5] composite descriptor. We also adopt the definition of simple descriptor used in [5], which is composed by a pair consisting of an extraction function and a distance function.

Optimum-path forest (OPF) is a classification method that represents each class of objects by one or multiple optimum-path trees rooted at key samples, called prototypes [11]. The training samples are nodes of a *complete graph.* In our work, the arcs are weighted by the distance provided by the composite descriptor of their nodes. The use of OPF for relevance feedback considers two classes: relevant subimages (chosen by the user) and irrelevant ones. The prototypes computed by the OPF classifier are used to rank database images according to the user's selection.

Algorithm 1 illustrates how *GOPF* is used in the classification system. Let $\hat{I}$ be an RSI divided into a set of subimages (block regions). The distance $d(s,t)$ between two subimages s and t is the distance between their corresponding feature vectors combined by a function (composite descriptor). For an initial query point s, the proposed method returns the N closest subimages in $\hat{I}$ to s (query by similarity). Due to the semantic gap problem, the closest subimages to s may not be the most relevant for a given user. By marking the relevant subimages among the returned ones, the user creates two sets: a set $\mathcal{I} \subset \hat{I}$ of irrelevant subimages and a set $\mathcal{R} \subset \hat{I}$ of relevant subimages. The method then uses sets R and I to compute relevant (set A) and irrelevant (set B) sets of prototypes and two optimum-path forests rooted at them. Each subimage $t \in \hat{I} \backslash \mathcal{I} \cup \mathcal{R}$ is then classified according to the root's label of the forest (relevant/irrelevant) that offers to t the optimum path from $\mathcal{A} \cup \mathcal{B}$. Only the N closest images labeled as relevant will be returned (in a set $\mathcal{C}$) to the user in the next iteration. Relevant prototypes ($\mathcal{A}$) and irrelevant ones ($\mathcal{B}$), computed in the previous step, are then used to sort the subimages in $\mathcal{C}$ for the next iteration. The method computes the average distance $\bar{d}_{\mathcal{A}}(t,\mathcal{A})$ between each subimage $t \in \mathcal{C}$ and subimages in the set of relevant prototypes $\mathcal{A}$. It also computes the average distance $\bar{d}_{\mathcal{B}}(t,\mathcal{B})$ between t and subimages in the set of irrelevant prototypes $\mathcal{B}$. Finally, a distance $\bar{d}(t,\mathcal{A},\mathcal{B})$ is computed as a normalized mean between relevant and irrelevant prototypes according to the composite descriptor: $\bar{d}(t,\mathcal{A},\mathcal{B}) = \dfrac{\bar{d}_{\mathcal{A}}(t,\mathcal{A})}{\bar{d}_{\mathcal{A}}(t,\mathcal{A}) + \bar{d}_{\mathcal{B}}(t,\mathcal{B})}$. After classifying each subimage in $\hat{I} \backslash \mathcal{I} \cup \mathcal{R}$, the method returns to the user a new set of N relevant subimages, which contains the lowest values of $\bar{d}(t,\mathcal{A},\mathcal{B})$ This process is then repeated for a few iterations T and, finally, the system returns all relevant subimages obtained so far. The complete approach, which we call *GOPF* (GP+OPF), is illustrated in Figure 1 (a).

The performance of the classifier is directly dependent on the good description of the objects involved in the classification. In order to combine spectral and texture distances from various descriptors, the distance $d(s,t)$ between two images s and t used by OPF classifier is provided by a GP-based composite descriptor [5]. The GP framework requires a training set to find a good combination function. As the method is interactive we propose training GP with the prototypes ($\mathcal{A}$ and $\mathcal{B}$) provided by the OPF since they are very informative subimages. The GP module starts with a population of combination functions created randomly. This population evolves generation by generation through genetic operations (e.g., crossover, mutation, reproduction). A fitness function is used to assign the fitness value for each individual based on the ranking of the training set. This value is used to select the best individuals. Next, genetic operators are applied to this population aiming to create more diverse and better performing individuals.

Algorithm 1. The subimage recognition process in the *GOPF*.

```
Compute the distance $d(s,t)$ from the descriptors for every subimage $t \in \hat{I}$.
Create an ordered list $L$ of the $N$ closest subimages $t$ to $s$ based on $d(s,t)$.
Set $\mathcal{I} \leftarrow \emptyset$ and $\mathcal{R} \leftarrow \emptyset$.
for each learning iteration $i = 1, 2, \dots, T$ do
    Set $\mathcal{C} \leftarrow \emptyset$.
    The user marks the relevant subimages in $L$, which are inserted into $\mathcal{R}$ and the
      irrelevant ones are inserted into $\mathcal{I}$.
    if $|\mathcal{R}| < N$ then
        Compute OPF using sets $\mathcal{I}$ and $\mathcal{R}$, resulting also $\mathcal{A}$ and $\mathcal{B}$ (prototypes).
        for each subimage $t \in \hat{I} \backslash \mathcal{I} \cup \mathcal{R}$ do
            if $t$ is labeled as relevant by OPF then
                Insert $t$ into the set $\mathcal{C}$ of images classified as relevant.
            end if
        end for
    else
        Show the final ordered list $L$ with the $N$ most relevant subimages in $\mathcal{R}$, as
      defined by the user selection.
    end if-else
    Create an ordered list $L$ with the $N$ most relevant subimages in $\mathcal{C}$, in increasing
      order of $\bar{d}(t, \mathcal{A}, \mathcal{B})$.
    Apply $GP$ to find the distance combination function $f(d_i(s,t))$ by using $\mathcal{A}$ and $\mathcal{B}$
      as training set.
    Recombine the subimages distances by using the best $GP$ function $f(d_i(s,t))$
end for
Return the final ordered list $L$ with the $N$ most relevant subimages in $\mathcal{R}$, completing
      it with the $N - |\mathcal{R}|$ relevant subimages in $\mathcal{C}$ in the increasing order of $\bar{d}(t, \mathcal{A}, \mathcal{B})$.
```

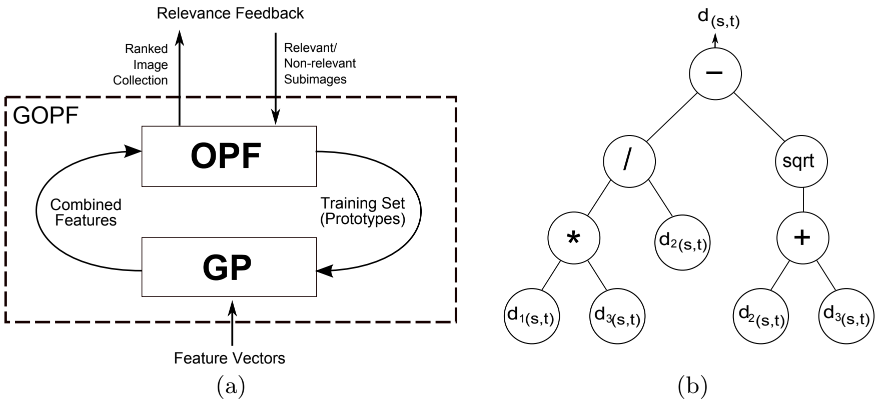
The last step consists in determining the best individual to be applied to the test set. The commonest choice is the individual with the best performance in the training set (e.g., the first function of the last generation).

In this work we adopt the same notion of descriptor proposed in [5]. In this case, a *GP* individual is a function used to combine the distances provided by a set of single descriptors concerning the features extracted from two subimages. The GP individual configuration in the *GOPF* is the same as that used for GP_{SR} method [6]. Figure 1 (b) shows an example of GP individual as a function to combine descriptors from two subimages. This individual corresponds to the function $f(d_1(s,t), d_2(s,t), d_3(s,t)) = \dfrac{d_1(s,t) * d_3(s,t)}{d_2(s,t)} - \sqrt{d_2(s,t) + d_3(s,t)}$.

4 Experiments

This section describes the experiments performed to validate our method. The experiment Setup is described as the following:

GP Parameters: Population size was 60; the number of generations, 10; initial population, *half-and-half*; initial tree depth, between 2 and 5; maximum tree



(a) (b)

Fig. 1. The proposed interactive classifier (GOPF)

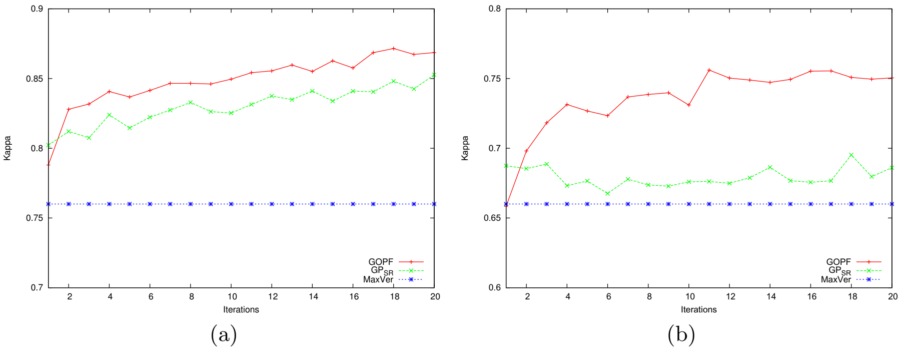
depth, 5; selection method, tournament with size 2; crossover rate, 0.8; mutation rate, 0.2; and functions set is composed by the operators $+, *, \sqrt{}, d^{const}$. **Descriptors:** We used the same set of descriptors in [6]: BIC, Color Histograms, Color Moments, Gabor Wavelets, and Spline Wavelets. **Baseline:** We compare our method against Maximum Likelihood Classification (MaxVer) [12] and GP_{SR} [6]. PASTURE image was classified by *MaxVer* with probability threshold 0.8 and using 20,580 points of pasture sample. COFFEE image was classified with probability threshold 0.98 and using 43,630 points of coffee sample. **Effectiveness measure:** We use *kappa–iterations* curves as effectiveness measure. Kappa [4] is an effective index to compare classified images, commonly used in RSI classification. Experiments in different areas show that kappa could have various interpretations and these guidelines could be different depending on the application. However, Landis and Koch [7] characterize Kappa values over 0.80 as "almost perfect agreement", 0.60 to 0.79 as "substantial agreement", 0.40 to 0.59 as "moderate agreement", and below 0.40 as "poor agreement". Kappa negative means that there is no agreement between classified and verification data. **Data:** Two RSIs were used to validade our method. Information about used RSIs is showed in Table 1. In this paper, we call Image 1 PASTURE and image 2 COFFEE. These images were divided into subimages of 20 × 20 and 30 × 30 pixels, respectively. Thus, the image PASTURE is composed by 5900 while COFFEE is composed by 6400 subimages. A total of 100 different subimages were used as initial query point s. The results presented are the average for each inital pattern s.

Figure 2 (a) illustrates the curve kappa-iterations of the $GOPF$ in comparison with GP_{SR} and the value found by the $MaxVer$ for the PASTURE image. Note that the value of kappa is always greater for the proposed method when compared to MaxVer, along iterations. Note also that the hybrid approach, $GOPF$, yields better results than GP_{SR}.

Table 1. Remote Sensing Images used in the experiments

	Image 1	**Image 2**
Region of interest	pasture	coffee
Terrain	plain	mountainous
Satellite	CBERS	SPOT
Spatial resolution	20 meters	2.5 meters
Bands composition	R-IR-G (342)	IR-NIR-R (342)
Acquisition date	08–20–2005	08–29–2005
Location	Laranja Azeda Basin, MS	Monte Santo County, MG
Dimensions (px)	1310 × 1842	2400 × 2400

Figure 2 (b) illustrates the curve-kappa iterations of *GOPF* in comparison with the kappa value found by GP_{SR} and *MaxVer* for the COFFEE image. The kappa values for the proposed methods were better than the MaxVer score. Another remarkable result is concerned with the superiority of *GOPF* when compared to GP_{SR}.



(a) (b)

Fig. 2. Kappa X Iterations comparing the results of the proposed method, GP_{SR} and the MaxVer to the PASTURE image (a) and to the COFFEE image.

5 Conclusions

We have proposed a hybrid framework for recognition of regions of interest in remote sensing images which combines *OPF* [11] classifier and *GP* [5] composite descriptor. The system uses image descriptors to encode the spectral and texture regions of the RSIs and exploits user's relevance feedback. *GOPF* has presented good results with respect to the identification of pasture and coffee crops, overcoming the results obtained by GP_{SR} [6] and the MaxVer algorithm. As future works, we plan to evaluate more image descriptors; to allow user to define multiple regions as query pattern; and to compare the method with other baselines.

Acknowledgments. The authors are grateful to CAPES, FAPESP, and CNPq for financial support.

References

1. Bandyopadhyay, S., Maulik, U., Mukhopadhyay, A.: Multiobjective Genetic Clustering for Pixel Classification in Remote Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing 45, 1506–1511 (2007)
2. Bazi, Melgani: Toward an Optimal SVM Classification System for Hyperspectral Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 44, 3374–3385 (2006)
3. Blaschke, T.: Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65(1), 2–16 (2010)
4. Congalton, R.G., Green, K.: Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. Lewis Publishers, Washington, DC (1977)
5. da Torres, R.S., Falcão, A.X., Gonçalves, M.A., Papa, J.P., Zhang, B., Fan, W., Fox, E.A.: A genetic programming framework for content-based image retrieval. Pattern Recognition 42(2), 217–312 (2009)
6. dos Santos, J.A., Ferreira, C.D., da Torres, R.S., Gonçalves, M.A., Lamparelli, R.A.C.: A relevance feedback method based on genetic programming for classification of remote sensing images. Information Sciences 181(13), 2671–2684 (2011)
7. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics 33(1), 159–174 (1977)
8. Li, N., Huo, H., Fang, T.: A novel texture-preceded segmentation algorithm for high-resolution imagery. IEEE Transactions on Geoscience and Remote Sensing (99), 1–11 (2010)
9. Lu, D., Weng, Q.: A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28(5), 823–870 (2007)
10. Munoz-Mari, J., Bruzzone, L., Camps-Valls, G.: A Support Vector Domain Description Approach to Supervised Classification of Remote Sensing Images. IEEE Transactions on Geoscience and Remote Sensing 45, 2683–2692 (2007)
11. Papa, J.P., Falcão, A.X., Suzuki, C.T.N.: Supervised pattern classification based on optimum-path forest. International Journal of Imaging Systems and Technology 19(2), 120–131 (2009)
12. Showengerdt, R.: Techniques for Image Processing and Classification in Remote Sensing. Academic Press, New York (1983)
13. Tseng, M.-H., Chen, S.-J., Hwang, G.-H., Shen, M.-Y.: A genetic algorithm rule-based approach for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing 63(2), 202–212 (2008)
14. Wilkinson, G.G.: Results and implications of a study of fifteen years of satellite image classification experiments. IEEE Transactions on Geoscience and Remote Sensing 43(3), 433–440 (2005)

A Dynamic Niching Quantum Genetic Algorithm for Automatic Evolution of Clusters

Dongxia Chang* and Yao Zhao

Institute of Information Science, Beijing jiaotong University, Beijing, 100044, China
Beijing Key Laboratory of Advanced Information Science and Network Technology,
Beijing, 100044, China
chang_dongxia@hotmail.com, yzhao@bjtu.edu.cn

Abstract. This paper proposes a novel genetic clustering algorithm, called a dynamic niching quantum genetic clustering algorithm (DNQGA), which is based on the concept and principles of quantum computing, such as the qubits and superposition of states. Instead of binary representation, a boundary-coded chromosome is used. Moreover, a dynamic identification of the niches is performed at each generation to automatically evolve the optimal number of clusters as well as the cluster centers of the data set. After getting the niches of the population, a Q-gate with adaptive selection of the angle for every niches is introduced as a variation operator to drive individuals toward better solutions. Several data sets are used to demonstrate its superiority. The experimental results show that DNQGA clustering algorithm has high performance, effectiveness and flexibility.

Keywords: Clustering, K-means, Evolutionary computation, quantum genetic, quantum rotation gate.

1 Introduction

Clustering analysis is a common technique for statistical multivariate analysis and has been used in a wide variety of engineering and scientific disciplines. Many clustering algorithms have been proposed in the literature. Generally, they may be broadly divided into two main categories: hierarchical and partitional. However, most hierarchical and partitional clustering methods have a drawback that the number of clusters need to be specified a priori. Since apriori knowledge is generally not always available, estimation of the number of clusters from the data set under review is required under some circumstances. The classical approach of determining the number of clusters involves the use of some validity measures [1,2].

Since the global optimum of the validity function would correspond to the most "valid" solutions with respect to the functions, stochastic clustering algorithms based on simple genetic algorithm or its variants have been developed [3,4,5,6]. In these GA-based algorithms, the validity functions are regarded

[*] Corresponding author.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 308–315, 2011.

as the fitness function to evaluate the fitness of the individual. And all these algorithms are characterized by the representation of the individual, the evaluation function, the population size, genetic operators, parent selection, survival competition methods, etc. To have a good performance, all these components should be designed properly. In order to represent the individual effectively to explore the search space and to exploit the global solution in the search space within a short time, some concepts of quantum computing are adopted in the proposed genetic clustering algorithm.

Some quantum genetic algorithms which inspired by certain concept and principles of quantum computing were proposed [7,8,11]. These algorithms denotes chromosome using quantum bit encoding, carries out evolutionary search through the action of quantum gates. Quantum genetic algorithm uses a Q-bit as a probabilistic representation, defined as the smallest unit of information. The Q-bit individual has the advantage that it can represent a linear superposition of states in search space probabilistically. Thus, the Q-bit representation has a better characteristic of population diversity than other representations [11]. In fact, there are two drawbacks for the classical quantum genetic algorithm. First, chromosomes are generally represented by binary qubit. So, the bigger of range of variables and the higher of precision of the problem, the chromosomes will become longer and longer run time is needed to obtain the solutions. Second, the angle parameters used for the rotation gate are usually obtained by the table in ref. [11] or its variants.

In order to solve these problems, a novel clustering algorithm based on dynamic niching quantum(DNQGA) is presented in this paper. Within the DNQGA, a dynamic niching is developed to preserve the diversity of the population. A simpler representation with boundary-coded is adopted, whereby each individual represents a single cluster center. All the niches presented in the population at each generation are automatically and explicitly identified. Then, the application of Q-gate with adaptive selection of the angle is limited to individuals belonging to the same niche.

The rest of this paper is organized as follows. Section 2 describes the dynamic niching quantum clustering algorithm. Experimental results are provided for several real-world data sets are given in Section 3. Experimental results demonstrate the efficiency of the DNQGA clustering algorithm. Finally, conclusions are drawn in Section 4.

2 The Dynamic Niching Quantum Genetic Clustering

In the traditional GA, a population of individuals evolve according to the transition operators. At the end of the evolution process, the population consists of a single fittest individual, representing the best solution found by the algorithm. There are many cases, however, when the desired solution is not necessarily the best one, but rather a collection of best. In order to deal with this class of problem, niching has been suggested as a viable mean to simultaneously evolve subpopulations exploiting different niches. In this section, a niching quantum genetic algorithm is proposed.

Before describing the DNQGA clustering algorithm, the basics of quantum computing are addressed briefly in the following. The smallest unit of information stored in a two-state quantum computer is called a quantum bit or qubit [10]. A qubit may be in the "1" state, in the "0" state, or in any superposition of the two. A qubit can be represented as $|\Psi\rangle = \alpha\,|0\rangle + \beta\,|1\rangle$, where α and β are complex numbers that specify the probability amplitudes of the corresponding states and $|\alpha|^2 + |\beta|^2 = 1$. The state of a qubit can be changed by the operation with a quantum gate [10]. Inspired by the concept of quantum computing, DNQGA clustering is designed with a novel Q-bit representation, a Q-gate with an adaptive selection of the angle as a variation operator. The representation and the proposed algorithm are presented in the following.

2.1 Chromosome Representation and Initialization

In order to make the representation more effective and intuitive, the following representation is used

$$\left|\Psi\right\rangle = \alpha\left|x^l\right\rangle + \beta\left|x^u\right\rangle \tag{1}$$

where x^l and x^u are the lower and the upper bound of the variable x, respectively. Obviously, a qubit may be in the x^l state, in the x^u state, or in any superposition of the two. The $|\alpha|^2$ and $|\beta|^2$ respectively give the probability that the qubit will be found in x^l state and in x^u state, and $|\alpha|^2 + |\beta|^2 = 1$. The chromosome of our algorithm can be represented by qubit as follow

$$\mathbf{q}^t = \begin{bmatrix} \alpha_{1,1}^t & \alpha_{1,2}^t & \cdots & \alpha_{N,1}^t & \alpha_{N,2}^t \\ \beta_{1,1}^t & \beta_{1,2}^t & \cdots & \beta_{N,1}^t & \beta_{N,2}^t \end{bmatrix} \tag{2}$$

where N is the dimension of the vector $\mathbf{x}$. In fact, a qubit chromosome will "collapse" into a boundary-coded chromosome. For any qubit $\left[\alpha_{k,j}^t, \beta_{k,j}^t\right]^T$, $k = 1, 2, \cdots, N$, $j = 1, 2$, we generate a random number $r_{k,j} \in [0, 1]$. If $r_{k,j} < \left|\alpha_{k,j}^t\right|^2$, the qubit will be found in the x^l state, otherwise, the qubit will be found in the x^u state. Therefore, the qubit chromosome collapses into $[x^i, x^j]$, where $\{i, j\} \in \{l, u\}$. And each dimension of the chromosome will be one of the four states: $[x^l, x^l]$, $[x^l, x^u]$, $[x^u, x^l]$ and $[x^u, x^u]$.

In order to decoding the chromosome into real-valued, a decoding rule is introduced in Table 1. Here, $\Delta x_i = (x^u - x^l)/4$ and r is a random number in $[0, 1]$. If the boundary-coded chromosome is $[x^l, x^l]$, x_i will take a small value inclining to the lower bound. If the boundary-coded chromosome is $[x^u, x^u]$, x_i will take a small value inclining to the upper bound. Through this decoding criterion, a real-valued chromosome is obtained. An initial population of size P for DNQGA clustering algorithm is usually chosen at random. After getting the boundary-coded chromosomes, the real-valued chromosomes can be obtain by the decoding criterion described in Table 1.

Table 1. Decoding rules of the boundary-coded chromosome

boundary-coded chromosome	chromosome after decoding
$[x^l, x^l]$	$x_i = x^l + r \cdot \frac{\Delta x_i}{4}$
$[x^l, x^u]$	$x_i = x^l + (1+r) \cdot \frac{\Delta x_i}{4}$
$[x^u, x^l]$	$x_i = x^u - (1+r) \cdot \frac{\Delta x_i}{4}$
$[x^u, x^u]$	$x_i = x^u - r \cdot \frac{\Delta x_i}{4}$

2.2 Fitness Function

The fitness function is used to define a fitness value to each candidate solution. Here, the fitness function of the chromosome, f, is defined as

$$f(\mathbf{c}) = \tilde{J}_s(\mathbf{c}) = \sum_{j=1}^{n} \left(\exp - \frac{\|\mathbf{x}_j - \mathbf{c}\|^2}{\beta} \right)^{\gamma}, j = 1, 2, \cdots, n \tag{3}$$

where $\mathbf{x}_j$, $j = 1, 2, \cdots, n$ are all data points in the data set to be clustered, $\beta = \frac{\sum_{j=1}^{n} \|\mathbf{x}_j - \bar{\mathbf{x}}\|^2}{n}$, $\bar{\mathbf{x}} = \frac{\sum_{j=1}^{n} \mathbf{x}_j}{n}$, and γ is estimated by CCA [12].

2.3 Niching

In order to preserve the population diversity which prevents GAs being trapped by a single local optimum, a dynamic niching algorithm is presented in Table 2, where Pop_t denotes the real-valued population of individuals at generation t.

After the dynamic identification of the niche master candidates of the population Pop_t at generation t, the individuals belonging to the same master candidate can be defined as a subset $S_t^i \neq \emptyset$ in the population Pop_t which have a distance from the master candidate less than the niche radius and do not belong to other niches. If the number of the individuals in S_t^i is larger than 1, then this subset is assumed as an actual niche; otherwise, the single individual in the subset is considered as an isolated individual and all the isolated individuals form the subset S_t^*. Then, the population Pop_t at the generation t is partitioned into $v(t)$ groups, say $S_t^1, S_t^2, \cdots, S_t^{v(t)}$, and a number of isolated individuals

$$Pop_t = \left(\bigcup_{i \in \{1,2,\cdots,v(t)\}} S_t^i \right) \cup S_t^* \tag{4}$$

where S_t^* represents the set of all the isolated individuals.

After getting the niches of the population, the individuals belonging to the same master candidate will update under the quantum rotation gate [11]. In our algorithm, quantum rotation gate for the substring of a qubit chromosome is described as

$$U_{k,i}(\theta^{t,q}) = \begin{bmatrix} \cos(\theta_{k,i}^{t,q}) & -\sin(\theta_{k,i}^{t,q}) \\ \sin(\theta_{k,i}^{t,q}) & \cos(\theta_{k,i}^{t,q}) \end{bmatrix}, k = 1, 2, \cdots, N, \quad i = 1, 2 \tag{5}$$

Table 2. The dynamic niching algorithm

```
Input: $Pop_t$ the real-valued population at generation $t$
       $P$ population size
       $\sigma$ the niche radius.
Sort the current population according to the their fitness
$v(t) = 0$ (the number of niches at generation $t$)
$u(t) = 0$ (the number of niche master candidates)
For $i = 1$ to $P$ do
  If the $i$th individual is not marked then
    $u(t) = u(t) + 1$
    $N(u(t)) = 1$ (number of individuals in the $u(t)$th niche candidate)
    For $j = i + 1$ to $P$ do
      If $(d(i, j) < \sigma)$ and ($u(t)$th individual is not marked)
        insert the $j$th individual into the $u(t)$th niche masters candidate
        $N(u(t)) = N(u(t)) + 1$
      End If
    End For
    If $(N(u(t)) > 1)$ then
      $v(t) = v(t) + 1$
      mark the $i$th individual as the niche master of the $v(t)$th niche
    End If
  End If
End For
```

In order to make the qubit chromosomes effectively converge to the fitter states, we put forward an adaptive rotation angles computing method for the actual niche, which is defined as:

$$\theta_{k,i}^{t,q} = \mathrm{sign}\{\alpha_{k,i}^{t,q} \cdot \beta_{k,i}^{t,q} \cdot (f(\mathbf{c}_l^t) - \bar{f}^q)\} \cdot \frac{f(\mathbf{M}_k^{t,q}) - f(\mathbf{c}_l^t)}{f(\mathbf{M}_k^{t,q}) - \bar{f}^q} \times 0.05\pi \tag{6}$$

where $\mathbf{M}_k^{t,q}$ is the master of the qth niche, $\mathbf{c}_l^t$ is the individual in the qth niche, $f(\mathbf{M_k}^{t,q})$, $f(\mathbf{c}_l^t)$ and $\bar{f}^q$ are the fitness of $\mathbf{M}_k^{t,q}$, $\mathbf{c}_l^t$, and the average fitness value of the individual in the qth niche, respectively. The value of $\theta_{k,i}^{t,q}$ increases when the individual is quite poor. In contrast when the individual is a good solution, $\theta_{k,i}^{t,q}$ will be low so as to reduce the likelihood of disrupting a good solution by the rotation. sign$(\cdot)$ is described as

$$\mathrm{sign}(\,\cdot\,) = \begin{cases} +1 & \text{if} \quad \alpha_{k,i}^t \beta_{k,i}^t (f(\mathbf{c}_l^t) - \bar{f}^q) > 0 \\ -1 & \text{if} \quad \alpha_{k,i}^t \beta_{k,i}^t (f(\mathbf{c}_l^t) - \bar{f}^q) < 0 \\ \pm 1 & \text{if} \quad \alpha_{k,i}^t = 0 \quad \text{and} \quad (f(\mathbf{c}_l^t) - \bar{f}^q) < 0 \\ \pm 1 & \text{if} \quad \beta_{k,i}^t = 0 \quad \text{and} \quad (f(\mathbf{c}_l^t) - \bar{f}^q) > 0 \\ 0.05pi & \qquad otherwise \end{cases} \tag{7}$$

For the individuals in the isolated individuals set S_t^*, the rotation angle is defined as

$$\theta_{k,i} = \begin{cases} 0 & \text{if} \quad f(\mathbf{c_l}) = \mathbf{f}(\mathbf{c^t_{best}}) \\ 0.05\pi & \quad otherwise \end{cases} \tag{8}$$

where $\mathbf{c^t_{best}}$ is the best individual at generation t.

2.4 Description of the Algorithm

The DQNGA clustering algorithm is described as follows:

1. Initialize a group of cluster centers, described by the boundary-coded qubit, with size of P.
2. Decoding the boundary-coded chromosomes into real-valued chromosomes.
3. Apply the dynamic niching algorithm and copy the niche masters in a separate location.
4. If the termination condition is not reached, go to Step 5. Otherwise, select the niche masters from the population as the final cluster centers.
5. Apply the quantum gate operator for each niche.
6. Evaluate the newly generated candidates.
7. Apply the elitist strategy.
8. Go back to Step 3.

3 Experiments Results

In this section, the performances of the GA-clustering [3], KGA-clustering [4], GAGR [6] and DQNGA algorithms are compared through the experiments based on the three real-world data sets from UCI Machine Learning Repository. For the GA-clustering, KGA-clustering and GAGR, the number of clusters is set as the actual number of clusters present in the data sets.

In the experiments, the population size is taken as 50. The crossover and mutation probabilities for GA-clustering KGA-clustering and GAGR algorithm are $p_c = 0.8$ and $p_m = 0.001$, respectively. At first 100 independent runs are taken to judge the automatic clustering efficiency of DQNGA, and the mean number of clusters found as well as the percentage of successful runs (those yielding the correct number of clusters according to the algorithm) are given in Table 3. Here AC and OC denote the actual number of clusters present in the data set and the obtained number of clusters by DQNGA, respectively. It shows the percentage of runs that managed to yield the correct number of classes for each data set. In the following, we compare the speed of convergence of the four algorithms. For each data set we have conducted the experiment 20 independent time. The characteristic of the computation time is shown in Table 4. Table 4 gives the computation time (the experiments were implemented on a machine running Windows XP, Intel (R) Xeon (R) CPU, 2.33 GHz) needed for convergence of the four algorithms. As seen from Table 4, the DQNGA clustering algorithm converges in relatively shorter computation time. In order to compare the qualities of the final clustering results obtained, three statistical score functions (Overall accuracy, Kappa index and Adjusted rand index) are used. Table 5 gives the mean

Table 3. Mean number of clusters obtained (with standard deviation) and percentage of successful runs obtained by DQNGA over 100 independent runs

Data	Iris	Breast	Wine
AC	3	2	3
OC	2.86(0.4494),72	2(0.0408),98	3.0769(0.4130),65

Table 4. The average computation time(s) to convergence of the four algorithms for 20 different runs for the three real-life data sets

Data	GA-clustering	KGA-clustering	GAGR clustering	DQNGA
Iris	45.295	0.4494	0.3741	0.2040
Breast	235.368	0.4539	0.3372	0.2496
Wine	46.478	0.4960	0.4011	0.3443

Table 5. Mean and standard deviation (in parentheses) of three statistical validity measures produced by the four algorithms

Data	Validity Measures	Mean and std. dev. of the validity measures over the final clustering results of the successful runs			
		GA-clustering	KGA-clustering	GAGR	DQNGA
Iris	Overall accuracy(%)	80.62(2.2011e-5)	82.39(0.0270)	83.67(0.0190)	85.36(0.0125)
	Kappa index(%)	75.93(4.9524e-5)	76.17(0.0523)	76.65(0.01132)	78.31(0.0185)
	Adjusted rand index	0.7149(8.8253e-5)	0.7009(0.0015)	0.7163(0.0058)	0.7652(0.0065)
Breast	Overall accuracy(%)	90.31(0.0011)	93.22(0.0024)	94.25(0.0016)	95.10(5.440e-5)
	Kappa index(%)	84.95(0.0020)	85.71(0.0087)	87.71(0.0057)	89.65(0.0062)
	Adjusted rand index	0.8225(0.0139)	0.8536(0.0021)	0.8665(0.0013)	0.8891(3.120e-6)
Wine	Overall accuracy(%)	80.58(0.0177)	82.25(0.0265)	83.64(0.0135)	88.21(0.0322)
	Kappa index(%)	75.92(0.0276)	76.49(0.0356)	80.53(0.0216)	85.32(0.0287)
	Adjusted rand index	0.7236(0.0237)	0.7398(0.0129)	0.7530(0.0058)	0.8136(0.0152)

value (and standard deviations)of the three statistical functions. From Table 5 one may observe that our approach outperforms GA-clustering, KGA-clustering, GAGR in a statistically significant manner. Not only does the method find the optimal number of clusters, but also it manages to find classifications closest to the ground truth as evident from the higher values of the overall accuracy, kappa index and the adjusted rand indices in Table 5.

4 Conclusions

In this paper, a novel clustering algorithm based on dynamic quantum niching genetic clustering algorithm (DQNGA) has been developed for clustering problem with unknown cluster number. The DQNGA algorithm can find the optimal number of clusters and the cluster centers automatically. As the number of clusters is not known a priori in most practical circumstance, DQNGA algorithm can be used more widely. In the DQNGA algorithm, each chromosome is encoded a center of a cluster by a boundary-coded qubit. The dynamic niching is accomplished without assuming any a priori knowledge on the number of niches. And an adaptive selection of the rotation angle used by the quantum rotation gate is introduced. The superiority of the DQNGA over GA-clustering, KGA-clustering and GAGR has been demonstrated by the experiments on three real-world data sets. All the experiment results have shown that our algorithm has high performance, effectiveness and flexibility.

Acknowledgments. This paper was supported by the Fundamental Research Funds for the Central Universities of China (2011JBM026); the China Postdoctoral Science Foundation (00480190); the Beijing Municipal Natural Science Foundation (4113075).

References

1. Glenn, W.M., Martha, C.C.: An examination of procedures for determining the number of clusters in a data set. Psychometrika 50, 159–179 (1985)
2. Xuanli, L.X., Gerardo, B.: A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell. 13, 841–847 (1991)
3. Murthy, C.A., Chowdhury, N.: In search of optimal clusters using genetic algorithms. Pattern Recognition Letters 17, 825–832 (1996)
4. Bandyopdhyay, S., Maulik, U.: An evolutionary technique based on K-Means algorithm for optimal clustering in RN. Information Sciences 146, 221–237 (2002)
5. Sanghamitra, B., Sriparna, S.: GAPS: A clustering method using a new point symmetry-based distance measure. Pattern Recogn. 40, 3430–3451 (2007)
6. Chang, D.X., Zhang, X.D., Zheng, C.W.: A genetic algorithm with gene rearrangement for K-means clustering. Pattern Recogn. 42, 1210–1222 (2009)
7. Ajit, N., Mark, M.: Quantum inspired genetic algorithm. In: Proc. of the Third IEEE International Conference on Evolutionary Computation, pp. 61–66 (1996)
8. Kuk, H.H.: Genetic quantum algorithm and its application to combinatorial optimization problem. In: Proc. 2000 Congress on Evolutionary Computation, pp. 1354–1360 (2000)
9. Miin, S.Y., Kuo, L.W.: A similarity-based robust clustering method. IEEE Trans. Pattern Anal. Mach. Intell. 26, 434–448 (2004)
10. Tony, H.: Quantum computing: an introduction. Computing & Control Engineering Journal, 105–112 (1999)
11. Kuk, H.H., Jong, H.K.: Quantum-inspired evolutionary algorithm for a class of combinatorial optimization. IEEE Trans. Evol. Comput. 6, 580–593 (2002)
12. Yang, M.S., WuK, L.: A similarity-based robust clustering method. IEEE Trans. Pattern Anal. Mach. Intell. 26(4), 434–448 (2004)

Spatio-Temporal Fuzzy FDPA Filter

Marek Szczepański

Faculty of Automatic Control, Electronics and Computer Science,
Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland
Marek.Szczepanski@polsl.pl

Abstract. An overview of the new spatio-temporal video filtering technique was presented in this paper. The extension of standard techniques based on temporal Gaussian combined with Fast Digital Paths Approach [9] with fuzzy similarity function was presented. Presented technique provides excellent noise suppression ability especially for low light sequences.

1 Introduction

The widespread use of webcams, camcoders, digital cameras embedded in mobile phones allows capturing of images and videos in different situations especially in low light environment. Unfortunately, the high level of miniaturization of sensors entails very poor quality of recorded material. In addition, increasing the number of megapixels packed into such a small area in order to improve image quality imposes strong noise artifacts in resulting images. Is therefore necessary to use algorithms that improve the quality of such images. One of the most challenging tasks is shot noise removal which is dominant in low light images [5].

In this paper, a novel noise spatio-temporal filter, which combines Temporal Gaussian smoothing with spatial Fast Digital Path Approach (*FDPA*) [9] has been proposed. Temporal Gaussian smoothing effectively removes shot noise or Gaussian artifacts but introduces ghosting artifacts in dynamic sequences, thus in such regions spatial filtering is preferred. The impact of spatial and temporal elements is determined using fuzzy membership function calculated using interframe differences.

The paper is organized as follows. In Section 2 the general concept of the digital paths applied to the spatial filters is introduced. Section 3 introduces concept of our spatio-temporal filter, while Section 4 presents simulation results. Finally, Section 5 summarizes our paper.

2 Fast Digital Paths Approach Spatial Filter (*FDPA*)

In this work general fuzzy filtering structure proposed in [6] will be used. The general form of the fuzzy adaptive filters proposed in this work is defined as weighted average of input vectors inside the processing window W:

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 316–323, 2011.

$$\hat{\mathbf{F}}_{\mathbf{0}} = \sum_{i=0}^{k-1} w_i \mathbf{F}_i = \frac{\sum\limits_{i=0}^{k-1} \mu_i \mathbf{F}_i}{\sum\limits_{i=0}^{k-1} \mu_i}, \tag{1}$$

where $\mathbf{F}_i$ and $\hat{\mathbf{F}}_{\mathbf{0}}$ denotes filter inputs and output respectively. The relationship between the pixel under consideration (window center) and each pixel in the window should be reflected in the decision how to define the filter weights. In our case weights will be calculated using similarity functions calculated over digital paths included in the processing window W.

Let a digital lattice $\mathcal{H} = (\mathbf{F}, \mathcal{N})$ be defined by $\mathbf{F}$, which is the set of all points of the plane (pixels of a color image) and a neighborhood relation $\mathcal{N}$ between the lattice points [7]. In the case of the ranked-type non-linear filters the processing window W forms a lattice where $\mathcal{N}$ is defined through the window size.

A digital path $P = \{p_i\}_{i=0}^{n}$ defined on the lattice $\mathcal{H}$ is a sequence of neighboring points $(p_{i-1}, p_i) \in \mathcal{N}$. The length $L(P)$ of the digital path $P\,\{p_i\}_{i=0}^{n}$ is simply $\sum_{i=1}^{n} \rho^{\mathcal{H}}(p_{i-1}, p_i)$, where $\rho^{\mathcal{H}}$ denotes the distance between two neighboring points of the lattice $\mathcal{H}$. An $\mathcal{N}_8$-neighborhood system is considered in this work with a topological distance of 1 assigned between two neighboring points.

Let us adopt the following notation, which will help us define the distance functions defined over digital paths. The starting point of a path will be denoted as $p_0 = (x_0, y_0)$. Its neighbors will be denoted as $p_1 = (x_{u_1}, y_{v_1})$, which means that the neighbors are the second points of all digital paths originating at p_0. Then the third point of a digital path starting at p_0 will be $p_2 = (x_{u_2}, y_{v_2})$ and so on, till the path reaches in n steps the ending point $p_n = (x_{u_n}, y_{v_n})$.

The set of all possible digital paths contained in W joining two points $a, b \in W$ will be denoted as $\Phi^W(a, b)$. Two pixels a and b will be called connected (hereafter denoted as $a \leftrightarrow\ b$), if there exists a digital path $P^W(a, b)$ contained in the set W starting from a and ending at b.

If two pixels p_0 and p_n are connected by a digital path $P^{W,n}\ \{p_0,\ \ p_1, \ldots, p_n\}$ of length n then let $\Lambda^{W,n}\{p_0, p_1,\ \ldots, p_n\}$ be a function which measures the connection cost defined over the digital path linking the starting point p_0 and ending point p_n. The connection cost over the digital path $\Lambda^{W,n}$ will be defined as a measure of dissimilarity between color image pixels p_0, p_1, …, p_n forming a specific path linking p_0 and p_n [3,10]:

$$\begin{gathered}\Lambda^{W,n}\left\{p_0, p_1, p_2, \ldots, p_n\right\} = f\left\{\mathbf{F}\left(p_0\right), \mathbf{F}\left(p_1\right), \mathbf{F}\left(p_2\right), \ldots, \mathbf{F}\left(p_n\right)\right\} = \\ \sum_{i=1}^{n} \|\mathbf{F}(p_1) - \mathbf{F}(p_{i-1})\|.\end{gathered} \tag{2}$$

Let us now define a similarity function, analogous to a membership function used in fuzzy systems, between the starting point $a = p_0$ and point $b = p_1$ crossed by the digital path connecting pixel p_0, its neighbor p_1 with all possible points p_n which can be reached in n steps from p_0.

The aim of taking into account the points p_2, ..., p_n when calculating the similarity between p_0 and p_1 is to explore not only the direct neighborhood of p_0 but also to use the information on the local image structure. This can be done by acquiring the information on the local image features investigating the connection costs of digital paths originating at p_0, passing p_1 and then visiting successive points, till the path reaches length n. In this case the similarity function takes the form:

$$\mu^{W,n}(a,b) = \mu^{W,n}(p_0,p_1) = \sum_{m=1}^{\omega} g\left(\Lambda_m^{W,n}(p_0,p_1)\right), \tag{3}$$

where ω denotes number of all possible paths $P\{p_0, p_1, p_2^*, \ldots, p_n^*\}$ of length n originating at $a = p_0$ and crossing $b = p_1$ totally included in W, $\Lambda_m^{W,n}\{\cdot\}$ is a dissimilarity value along a specific path and $g(\cdot)$ is a smooth function of $\Lambda_m^{W,n}$.

In this work we assume that $g(\cdot)$ is the exponential function [9] so our similarity function takes the form:

$$\mu^{W,n}(a,b) = \mu^{W,n}(p_0,p_1) = \sum_{m=1}^{\omega} \exp\left[-\beta \cdot \Lambda_m^{W,n}(p_0,p_1)\right] \tag{4}$$

where β is the filter design parameter. A normalized form of the similarity function can be defined as follows:

$$\psi^{W,n}(a,b) = \psi^{W,n}(p_0,p_1) = \frac{\mu^{W,n}(p_0,p_1)}{\sum_{p_1^*} \mu^{W,n}(p_0,p_1^*)}, \tag{5}$$

where p_1^* denotes all p_0 neighbors.

Assuming that the pixel $a = p_0$ is the pixel under consideration, with $\mathbf{F}(b)$ representing the pixel $b = p_1$ the filter output $\hat{\mathbf{F}}(a)$ is given as follows:

$$\hat{\mathbf{F}}(a) = \sum_{b \sim \mathcal{N}_8(a)} \psi^{W,n}(a,b) \cdot \mathbf{F}(b) = \sum_{p_1} \psi^{W,n}(p_0,p_1) \cdot \mathbf{F}(p_1). \tag{6}$$

The performance of the new filters strongly depends on the type of digital paths selected. Different models of paths result to application-specific filters, which are able to suppress certain types of noise. In this paper we concentrate on the *"Self-Avoiding Path model"* (SAP), which provides a model suitable for image processing applications [8,9].

The *Self-Avoiding Path* (SAP) is a special type of path taken along the image lattice so that adjacent pairs of edges in the sequence share a common vertex of the lattice. In the SAP approach no vertex is visited more than once resulting in a trajectory that never intersects itself. In other words the *Self-Avoiding Path* is a path that does not pass through the same lattice point twice.

In order to reduce filter complexity the FDPA filter uses fixed size of the supporting window W is set to (3×3) independently of the path's length.

3 Spatio-Temporal Fuzzy FDPA Filter (*STFFDPA*)

Noise introduced by CCD and CMOS sensors significantly reduce the quality of the recorded material and cause considerable losses during compression. Because we are dealing with a sequence of images, rather than a single frame, we can eliminate the noise using inter frame relations. It is important that the sensor noise is characterized by a low correlation between individual frames (with exception of hot pixels), while parts of the image, even the fast-changing, they are strongly correlated.

These properties are used during temporal filtering using different variants of averaging the values of individual pixels in successive video frames. The simplest temporal filter is the Temporal Arithmetic Mean Filter (*TAMF*), the output of that filter can be represented by the following relationship:

$$\hat{\mathbf{F}}\left(x,y,t\right)=\frac{1}{2n+1}\sum_{\varDelta t=-n}^{n}\mathbf{F}\left(x,y,t+\varDelta t\right), \tag{7}$$

where $\mathbf{F},\hat{\mathbf{F}}$ denote the input and output frames and n determines temporal window size.

This method, although the simplest and quickest is only suitable for static sequences, because averaging frames where there are objects in motion creates "ghosting" effects in output sequence. One way to reduce the ghosting effect is to use temporal Gaussian filtering instead of simply averaging:

$$\hat{\mathbf{F}}_G\left(x,y,t\right)=\sum_{t=-n}^{t=n}g(\sigma,t)*\mathbf{F}(x,y,t). \tag{8}$$

Some algorithms utilize motion compensation to reduce blurring effect such as works presented by Dubois and Sabri [4].Another solution to reduce ghosting artifacts is to use a temporal version of bilateral filter used as a element of ASTA filter [2].

As can be seen in Fig. 2 b) in the case of static scenes get excellent results using a simple averaging or Gaussian smoothing over time, which, however, completely fails when the scene contains moving objects, then we should use the spatial filtering algorithms. So we can define filter output as weighted spatial and temporal filters denoted as $\mathbf{F}_S$ and $\mathbf{F}_T$:

$$\hat{\mathbf{F}}\left(x,y,t\right)=w\mathbf{F}_T+\left(1-w\right)\mathbf{F}_s \tag{9}$$

with fuzzy weight w calculated using cumulated lightness differences $\varDelta L$ over full temporal window:

$$\varDelta L\left(x,y,t\right)=\sum_{t=-n}^{t=n}dist\left(\mathbf{F}\left(x,y,t\right),\mathbf{F}\left(x,y,t+\varDelta t\right)\right), \tag{10}$$

where

$$dist\left(\mathbf{F}_1,\mathbf{F}_2\right)=\left|\frac{1}{3}\left[\left(F_{1R}+F_{1G}+F_{1B}\right)-\left(F_{2R}+F_{2G}+F_{2B}\right)\right]\right|, \tag{11}$$

so now we can define our similarity function:

$$w\left(x,y,t\right)=\exp\left(-\gamma\cdot\mathit{\Delta}L\left(x,y,t\right)\right). \tag{12}$$

Finally filter output takes form:

$$\begin{aligned}\hat{\mathbf{F}}\left(x,y,t\right)&=\exp\left(-\gamma\cdot\mathit{\Delta}L\left(x,y,t\right)\right)\cdot\sum_{t=-n}^{t=n}g(\sigma,t)*\mathbf{F}(x,y,t)+\\&+\left(1-\exp\left(-\gamma\cdot\mathit{\Delta}L\left(x,y,t\right)\right)\right)\cdot\mathbf{F}_{FDPA}\left(x,y,t\right).\end{aligned} \tag{13}$$

4 Simulation Results

Several filters capable of real-time video processing were examined on numerous video sequences. Subjective results were obtained from original noisy video sequences,however some synthetic tests with artificial noise were also performed.

Objective quality measures such as the *Root Mean Squared Error* (RMSE), the *Signal to Noise Ratio* (SNR), the *Peak Signal to Noise Ratio* (PSNR), the *Normalized Mean Square Error* (NMSE) and the *Normalized Color Difference* (NCD) were used for the analysis. All those objective quality measures were calculated for the sequence of the filtered images.

The performance of the following filters was evaluated:

- Temporal Gaussian Filter *TGauss* (with time window $n = 5$ and $\sigma = 5$),
- Spatial Vector Median Filter (with window 3×3 and L_1 norm)[1],
- Spatial Fast Digital Paths Approach *FDPA*,
- Spatio-temporal Vector Median Filter - *VMF3D* (with window $3\times3\times3$ and L_1 norm),
- Spatio-Temporal Fuzzy FDPA Filter (*STFFDPA*) ($n = 5, \sigma = 5, \gamma = 4, \beta = 15$)

For both digital paths filters, the path size was limited to two steps.

Figure 1 shows the frame from noisy sequence captured in low light conditions containing small toy-car moving rapidly. It can bee seen that temporal methods produces perfect background while moving object is blurred, static techniques can't clear the noise effectively. Only combination of spatial and temporal techniques gives satisfactory results.

Figure 2 depicts filtering results of sample frame from standard video sequence *Hall Monitor* corrupted with Gaussian noise ($\sigma = 10$).

Objective quality measures for sequence *Foreman* corrupted with Gaussian noise ($\sigma = 20$) are presented in Table 1 while Table 2 combines results of filtering *Hall monitor* sequence with Gaussian noise ($\sigma = 10$). Average values and their standard deviation were evaluated.

The Figure 3 depicts PSNR values of subsequent frames filtered with temporal Gaussian, Spatial FDPA and the new spatio-temporal filters, it can be clearly seen when temporal filter produces ghosting artifacts.

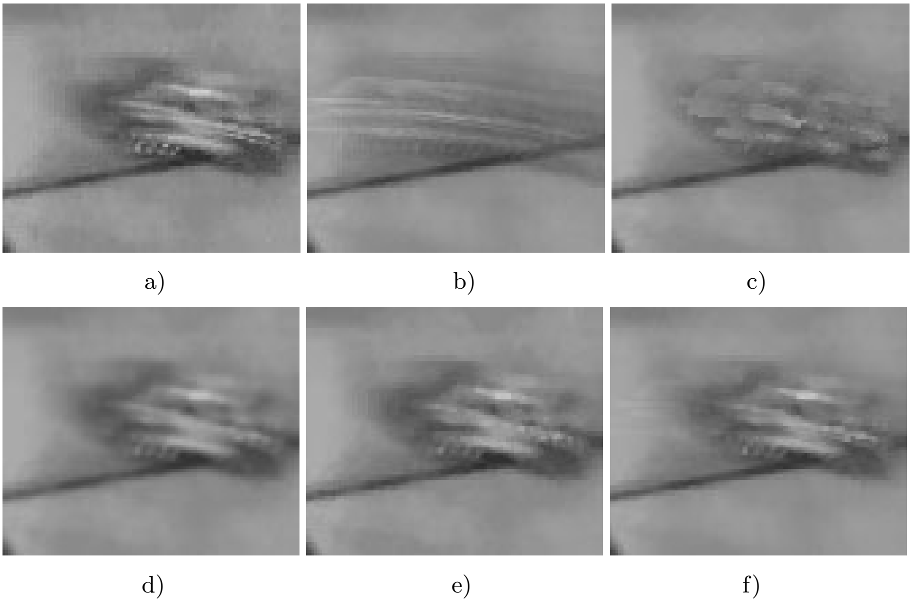


Fig. 1. a) Frame from the noisy video seqence - *Toy Car* and results of filtering with b) Temporal Gaussian, c) *VMF3D*, d)Spatial *FDPA*, e) *VMF* and f)Spatio-Temporal Fuzzy FDPA Filter *STFFDPA*

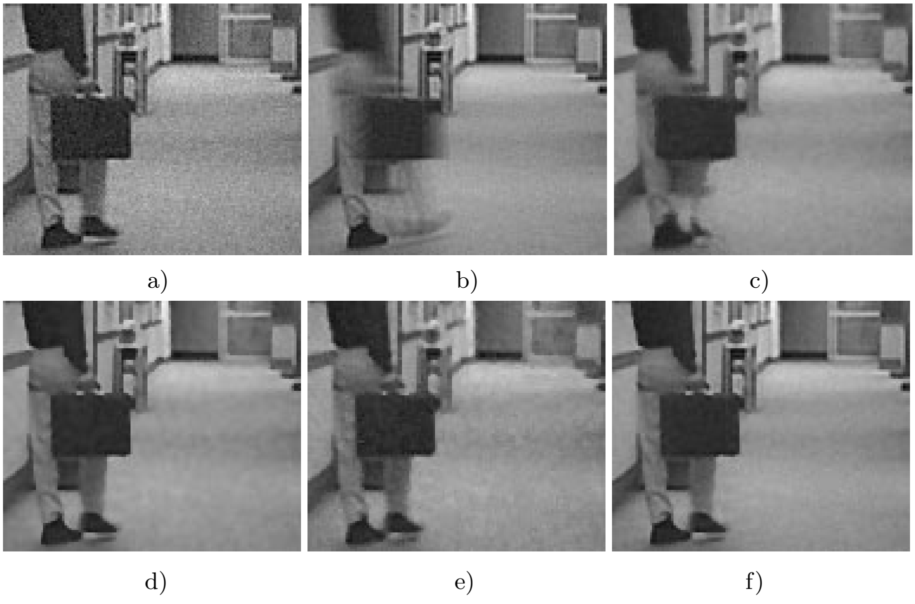


Fig. 2. a) Frame from the noisy video sequence *Hall Monitor* with Gaussian noise and results of filtering with b) Temporal Gaussian, c) *VMF3D*, d)Spatial *FDPA*, e) *VMF* and f)Spatio-Temporal Fuzzy FDPA Filter *STFFDPA*

Table 1. Comparison of the filtering algorithms applied for *Foreman* sequence corrupted with Gaussian noise ($\sigma = 20$)

Filter	**PSNR** [dB]	σ_{PSNR} [dB]	**NCD** [10^{-3}]	σ_{NCD} [10^{-3}]
None	22.22	0.03	185.8	24.7
VMF	25.7	0.72	109.4	16.5
VMF3D	26.6	1.81	80.0	11.6
FDPA	**28.5**	**0.72**	**77.1**	**11.2**
TGauss	25.7	2.23	90.2	11.9
STFFDPA	**29.6**	**0.70**	**65.3**	**9.0**

Table 2. Comparison of the filtering algorithms applied for *Hall monitor* sequence corrupted with Gaussian noise ($\sigma = 10$)

Filter	**PSNR** [dB]	σ_{PSNR} [dB]	**NCD** [10^{-3}]	σ_{NCD} [10^{-3}]
None	28.3	0.01	95.2	1.0
VMF	28.8	0.57	62.2	0.7
VMF3D	29.4	0.14	51.0	1.1
FDPA	**32.4**	**0.05**	**38.4**	**0.4**
TGauss	32.4	0.79	51.2	1.3
STFFDPA	**34.4**	**0.12**	**38.9**	**0.6**

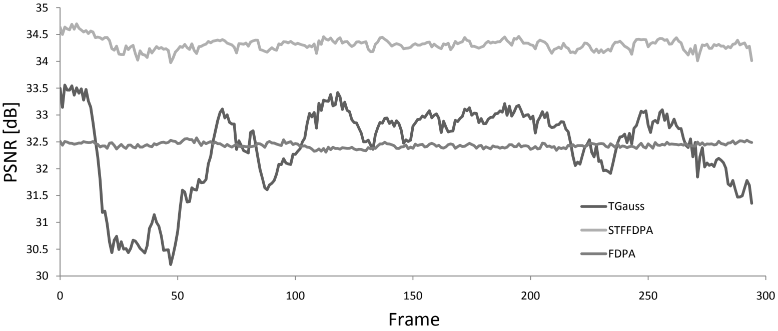


Fig. 3. PSNR coefficients of noisy *Hall monitor* sequence with TGauss, FDPA and STFDPA filters

5 Conclusions

From several years we can observe increasing interest in video processing. Video noise reduction without structure degradation is perhaps the most challenging video enhancements task. Several techniques have been proposed over the years. Among them are standard noise reduction techniques, the so-called spatial filters, applied to subsequent frames of the video stream. However, standard image processing techniques cannot utilize all available information i.e. similarities in neighboring frames, so modern video denoising algorithms utilize also temporal information. The new approach utilizing temporal Gaussian filtering and spatial digital path filter was presented in this paper. Presented technique provides excellent noise suppression, especially for low light video sequences.

Acknowledgements. This work has been supported by the Polish Ministry of Science and Higher Education under SUT internal grant BK-214/RAu1/2011.

References

1. Astola, J., Haavisto, P., Neuovo, Y.: Vector median filters. IEEE Proc. 78, 678–689 (1990)
2. Bennett, E.P., McMillan, L.: Video enhancement using per-pixel virtual exposures. ACM Trans. Graph. 24(3), 845–852 (2005)
3. Cuisenaire, O.: Distance transformations: fast algorithms and applications to medical image processing. PhD thesis, Université Catholique de Louvain (October 1999)
4. Dubois, E., Sabri, S.: Noise reduction in image sequences using motion-compensated temporal filtering. IEEE Transactions on Communications 32(7), 826–831 (1984)
5. Lee, S., Maik, V., Jang, J., Shin, J., Paik, J.: Noise-adaptive spatio-temporal filter for real-time noise removal in low light level images. IEEE Transactions on Consumer Electronics 51, 648–653 (2005)
6. Plataniotis, K.N., Androutsos, D., Vinayagamoorthy, S., Venetsanopoulos, A.N.: Color image processing using adaptive multichannel filters. IEEE Trans. on Image Processing 6(7), 933–950 (1997)
7. Schmitt, M.: Lecture notes on geodesy and morphological measurements. In: Proceedings of the Summer School on Morphological Image and Signal Processing, Zakopane, Poland, pp. 36–91 (1995)
8. Smolka, B., Wojciechowski, K.: Random walk approach to image enhancement. Signal Processing 81(3), 465–482 (2001)
9. Szczepanski, M., Smolka, B., Plataniotis, K.N., Venetsanopoulos, A.N.: On the geodesic paths approach to color image filtering. Signal Processing 83(6), 1309–1342 (2003)
10. Toivanen, P.J.: New geodesic distance transforms for gray scale images. Pattern Recognition Letters 17, 437–450 (1996)

Graph Aggregation Based Image Modeling and Indexing for Video Annotation

Najib Ben Aoun[1], Haytham Elghazel[2], Mohand-Said Hacid[3], and Chokri Ben Amar[1]

[1] University of Sfax, National School of Engineers (ENIS),
REGIM: REsearch Group on Intelligent Machines, BP 1173, 3038, Sfax, Tunisia
{najib.benaoun,chokri.benamar}@ieee.org
[2] University of Lyon, University of Lyon 1,
GAMA laboratory, 69622, Villeurbanne, France
haytham.elghazel@univ-lyon1.fr
[3] University of Lyon, University of Lyon 1,
LIRIS laboratory, 69622, Villeurbanne, France
mohand-said.hacid@univ-lyon1.fr

Abstract. With the rapid growth of video multimedia databases and the lack of textual descriptions for many of them, video annotation became a highly desired task. Conventional systems try to annotate a video query by simply finding its most similar videos in the database. Although the video annotation problem has been tackled in the last decade, no attention has been paid to the problem of assembling video keyframes in a sensed way to provide an answer of the given video query when no single candidate video turns out to be similar to the query. In this paper, we introduce a graph based image modeling and indexing system for video annotation. Our system is able to improve the video annotation task by assembling a set of graphs representing different keyframes of different videos, to compose the video query. The experimental results demonstrate the effectiveness of our system to annotate videos that are not possibly annotated by classical approaches.

Keywords: Graph aggregation, graph modeling, content-based video annotation, Region Adjacency Graph.

1 Introduction

Graph data modeling has received a big interest last decade. This is reinforced with the development of new graph mining techniques, making it a very promising field. That's why, graphs have become increasingly important in modeling complicated structures and schemaless data in many application domains such as molecular structure for chemical compounds in chemistry domain, network connections in telecommunication domain, web pages links in web domain, bioinformatics, robotics, etc [4],[5],[6],[8],[9]. Moreover, graphs are very powerful in the computer vision domain since images regions can be efficiently modeled using graphs [2].

In computer vision, video annotation aims to assign to one video a description that characterizes its semantic content, including visual content, audio content, textual

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 324–331, 2011.

content [1] or even the events and the actions presented on it. In this paper, we are focused on the spatial visual content to annotate video data. For that, motivated by the emergence of graphs to be the most sophisticated and general form of structure able to represent any kind of data, graphs are used to model video frames where vertices represent video regions and edges concern the relationships between them. This allows us to give a rich frame description that is effectively used in the frame-indexing phase.

The video annotation method proposed in this paper uses graph for image modeling and indexing to annotate video and improve this annotation using a graph aggregation scheme. Thus, after modeling every video frames using graphs, an efficient graph based search method will be applied to compose the video from the whole database and then to annotate it.

The remaining of this paper is organized as follows: in Sec. II we will present the video annotation task. Then, in Sec. III, graph based images modeling is introduced. In Sec. IV, we describe our proposed video annotation system that initially consists of a graph based image indexing phase for video annotation. In Sec. V, graph aggregation scheme is offered to improve the later video annotation. Experimental results are shown in Sec. VI to assess the performance of our system. Finally, we will conclude our paper by summarizing the key findings of the proposed system and suggesting possible extensions for our system.

2 Video Annotation

Content based video annotation (CBVA) aims to annotate a query video by matching its features with those of the database. Consequently, CBVA systems can be divided to two phases: video indexing phase which extract video features and features matching phase intended to find videos in the database which are similar to the query video that will be used to annotate it.

Currently, there are many problems in the video databases precisely related to video annotation. However, in many cases, videos are not annotated or suffer from a lack of annotation. That's why video annotation is a very important task not only to annotate the non-annotated videos but also to verify the annotations of videos that are already annotated and to enrich their annotations. Besides, video annotation system must deal with the problems of transformations and changes in videos such as the change in luminosity.

Conventional video annotations methods apply video features extraction and features matching to annotate video. In [1], a video annotation system was developed which exploits the textual content existing in the video to describe it and match the query video feature with the database features to find the similar videos. Others systems index video based on the existing events such as [2] which try to index video by detecting person running or walking events. These existing systems have many limits since they (1) didn't include the regions relationship giving the spatial behavior that is powerful information describing an image and (2) treat the combination of images to annotate a video query. Motivated by this, we have developed a graph based video annotation system to overcome these limits. Since finding similar videos to one query video seems to be not possible in some cases, a graph aggregation scheme is used in our system to compose query graph by several sub-graphs from the database. The query video is then annotated using these composed sub-graphs.

3 Graph Based Image Modeling

Video indexing is a fundamental part in video annotation system. As we have presented previously, video can be characterized with its visual, audio or textual content or even the events or the actions which contains. In our system, we have built a video annotation system based on the spatial visual content. This spatial visual content is representing with a graph modeling for video frames. Thus every video frame is model with a graph. Graph vertices represent the frame regions which are recognized using color segmentation in the CIELab space with the Hill Climbing algorithm [3]. Graph edges represent the relationship between frame regions and precisely the adjacent regions to have the Region Adjacency Graph (RAG) [2].RAG has proved that is the most efficient representation compared with others graph form such as Attributed Relational Graph (ARG) and the tree form such as the Quadtree.

It should be mentioned here that many works have model image with graph such as [4] in which ARG is used as an image content representation to search for similarity in a medical image databases to annotate unlabelled images regions. Thereafter, authors in [5] have model image with RAG to build a graph based image indexing engine used search images in the image databases. Recently, [6] have proposed a Quadtree representation based image indexing method to classify the magnetic resonance images according to the Corpus Callosum.

A graph is represented by $\{V,E,L_v,L_e,F_v,F_e\}$ where **V** is the set of the graph vertices, **E** is the set of the graph edges connecting two different vertices, $\mathbf{L_v}$ is the set of the vertices labels, $\mathbf{L_e}$ is the set of the edges labels, and $\mathbf{F_v}$ is the function labeling the vertices, and $\mathbf{F_e}$ is the function labeling the edges. There are two categories of labeled graphs: directed graphs assigning a direction for the edges labels which make two edges connecting every two different vertices with an opposed directions, and an undirected graphs assigned only one edge connecting every two vertices. Depending on Fe, the graph category is chosen: if we choose the distance between two vertices as an edge label, then we will have an undirected graph and that is what we have used in our system. But, if we choose the angle from the line connecting the two vertices to the horizontal line as an edge label, we will have a directed graph.

After segmenting the frame into different regions, we will extract a visual characterization from every region forming its descriptors. To obtain this characterization, the Gabor filter and the Gray Scale Cooccurrence Matrix (GLCM) are performed for texture information and the color histogram is extracted for color information [11]. Based on these descriptors, a label is assigned to every region by clustering the different video-frames regions in the database using K-means method with a predefined number of classes. Each region class will be used to label its corresponding vertices (regions) in the graph representation.

Consequently, each video frame is well described based on its visual content including shape information, color content, texture information and also the intra-regions information. In addition, our approach is proposed to treat real color images unlike classical methods working only on gray scale images such as in [4], [5] or [6] where only color information has been used to describe image regions. Moreover, instead of treating directly the frame, we model it with graph, which significantly reduces the computational cost without losing its visual content.

4 Proposed Video Annotation System

Modeling image with graph will lead to a faster video annotation system since the quantity of data to be treated will be significantly reduced. In addition, we extracted keyframes from the video (we refer to them as images in the rest of the paper), using motion estimation [12], since they are the most representative video frames. Video annotation task will be then refined to keyframes annotation that will be indexed based on their graph representation.

Our system consists of two phases: graph based image (keyframe) indexing and image feature matching to find the videos similar to a given video query.

4.1 Graph Modeling Based Image Indexation

By modeling the query image and every images in the database with RAG, we will have a query graph **q** and a graphs database **D**= {g1, g2, g3, ... , gn}. Afterwards, frequent sub-graphs are extracted from the graph database. There frequent subgraphs, called also graph features, must obey to two conditions: graph features should be connected and its support, which is the occurrence in the graph database, exceeds a prefixed minimum threshold **s**. These graph features expose the intrinsic propriety of the graph database. As well as the k-means classes number and as shown in Table.1, choosing the threshold s is very crucial since with a great value, few graph features will be selected and the database will be represented only with some graphs. Contrarily, setting this threshold to a small value will result on having a big number of graph features which will be hard to handle with.

These graph features are obtained following several methods [5] such as AGM, FSG, and gSpan [7] which outperforms the previous methods and proved its power and efficiency since it is capable to mine large graph features in a bigger graph set with small threshold s. gSpan is a depth-first search (DFS) based graph mining algorithm by constructing a DFS code for every graph as its canonical label, building new lexicographic ordering among these codes and conducting DFS to discover graph features in large graph databases. Based on the graph features extracted by gSpan, we index every graph as well as the query graph. Consequently, every image will be indexed with the frequent graph features that contains.

4.2 Image Feature Matching

Applying an exhaustive search to find the query graph in a graphs database is a very time consuming process and an NP-complete problem since we have not only to survey the entire graphs database but also to investigate the subgraph isomorphism [8],[9]. That's why; we have applied a searching technique based on the extracted frequents subgraphs from the graph database. That's why we have extracted graph features from D with gSpan and used them to index the graphs in the databases as well as the query graph q.

The main goal here is to find all $g_i \in D$ where q is isomorphic (equal) to g_i. To do this, we have followed the commonly used filtering-and-verification framework [8],[9]. In the beginning, we must identify the graph features presented in q. This is done by determining, for every graph feature, the maximum common subgraph (MCS) between it and the query graph [10]. Finding MCS is obtained by computing the maximum clique, which is the largest set of connected vertices, in the association graph between the graph feature and q [9]. So, graph feature exist in q if it have the same size as the maximum clique in the association graph. Thereafter, since we know the graph features existing in q, we conduct filtering phase which consist on identifying the graphs that contains the same graph features as the query graph. These graphs will form a candidate set **C** that will be inputted to the verification phase. In the verification phase, graph isomorphism is performed on the graph database in order to retrieve similar graphs to the query graph.

In this way, query graph will be annotated based on its similar found graphs. Applying this image feature matching method on query video keyframes, the video annotation will be that of the major likewise annotated keyframes.

5 Video Annotation Improvement by Graph Aggregation

As we have stated beforehand, video annotation investigate video database attempting to get videos similar to the video query which will be annotated based on them. Unfortunately, this is not always feasible since it is not possible to return output result for many videos. Consequently, we have thought to apply a graph aggregation method to have an answer for the non annotated videos by the basic annotation method described in section 4, especially with the encouraging results of the graph aggregation system for databases querying developed in [9].

Once the filtering-and-verification process performed as a basic video annotation system and no similar graph are returned, we will use the graphs candidate set **C** to compose the query graph using their set of subgraphs **S**. To accelerate the aggregation procedure, the graph database is sorted according to the similarity to the query graph q based on the regions visual features of its corresponding image.

However, graph aggregation can be applied following several steps: the first graph in the sorted list is initially used to form the query graph q. So, we will find the MCS between it and q, adding this MCS to S and repeating this process for the rest of q until we will no more query graph vertices to treat or all the graphs in C are investigated. As a result, joining the subgraphs in S will form the query graph q. Because the subgraphs in S can have different annotation, we will apply the majority vote on the half of subgraphs in S, which are the subgraphs that have the biggest size, to return the final annotation for q. Many aggregation set can be build which insure a better annotation result. Even if we annotate the video query with the basic annotation system, we can enrich this annotation with the graph aggregation results. As shown in the example in figure.3, the query video keyframe q is composed of the two video keyframes VK_2 (containing the girl and the background) and VK_4 (containing the man) from the video keyframes database and it will be annotated as their annotation since they belong to the same video.

Fig. 1. Query video keyframe

Fig. 2. Video keyframes database samples

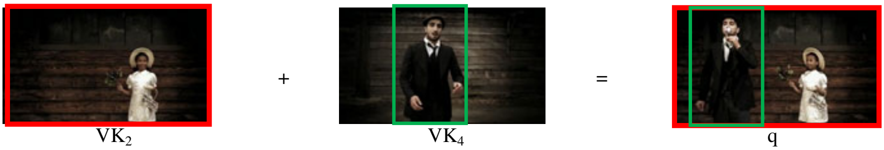


Fig. 3. An example of video keyframes aggregation

Graph aggregation for video annotation is a very efficient method since video sequences similar to the query video sequence are not always present in the video databases but there are, frequently, video sequences from the same video or the same movie. In this case, we will have a right annotation.

6 Experimental Results

To better evaluate our video annotation system, we have collected varieties of labeled real world video sequences to make the video database. After extraction keyframes from video database and query video sequences, we have obtained 1091 frames from the database and 242 keyframes from the query video sequences. However, 116 of these query keyframes are from a video sequences which didn't exist in the video database. But, for many query video sequences, the video database contains video sequences belonging to the same video. By conducting gSpan, with different minimum threshold, to the graph database gotten after modeling and indexing every database keyframes, we had the frequent graph features number illustrated in the table below. The graph features number decrease with the decreasing of the prefixed minimum threshold as soon as the k-means classes number using for image regions labeling. In our system, we have used 10 classes to label image regions and we have fixed the minimum threshold to 5% since they give a sufficient graph features as well as better result.

Table 1. Frequent subgraphs number extracted with gSpan

Threshold s	K-means classes number		
	5	*10*	*20*
1%	517163	63786	4770
2%	35214	4333	236
5%	3014	367	54
10%	632	82	16

By applying the basic video annotation and the graph aggregation based improved video annotation method, we have gotten the results shown in Table.2.

Table 2. Video annotation accuracy (with minimum threshold=5%)

Labels	Basic method	Graph aggregation based method
5	52%	78%
10	58%	**91%**
20	56%	84%

It is clear here that the second method have given a superior accuracy since it annotated the videos sequences that don't exist in the video database. Having small labels number doesn't only increase the computational time but also decrease the annotation accuracy since it allows wrong graph matching. It is well confirmed here that graph aggregation based method outperform basic method since it annotate the videos which are not in the database.

7 Conclusions

Motivated by the advantages of graph based modeling for image indexing, we have presented, in this paper, our graph based video annotation system which is improved by a graph aggregation scheme. Our system has proved its efficiency and power since it brings answers for videos that are not possibly annotated by classical methods.

In the future work, we will use a largest database, so that our system will be more evaluated. Furthermore, we will ameliorate our system by using more efficient technique for image segmentation, more robust visual features for image regions description and more powerful graph mining techniques. Moreover, we will extend our approach to others applications such as satellite surveillance.

Acknowledgement. The authors would like to acknowledge the financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program. Moreover, this work is partially supported by ANR (Agence Nationale de la Recherche) project ANR-08-CORD-009.

References

1 Pramod Sankar, K., Meshesha, M., Jawahar, C.V.: Annotation of Images and videos based on Textual Content without OCR. In: Workshop on Computation Intensive Methods for Computer Vision (in conjunction with ECCV 2006) (2006)

2 Ben Aoun, N., Elghazel, H., Ben Amar, C.: Graph modeling based video event detection. In: International Conference on Innovations in Information Technology, Abu Dhabi, United Arab Emirates (2011)

3 Sukhwinder Bir, S., Kaur, A.: Color Image Segmentation in CIEab Space Using Hill Climbing Algorithm. The International Journal of Computer Applications 7(3), 48–53 (2010)

4 Petrakis, E., Faloutsos, C.: Similarity Searching in Medical Image Databases. IEEE Transactions on Knowledge and Data Engineering 9(3), 435–447 (1997)

5 Iváncsy, G., Iváncsy, R., Vajk, I.: Graph Mining-based Image Indexing. In: 5th International Symposium of Hungarian Researchers on Computational Intelligence, Budapest, Hungary, pp. 313–323 (2004)

6 Elsayed, A., Coenen, F., Jiang, C., García-Finana, M., Sluming, V.: Corpus Callosum MR image classification. Knowledge-Based Systems 23, 330–336 (2010)

7 Yan, X., Han, J.: gSpan: graph-based substructure pattern mining. In: The Proceeding of the IEEE International Conference on Data Mining (ICDM 2002), Maebashi City, Japan, pp. 721–724 (2002)

8 Shang, H., Zhang, Y., Lin, X., Yu, J.X.: Taming verification hardness: an efficient algorithm for testing subgraph isomorphism. In: International Conference on Very Large Data Bases, pp. 364–375 (2008)

9 Elghazel, H., Hacid, M.: Aggregated Search in Graph Databases: Preliminary Results. In: 8th IAPR-TC-15 Workshop on Graph-based Representations in Pattern Recognition (GbR 2011), Munster, Germany (2011)

10 Raymond, J.W., Willett, P.: Maximum common subgraph isomorphism algorithms for the matching of chemical structures. Journal of Computer-Aided Molecular Design 16(7), 521–533 (2002)

11 Wali, A., Ben Aoun, N., Karray, H., Ben Amar, C., Alimi, A.M.: A new system for event detection from video surveillance sequences. In: Blanc-Talon, J., Bone, D., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2010, Part II. LNCS, vol. 6475, pp. 110–120. Springer, Heidelberg (2010)

12 Ben Aoun, N., El'Arbi, M. Ben Amar, C.: Multiresolution motion estimation and compensation for video coding. In: The 10th IEEE International Conference on Signal Processing (ICSP 2010), Beijing, China, pp. 1121–1124 (2010)

Violence Detection in Video Using Computer Vision Techniques

Enrique Bermejo Nievas[1], Oscar Deniz Suarez[1], Gloria Bueno García[1], and Rahul Sukthankar[2]

[1] E.T.S.I.Industriales, Universidad de Castilla-La Mancha
Avda. Camilo Jose Cela s/n, 13071 Ciudad Real, Spain
Oscar.Deniz@uclm.es,
http://visilab.etsii.uclm.es/

[2] Intel Labs Pittsburgh and Robotics Institute, Carnegie Mellon, USA

Abstract. Whereas the action recognition community has focused mostly on detecting simple actions like clapping, walking or jogging, the detection of fights or in general aggressive behaviors has been comparatively less studied. Such capability may be extremely useful in some video surveillance scenarios like in prisons, psychiatric or elderly centers or even in camera phones. After an analysis of previous approaches we test the well-known Bag-of-Words framework used for action recognition in the specific problem of fight detection, along with two of the best action descriptors currently available: STIP and MoSIFT. For the purpose of evaluation and to foster research on violence detection in video we introduce a new video database containing 1000 sequences divided in two groups: fights and non-fights. Experiments on this database and another one with fights from action movies show that fights can be detected with near 90% accuracy.

Keywords: action recognition, fight detection, video surveillance

1 Introduction

In the last years, the problem of human action recognition at a distance has become tractable by using computer vision techniques. Although the first approaches obtained good results, they have some limitations too. There are, for example, aperture problems and discontinuities in optical flow based approaches [8], and illumination and reinitialization problems in feature tracking approaches [2]. More recently, the use of feature descriptors around interest points has become popular within the action recognition community, see the recent survey [16]. This approach analyzes actions by considering the video sequence as a space-time volume and using gradients, intensities, flows or other local features. This approach has shown better tolerance to posture, occlusion, illumination or deformation. On the other hand, current methods usually involve spatio-temporal analysis of 3D descriptors at multiple scales in high resolution videos, so they require high computational costs.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 332–339, 2011.

Approaches based on feature descriptors typically use the well-known bag-of-words framework [14,7]. In this case the output is simply a histogram that reflects *word* distribution as frequencies. In order to obtain the histogram, the bag-of-words representation creates a vocabulary using for example k-means clustering. The complete procedure is described in Section 4.

The goal of this paper is to assess the performance of modern action recognition approaches for the recognition of fights in videos, movies or video-surveillance footage. Most of previous work on action recognition focuses on simple human actions like walking, jumping or hand waving [13]. Despite its potential usefulness, violent action detection has been less studied. Whereas there is a number of well-studied datasets for action recognition, significant datasets with violent actions have not been made available. In this work we introduce a fight dataset and use two of the best action recognition methods currently available (STIP [12] and MoSIFT [4]) to assess the performance in the fight detection problem.

A violence detector has immediate applicability both in the surveillance domain and for rating/tagging online video content. The primary function of large-scale surveillance systems deployed in institutions such as schools, prisons and elder care facilities is for alerting authorities to potentially dangerous situations. However, human operators are overwhelmed with the number of camera feeds and manual response times are slow, resulting in a strong demand for automated alert systems. Similarly, there is increasing demand for automated rating and tagging systems that can process the great quantities of video uploaded to websites. The primary contribution of this paper are two-fold. First, we show that one can construct a versatile and accurate fight detector using a local descriptors approach. Second, we present a new dataset of hockey video containing fights and demonstrate that our proposed approach can reliably detect violence in sports footage, even in the presence of camera motion.

The paper is organized as follows. Section 2 analyzes previous work on violence recognition. Next, we describe the new hockey fights dataset in Section 3. Section 4 presents the two descriptors we use for activity recognition. Then, we describe the bag-of-words approach and the discriminative classifier. Section 5 details our evaluation methodology and summarizes our experimental results on the hockey fights dataset. Finally, in Section 6 we summarize key conclusions.

2 Related Work

One of the first proposals for violence recognition in video is Nam *et al.* [18], which proposes recognizing violent scenes in videos using flame and blood detection and capturing the degree of motion, as well as the characteristic sounds of violent events. Cheng *et al.* [5] recognizes gunshots, explosions and car-braking in audio using a hierarchical approach based on Gaussian mixture models and Hidden Markov models (HMM). Giannakopoulos *et al.* [10] also propose a violence detector based on audio features. Clarin *et al.* [6] present a system that uses a Kohonen self-organizing map to detect skin and blood pixels in each frame and motion intensity analysis to detect violent actions involving blood. Zajdel *et al.* [19], introduce the CASSANDRA system, which employs motion features related to articulation in video and scream-like cues in audio to detect aggression in surveillance videos.

More recently, Gong *et al.* [11] propose a violence detector using low-level visual and auditory features and high-level audio effects identifying potential violent content in movies. Chen *et al.* [3] use binary local motion descriptors (spatio-temporal video cubes) and a bag-of-words approach to detect aggressive behaviors. Lin and Wang [15] describe a weakly-supervised audio violence classifier combined using co-training with a motion, explosion and blood video classifier to detect violent scenes in movies. Giannakopoulos *et al.* [9] present a method for violence detection in movies based on audio-visual information that uses a statistics of audio features and average motion and motion orientation variance features in video combined in a k-Nearest Neighbor classifier to decide whether the given sequence is violent.

In summary, a number of previous works require audio cues for detecting violence or rely on color to detect cues such as blood. In this respect, we note that there are important applications, particularly in surveillance, where audio is not available and where the video is greyscale. Finally, while explosions, blood and running may be useful cues for violence in action movies, they are rare in real-world surveillance video. In this paper, we focus on reliable cues for early detection of violence in such settings.

3 Dataset

The majority of widely used, publicly-available datasets in action recognition, such as KTH [13], focus on single actors performing a simple action like walking, jumping or waving against an uncluttered background; these are clearly unsuitable for evaluating violence detection. Datasets such as INRIA IXMAS, which show an individual kicking or punching could be used to train (but not evaluate) fight detection systems. Some datasets like CAVIAR, BEHAVE or CareMedia contain some instances of people engaged in aggressive behaviors, but that is not their primary focus.

Our intention is to introduce a new video dataset created specifically for evaluating violence detection systems, where both normal and violent activities occur in similar, dynamic settings. To this end, we collected 1000 clips of action from hockey games of the National Hockey League (NHL), as shown in Fig. 1. Each clip consists of 50 frames of 720×576 pixels and is manually labeled as "fight" or "non-fight". This dataset enables us to easily and robustly measure the performance of a variety of violence recognition approaches, as shown in Section 5. Our fight dataset is available by request from the authors.

4 Activity Recognition

Local image features or interest points provide compact and abstract representations of patterns in an image. Analogously, with local spatio-temporal features

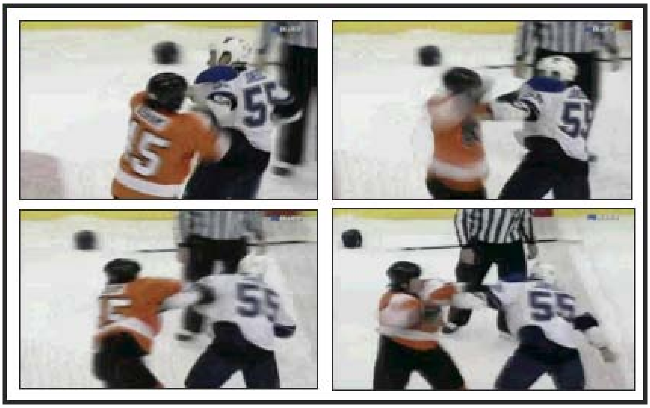


Fig. 1. Sample of a fight clip from our 1000-video database

it is possible to obtain compact and descriptive representations of motion. In this respect, two prominent spatio-temporal descriptors are Space-Time Interest Points (STIP) and Motion SIFT (MoSIFT).

As described in [12], Space-Time Interest Points (STIP) is an extension of the Harris corner detection operator to space-time. The detected interest points are characterized by a high variation of the intensity in space, and non-constant motion in time. These salient points are detected at multiple spatial and temporal scales. Then, HOG (Histograms of Oriented Gradients), HOF (Histograms of Optical Flow) and a combination of HOG and HOF termed HNF feature vectors are extracted for 3D video patches in the neighborhood of the detected STIPs. These features can be used for recognizing motion events with high performance and they are robust to scale, frequency and velocity variations of the pattern.

MoSIFT [4] is an extension of the popular SIFT [17] image descriptor for video. The standard SIFT extracts histograms of oriented gradients in the image. The 256-dimensional MoSIFT descriptor consists of two portions: a standard SIFT image descriptor and an analogous histogram of optical flows, which represents local motion. These descriptors are extracted only from regions of the image with sufficient motion. The MoSIFT descriptor has shown better performance in recognition accuracy than other state-of-the-art descriptors [4] but the approach is significantly more computationally expensive than STIP.

On the other hand, the Bag-of-Words (BoW) approach, adopted from the text retrieval community [14], has recently become popular for image [7] and video understanding [16]. The approach represents each video sequence as a histogram over a set of *visual words* to generate a fixed-dimensional encoding that can be processed using a standard classifier. In a learning phase, the vocabulary of visual words is typically defined as the cluster centers obtained from k-means clustering over a large collection of sample low-level descriptors (STIP or MoSIFT descriptors, see above). In our study, we evaluated vocabularies with 50, 100, 150, 200, 300, 500 and 1000 cluster centers.

Given a vocabulary, the next step is to quantize each descriptor extracted from the given video to the closest *visual word*, thus generating histograms of word occurrence. The final step of this BoW approach is the classification of the histograms. These histograms are high-dimensional vectors that can be classified using a standard classifier, typically a Support Vector Machine (SVM). It is well-known that the choice of SVM kernel can significantly affect performance; in our experiments, we explore the following popular kernels: the histogram intersection kernel [1] (HIK), radial basis function (RBF) and Chi-Squared kernel, which is a variant of RBF that uses χ^2 distance. Kernel parameters are tuned using 5-fold cross-validation.

5 Experimental Results

The BoW approach using the STIP and MoSIFT descriptors was evaluated on the 1000-clip hockey fight dataset. In order to assess the impact of vocabulary size, we generated vocabularies of 50, 100, 150, 200, 300, 500 and 1000 words. Table 1 presents the accuracy of fight detection, averaged over 5-fold cross-validation. For space reasons, we only show here results obtained with the histogram intersection kernel (HIK) since it consistently outperformed RBF and χ^2. We see that the BoW variants all achieve accuracies near 90%, with a slight improvement with increasing vocabulary size. On this dataset, STIP(HOG) and MoSIFT perform comparably. The ROC curve for the best of those runs is shown in Figure 2. Note that this result (MoSIFT on 500-word vocabulary) does not correspond to the highest result in Table 1 since the latter shows accuracies averaged over all folds.

Table 1. Accuracy of fight detection on 1000-clip hockey dataset (5-fold CV)

Vocabulary	STIP (HOG) + HIK	STIP (HOF) + HIK	MoSIFT + HIK
50	87.8%	83.5%	87.5%
100	89.1%	84.3%	89.4%
150	89.7%	85.9%	89.5%
200	89.4%	87.5%	90.4%
300	90.8%	87.2%	90.4%
500	91.4%	87.4%	90.5%
1000	**91.7**%	**88.6**%	**90.9**%

Hockey fights contain useful information for learning fight patterns. Still, can those patterns be translated to other scenarios? To explore the generalization capacity of the studied approaches, we also evaluated the fight recognition system on a second dataset consisting of 200 video clips obtained from action movies (see Figure 3 for examples), of which 100 contained a fight. Unlike the hockey dataset, which was relatively uniform both in format and content, these videos depicted a wider variety of scenes and were captured at different resolutions. Table 2

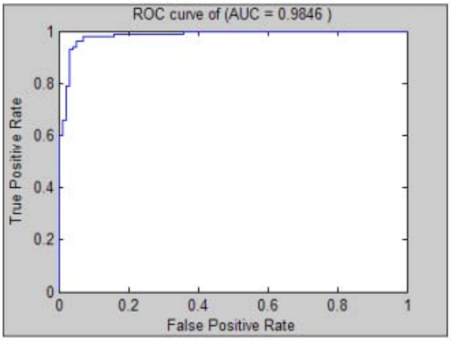


Fig. 2. ROC curve of fight detection on 1000-video Hockey dataset using MoSIFT with 500-word vocabulary and histogram intersection kernel

summarizes the results. In this case, STIP (HOF) outperformed STIP (HOG), but STIP's overall performance is very poor as compared to MoSIFT (59.0% vs. 89.5%). Note that increasing vocabulary size does not uniformly improve recognition accuracy.

We make several observations based on these experiments. First, detecting fights in televised hockey footage is easier than detecting fights in action movies, despite the fact that the former contains very similar footage for both classes. This could partially be attributed to the fact that fights in movies are more varied in appearance and cinematography while sports footage is relatively consistent. However, it also indicates that televised hockey fights may exhibit more reliable cues that a supervised classifier can exploit — for instance, the camera tends to zoom in to a hockey fight while showing more wide-angle shots during non-fight segments of the hockey game. Second, we see that STIP and MoSIFT are similar in performance on the former task but MoSIFT is dramatically superior

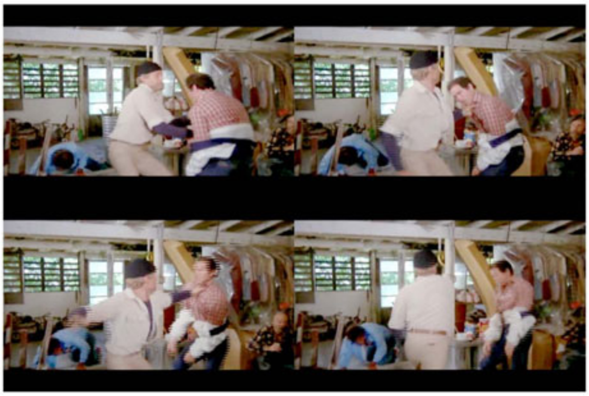


Fig. 3. Example of a fight sequence from an action movie

Table 2. Accuracy of fight detection on action movie dataset (5-fold CV)

Vocabulary	STIP (HOG) + HIK	STIP (HOF) + HIK	MoSIFT + HIK
50	44.5%	51.2%	76.0%
100	45.0%	56.5%	79.5%
150	**49.0**%	**59.0**%	80.0%
200	46.5%	53.5%	80.0%
300	44.5%	52.5%	87.5%
500	44.5%	50.5%	**89.5**%
1000	38.5%	52.5%	89.0%

on the action movie dataset (retaining 90% accuracy levels). This indicates that the MoSIFT representation, though more computationaly expensive than STIP, does make a difference.

6 Conclusions

Recognizing fights and aggressive behavior in video is an increasingly important application area. Such capability may be extremely useful in video surveillance scenarios like in prisons, psychiatric or elderly centers. Action recognition techniques that have focused largely on individual actors and simple events can be extended to this specific application. This paper evaluates how state-of-the-art video descriptors can perform fight detection on two new datasets: a 1000-video collection of NHL hockey games and a smaller 200-clip collection of scenes from action movies. Experiments show that the popular bag-of-words approach can accurately recognize fight sequences with approximately 90% accuracy. For the hockey dataset, we observed that accuracy was insensitive to the choice of low-level feature descriptor and vocabulary size; however, on the second dataset, the choice of descriptor was critical, with MoSIFT dramatically outperforming the best STIP under all conditions. The promising performance of action recognition methods on this challenging task shows that a versatile marketable fight detector may be feasible.

Acknowledgements. This work was partially funded by Project PII2I09-0043-3364 of Castilla-La Mancha Regional Government. We thank Drs. Alex Hauptmann and Ming-yu Chen for providing the MoSIFT descriptor and their helpful suggestions.

References

1. Barla, A., Odone, F., Verri, A.: Histogram intersection kernel for image classification. In: Proceedings of ICIP, pp. 513–516 (2003)
2. Bregler, C.: Learning and recognizing human dynamics in video sequences. In: Proceedings of Computer Vision and Pattern Recognition (1997)

3. Chen, D., Wactlar, H., Chen, M., Gao, C., Bharucha, A., Hauptmann, A.: Recognition of aggressive human behavior using binary local motion descriptors. In: Engineering in Medicine and Biology Society, pp. 5238–5241 (20-25 2008)
4. Chen, M., Hauptmann, A.: MoSIFT: Recognizing human actions in surveillance videos. Tech. rep., Carnegie Mellon University, Pittsburgh, USA (2009)
5. Cheng, W.H., Chu, W.T., Wu, J.L.: Semantic context detection based on hierarchical audio models. In: Proceedings of the ACM SIGMM workshop on Multimedia information retrieval, pp. 109–115 (2003)
6. Clarin, C., Dionisio, J., Echavez, M., Naval, P.C.: DOVE: Detection of movie violence using motion intensity analysis on skin and blood. Tech. rep., University of the Philippines (2005)
7. Csurka, G., Dance, C., Fan, L.X., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision (2004)
8. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: IEEE International Conference on Computer Vision, pp. 726–733 (2003)
9. Giannakopoulos, T., Makris, A., Kosmopoulos, D., Perantonis, S., Theodoridis, S.: Audio-visual fusion for detecting violent scenes in videos. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 91–100. Springer, Heidelberg (2010)
10. Giannakopoulos, T., Kosmopoulos, D., Aristidou, A., Theodoridis, S.: Violence content classification using audio features. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 502–507. Springer, Heidelberg (2006)
11. Gong, Y., Wang, W., Jiang, S., Huang, Q., Gao, W.: Detecting violent scenes in movies by auditory and visual cues. In: Proceedings of the 9th Pacific Rim Conference on Multimedia, pp. 317–326. Springer, Heidelberg (2008)
12. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64, 107–123 (2005)
13. Laptev, I., Lindeberg, T.: Space-time interest points. In: Proceedings of International Conference on Computer Vision, pp. 432–439 (2003)
14. Lewis, D.: Naive Bayes at Forty: The independence assumption in information retrieval. In: European Conference on Machine Learning, pp. 4–15 (1998)
15. Lin, J., Wang, W.: Weakly-supervised violence detection in movies with audio and video based co-training. In: Muneesawang, P., Wu, F., Kumazawa, I., Roeksabutr, A., Liao, M., Tang, X. (eds.) PCM 2009. LNCS, vol. 5879, pp. 930–935. Springer, Heidelberg (2009)
16. Lopes, A.P.B., do Valle Jr., E.A., de Almeida, J.M., de Albuquerque Araújo, A.: Action recognition in videos: from motion capture labs to the web. CoRR abs/1006.3506 (2010)
17. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(91) (2004)
18. Nam, J., Alghoniemy, M., Tewfik, A.: Audio-visual content-based violent scene characterization. In: Proceedings of ICIP, pp. 353–357 (1998)
19. Zajdel, W., Krijnders, J., Andringa, T., Gavrila, D.: CASSANDRA: audio-video sensor fusion for aggression detection. In: IEEE Conference on Advanced Video and Signal Based Surveillance, AVSS 2007, pp. 200–205 (2007)

Speckle Denoising through Local Rényi Entropy Smoothing

Salvador Gabarda and Gabriel Cristóbal

Instituto de Optica (CSIC), Serrano 121, 28006 Madrid, Spain
{salvador,gabriel}@optica.csic.es

Abstract. Quality enhancement of radar images is highly related to speckle noise reduction. There are plenty of such techniques that have been developed by different authors. However, a definitive method has not been already attained. Filtering methods are popular to reduce speckle noise. This paper introduces a new method based on filtering a smoothed local pseudo-Wigner distribution using a local Rényi entropy measure. Results are compared to other well-known noise reduction filtering methods for artificially degraded speckle images and real world image examples. Experimental results confirm that this method outperforms other classical speckle denoising methods.

Keywords: SAR imaging, speckle noise, image filtering, denoising.

1 Introduction

Speckle is a form of multiplicative noise that is exhibited by coherent imaging which appears in many applications such as ultrasound imaging, synthetic aperture radar (SAR) images or optical coherent tomography (OCT) to name a few. Speckle noise in radar (Radio Detection and Ranging) has its origin in the backscatter wave interference that generates characteristic bright and dark pixels in remote imaging systems operating with coherent radiation. Speckle is a natural component of radar images giving to them a characteristic granular or mottled appearance. The waves emitted by the transmitter travel in phase, but after the interaction with the target they are no longer in phase due to the different distance they travel back to the detector and to the scattering effect they undergo. The result is that radar waves interfere originating multiplicative or speckle noise. In order to enhance the quality of the images created by radar technology, speckle has to be suppressed or reduced. Speckle is a kind of correlated noise and therefore it can be hardly completely eliminated although it can be significantly reduced by denoising techniques. Different filtering methods have been proposed in the literature to reduce speckle noise. One posible side effect is that filtering algorithms eliminate part of the original image information along with the noise, especially the high-frequency information related to image edges or details. In general, denoising methods can be broadly classified as adaptive and non-adaptive filtering algorithms. Non-adaptive filters are faster and easily

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 340–347, 2011.

to be implemented. They use the same smoothing weights for the whole image, ignoring differences in image contrast or texture. Examples of non-adaptive filters are the mean and median filters. On the other side, adaptive speckle filtering methods preserve edges and high-textured details. Among the best known adaptive filtering methods used for speckle denoising we can cite the Lee's [1] , Frost's [2] and Kuan's [3] methods. A new kind of adaptive filter is proposed here. The aim of this technique is to deal with the problem of speckle noise reduction in SAR imaging. This filter operates in the space-frequency domain and it is based on smoothing a local pseudo-Wigner distribution of the image, according to a local Rényi entropy measure. Elsewhere, a comparison with other existing speckle denoising techniques has been performed.

This paper is organized as follows. Section 2 presents the mathematical background. Section 3 describes the method of speckle noise reduction in SAR imaging based on the previously described theory. Section 4 illustrates the performance of this new method by means of simulated and real world image examples. Finally, conclusions are drawn in Section 5.

2 Mathematical Background

The speckle noise reduction method described here is based on smoothing the coefficients of one of the most frequently used space-frequency representations (SFR) namely the Wigner-Ville distribution [5]. Although the SFR of a signal can be achieved by diverse existing functions, we have selected the Wigner-Ville distribution (WVD) because, in addition to its excellent properties, it is considered as a master form function from which the other existing representations can be derived [6]. For a common analytical SFR framework see e.g. Cohen [4]. The WVD has been approximated to the discrete case by different authors as, for example, by Claasen and Mecklembräuker [7] and Brenner [8] by means of the following equation

$$W(n,k) = 2 \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} z(n+m) z^{*}(n-m) e^{-i2\pi k\left(\frac{2m}{N}\right)} \tag{1}$$

The discrete WVD approximations have been referred as pseudo-Wigner distributions (PWD) to take into account that they do not exactly match the properties of the continuous version. In Eq.(1), $z[n]$ is a 1-D sequence of data from the input image, n and k represent the space and frequency discrete variables respectively, and m is a shifting parameter, which is also discrete. $W(n,k)$ is a matrix where every row is a vector representing the pixel-wise PWD corresponding to pixel n. By scanning the image with a 1-D window of N data, i.e., by shifting the window to all possible positions along the signal, the full pixel-wise PWD of the image is produced. The use of a 1-D distribution for images can be justified by the fact that we can scan the image matrix in any desired direction (see figure 1). Hence, this 1-D PWD can be considered as a directional space-frequency representation of the image. Also as the correlation length of the image is finite, we can use an analysis window restricted to a few pixels around position n to calculate de PWD, saving in this way a great amount of computational time.

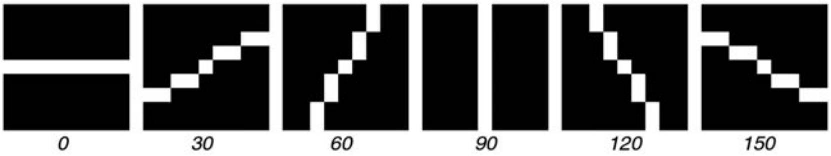


Fig. 1. Different configuration of pixels for calculating a directional PWD by means of Eq.(1), with $N = 8$ pixels and six equally spaced orientations from 0 to 150 degrees (note that due to π periodicity, 180 degree orientation coincides with 0)

Given a position n and the values of the signal in a neighborhood of n, Eq.(1) gives a vector whose elements contain information of the strength of the discrete frequencies existing in such spatial neighborhood (see figure 2).

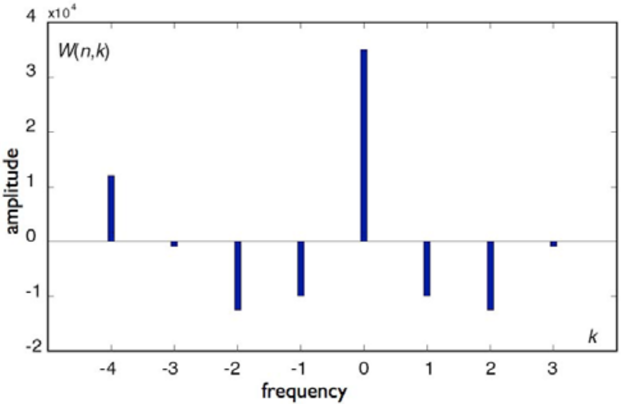


Fig. 2. Graphical representation of the PWD of a given signal at position n for a neighborhood of 8 pixels, according to Eq.(1)

Namely, Eq.(1) represents the discrete Fourier transform (DFT) of the product $r(n,m) = z(n+m)z^*(n-m)$. Here z^* indicates the complex-conjugate of signal z. This equation is limited to a spatial interval $n \in [-\frac{N}{2}, \frac{N}{2}]$. Note that unless pixel $n = \frac{N}{2}$ is not included in the span of the sum in Eq.(1) (the period of the PWD), this position is required to fulfill the calculation of the first product inside the integral. Even more, to keep the real character of the WD, based on the periodicity of the data, it is required to consider that $z(-\frac{N}{2}) = z(\frac{N}{2})$. This restriction may be relaxed when dealing with real-valued images (pixel $z(\frac{N}{2})$ may have indeed any real value). This is because the lack of periodicity in real functions does not interfere with the real character of the PWD. On the other hand, the PWD makes posible a pixel by pixel analysis of the image by means of a pixelwise entropic measure indicating the noisiness degree of the signal in a given pixel position. Once the coefficients have been filtered, the PWD can be inverted to recover the filtered image. To do that, we have to perform an inverse DFT for recovering the product $r(n, m) = z(n+m)z^*(n-m)$. According to the inversion property [7] the even samples can be recovered from [20].

$$r(n, n) = z(2n)z^*(0) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} W(n, k)e^{-i2\pi n\left(\frac{2k}{N}\right)} \tag{2}$$

and the odd samples from

$$r(n, n-1) = z(2n-1)z^*(1) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} W(n, k)e^{-i2\pi n\left(\frac{2k}{N}\right)} \tag{3}$$

Expression (2) is true when calculating $r(n, n)$ in $r(n, m)$ and similarly, equality (3) is true when calculating $r(n, n-1)$ in the product function $r(n, m)$. To recover the exact values of the samples, we have to divide these values by $z^*(0)$ and $z^*(1)$, respectively. In practice, when the PWD is applied by means of an sliding window over the image, the inverse value to be recovered is always the central value of the inverse DFT of the PWD in each pixel. That is, $z(n) = \sqrt{r(n, \frac{N}{2}+1)}$. Although the sign of the samples are indetermined due to the product sign rule, they can always be considered positive, because we are dealing with digital images which have real positive gray level values.

For speckle images, we have selected a Rényi entropy extracted from a joint spatial frequency representation such as the PWD as a measure for denoising. In this case we identify the outcome from Eq.(1) as a probability distribution given by $W(n) = (k_1, ..., k_N)$ for each pixel n. By considering PWD as a probability distribution, diverse measures of entropy may be defined, according with the type of normalisation applied. Using the formulation proposed by Rényi [9], such measure can be expressed as

$$R_\alpha(n) = \frac{1}{1-\alpha} log_2 \left(\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} \widetilde{W}^\alpha(n, k) \right) \tag{4}$$

for the distribution given by Eq.(1). We introduce the notation $\widetilde{W}(n, k)$ to indicate that the distribution has been normalised according to some criterion that we will describe hereafter, for providing a real-positive distribution.

The Rényi entropy is a generalisation of the Shannon entropy [10] that reduces to it when $\alpha \to 1$. Although the Rényi measures from time-frequency distributions formally resemble the original entropies, they do not have the same properties, conclusions and results derived in classical information theory. For instance, the positivity will not be always preserved (see figure 2), along with the unity energy condition. In order to reduce a distribution to the unity signal energy case, some kind of normalisation must be done. The normalisation can be performed in various ways, leading to a variety of possible measurement definitions e.g. due to: Stankovic [11], Sang [12] or Williams [13], with a significant contribution from Flandrin et al. [14] in establishing the properties of these measures. The use of entropic measures in positive time-frequency representations was done by Pitton et al. [15].

We have chosen to normalize $W(n,k)$ in the same way that a wave function is normalised in Quantum Mechanics [16], that is

$$\widetilde{W}(n,k) = \frac{W(n,k)W^*(n,k)}{\sum_k (W(n,k)W^*(n,k))} = \frac{W^2(n,k)}{\sum_k W^2(n,k)} \tag{5}$$

This normalisation has shown to be most suitable than other tested in our experiments. The values of $R_\alpha(n)$ depend upon the value of α and the size N of the window used in Eq.(1). Note that integer orders $\alpha > 2$ are recommended values for time-frequency distribution measures [14]. Parameter α gives an interesting flexibility for entropic measures. When α approaches to infinity, this entropy seems to consider only events with the highest probability. Oppositely, small values of α, tend to consider events more equally, regardless of their probabilities. Hence, α can be used as a parameter that determines the sensibility to entropy variations. We can expect some differences in the results by using different α values, but preserving the relative order of the measures. We have set α according with the results of our experiments. It can be shown that $0 \leq R_\alpha(n) \leq log_2(N)$. Hence, the measure can be normalised by applying $\widetilde{R}_\alpha(n) = R_\alpha(n)/log_2 N$.

3 Description of the Method

The rationale of the method that we are describing for speckle noise reduction is based on considering speckle as a high frequency noise phenomenon that interferes the original image and that it can be reduced by the use of smoothing filters. We will show that our technique is able to eliminate much of the high frequency content of the noise while simultaneously preserving the high frequency components of image edges. Eq.(1) is used for determining de PWD of the image. The parameter N is set to 8 pixels and the image is scanned in six different orientations denoted by $W_\theta(n,k)$ with $\theta \in \{0, 30, 60, 90, 120, 150\}$ degrees. By means of Eq.(4) we get the pixelwise normalised directional scalar entropies of the image, denoted by $R_\theta(n)$, where we have chosen $\alpha = 3$ and θ indicates the directionality. For each pixel n we determine the argument $\varphi(n) = \arg\min_\theta R_\theta(n)$ for which the entropy attains a minimum value among all the directionalities. The entropy corresponding to this orientation is used for smoothing the coefficients of the PWD as follows

$$\widehat{W}(n,k) = W_{\varphi(n)}(n,k)e^{-(1-\widetilde{R}_{\varphi(n)}(n))k^2} \times \frac{\|W_{\varphi(n)}(n,k)\|}{\|W_{\varphi(n)}(n,k)e^{-(1-\widetilde{R}_{\varphi(n)}(n))k^2}\|} \quad (6)$$

where the exponential function attenuates the high frequency coefficients of the PWD and the fractional factor keeps the strength of the signal normalized. When the resulting function, $\widehat{W}(n,k)$, is inverted to give $\widehat{z}(n) = \widehat{W}^{-1}(n,k)$, a denoised version of the original image is recovered. The algorithm is iterated until no significant change in the PSNR between the two last denoised images is observed. In each iteration a new set of directionalities is chosen as $\theta \in \{\psi, \psi + 30, \psi + 60, \psi + 90, \psi + 120, \psi + 150\}$, where ψ is a random argument introduced for decorrelating the iterations.

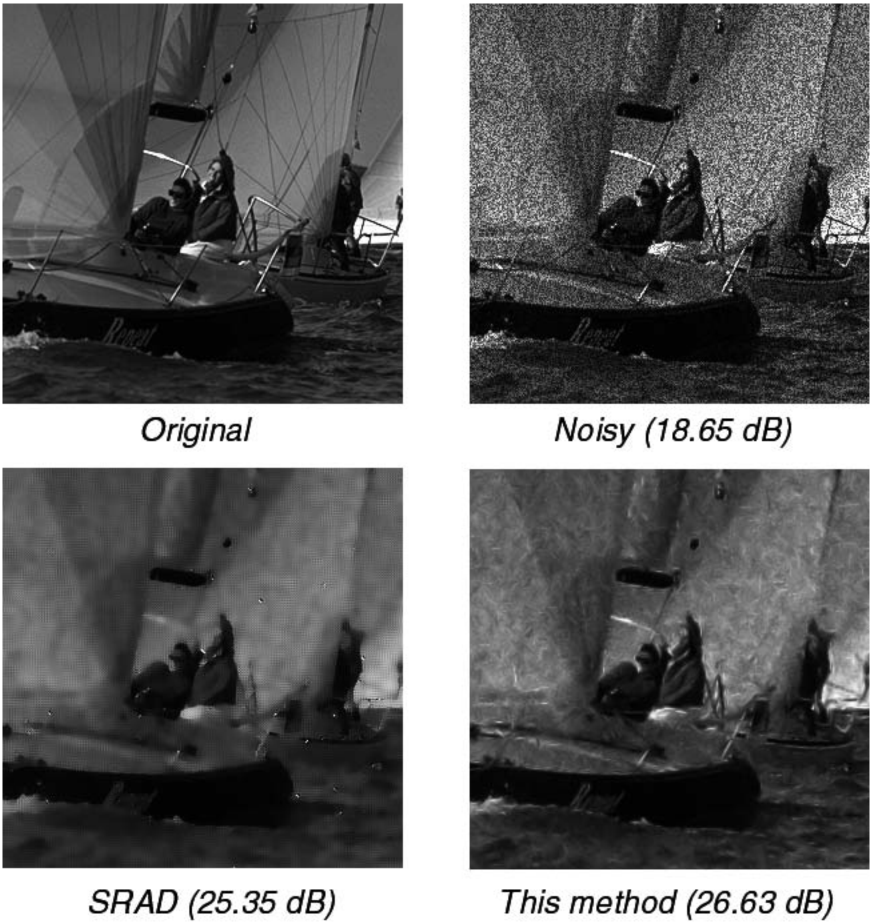
4 Experimental Results

A set of 25 images included as reference in the TID2008 database [17] has been used to perform a comparative test by adding speckle noise to the original images. Then the images have been processed through five different denoising methods, i.e.: Frost [2], Kuan [2], relaxed median filters [18], SRAD [19] and the method described here. The original color images of 512 × 384 pixel size from TID2008 database have been previously converted to 8-bit gray-level images to conform the reference database for our test. Table 1 shows the average values of the PSNR after the 25 experiments. The PSNR has been calculated by taking the corresponding original image as reference in each case.

Table 1. Speckle denoising comparative quality (Q) measures in dB

	input Q	**output Q**	**Q gain**
Frost	20.27	23.68	3.41
Kuan	20.27	24.40	4.13
Rmedian	20.27	23.53	3.26
SRAD	20.27	25.02	4.75
This method	20.27	26.10	5.83

Fig.(3) shows a visual comparative example with a detail of 300 × 300 pixels from one of the 25 images. From the above examples, one can conclude that the described method provides better performance both mathematically (see Table 1) and from the visual point of view (see Fig.(3)).



Original

Noisy (18.65 dB)

SRAD (25.35 dB)

This method (26.63 dB)

Fig. 3. Comparative image speckle denoising results

5 Conclusions

A new method for speckle noise reduction has been introduced in this paper. The method is iterative, adaptive and works at pixel level. A smoothed directional entropy is used for determining the filtering algorithm. A distinctive part of this algorithm is that the set of selected orientations is changed randomly in each iteration for decorrelating intermediate results. Experimental results confirm that this method outperforms other classical speckle denoising methods. Further work will require validating this method by comparing the test results with perceptual mean opinion scores (MOS) given by observers.

Acknowledgments. This research has been partially funded by the Spanish Ministry of Science and Innovation from the projects TEC2010-20307 and TEC2010-09834-E.

References

1. Lee, J.S.: Digital image enhancement and noise filtering by use of local statistics. IEEE Trans. Pattern Anal. Machine Intell. PAMI-2, 165–168 (1980)
2. Frost, V.S., Stiles, J.A., Shanmugan, K.S., Holtzman, J.C.: A model for radar images and its application to adaptive digital filtering of multiplicative noise. IEEE Trans. Pattern Anal. Machine Intell. PAMI-4(2), 157–166 (1982)
3. Kuan, D.T., Sawchuck, A.A., Strand, T.C., Chavel, P.: Adaptive restoration of images with speckle. IEEE Trans. Acoustics, Speech Signal Processing, PAMI-4 ASSP-35(3), 373–383 (1987)
4. Cohen, L.: Generalized phase-space distribution functions. J. Math. Physics 7, 781–786 (1966)
5. Wigner, E.: On the quantum correction for thermodynamic equilibrium. Phys. Rev. 40, 749–759 (1932)
6. Jacobson, L.D., Wechsler, H.: Joint spatial/spatial-frequency representation. Signal Process 14, 37–68 (1988)
7. Claasen, T.A.C.M., Mecklenbrauker, W.F.G.: The Wigner distribution–A Tool for Time Frequency Analysis, Parts I-III. Philips J. Research 35, 217–250, 276-300, 372-389 (1980)
8. Brenner, K.H.: A discrete version of the Wigner distribution function. In: Proc. EURASIP, Signal Processing II: Theories and Applications, pp. 307–309 (1983)
9. Rényi, A.: Some fundamental questions of information theory. In: Turán, P. (ed.) Selected Papers of Alfréd Rényi, vol. 3, pp. 526–552, Akadémiai Kiadó, Budapest, Originally MTA III, Oszt. Kazl, 10 (1960) pp. 251-282 (1976)
10. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. The University of Illinois Press, Urbana (1949)
11. Stankovic, L.: A measure of some time-frequency distributions concentration. Signal Processing 81, 621–631 (2001)
12. Sang, T.H., Williams, W.J.: Rényi information and signal dependent optimal kernel desig. In: Proceedings of the ICASSP, vol. 2, pp. 997-1000 (1995)
13. Williams, W.J., Brown, M.L., Hero, A.O.: Uncertainity, information and time-frequency distributions. In: SPIE Adv. Signal Process. Algebra Arch. Imp., vol. 1566, pp. 144–156 (1991)
14. Flandrin, P., Baraniuk, R.G., Michel, O.: Time-frequency complexity and information. In: Proceedings of the ICASSP, vol. 3, pp. 329–332 (1994)
15. Pitton, J., Loughlin, P., Atlas, L.: Positive time-frequency distributions via maximum entropy deconvolution of the evolutionary spectrum. In: Proc. ICASSP, vol. IV, pp. 436-439 (1993)
16. Eisberg, R., Resnick, R.: Quantum Physics. Wiley, Chichester (1974)
17. Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., Battisti, F.: TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. Advances of Modern Radioelectronics 10, 30–45 (2009)
18. Hamza, P.B., Luque-Escamilla, P.L., Martínez-Aroza, J., Román-Roldán, R.: Removing Noise and Preserving Details with Relaxed Median Filters. Journal of Mathematical Imaging and Vision 11(2), 161–177 (1999)
19. Yu, Y., Acton, S.T.: Speckle reducing anisotropic difusion. IEEE Transactions on Image Processing 11(11), 1260–1270 (2002)
20. Gonzalo, C., Bescos, J., Berriel-Valdos, L.R., Santamaria, J.: Spatial-variant filtering through the Wigner distribution function. Appl. Opt. 28 (1989)

Multiresolution Optical Flow Computation of Spherical Images

Yoshihiko Mochizuki[1,⋆] and Atsushi Imiya[2]

[1] School of Science and Technology, Chiba University, Japan
Yayoicho 1-33, Inage-ku, Chiba, 263-8522, Japan
[2] Institute of Media and Information Technology, Chiba University, Japan
Yayoicho 1-33, Inage-ku, Chiba, 263-8522, Japan

Abstract. As an application of image analysis on Riemannian manifolds, we develop an accurate algorithm for the computation of optical flow of omni-directional images. To guarantee the accuracy and stability of image processing for spherical images, we introduce the Gaussian pyramid transform, that is, we develop variational optical flow computation with pyramid-transform-based multiresolution analysis for spherical images.

1 Introduction

In this paper, we introduce a pyramid-transform on the sphere for the optical flow computation of a spherical image sequence. A spherical image is a non-negative function on the sphere, which is obtained by omnidirectional cameras.

Omnidirectional camera systems have been developed to observe a 360-degree field of view. The well-established omnidirectional imaging systems are catadioptric and dioptric camera systems. The catadioptric camera system is constructed as a combination of a quadric mirror and a conventional camera [4]. It is possible to transform the images acquired by omeni-directional camera systems into spherical images [2] if the appropriate factors of the camera systems, such as the parameters of the quadric surface in the catadioptric system and the refraction angle in the dioptric system, are known. Since the omnidirectional imaging system is widely used in mobile robots [5] the analysis of images on a sphere is required in robot vision. Furthermore, in biological vision, spherical views are fundamental tools for ego-motion estimation of insects in environments [6].

There are two typical methods for optical flow estimation for pinhole images-the Lucas-Kanade method (LK) and the Horn-Schunck method (HS), which are the template matching-based method and the variational-based method, respectively. The image pyramid technique is commonly used to refine the accuracy and stability of optical flow. The image pyramid is separated into filtering of images by Gaussian kernel and resizing of images by downsampling. The LK method with pyramid-based multiresolution optical flow computation (LKP) is

[⋆] Present Address: Faculty of Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo, 169-8555, Japan.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 348–355, 2011.

used to guarantee the accuracy and stability of the solution for image sequence observed by a conventional pinhole cameras, since filtering removels discontinuity of image intensity and dawonsampling preserves the global properties of image features.

Smoothing by the Gaussian kernel of the pyramid transform is computed using a discretized small kernel for an planar image, for example, the 5×5 window is a typical selection for the kernel assuming that the image is planar in this region. For the spherical coordinate system to represent spherical images, since the grid points of the spherical coordinate are not distributed uniformly on the sphere, the LKP is not suitable for optical flow computation on the sphere. However, since variational method such as the HS only involves the differentialsof a function, the method does not require uniform grid for numerical computation. Therefore, variational method is suitable for the optical flow computation on the spherical coordinates [1].

To guarantee the accuracy and stability of the optical flow computation on the sphere, we develop the Gaussian pyramid transform, that is, we develop variational optical flow computation with pyramid-transform-based multiresolution analysis. The Gaussian pyramid transform on the plane is achieved by downsampling of the convolution between an image and a kernel function. Since the convolution with the Gaussian kernel is the solution of the linear diffusion equation, the Gaussian pyramid is achieved by applying downsampling to the solution of linear diffusion equation. We extend this idea to the spherical images, that is, we construct the Gaussian pyramid of a spherical image by using spherical harmonic transform as a Gaussian filter on the sphere.

2 Gaussian Scale on the Sphere

On the unit sphere $\mathbb{S}^2$, centred at the origin in three-dimensional Euclidean space $\mathbb{R}^3$, the vector $\omega \in \mathbb{S}^2$ is expressed as

$$\omega = \omega(\phi, \theta) = (\sin\theta\cos\phi,\ \sin\theta\sin\phi,\ \cos\theta) \tag{1}$$

with $\phi \in [0, 2\pi)$, $\theta \in [0, \pi]$. The vector $\omega(\phi, \theta)$ satisfies the relation $\omega(\phi + \pi, \pi - \theta) = \omega(\phi, \theta)$.

The scale image of $f(\phi, \theta, \tau)$ of the image $f(\phi, \theta) : \mathbb{S}^2 \to \mathbb{R}$, is defined as the solution of the linear heat equation

$$\frac{\partial}{\partial \tau} f(\phi, \theta, \tau) = \Delta_{\mathbb{S}^2} f(\phi, \theta, \tau), \;\; f(\phi, \theta, 0) = f(\phi, \theta), \tag{2}$$

where

$$\Delta_{\mathbb{S}^2} := \frac{\partial^2}{\partial \theta^2} + \frac{1}{\tan\theta}\frac{\partial}{\partial \theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial \phi^2}, \tag{3}$$

On $\mathbb{S}^2$, $f(\phi, \theta)$ is expressed by the spherical harmonic series as

$$f(\phi, \theta) = \sum_{l=0}^{\infty} \sum_{m=0}^{l} c_l^m Y_l^m(\phi, \theta) \tag{4}$$

where

$$c_l^m = \int_{\mathbb{S}^2} f(\phi, \theta)\overline{Y_l^m(\phi, \theta)} \sin\theta d\phi d\theta. \tag{5}$$

The Gaussian scale image $f(\phi, \theta, \tau)$ of the scale τ is expressed as

$$f(\phi, \theta, \tau) = \sum_{l=0}^{\infty} \sum_{m=0}^{l} \left(c_l^m e^{-l(l+1)\tau}\right) Y_l^m(\phi, \theta). \tag{6}$$

As generalisation of the Gaussian pyramid transform (See Appendix), we define the pyramid transform on the sphere.

Definition 1. [1] *The Gaussian pyramid transform with the factor σ on the sphere is*

$$R_\sigma f(\phi, \theta) = f(\sigma\phi, \sigma\theta, \tau), \tag{7}$$

where $0 \leq \sigma\theta \leq \pi$ and $0 \leq \sigma\phi \leq 2\pi$, for an appropriate positive constant τ.

Fig. 1(a) shows that the mapping from the exterior sphere to the interior sphere defines the spherical pyramid transformation. Figs. 1(b) and 1 (c) show the fine and coarse resolution grids on the sphere, respectively, The image on the coarse resolution grid is generated from the image on the fine resolution grid by smoothing and downsampling

3 Spherical Optical Flow

The vector expressionof the spatial gradient on the unit sphere is $\nabla_{\mathbb{S}^2} = \left(\frac{\partial}{\partial\theta}, \frac{1}{\sin\theta}\frac{\partial}{\partial\phi}\right)^\top$. For the temporal image $f(\theta, \phi, t)$ on the unit sphere $\mathbb{S}^2$, the total derivative is

$$\frac{d}{dt}f = \frac{\partial}{\partial\theta}f + \frac{1}{\sin\theta}\frac{\partial}{\partial\phi}f + \frac{\partial}{\partial t}f \tag{8}$$

The solution $\dot{\omega} = \boldsymbol{v} = (\dot{\theta}, \dot{\phi})^\top$ of the equation

$$\boldsymbol{v}^\top \nabla_{\mathbb{S}^2} f + f_t = 0 \tag{9}$$

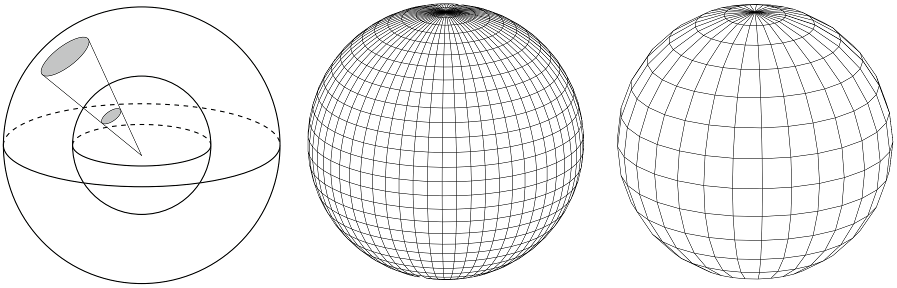
is optical flow of image f on the unit surface $\mathbb{S}^2$. The computation of optical flow from Eq. (9) is an ill-posed problem. Horn-Schunck criterion for the computation of optical flow [3] on the unit sphere is expressed as the minimisation of the functional

$$J(\dot{\theta}, \dot{\phi}) = \int_{S^2} \left\{(\boldsymbol{v}^\top \nabla_{\mathbb{S}^2} f + f_t)^2 + \alpha(||\nabla_{\mathbb{S}^2}\dot{\theta}||_2^2 + ||\nabla_{\mathbb{S}^2}\dot{\phi}||_2^2)\right\} \sin\theta d\theta d\phi, \tag{10}$$

[1] Setting $\sigma\boldsymbol{x} = (\sin\sigma\theta\cos\sigma\phi, \sin\sigma\theta\sin\sigma\phi, \cos\sigma\theta)$, the operation is expressed as

$$R_\sigma f(\boldsymbol{x}) = \frac{1}{2\pi}\int_{\mathrm{SO}(3)} f(\boldsymbol{R}\sigma[\boldsymbol{x}])G(\boldsymbol{R}^{-1}\boldsymbol{y}, \tau)d\boldsymbol{R}, \ |\boldsymbol{y}| = 1,$$

for the spherical Gaussian kernel $G(\boldsymbol{x}, \tau)$.


(a) Mapping from the exterior sphere to the interior sphere

(b) Fine resolution grid

(c) Coarse resolution grid

Fig. 1. Pyramid transformation. (a) The mapping from the exterior sphere to the interior sphere defines the spherical pyramid transformation. (b) The exterior fine resolution grid on the sphere. (c) The interior coarse resolution grid on the sphere. The interior and exterior spheres are expressed using the same radii.

where L_2 norm on the unit sphere is

$$||f(\theta,\phi)||_2^2 = \frac{1}{4\pi^2}\int_{\mathbb{S}^2} |f(\theta,\phi)|^2 \sin\theta d\theta d\phi.$$

4 Discretisation and Algorithm

For optical flow computation, we use the semi-implicite discritisation of the associated diffusion equation of the Eular-Lagrange equation of eq. (10), such that,

$$(\boldsymbol{I} + \frac{\varDelta\tau}{\alpha}\boldsymbol{S}_{\mathbb{S}^2})\boldsymbol{v}^{(n+1)} = (\boldsymbol{I} + \varDelta\tau\nabla_{\mathbb{S}^2}^\top \cdot \nabla_{\mathbb{S}^2})\boldsymbol{v}^{(n)} + \frac{\varDelta\tau}{\alpha} f_t \nabla_{\mathbb{S}^2} f, \qquad (11)$$

for ***upsilon*** $= (\dot\theta, \dot\phi)^\top$, where $\boldsymbol{S}_{\mathbb{S}^2} = \nabla_{\mathbb{S}^2} f \nabla_{\mathbb{S}^2} f^\top$ is the structure tensor of the spherical function with the condition $\nabla_{\mathbb{S}^2}\dot\theta|_{\theta=0,\pi}$.

Furthermore, on $\mathbb{S}^2$, sampling $I(i,j)$ of $f(\phi,\theta)$ is defined as

$$I(i,j) = f(i\triangle_\phi, j\triangle_\theta), \quad 0 \le i \le 2N-1, \quad 0 \le j \le N-1 \qquad (12)$$

where $\triangle_\phi = \triangle_\theta = \pi/N$ for a positive integer N. The downsampling operation of the factor 2 on the unit sphere is

$$I(i,j) = f(i(2\triangle_\phi), j(2\triangle_\theta)), \quad 0 \le i \le N-1, \quad 0 \le j \le [\frac{N}{2}] - 1, \qquad (13)$$

where $\triangle_\phi = \triangle_\theta = \pi/N$. The image pyramid is the sequence of images $I^0, I^2, \ldots, I^n$, where $I^0 = I$. I^i is a reduced image of I^{i-1}.

Algorithm 1. PYRAMID_OPTICALFLOW(I, J, n)

```
Input: $I$: an frame image
Input: $J$: the next frame of $I$
Input: $n$: the number of levels of pyramid
Result: optical flow field
begin
    Compute pyramid images $\{I^i\}$ and $\{J^i\}$ for $i = 0, \dots, (n-1)$ from $I$ and $J$ respectively.
    $\boldsymbol{v}^n \leftarrow \mathbf{0}$
    $i \leftarrow n - 1$
    repeat
        $\boldsymbol{v}^i \leftarrow$ OPTICALFLOW($I^i$, $J^i$, EXPAND($\boldsymbol{v}^{i+1}$))
        $i \leftarrow i - 1$
    until $i \geq 0$
    return $\boldsymbol{v}^0$
```

For a pair of image frames $I := f(\phi, \theta, t)$, $J := f(\phi, \theta, t+1)$, setting $f_t := I - J$, Algorithm 1 is optical flow computation on the unit sphere with pyramid transform.

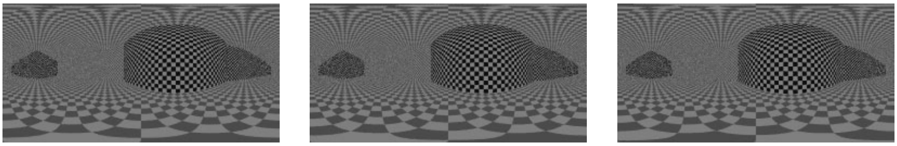
5 Numerical Examples

Figure 2 shows a sequence of input images. This is a sequence of panoramic views of spherical images captured using an omnidirectional camera mounted on a mobile robot moving in a synthetic enviroment. The robot is passing through a corridor.

Figure 3 shows results for the translation motion. The size of the original image is 256×128 pixels in the equirectangular projection map. Therefore, the sizes of the images in the first and second layers are 128 and 64×32 pixels, respectively. The scale parameters for the first and second transforms are $\tau_1 = 0.0001$ and $\tau_2 = 10 \times \tau_1$, respectively. The first and second results are for small and large displacement motions, respectively. The first result is computed from the frames t and $t+1$ and the second result is computed from the frames t and $t+3$. The results show that the pyramid-based optical flow computation can be used to compute both small and large displacement motions. In the experiments, we set the maximum order of the spherical harmonic series $l_{\max} = 127$.

Figure 4 shows the first three frames in panoramic images of a spherical image sequence captured using a moving omni-directionl camera system. The translation is in the $\phi = 0°$ direction with 5cm par frame.

Figure 5 shows the computed optical flow with and without the pyramid method for two frame intervals. The left column shows the optical flow fields between #0 and #1, which yields a small displacement field. The right column shows the optical flow fields between #0 and #1, which are small displacement image sequence, and, between #0 and #2, which yields a large displacement image sequence.



(a) Frame #0 (b) Frame #1 (c) Frame #3

Fig. 2. Input images. We use the three frames for optical flow computation. This is a sequence of panoramic views of spherical images captured by an omnidirectional camera mounted on a mobile robot moving in a synthetic enviroment. The robot is passing through a corridor.

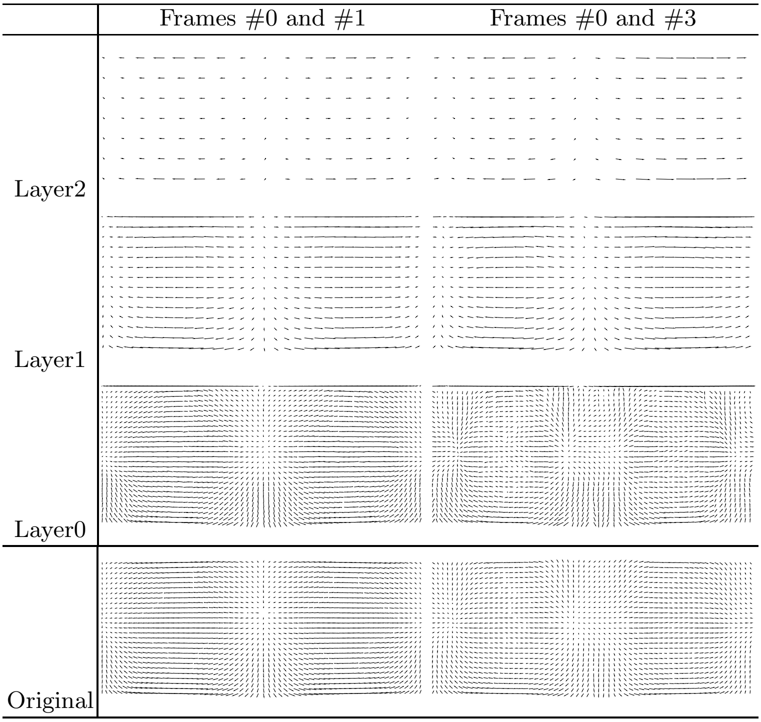
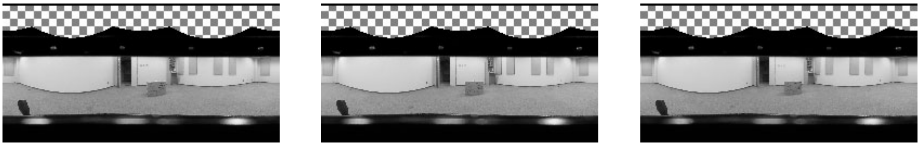


Fig. 3. Optical flow results by pyramid. We used the pyramid transform of the order 2. Scale parameters for the first and second transforms are $\tau_1 = 0.0001$ and $\tau_2 = 10 \times \tau_1$, respectively. The results on the sphere are shown in the equirectangular projection.



(a) frame #0 (b) frame #1 (c) frame #2

Fig. 4. Real image sequence. The size is 256 × 128. This sequence is translational motion for $\phi = 0°$ with 5cm par frame.

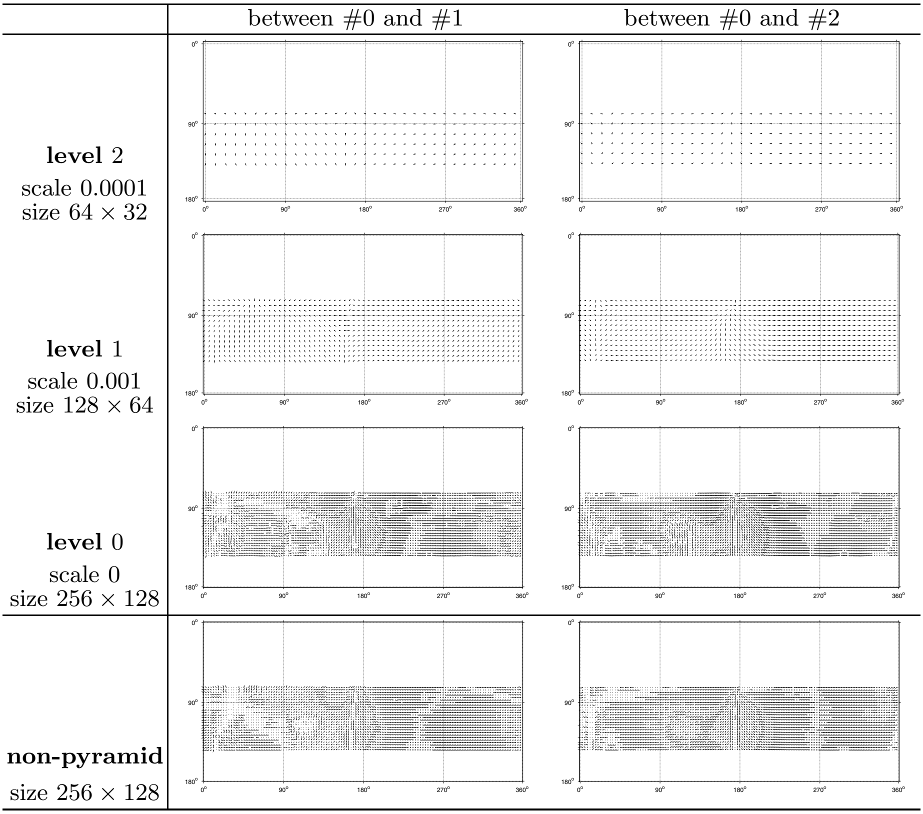


Fig. 5. Spherical optical flow computed using the spherical pyramid transform. The first column shows the level of the pyramid. The first row means the interval between two frames for optical computation. From top to bottom, the results of optical flow are propagated to a smaller level. The last row shows the results of optical flow without using the pyramid method.

6 Conclusions

By introducing the Gaussian pyramid transform on the sphere, we developed an accurate algorithm for the computation of optical flow of omni-directional images. The Gaussian pyramid transform on the plane is achieved by downsampling to the scale space image. We extend this idea to the spherical images, that is, we construct the Gaussian pyramid of a spherical image by using spherical harmonic transform as a Gaussian filter on the sphere. Numerical examples showed the efficiency of the algorithm.

References

1. Imiya, A., Sugaya, H., Torii, A., Mochizuki, Y.: Variational analysis of spherical images. In: Gagalowicz, A., Philips, W. (eds.) CAIP 2005. LNCS, vol. 3691, pp. 104–111. Springer, Heidelberg (2005)
2. Sturm, P., Ramalingam, S.: A generic concept for camera calibration. In: ECCV 2004. LNCS 3021-3024, vol. 2, pp. 1–13. Springer, Heidelberg (May 2004)
3. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artificial Intelligence 17, 185–203 (1981)
4. Nayar, S.K.: Catadioptric omnidirectional cameras. In: CVPR 1997, pp. 482–488 (1997)
5. Gaspar, J., Winters, N., Santos-Victor, J.: Vision-based navigation and environmental representations with an omnidirectional camera. IEEE Transactions on Robotics and Automation 16(6), 890–898 (2000)
6. Franz, M.O., Chahl, J.S., Krapp, H.G.: Insect-inspired estimation of egomotion. Neural Computation 16(11), 2245–2260 (2004)
7. Berger, M.: Geometry I & II. Springer, Heidelberg (1987)

An Improved SalBayes Model with GMM

Hairu Guo[1,2], Xiaojie Wang[1], Yixin Zhong[1], and Song Bi[1]

[1] Center for Intelligence Science and Technology,
Beijing University of Posts and Telecommunications, 100876, Beijing, China
guohr@bupt.edu.cn
[2] College of Computer Science and Technology,
Henan Polytechnic University, 454000, Jiaozuo, China

Abstract. SalBayes is an efficient visual attention model. We describe an improved SalBayes model with Gaussian Mixture Model (GMM) which can fit the object with various transformations better. The improved model learns the probability of an object's visual appearance within a particular feature map, and the Probability Distribution Function (PDF) is modeled using a Mixture Gaussian distribution for each individual feature. The results tested on Amsterdam Library of Object Images (ALOI) shows the better performance than that with the original model.

Keywords: SalBayes, GMM, saliency, visual attention, object recognition.

1 Introduction

Visual attention is regarded as an essential psychological adjustment mechanism of human visual system to realize the selectivity of visual perception[1]. Human vision relies on visual attention mechanism to select the relevant parts of scene rapidly, on which higher level tasks can be processed. Saliency model which simulates the human visual attention mechanism, has a widely application and has achieved good results in the fields of image processing and understanding, such as image retrieval, object recognition, object classification, remote sensing image processing[2-4].

Based on the feature integration theory of Treisman[5] and research work of Koch[6], a biologically inspired bottom-up saliency model was proposed by Itti et al[7]. Through extracting a number of features which consist of intensity, color opponency and orientations from the image and using multi-scale feature fusion mechanism and the prohibition of return to WTA (Winner Take All) method, the visual saliency map which simulates the human visual attention was computed. To provide efficient visual search for the learned objects, a novel model named SalBayes was proposed by Elazary et al[8]. SalBayes model denotes the system's marriage of both saliency and Bayesian modeling. Its core lies in that the model learns the probability of an object's visual appearance, which has a range of values within a particular feature map, and the Probability Distribution Function (PDF) is modeled using a Gaussian distribution for each individual feature. In an object recognition task, the model influences the various feature maps by computing the probability of a

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 356–363, 2011.

given target object for each detector within a feature map. As a result, locations in the maps with the highest probability will be searched first, as they indicate likely positions for the target object[8]. When it is tested in Amsterdam Library of Object Images (ALOI)[9], robust machine vision performance is achieved by this model[8].

It is found that the images in the ALOI dataset are often clustered to three clusters (which vary in viewing angle, illumination angle, and illumination color) [9]. Therefore it is not appropriate to model the feature distribution simply using a single Gaussian model. In this paper, a mixture of Gaussian function was used for a better model of the probability distribution. The various transformations of the objects in the ALOI were modeled respectively, and then a single Gaussian was used to describe a particular part of an object, and the mixture was used to describe all the parts. In testing it in ALOI, we find that this approach delivers better performance as well as high efficiency for object recognition task than SalBayes model.

Section 2, 3 and 4 describe the improved SalBayes model with Gaussian Mixture Model (GMM)[10] and its performance, in which the methodologies, the experiment results and discussions of the model are provided respectively.

2 SalBayes Model with GMM

The model proposed in this paper draws its inspiration from SalBayes model proposed by Elazary et al[8] as well as from Gaussian Mixture Model theory[10]. Based on the fact that the images from the ALOI dataset are systematically varied by different types of properties, the Gaussian mixture model is preferred for its better performance in reflecting the characteristics of the expression of the consistency of object than a single Gaussian model in the SalBayes model.

2.1 The Necessity and Sufficiency of GMM

Sufficiency

While the Gaussian distribution has some important analytical properties, it suffers from significant limitations when it comes to modeling real datasets. Consider the example shown in Fig. 1, it can be seen that the dataset forms two dominant clumps, and that a simple Gaussian distribution is unable to capture this structure, whereas a linear superposition of two Gaussians gives a better characterization of the dataset.

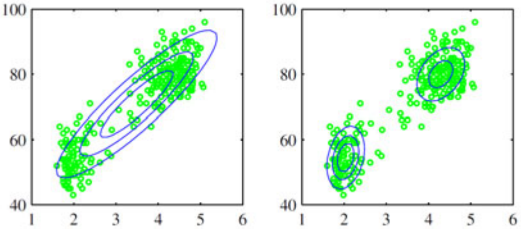


Fig. 1. An example of a Gaussian mixture distribution

Necessity

As is known that the ALOI dataset is divided into three categories (varying in viewing angle, illumination angle, and illumination color). To describe the dataset more suitablely, it is necessary to make use of a mixture of Gaussian function to model the feature distribution for the various variations of the objects in the ALOI.

2.2 Feature Extraction and Representation

To uniquely describe the appearance of an object, a number of feature maps are computed from the biologically inspired bottom-up saliency model proposed by Itti et al[7,8]. The feature map domains consist of intensity, color opponency (red–green, blue–yellow) and four orientations (0°, 45°, 90°, 135°).

Totally, 42 feature maps are computed: 6 for intensity, 12 for color, and 24 for orientation[7,8].

2.3 The SalBayes Model with GMM

Gaussian Mixture Model (GMM)[10] is a statistical model, which describes the feature vector distribution of the probability space using a number of weighted Gaussian probability density functions (PDFs), fitting data better than a single Gaussian.

GMM constitutes a widely used approach for unsupervised learning problems. The process of fitting data in GMM can be interpreted as identifying clusters with the mixture components. The estimation of the parameters of GMM with a predefined number of components is usually achieved through likelihood maximization using the EM algorithm. The drawbacks of EM algorithm, however, are that it is an iterative method and its computation cost is high.

To avoid this shortcoming, this paper takes a similar approach to estimate the parameters of the GMM. First, cluster the data, extract the means, and learn a single Gaussian on the cluster. Then, the multiple clusters would yield to the mixture model.

Training

During training, the object model descriptor is built by computing the likelihood probability distributions of the 42 features resulting from each feature map. This PDF is modeled using a Mixture of three Gaussian distributions for each individual feature type, where the mean and the covariance as well as the proportion of mixture are learned. That is to say, the algorithm learns 42 separate Mixture of three Gaussian distributions for each object. The choice of the Mixture distribution is due to that the images of ALOI are made in three situations, and GMM can fit the data better than a single Gaussian model.

Given a region of interest patch q with N pixels from a particular location (this location will correspond to the image being trained with) from within a given feature map (from the 42 feature maps computed above), the spatial competition method $N(.)$ (the nonlinear normalization method described in [7]) is applied to this patch to form a new set of patch values q'. A feature vector F is then built using the value of q from the location at which q' has a maximum response. This value then forms the

j th component of the feature vector F , and is denoted F_j . In other words we select the center-surround feature that has the highest value in the spatial competition layer (the most unique feature in that map).

$$F_j = q[\arg\max(q'_i)_{i=1,2,...,N}]\forall j \in F \tag{1}$$

Where i represents the pixel position within the patch, F_j is the particular feature value from feature map j and F is the set of feature maps.

Due to that the images of ALOI are systematically varied from three transformations, k-means method is used to cluster F_j into k clusters, which is denoted as F_{jk} where $k \in \{1,2,3\}$.

The Normal distribution is then used to estimate the likelihood, $p(F_{jk} \mid \theta_j)$, of each cluster of observing feature F_{jk} given a particular object class parameter for this feature θ_j .

$$p(F_{jk} \mid \theta_j) \propto N(F_{jk};\mu_{jk};\sigma_{jk}) = \frac{1}{\sigma_{jk}\sqrt{2\pi}} e^{-(F_{jk}-\mu_{jk})^2/2\sigma_{jk}^2} \tag{2}$$

The final model (θ) is then a set of n parameters (θ_j), with each composed of three mean (μ_{jk}) , three variance (σ_{jk}^2) and three proportion of mixture (π_{jk}) for each individual feature map. This enables the model to simply compute the model parameters (θ) mean (μ_{jk}), variance (σ_{jk}^2) and proportion of mixture (π_{jk}) from the training views of the object within each feature map, and to use a mixture Gaussian distribution to estimate the likelihood.

$$p(F_j \mid \theta_j) = \sum_{k=1}^{3} \pi_{jk} p(F_{jk} \mid \theta_j) \tag{3}$$

Testing

To classify particular features obtained from the feature maps, a naive Bayesian network is used.

Once a set of features (F) is collected from a given salient location within the feature maps (as described above), the classification is performed using Bayes formula:

$$p(\theta_j \mid F) = \frac{p(F \mid \theta_j)p(\theta_j)}{p(F)} \tag{4}$$

To make a decision as to the type of classification assigned to an object, i can be iterated over all known objects and the object with the greatest posterior can be chosen as the best match. This method is known as Maximum a Posteriori (MAP).

In our experiments, the prior is taken to be $1/C$, where C equals 1000 which is the number of classes.

Since the probability of the evidence can be viewed as a normalizing constant (used to ensure that probabilities all add up to unity), it can be dropped from the equation.

$$p(\theta_i \mid F) = \begin{cases} 1 & \arg\max_i \left(p(\theta) \prod_{j=1}^{n} p(F_j \mid \theta_{ij}) \right) \\ 0 & others \end{cases} \tag{5}$$

Additionally, taking the product of many probabilities, some of which may be very small, can give rise to numerical instability. As a result, an underflow often occurs in a straightforward implementation of (4) when using more than a few features. A solution to this problem is to take the log of the likelihood which will convert the probabilities from being less than one to negative numbers greater than one. This also greatly simplifies the computations in our practical implementation, as likelihood products are transformed into likelihood summations. Also, the decision to select a suitable classification is not affected, since only the maximum of these values is considered. As a result of these various techniques, (4) can be described by the following formula:

$$p(\theta_i \mid F) = \begin{cases} 1 & \arg\max_i \left(p(\theta) \sum_{j=1}^{n} \log(p(F_j \mid \theta_{ij})) \right) \\ 0 & others \end{cases} \tag{6}$$

The enhanced version of the SalBayes with the GMM used for object recognition can be seen visually in Fig. 2.

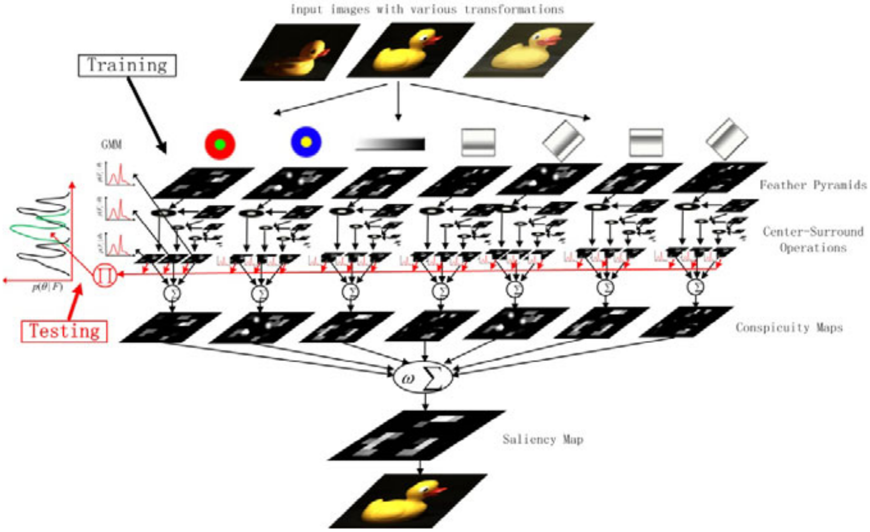


Fig. 2. SalBayes Model enhanced with GMM

3 Experiments and Results

Object recognition testing was performed on the large standard database ALOI to test the proposed algorithm. The ALOI dataset contains photographs of 1000 objects placed on a turntable and subjected to various transformations. These transformations include 12 illumination colors, 24 illumination directions, and 72 viewpoints (each object was rotated in steps of 5°). All photographs were first scaled down to a 256×256 pixel image. The dataset are systematically broken into training and testing sets composed of the various images in the dataset. These sets include 6.25% for training and 93.75% for testing, 12.5% training 87.5% testing, 25% training 75% testing, half training half testing and all training all testing.

The main source code of the implementation of SalBayes in Matlab was obtained from the website (http://www.saliencytoolbox.net/index.html)[11]. However, the software was modified due to the use of GMM.

3.1 Experiment 1

Because the ALOI dataset contains the most systematic transformations, a test under each type of transformation was done to analyze the impact of transformations on the original SalBayes algorithm.

Table 1. Recognition rate under various transformations

Images Number		Recognition Rate (%)	
(Train/Test)	Illumination color	Illumination direction	Rotation
100%/100%	100%	81.7%	68.3%
50%/50%	99.6%	74.4%	66.9%
25%/75%	97.2%	60.0%	65.4%
12.5%/87.5%	-	21.8%	60.1%
6.25%/93.75%	-	-	37.1%

Look at Table 1, it is seen that SalBayes does exceptionally well on the illumination color task, and worse on the rotation task. Additionally, the recognition rate of the model can achieve 74.2% accuracy rate on the average at 25% for training and 75% for testing. This shows the model's robustness against illumination color and high-performance on few training samples.

3.2 Experiment 2

Comparative experiment was made with the original SalBayes model and improved SalBayes model with GMM.

The results in the Table 2 show that by using the GMM, the SalBayes model is able to learn the objects more successfully and recognize them more correctly.

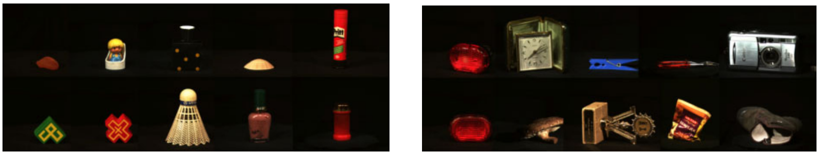
The 60% recognition rate on average of the SalBayes algorithm is due to the random selection of the training images which may not be representative for the object to be identified. Some methods can be used to improve the recognition rate, for example, choosing the images every n degree of rotation from 20 degree to 60 degree on the basis of the specific situation, seems a feasible method. Because through this method, the characteristics of a variety of directions will be fitted in the model, then the model would fit the object better.

Table 2. Recognition rate under the various model

Images Number	Recognition Rate (%)	
(Train/Test)	SalBayes	SalBayes with GMM
100%/100%	62.8%	70.2%
50%/50%	61.8%	68.4%
25%/75%	60.3%	65.5%
12.5%/87.5%	59.6%	62.9%
6.25%/93.75%	52.0%	60.7%

3.3 Experiment 3

The top 10 and the bottom 10 objects on recognition rate in Experiment 2 are shown in Fig. 3.



(a) Top 10 objects on recognition rate

(b) Bottom 10 objects on recognition rate

Fig. 3. Top 10 and Bottom 10 objects on recognition rate

As seen, the Top 10 objects on recognition rate is generally symmetric in the visual sense, while the Bottom 10 objects on recognition rate is random. This shows that the conclusion in Experiment 1, that the rotation of the object to be recognized is very essential with the SalBayes-based model, is declared again. However, the recognition rate is not satisfactory. One suggested way to enhance it would be to compute the two of the most saliency location, and then learn two representation of the object to be recognized. The double representations would then yield to a mixture model which would fit the object better for using more information of the object itself.

4 Conclusions and Discussions

In this paper, the SalBayes model is developed with GMM for recognition task on ALOI, and performs informed recognition better than the previous related efforts under the same software and hardware condition. As shown in the results, the rotation view of the object plays more key role than illumination direction and illumination color for recognition task on ALOI with SalBayes-based models. On the other hand, the GMM used in this paper was more able to fit the object's probability distribution than a single Gaussian model used in the original SalBayes algorithm.

Note that the recognition rate is not satisfactory. We can improve the SalBayes algorithm by increasing the number of the saliency location and using more suitable probability distribution of the object to be recognized.

Additionally, it is important to note that top-down information is as important as bottom-up feature extraction for recognition tasks. In this paper, we concentrated on the bottom-up feature extraction. However, a hybrid SalBayes-based model with top-down information could be used so that the object can be recognized more efficiently. In particular, it could help achieve greater performance under certain conditions by using some of the knowledge of the object.

Acknowledgment. This research was supported by the Major Research Plan of the National Nature Science Foundation of China (90920006).

References

[1] Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y.H., Davis, N., Nuflo, F.: Modelling Visual Attention via Selective Tuning. Artificial Intelligence 78(1-2), 507–545 (1995)

[2] Marques, L.M., Mayron, G.B., Borba, H.R.: Using Visual Attention to Extract Regions of Interest in the Context of Image Retrieval. In: Proceedings of the 44th Annual Southeast Regional Conference (ACM-SE 44), pp. 638–643 (2006)

[3] Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C.: Attentional Selection for Object Recognition - A Gentle Way. In: Bülthoff, H.H., Lee, S.-W., Poggio, T.A., Wallraven, C. (eds.) BMCV 2002. LNCS, vol. 2525, pp. 472–479. Springer, Heidelberg (2002)

[4] Peng, Z., Run-sheng, W.: An Approach to the Remote Sensing Image Analysis Based on Visual Attention. Journal of Electronics and Information Technology, 1855–1860 (2005) (in Chinese)

[5] Treisman, A.M., Gelade, G.: A Feature-Integration Theory of Attention. Cognitive Psychology 12(1), 97–136 (1980)

[6] Koch, C., Ullman, S.: Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. Human Neurobiology 4, 219–227 (1985)

[7] Itti, L.: Models of Bottom-Up and Top-Down Visual Attention. California Institute of Technology (2000)

[8] Elazary, L., Itti, L.: A Bayesian Model for Efficient Visual Search and Recognition. Vision Research 50(14), 1338–1352 (2010)

[9] Geusebroek, J.M., Burghouts, G.J., Smeulders, A.W.M.: The Amsterdam Library of Object Images. Int. J. Comput. Vision 61(1), 103–112 (2005)

[10] Bishop, C.M.: Pattern Recognition and Machine Learning. Ch. 9, pp. 423–455. Springer, Heidelberg (2006)

[11] Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks 19, 1395–1407 (2006)

Exploring Alternative Spatial and Temporal Dense Representations for Action Recognition

Pau Agustí[1], V. Javier Traver[1],
Manuel J. Marin-Jimenez[2], and Filiberto Pla[1]

[1] iNIT and DLSI, Universidad Jaume I, Castellón Spain
{pagusti,vtraver,pla}@uji.es
[2] Universidad de Córdoba Spain
mjmarin@uco.es

Abstract. The automatic analysis of video sequences with individuals performing some actions is currently receiving much attention in the computer vision community. Among the different visual features chosen to tackle the problem of action recognition, local histogram within a region of interest is proven to be very effective. However, we study for the first time whether spatiograms, which are histograms enriched with per-bin spatial information, can be alternatively effective for action characterization. On the other hand, the temporal information of these histograms is usually collapsed by simple averaging of the histograms, which basically ignores the dynamics of the action. In contrast, this paper explores a temporally holistic representation in the form of recurrence matrices which capture pair-wise spatiograms relationships on a frame-by-frame basis. Experimental results show that recurrence matrices are powerful for action classification, whereas spatiograms, in its current usage, are not.

Keywords: Human action recognition, recurrence matrices, spatiograms, spatio-temporal representations.

1 Introduction

At present, the analysis of video sequences in which individuals are performing an action has become one of the most emerging investigation fields in computer vision [15]. That new challenge is due to its high impact on large technical and social applications: surveillance, human-machine interfaces, automatic diagnostics of orthopedic patients, analysis and optimization of athletes' performances, etc. This new area, where the aim is to recognize individuals performing some action, is commonly called human action recognition. In particular, we are interested in full-body human action recognition.

In this paper two image representations will be distinguished: a dense one and a no-dense one. Dense representations could be considered as a design where all the points in the image are considered for the computation. Following this kind of representation several techniques as dense optical flow, space-time volumes, etc. are used. On the other hand, no-dense representations compute features

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 364–371, 2011.

only from some points. In this group the used techniques are bag-of-features [13], histograms of oriented rectangles [8], shape-context [14,7], etc.

In this field, the spatial information is commonly represented by means of a grid and the computation of histograms for each region of the grid [1]. A related work proposed in [4] suggests the use of spatiograms. The spatiograms are histograms but with a normal distribution of the contributing pixels coordinates in the image for each bin. This technique has been used in tracking [4] but it has not been applied to human action recognition. In this work, an evaluation of the potential improvement on the degree of discrimination of the features will be done. On the other hand, the temporal representation is usually represented by accumulating features along the time [1], which basically ignores the dynamics of the action. In this work, the recurrence matrices to represent the temporal information will be used in order to overcome this lack of temporal information. This technique allows to represent the distance of a frame against all the rest in a matrix.

Hence, the main contributions of this work are the study of the application of alternative spatial and temporal dense representation for human action recognition and compare them with another no-dense representation and with some novel techniques from the state-of-art.

The rest of this paper is organized as follows: In Section 2 the spatiograms and the recurrence matrices are explained. Section 3 describes the method for obtaining the feature vector. Section 4 shows the experimental setup and the obtained results. In Section 5, the discussion of the results and overall conclusions are presented.

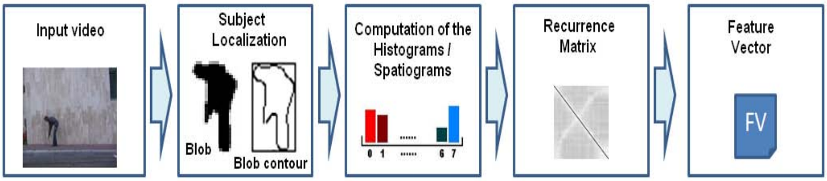


Fig. 1. Overview of the proposed method

2 Background

In this section the theory of the novel techniques used in this paper is explained. First an explanation of the spatiograms is given and how they are compared. Finally, the recurrence matrix is introduced.

2.1 Spatiograms

The concept of *spatiogram* is introduced in [4] as an extension of the well-known histograms. Spatiogram model is like the histogram concept but adding the spatial information in a normal distribution. In particular, it stores the mean and the covariance matrix of the position of all pixels that contribute each histogram bin.

In order to compute the mean and the covariance matrix on-line, i.e. during the process instead of at the end of it, Equation 1 and 2 are used:

$$\mu_n(x,y) = [\mu^x_{n-1} + \frac{x-\mu^x_{n-1}}{n}, \mu^y_{n-1} + \frac{y-\mu^y_{n-1}}{n}] \tag{1}$$

$$\Sigma_n(x,y) = \begin{pmatrix} \frac{n-1}{n}\sigma^{xx}_{n-1} + \frac{(x-\mu^x_n)^2}{n} & \frac{n-1}{n}\sigma^{xy}_{n-1} + \frac{(x-\mu^x_n)(y-\mu^y_n)}{n} \\ \frac{n-1}{n}\sigma^{yx}_{n-1} + \frac{(y-\mu^y_n)(x-\mu^x_n)}{n} & \frac{n-1}{n}\sigma^{yy}_{n-1} + \frac{(y-\mu^y_n)^2}{n} \end{pmatrix} \tag{2}$$

where n ($n \geq 1$) is the time step of the on-line calculation, (x, y) are the Cartesian coordinates, μ_n is the mean, Σ is the covariance matrix and σ_n is the variance. At the beginning μ_n and σ_n are 0.

Having two spatiograms S and S' the difference used it is taken from [4] but as the bins are normalized it is not need to use the complete Bhattacharyya coefficient. The difference between them is defined as Equation 3:

$$d(S,S') = \sum_{b=0}^{B-1} \sqrt{N(b)N'(b)} e^{-\frac{1}{2}(\mu_b-\mu'_b)^T(\Sigma_b^{-1}+\Sigma_b^{-1\prime})(\mu_b-\mu'_b)} \tag{3}$$

where N are the contributions for the bin b and d is the distance.

Note that a histogram is a spatiogram but without spatial information, so the difference used for histograms is the same but without the spatial information distance: $d(H, H') = \sum_{b=0}^{B-1} \sqrt{N(b)N'(b)}$, where H and H' are the histograms.

2.2 Recurrence Matrix

The Recurrence Matrix (RM) [12] has several names in the bibliography because it is just a matrix containing information about the recurrence of some features along the time. We can find similarities between this kind of matrices and either the concurrence matrices or the self-similarities matrices. This idea has been widely used in research fields as texture recognition, music structure analysis and bioinformatics. The use of this technique is not so common in human action recognition, although it is used in [10]. In our case, each value of the RM is the distance between the spatiograms/histograms of one frame against all the others. It is a symmetric matrix and its size is F^2 with F being the number of frames.

As we use a grid-based method, we have more than one spatiogram/histogram in each frame. Considering M the number of spatiograms/histograms and F_1 and F_2 the frames, the final distance (d') stored into the matrix is calculated following Equation 4:

$$d'(F_1, F_2) = \frac{1}{M} \sum_{i=0}^{M-1} d(S_{i1}, S_{i2}) \tag{4}$$

where S_{i1} is the Spatiogram/Histogram of the frame 1 from the region i and S_{i2} the same of the frame 2.

3 Method

The recognition method is performed according to the following steps (see Figure 1):

1. *Subject localization*: the frame is segmented and the bounding box (BB) of the region of interest (ROI) is located. The BB is obtained by using a component connected algorithm. This ROI, which contains the person performing the action, is cropped from the segmented frame and resized to have the same BB size for all the frames.
2. *Blob contour*: once the BB containing the silhouette blob is resized, its contour is obtained by using Canny edge detection [5].
3. *Computation of the histograms/spatiograms*: this step has different methods. For each frame M histograms/spatiograms are built depending on the choice:
   - Density mask with histograms (see Subsection 3.1): two different grid geometries are used, the rectangular one and the polar one. For the rectangular grid, the M histograms generated correspond to the $A \times B = M$, where A are the column divisions and B the row divisions. For the polar grid, the M histograms generated correspond to the $R \times O = M$, where R are the radial divisions and O the angular divisions.
   - Density mask with spatiograms (see Subsection 3.1): the geometries used are the same above.
   - Shape context (see Subsection 3.2): here the M histograms generated correspond to the number of points where the shape context histogram is computed.
4. *Computation of the RM*: at the end of the process M histograms/spatiograms are available for each frame and the distance between the M histograms/spatiograms of each frame against the others is computed to produce the RM.
5. *Feature vector creation*: once the recurrence matrix has been completed, it is resized (same size defined for the cropped ROI) and the feature vector is created by following two different ways:
   - Raw data: original data is taken from the RM without any process, and only the superior triangle data is concatenated and used, since it is a symmetric matrix.
   - Analyzed data: eight descriptors extracted from the RM are used: *Density of two sections (DTS), Center of mass (CM), Density of quadrants (DQ)* from [17]; *Recurrence rate (RR), Determinism (DET), The averaged diagonal line length (LEN), The average length of the vertical lines (TT, Trapping Time)* from the recurrence quantitative analysis [12]; and, the *Velocity of the bounding box.*

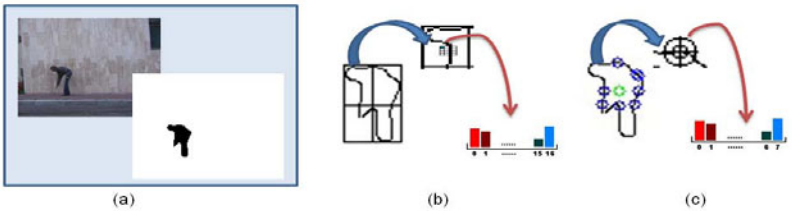


Fig. 2. (a) Frame and silhouette extraction for the example (b) Histogram/Spatiogram updating by using Density mask (c) Histogram updating by using Shape-Context

3.1 Density Mask (DM)

Once the person is located the blob contour is obtained. For each pixel inside the BB is calculated how many pixels belong to the contour between the neighbor pixels inside a windows (Figure 2(b)). This calculated number will be the bin to contribute into the histogram/spatiogram. The windows size will define the number of bins, i.e. for a windows size of $S \times S$ the number of bins will be $S \times S + 1$. The +1 is because it is possible in a windows neighborhood no pixel belongs to the contour.

3.2 Shape-Context (SC)

The method is adapted from [3] (Figure 2(c)). M points of the silhouette are chosen. From the center of mass of the person is traced rays each $\frac{360}{M}^{\circ}$ and the point of the silhouette crossed by the ray is one of the points chosen. In each of these M points it is calculated a histogram of $\rho \times \theta$ bins, where ρ is the number of radial divisions of the polar grid centered at the point and θ the number of angular divisions. Each part of this polar grid corresponds to a bin. The points that belong to the silhouette situated into this polar grid contribute to the bin numbered by the region where it is situated.

4 Experiments

For the experiments, the Weizmann action dataset [6] is used. Although this is certainly a simple dataset for the current state-of-the art standards, it is a good one to start with the exploration of new action descriptors or machine learning algorithms. Additionally, unlike other datasets, this one includes the segmented silhouette, which facilitates both the experimentation and the results reported by other authors. This dataset contains 10 actions (*running, walking, skipping, jumping-jacks, jumping forward on two legs, jumping in place on two legs, jumping sideways, waving with two hands, waving with one hand and bending*) performed by 9 different persons. A total of 90 video sequences are used. The size of the resized images is 60 × 60 but it is completely scalable. A leave-one-out-cross validation scheme is used, therefore 9 runs for each experiment are

Table 1. Classification results in % for the different configurations using the DM and extracting features from the RM raw data

	Rectangular grid					Polar grid		
	1 × 1	2 × 2	2 × 4	4 × 2	3 × 3	2 × 2	2 × 4	3 × 3
Histogram	96.6	96.6	**97.7**	95.5	96.6	95.5	96.6	95.5
Spatiogram	95.5	93.3	95.5	94.4	92.2	94.4	92.2	93.3

needed, reporting the average accuracy. A linear Support Vector Machine is used for training and test.

Table 1 shows the results of the experiments for the density mask either histograms and spatiograms, generating the feature vector from the raw data of the RMs.

The same configurations as in Table 1 are used for the next experiment. However, this time the feature vector is obtained from the analyzed data of the RMs. The results are exactly the same for all the configurations being 92.2%, due to the fact that the used descriptors are not discriminative enough.

After these experiments it is possible to answer the question: *Are the spatiograms contributing more information than histograms?* Results reveal that spatiograms do not bring an actual advantage over histograms. Despite of the good results obtained, spatiograms do not outperform histograms in any experiment, in its current usage.

Table 2 shows the results of the experiments for the SC. The same results are obtained independently from the way to obtain the feature vector (raw data and analyzed data of the RMs).

After concluding the previous experiments, we could try to answer the following questions: *Can the RMs capture properly the spatio-temporal information?* And, *is it necessary to have a complex method in order to improve the recognition rates?* The results show in Tables 1 and 2, using simple methods like Density mask and a simple Shape-Context let us think that a complex methodology is no needed. The RMs are quite effective in capturing the nature of an action and sufficient to distinguish between them by caching pair-wise spatiograms/histograms relationships on a frame-by-frame basis.

In Table 3, some of the latest (from 2007 until 2010) works using the Weizmann dataset are reported. They also use a leave-one-out-cross validation scheme, but in some of these experiments the *skip* action has not been used (w/o skip in

Table 2. Classification results in % for the different configurations using the Shape-Context + RM (where $\rho \times \theta$ (size of the radius))

Number of points	2 × 4(4)	2 × 8(4)	2 × 4(8)	2 × 8(8)
8	70.0	70.0	70.0	77.7
16	80.0	82.2	86.6	89.9
32	88.8	87.7	88.8	87.7
64	90.0	90.0	**93.3**	91.1

Table 3. Performance achieved by the state-of-the-art approaches

Method	w/o skip	w/ skip
Density mask + RM	*	97.7
Shape-Context + RM	*	93.3
3D SIFT [16]	*	92.6
Volumetric shape-action features [6]	100	100
Shape-Context + Gradients [14]	72.8	*
Hierarchy of spatio-temporal features [9]	97.0	*
Bag-of-features [13]	*	90
3D Shape-Context [7]	96.4	94.6
Histogram of oriented gradients [11]	*	84.3
Histogram of oriented rectangles [8]	*	100
Key poses [2]	*	92.6

the table). The result obtained by using the density mask reveals that a simple dense method outperforms works like [16,9,13,11], or [2] which is supposed to be a very simple method, only with the help of the RMs. However, other works like [6,8] are only slightly better than our approach. Our method relies on the simplicity of a density mask and the only drawbacks could be in the memory because it is need to store $M \times F$ histograms/spatiograms (where M is the number of histograms/spatiograms per frame and F the number of frames). This problem can be addressed by having spatiograms describing a few consecutive frames, which not only will save memory, but also time to compute the recurrence matrix. Furthermore, these spatiograms with a wider temporal support can be more robust and descriptive than those based on single frames. On the other hand, our no-dense method (the shape-context+RM) outperforms [14] and is almost the same as the 3D shape-context [7].

5 Conclusion

A simple method for action recognition, based on straightforward information derived from the silhouette, has been proposed in this paper. When tested on the Weizmann dataset, it has been shown to outperform a number of previous approaches, and to be slightly worse than the best reported recognition rate. Two novel dense representations have been explored. Regarding the spatial representation of actions, local spatiograms have been proposed as an alternative for local histograms. As for the temporal representation, recurrence matrices have been explored as a temporally more holistic way of capturing the action dynamics. Results reveal that spatiograms do not bring an actual advantage over histograms, but recurrence matrices seem to encode relevant temporal information. Further work is directed to improve these spatio-temporal representations, to test them on more datasets and to aim at on-line action recognition.

Acknowledgements. The authors acknowledge the Spanish research programme Consolider Ingenio-2010 CSD2007-00018, UJI-Fundació Caixa-Castelló under projects P1-1A2010-11 and P1-1B2010-27, and PROMETEO/2010/028 from Generalitat Valenciana.

References

1. Agustí, P., Traver, V.J., Montoliu, R., Pla, F.: An evaluation of the grid geometry for human action recognition. In: II Workshop de Reconocimiento de Formas y Análisis de Imágenes, pp. 9–16 (September 2010)
2. Baysal, S., Kurt, M.C., Duygulu, P.: Recognizing human actions using key poses. In: International Conference on, Pattern Recognition, pp. 1727–1730 (2010)
3. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. 24, 509–522 (2002)
4. Birchfield, S.T., Rangarajan, S.: Spatiograms versus histograms for region-based tracking. In: Proceedings of Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1158–1163 (2005)
5. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. 8(6), 679–698 (1986)
6. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Transactions on Pattern Analysis and Machine Intelligence 29(12), 2247–2253 (2007)
7. Grundmann, M., Meier, F., Essa, I.: 3D shape context and distance transform for action recognition. In: International Conference on Pattern Recognition, pp. 1–4 (2008)
8. Ikizler, N., Duygulu, P.: Histogram of oriented rectangles: A new pose descriptor for human action recognition. Image and Vision Computing 27(10), 1515–1526 (2009)
9. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: International Conference on Computer Vision, pp. 1–8 (2007)
10. Junejo, I., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. IEEE Trans. on Pattern Analysis and Machine Intelligence (2009)
11. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: British Machine Vision Conference (2008)
12. Marwan, N., Carmenromano, M., Thiel, M., Kurths, J.: Recurrence plots for the analysis of complex systems. Physics Reports 438(5-6), 237–329 (2007)
13. Niebles, J., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision 79, 299–318 (2008)
14. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
15. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing 28(6), 976–990 (2010)
16. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional SIFT descriptor and its application to action recognition. In: Proceedings of the 15th International Conference on Multimedia, pp. 357–360 (2007)
17. Serra-Toro, C., Montoliu, R., Traver, V.J., Hurtado-Melgar, I.M., Núnez-Redó, M., Cascales, P.: Assessing water quality by video monitoring fish swimming behavior. In: Proceedings of the International Conference on Pattern Recognition, pp. 428–431 (2010)

Image Denoising Using Bilateral Filter in High Dimensional PCA-Space

Quoc Bao Do, Azeddine Beghdadi, and Marie Luong

L2TI,University of Paris 13
{do.quocbao,azeddine.beghdadi,marie.luong}@univ-paris13.fr

Abstract. This paper proposes a new noise filtering method inspired by Bilateral filter (BF), non-local means (NLM) filter and principal component analysis (PCA). The main idea here is to perform the BF in a multidimensional PCA-space using an anisotropic kernel. The filtered multidimensional signal is then transformed back onto the image spatial domain to yield the desired enhanced image. The proposed method is compared to the state-of-art. The obtained results are highly promising.

Keywords: Denoising, Bilateral filter, Non-local means, High dimensional space, PCA.

1 Introduction

Image denoising is an important problem in image and signal processing. Many methods share the same basic idea: denoising each pixel is carried out by averaging other ones which are similar to it. These methods are based on the observation that any image often contains self-similarity and some spatial redundancy. If the noise is considered as an independent and identically distributed (i.i.d.) random signal, it could be smoothed out by averaging similar pixels. In [13], the authors propose Bilateral filter (BF) which takes into account both spatial and intensity information to define similar pixels for a given one. The relation between BF and anisotropic filtering has been investigated in [1,6]. Another adaptive filtering approach, called non-local means (NLM) [3,4], has been recently proposed. Instead of using pixel-based similarity as in BF, NLM proposes to use patch-based similarity which makes the method more robust in textured and contrasted regions. Many methods for improving the performance of NLM have been proposed. The fast NLM (FNLM) is presented in [12]. NLM in the wavelet domain is introduced in [11]. In [14], the authors propose a transform which maps each patch in the image domain to a point in a high dimensional space called patch-space and show that NLM algorithm is a variant of an isotropic filter in this new space. In this paper, we propose to use principal component analysis (PCA) to reduce the dimensionality of the patch-space and then form another one called high dimensional PCA-space (HDPCA) from the most significant components. Similar to the work in [14], FNLM can be drawn as a variant of an isotropic filter in the HDPCA-space. In order to improve the denoising

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 372–379, 2011.

performance, instead of using this isotropic filter, we propose to use BF, i.e. an anisotropic filter, in the HDPCA-space.

The paper is organized as follows: section 2 is devoted for a short review of NLM and related works, the HDPCA-space is presented in section 3. The proposed method is described in section 4 followed by experimental results in section 5. The conclusions are finally given in section 6.

2 Related Works

Let us define a 2D noise-free image $u : R^2 \rightarrow R$. Its noisy version v at pixel (k, l) defined as $v(k, l) = u(k, l) + n(k, l)$ where n is identical, independent Gaussian noise. The aim of denoising is to estimate u from v. Both BF and NLM restore noisy pixel by averaging the neighboring pixels. An unifying formula for these methods could be expressed as follows:

$$\widehat{u}(k,l) = \frac{\sum_{(i,j)\in\Omega} w(k,l,i,j)v(i,j)}{\sum_{(i,j)\in\Omega} w(k,l,i,j)} \tag{1}$$

where Ω is the image domain and $w(k, l, i, j)$ stands for the weight which corresponds to the similarity between pixel $v(k, l)$ and $v(i, j)$. Indeed, each method proposes a kernel to estimate this weight. According to BF [13], the kernel is defined as follows:

$$w_{BF}(k,l,i,j) = exp\left(\frac{-\left(\|k-i\|^2 + \|l-j\|^2\right)}{h_s^2}\right) exp\left(\frac{-\|v(k,l)-v(i,j)\|^2}{h_r^2}\right) \tag{2}$$

where h_s and h_r are space and range parameters, respectively. Note that this filter takes into account both spatial and intensity information. In [1,6], it has been proven that BF is an anisotropic filter with a special regularization term. In the case of NLM, its weight is given by:

$$w_{NLM}(k,l,i,j) = exp\left(\frac{-G_a * \|\boldsymbol{N}(k,l) - \boldsymbol{N}(i,j)\|^2}{h_r^2}\right) \tag{3}$$

where $\boldsymbol{N}(k, l), \boldsymbol{N}(i, j)$ are two small patches of size $r \times r$ around the pixel (k, l) and (i, j), respectively (r is usually equal to 7), G_a is a Gaussian kernel with standard deviation a of the same size as $\boldsymbol{N}(k, l), \boldsymbol{N}(i, j)$. Note that, the patch-similarity measure is weighted G_a to give more weight to pixels close to the patch center. The equation(3) can be rewritten as follows:

$$w_{NLM}(k,l,i,j) = exp\left(\frac{-\|\boldsymbol{N}_a(k,l) - \boldsymbol{N}_a(i,j)\|^2}{h_r^2}\right) \tag{4}$$

where the weighted patch $\boldsymbol{N}_a(k, l) = \sqrt{G_a}\boldsymbol{N}(k, l)$. Note that NLM considers only intensity information.

Due to computational burden, to denoise pixel (k, l), instead of considering all pixel (i, j) in the noisy image, Buades et al. propose to restraint a search window $\Omega(k,l)$ which is usually set equal to 21x21, i.e. there are 441 patch candidates. Indeed, further analyses [7] have shown that if the size of $\Omega(k,l)$ increase, more bias is introduced due to several mismatching patches which are taken into account. In [5], we used a strategy where only the best candidates are selected by exploring the entire image plane. Counter intuitively, this approach yields worse result in both term of subjective and objective measurement. In flat regions, the noise pattern of a given-patch will match well with that of the best candidates. Averaging these similar noise patterns cannot effectively remove the noise. We refer to this as "*best-worse paradox*" in the sense that if we consider only the *best* candidates, the result is *worse*. Based on these remarks, the semi-non local approach, i.e. restrained to a small window $\Omega(k,l)$, is used in this work.

In [12], fast NLM (FNLM) filter is proposed by approximating the distance $\|\boldsymbol{N}_a(k,l) - \boldsymbol{N}_a(i,j)\|^2$ in (4) by another one estimated from projections of $\boldsymbol{N}_a$ onto a subspace defined by PCA. It is well known that the eigenvectors $\{\boldsymbol{e}_m\}_{m=1}^{r^2}$ (sorted in order of descending eigenvalues) of the covariance matrix $\mathbf{M}$ estimated from a set of all weighted patch $\boldsymbol{N}_a$ form an orthonormal basis. Note $\boldsymbol{F}(k,l) = [f_1(k,l), f_2(k,l), ..., f_{r^2}(k,l)]^T$ is projected vector of $\boldsymbol{N}_a(k,l)$ onto this orthonormal basis, i.e. $f_m(k,l) =< \boldsymbol{N}_a(k,l), \boldsymbol{e}_m >$ where $<,>$ stands for inner product. As the signal energy concentrates on a few the most significant d components ($d \ll r^2$), Tasdizen[12] proposes to approximate the norm $\|\boldsymbol{N}_a(k,l) - \boldsymbol{N}_a(i,j)\|^2$ by using only these d components, i.e.

$$\|\boldsymbol{N}_a(k,l) - \boldsymbol{N}_a(i,j)\|^2 \approx \|\boldsymbol{F}^d(k,l) - \boldsymbol{F}^d(i,j)\|^2 = \sum_{m=1}^{d} \|f_m(k,l) - f_m(i,j)\|^2 \quad (5)$$

where $\boldsymbol{F}^d(k,l) = [f_1(k,l), f_2(k,l), ..., f_d(k,l)]^T$. The new weight is now defined as follows:

$$w^d_{NLM}(k,l,i,j) = exp\left(\frac{-\|\boldsymbol{F}^d(k,l) - \boldsymbol{F}^d(i,j)\|^2}{h_r^2}\right) \quad (6)$$

Finally, the FNLM is given by:

$$\widehat{u}^d(k,l) = \frac{\sum_{(i,j)\in\Omega} w^d_{NLM}(k,l,i,j)v(i,j)}{\sum_{(i,j)\in\Omega} w^d_{NLM}(k,l,i,j)} \quad (7)$$

where d is a parameter of the algorithm. Recall that when $d = r^2$, FNLM tends to the classical NLM. Indeed, the use of PCA has twofold: (i) the computational complexity is highly reduced, (ii) patch similarity measure improves robustness to noise.

In the next sections, we present a new high dimensional space called HDPCA in which spatial coordinates of each point corresponds to values of projected vector $\boldsymbol{F}^d$. We will show that FNLM and NLM are simply derived from an isotropic filter in this space.

3 High Dimensional PCA-Space

Mapping in the HDPCA-space: First, each small patch is passed through the PCA system to obtain the corresponding projected vector $\boldsymbol{F}^d$. We define d dimensional HDPCA-space noted $\Psi_d \in R^d$ where the coordinates $\mathbf{p}$ of each point in this space are the values of $\boldsymbol{F}^d$. In other words, the intensity values of $\boldsymbol{F}^d$ become spatial coordinates in the new HDPCA-space. Each value $\mathbf{V}$ of a point $\mathbf{p}$ in this space is defined as follows:

$$\mathbf{V}(\mathbf{p}) = (V_1(\mathbf{p}), V_2(\mathbf{p})) = (v(k,l), 1)\ if\ \mathbf{p} = \mathbf{F}^d$$

Note that $\mathbf{V}(\mathbf{p})$ contains two components:

- The first one $V_1(\mathbf{p}) = v(k,l)$ (the gray level of the center pixel (k,l) of the patch $\boldsymbol{N}(k,l)$).
- The second one $V_2(\mathbf{p})$ is always set equal to 1.

Back-projection on the image domain: Instead of filtering directly the pixel value in the image domain Ω, we alter the multi-values $\mathbf{V}(\mathbf{p})$ in the HDPCA-space to obtain $\widehat{\mathbf{U}}(\mathbf{p})$ (the filtering method in this space will be discussed in the next section). This filtered value is then transformed back onto the image domain Ω as follows:

$$\widehat{u}(k,l) = \frac{\widehat{U_1}(\mathbf{p})}{\widehat{U_2}(\mathbf{p})}\ if\ \mathbf{p} = \mathbf{F}^d \tag{8}$$

Note that HDPCA is a sparse space where only points corresponding to the projected vectors $\boldsymbol{F}^d$ are defined.

4 Proposed Method

To restore the pixel (k,l), in order to avoid *"best-worse paradox"* phenomenon, instead of projecting all patches in the image domain Ω onto the HDPCA-space, we project only the patches on the sub-domain $\Omega(k,l)$ (small windows of size 21×21 around the being processed pixel (k,l)) and carry out the filtering on these projected values. Since all values in projected vector $\boldsymbol{F}^d(k,l)$ become spatial coordinates $\mathbf{p}$ in the HDPCA-space, we can rewrite equation (7) of FNLM as follows:

$$\widehat{u}^d(k,l) = \frac{\widehat{U}_1(\mathbf{p})}{\widehat{U}_2(\mathbf{p})} = \frac{\sum_{\mathbf{q}\in\Psi_d} exp\left(-\frac{\|\mathbf{p}-\mathbf{q}\|^2}{h_r^2}\right) V_1(\mathbf{q})}{\sum_{\mathbf{q}\in\Psi_d} exp\left(-\frac{\|\mathbf{p}-\mathbf{q}\|^2}{h_r^2}\right) V_2(\mathbf{q})} \tag{9}$$

Note that both nominator and denominator of this equation can be interpreted as Gaussian filter in the HDPCA-space. Therefore, we can summarize FNLM in the two following steps:

- Step 1: Gaussian filtering in the HDPCA-space
- Step 2: Projection back onto the image space by using the division of two components of the filtered values (equation (8))

It is worth to notice that the Gaussian filter in the first step is an isotropic filter. Here, we propose to replace it by an anisotropic one. In the literature, there are many anisotropic diffusion methods such as Total Variation[9], Perona-Malik[10] which mimic physical processes by locally diffusing pixel values along the image structure. Since these methods are local-based, their adaptation to a such sparse HDPCA-space is rather a difficult task. However, as discussed above, BF acts as an anisotropic filter and it works in non-local manner therefore it could be used. The proposed method (called BF-HDPCA) consists of the two following steps:

- Step 1: Bilateral filtering in the HDPCA-space
- Step 2: Projection back onto the image space by using the division of two components of the filtered values (equation (8))

In the first step, the filtered values $\widehat{\mathbf{U}}(\mathbf{p})$ are given by:

$$\widehat{U_\eta}(\mathbf{p}) = \frac{\sum_{\mathbf{q}\in\Psi_d} w_\eta(\mathbf{p},\mathbf{q}) V_\eta(\mathbf{q})}{\sum_{\mathbf{q}\in\Psi_d} w_\eta(\mathbf{p},\mathbf{q})} \tag{10}$$

where subscript $\eta = 1, 2$, and according to BF's principle the weight $w_\eta(\mathbf{p},\mathbf{q})$ is estimated as follows:

$$w_\eta(\mathbf{p},\mathbf{q}) = exp\left(\frac{-\|V_\eta(\mathbf{p}) - V_\eta(\mathbf{q})\|^2}{h^2}\right) exp\left(\frac{-\|\mathbf{p}-\mathbf{q}\|^2}{h_r^2}\right) \tag{11}$$

where the first term is an intensity proximity measure and the second one stands for a geometric proximity measure, h is range parameter in the new HDPCA-space. Recall that this second term is the patch similarity measure defined in the image domain. The filtered values $\widehat{\mathbf{U}}(\mathbf{p})$ are finally projected back into the image domain Ω using the division in equation (8). Note that when $h = \infty$ the proposed method tends to FNLM.

5 Experimental Results

The experimental results are carried out on three natural images: Lena, Peppers and Fingerprint of size 512×512. The last one is typical of highly textured image whereas the second one contains mostly homogenous regions. The first image contains different types of features, texture, sharp edges and smooth regions. These images are perturbed by additive, independent Gaussian noise at two levels of standard deviation $\sigma = 10$ and $\sigma = 25$. The subspace $\Omega(k, l)$ is defined by small windows 21×21 around the being processed pixel (k, l). The patch size is equal to 7×7 which results of full dimension r^2 =49. The reduced dimension d is tested with 11 values: 1, 3, 6, 8, 10, 15, 20, 25, 30, 40, 49. h_r is set equal to $n_{h_r}\sigma$ where $n_{h_r} = [0.6 : 0.1 : 1.4]$ (here we use Matlab notation) and $h = n_h\sigma$ with $n_h = [1 : 1 : 10, \infty]$. Recall that when $n_h = \infty$ and $d = 49$, the proposed method tends to the classical NLM. A comparative evaluation using both objective and subjective measures has been performed to demonstrate the advantages of the proposed method over NLM and FNLM filters. To objectively

Table 1. Objective measures of Lena image

Metric	Method	n_{h_r}	n_h	d	Result
PSNR	NLM	0.9	∞	49	33.55
	FNLM	0.9	∞	15	33.63
	BF-HDPCA	0.9	4	**8**	**34.04**
PSNR$_W$	NLM	0.8	∞	49	16.61
	FNLM	0.8	∞	10	16.78
	BF-HDPCA	0.8	4	**8**	**17.18**
MAD	NLM	0.8	∞	49	1.88
	FNLM	0.7	∞	10	1.75
	BF-HDPCA	0.8	4	**8**	**1.66**

$\sigma = 10$

Metric	Method	n_{h_r}	n_h	d	Result
PSNR	NLM	1	∞	49	29.81
	FNLM	1	∞	30	29.81
	BF-HDPCA	1	20	**15**	29.80
PSNR$_W$	NLM	0.7	∞	49	12.19
	FNLM	0.7	∞	15	12.21
	BF-HDPCA	0.7	20	15	12.21
MAD	NLM	1	∞	49	7.00
	FNLM	1	∞	15	6.97
	BF-HDPCA	0.6	4	**3**	**6.71**

$\sigma = 25$

Table 2. Objective measures of Fingerprint image

Metric	Method	n_{h_r}	n_h	d	Result
PSNR	NLM	1	∞	49	29.93
	FNLM	1	∞	15	30.09
	BF-HDPCA	1.1	4	15	**31.02**
PSNR$_W$	NLM	0.9	∞	49	17.93
	FNLM	0.9	∞	15	18.19
	BF-HDPCA	1.1	2	**6**	**19.54**
MAD	NLM	0.8	∞	49	0.42
	FNLM	0.6	∞	6	0.29
	BF-HDPCA	0.7	4	6	**0.21**

$\sigma = 10$

Metric	Method	n_{h_r}	n_h	d	Result
PSNR	NLM	0.7	∞	49	26.64
	FNLM	0.6	∞	6	27.27
	BF-HDPCA	0.6	6	6	**27.40**
PSNR$_W$	NLM	0.6	∞	49	13.91
	FNLM	0.6	∞	6	14.48
	BF-HDPCA	0.6	4	6	**14.63**
MAD	NLM	0.6	∞	49	2.34
	FNLM	0.6	∞	6	1.70
	BF-HDPCA	0.6	4	6	**1.57**

$\sigma = 25$

Table 3. Objective measures of Peppers image

Metric	Method	n_{h_r}	n_h	d	Result
PSNR	NLM	0.9	∞	49	33.71
	FNLM	0.8	∞	10	33.89
	BF-HDPCA	0.9	4	**6**	**34.35**
PSNR$_W$	NLM	0.8	∞	49	17.09
	FNLM	0.7	∞	10	17.39
	BF-HDPCA	1	4	**6**	**17.86**
MAD	NLM	0.8	∞	49	3.07
	FNLM	0.7	∞	10	2.49
	BF-HDPCA	0.8	4	**6**	**2.20**

$\sigma = 10$

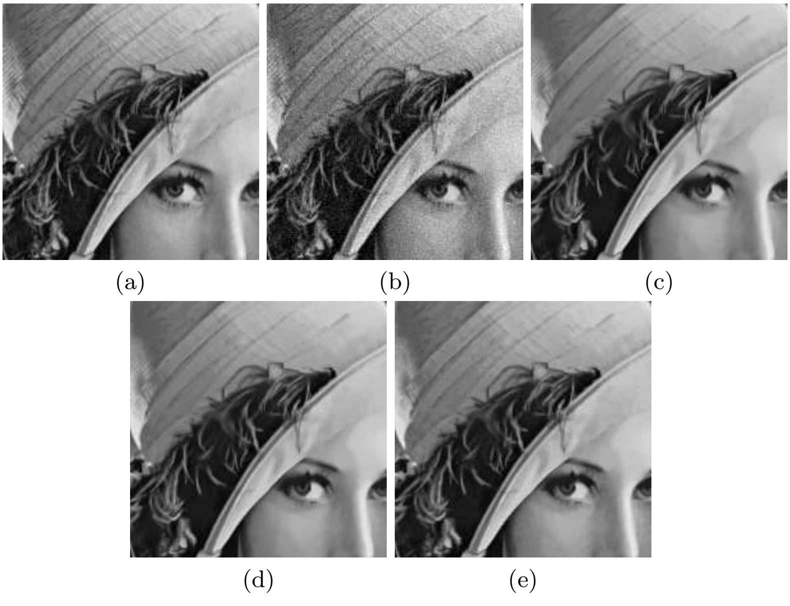
Metric	Method	n_{h_r}	n_h	d	Result
PSNR	NLM	1	∞	49	30.24
	FNLM	1	∞	15	30.28
	BF-HDPCA	0.6	4	**3**	**30.41**
PSNR$_W$	NLM	0.7	∞	49	12.81
	FNLM	0.7	∞	15	12.90
	BF-HDPCA	0.6	4	**3**	**13.05**
MAD	NLM	1	∞	49	8.14
	FNLM	1	∞	15	8.04
	BF-HDPCA	0.6	4	**3**	**7.75**

$\sigma = 25$

Table 4. Computational time in function of d (The program is written by C, runs on PC of 2GHz and 2G Ram)

d	1	3	6	8	10	15	20	25	30	40	49
Time (in second)	9.29	11.04	14.41	17.33	22.65	30.12	41.04	51.46	62.17	84.92	107.2

evaluate the results, beside PSNR, we use also two other metrics namely MAD [8] and PSNR$_W$ [2] which are based the human visual system (HVS). Note that while small value of MAD indicates high level of image quality, small value of PSNR$_W$, PSNR corresponds to a low level of image quality. *The best results* of three methods are reported in Tables 1, 2, 3 with the corresponding optimal parameters d, n_h, n_{h_r}. Note that, for each method, each metric results in different optimal parameters. For $\sigma = 10$, BF-HDPCA clearly outperforms the others methods for all images and it is confirmed by all metrics. For higher noise level $\sigma = 25$, this advantage is no longer true for Lena image, but still valid for images which contain many redundant structures such as Fingerprint and Peppers. It is also worth to note that, in many cases, BF-HDPCA can achieve better quality with smaller dimension d compared to FNLM (for example, for Pepper image, $\sigma = 25$, optimal d for FNLM is 15 whereas in our case, this value is 3 which makes BF-HDPCA 3 times faster than FNLM - see table 4). For all images and all metrics, the optimal h of the proposed method can be found from 2σ to 4σ. For the subjective comparison, due to limit of space, only Lena images are presented in Fig.1. As can be seen, the small details on the hat are well preserved with the proposed method whereas they are oversmoothed in NLM and FNLM. There is still some kind of natural grain in the forehead and cheek regions in our case, what seems more comfortable for our eyes than too flat region obtained with NLM and FNLM (*please use your monitor to view all images in this paper*).


(a) (b) (c)

(d) (e)

Fig. 1. Lena image: from left to right, top to bottom, (a) The original image, (b) The noisy image for $\sigma = 10$, (c) NLM's result (the best PSNR=33.55 for $h_r = 0.9\sigma$,$h = \infty$,$d = 49$), (d) FNLM's result (the best PSNR=33.63 for $h_r = 0.9\sigma$,$h = \infty$,$d = 15$), (e) The proposed method (the best PSNR=34.04 for $h_r = 0.9\sigma$,$h = 4\sigma$,$d = 8$)

6 Conclusions

A new nonlinear anisotropic filtering method based on BF and the HDPCA-space is proposed. Through this study, it has been shown that NLM and FNLM can be expressed as an isotropic filter in this space. A series of tests has been performed to assess the efficiency of the proposed method. The obtained results demonstrate the efficiency of the proposed filtering approach objectively and subjectively.

References

1. Barash, D., Comaniciu, D.: A common framework for nonlinear diffusion, adaptive smoothing, bilateral filtering and mean shift. Image Vision Comput., 73–81 (2004)
2. Beghdadi, A., Pesquet-Popescu, B.: A new image distortion measure based wavelet decomposition. In: Proc. ISSPA, pp. 485–488 (2003)
3. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Simul 4, 490–530 (2005)
4. Buades, A., Coll, B., Morel, J.M.: Nonlocal image and movie denoising. International Journal of Computer Vision, 123–139 (2008)
5. Do, Q.B., Beghdadi, A., Luong, M.: Combination of closest space and closest structure to ameliorate non-local means method. In: IEEE Symposium on Computational Intelligence for Multimedia, Signal and Vision Processing (2011)
6. Elad, M.: On the origin of the bilateral filter and ways to improve it. IEEE Transactions on Image Processing, 1141–1151 (2002)
7. Kervrann, C., Boulanger, J.: Optimal spatial adaptation for patch-based image denoising. IEEE Trans. on Image Processing 15(10), 2866–2878 (2006)
8. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of Electronic Imaging 19 (2010)
9. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear Total Variation Based Noise Removal Algorithms. Physica D 60, 259–268 (1992)
10. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell., 629–639 (1990)
11. Souidene, W., Beghdadi, A., Abed-Meraim, K.: Image denoising in the transformed domain using non local neighborhood. In: Proc. ICASSP (2006)
12. Tasdizen, T.: Principal neighborhood dictionaries for nonlocal means image denoising. IEEE Transactions on Image Processing, 2649–2660 (2009)
13. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images, pp. 839–846 (1998)
14. Tschumperlé, D., Brun, L.: Image denoising and registration by pde's on the space of patches (2008)

Image Super Resolution Using Sparse Image and Singular Values as Priors

Subrahmanyam Ravishankar[1], Challapalle Nagadastagiri Reddy[1],
Shikha Tripathi[1], and K.V.V. Murthy[2]

[1] Amrita Viswa Vidya Peetham University,
s_ravishankar@blr.amrita.edu,
dastagiri.c@gmail.com,
t_shikha@blr.amrita.edu
[2] IIT Gandhi Nagar,
kvvm@iitgn.ac.in

Abstract. In this paper single image superresolution problem using sparse data representation is described. Image super-resolution is ill - posed inverse problem. Several methods have been proposed in the literature starting from simple interpolation techniques to learning based approach and under various regularization frame work. Recently many researchers have shown interest to super-resolve the image using sparse image representation. We slightly modified the procedure described by a similar work proposed recently. The modification suggested in the proposed approach is the method of dictionary training, feature extraction from the trained data base images and regularization. We have used singular values as prior for regularizing the ill-posed nature of the single image superresolution problem. Method of Optimal Directions algorithm (MOD) has been used in the proposed algorithm for obtaining high resolution and low resolution dictionaries from training image patches. Using the two dictionaries the given low resolution input image is super-resolved. The results of the proposed algorithm showed improvements in visual, PSNR, RMSE and SSIM metrics over other similar methods.

Keywords: Method of Optimal Directions, Orthogonal Matching Pursuit, Singular Value Decomposition, Sparse representation.

1 Introduction

Image super resolution (SR) is an active area of research as physical constraints limit image quality in many imaging applications. These imaging systems yield aliased and under sampled images if their detector array is not sufficiently dense. Super resolution is also crucial in satellite imaging, medical imaging, and video surveillance[1]. There are several approaches for image enhancement such as Bi-cubic interpolation, B-spline interpolation etc[2]. These approaches produce overly smooth images which lack high frequency components and thus blur the edge information of the reconstructed image. Other conventional approaches like multi frame super resolution produce high resolution images employing more

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 380–388, 2011.

than one low resolution image[3]. These methods use multiple low resolution images of the same scene, which are aligned with sub pixel accuracy to generate high resolution image. Recently, single frame image super resolution algorithms have been successfully employed for image super resolution. These methods use only one low resolution (LR) test image for super resolution. Freeman et. al. proposed an example based super resolution technique[4]. They estimate missing high frequency details by interpolating the input low resolution image into a desired scale. The super resolution is performed by the nearest neighbor based estimation of high frequency patches corresponding to the patches of the low frequency input image. An excellent review on single image super resolution is given in[5]. Super resolution of images is an ill-posed inverse problem due to unknown blurring operators and insufficient number of low resolution images. Different regularization methods have been proposed for providing solutions to this problem[6,7,8]. Yang et. al. proposed a super resolution algorithm based on sparse representations[9]. The authors sparsely represent given low resolution image patches using low resolution dictionary and use these coefficients to construct high resolution patches from high resolution dictionary. It is also reported that the authors used first-order and second-order derivatives as feature for dictionary training. Roman et. al. have proposed another dictionary based technique for super resolution using sparse representation[10]. The authors used k-means Singular Value Decomposition (K-SVD) algorithm for dictionary training. In the proposed approach Method of Optimal Directions (MOD) algorithm have been used for dictionary training[11], and Orthogonal Matching Pursuit (OMP) algorithm for sparse representations[12]. Though the performance of MOD and K-SVD is similar for dictionary training, however we find that the algorithmic steps involved in MOD are less compared to K-SVD thus the computational time reduces. By using regularization the proposed method gives better results as compared to earlier methods . In the present approach for experimental purpose we created a data set of low resolution (LR) images and corresponding high resolution images (HR), and we extracted patches from the data set. MOD algorithm is applied on these patches to create low resolution dictionary (A_l) and high resolution dictionary (A_h). Low resolution (LR) test image patches are sparsely represented using A_l and using these sparse coefficients we obtained high resolution patches from A_h. We concatenate these patches to get high resolution image. Gradient descent optimization algorithm is used to minimize the cost function.

2 Image Formation Model

Let the original high resolution image be represented as $y_h \in R^N$, blur operator as $B : R^N \rightarrow R^N$ and down sampling operator as $D : R^N \rightarrow R^M$ where $M < N$, and $z_l \in R^M$ is defined as low resolution version of the original high resolution image. The observed image can be modeled as

$$z_l = BDy_h + v \tag{1}$$

where v is an i.i.d Gaussian noise of mean zero and co variance σ_v^2. The problem is to find high resolution estimate $\hat{y_h}$ for a given low resolution image z_l such that $\hat{y_h}$ is closer to y_h. Although $\hat{y_h}$ can be obtained by minimizing $\|BDy_h - z_l\|^2$, we cannot find exact solution to above problem since BD is rectangular matrix and there are infinitely many solutions for equation (1). The goal of super resolution technique is to recover the high frequency content that is lost during image acquisition process. This is an inverse problem wherein the original information is to be retrieved from observed data. There exists infinite number of High Resolution (HR) images which are consistent with the original data, thus the super resolution problem is an ill posed inverse problem. While solving ill posed inverse problems knowing the forward model alone is not sufficient to obtain satisfactory results. Some form of constraints on the space of solutions must be included. Procedure adopted to stabilize ill posed inverse problems is called regularization. Existing approaches for image super resolution uses various priors such as textures, edges, and curves etc; to regularize the above equation(1) as discussed in [13,14].

3 Sparse Representation

Sparse representation account for most or all information of a signal with a linear combination of a small number of elementary signals called atoms in an over complete dictionary. The coefficients are sparse which means most of the coefficients have zero values. Sparse representation is widely used in the areas of compressed sensing, image super resolution and face hallucination. The objective of sparse representation is that of representing some data $y \in R^m$ (for example a patch of an image) using a small number of non zero components in a sparse vector $x \in R^n$ under the linear generative model

$$y = Ax + v \tag{2}$$

where the full low rank dictionary $A \in R^{m \times n}$ may be over complete $(n > m)$, and the additive noise ǐs assumed to be Gaussian with zero mean and variance σ_v^2. Sparse coding techniques[15,16] have been proposed for representing an arbitrary signal as linear combination of basis functions called atoms which are learned from training signals. In sparse coding the atoms are learned from the set of training signals to maximize the sparsity of coefficients that are assigned to the atoms for the linear representation. Formally if $x \in R^n$ is a column vector, and the atom vectors are arranged as the columns of the dictionary $A \in R^{n \times m}$, the sparsity assumption is described by the following sparse approximation problem, for which we assume a sparse solution exists:

$$\hat{x} = argmin_x \ \|x\|_0^0 \ s.t \ \|y - Ax\|_2 \leq e \tag{3}$$

in this expression, $\hat{x}$ is the sparse representation of y, e is the error tolerance, and the function $\|.\|_0^0$ is the l^0-norm, which counts the non-zero elements of a vector.

4 Proposed Algorithm

The proposed algorithm has two parts. Part-1 of the algorithm (Training phase) constructs dictionaries of low resolution and high resolution data sets, and Part-2 of the algorithm (reconstruction phase) super resolves the given input low resolution test image using the results of part-1.

4.1 Training Phase

Following steps are performed in this phase

i. Training set construction: In this phase an Image data bank is formed that consists of several high resolution images $\{y_h^j\}_j$ and the corresponding low resolution images $\{z_l^j\}_j$ which are blurred, down sampled versions of the high resolution images. $\{z_l^j\}_j$ are further interpolated to the size of high resolution image y_h.

ii. Feature extraction: After constructing the training set, low frequency components of the high resolution image that correspond to low frequency non-aliased components of the low resolution image are removed by computing Discrete Cosine Transform (DCT) and retaining coefficients which corresponds to high frequency information. The reason for this feature extraction is to focus the training on characterizing the relation between the low resolution patches and the edges, texture content within the corresponding high resolution ones [10]. Patches of size nxn are formed from low and high resolution images resulting in the data set $\{p_h^k, p_l^k\}_k$.

iii. Dictionary Training: The construction of efficient dictionary is one of the most important steps to be incorporated in the sparse representation of signals. In the quest for the proper dictionary, one line of thought is to propose a learning method in the formation of dictionary. Assume that ϵ a model deviation parameter which is known to us, and our aim is the estimation of the dictionary A under the constrained optimization given by the following equation.

$$\underset{A,[x_i]_{i=1}^M}{min} \sum_{i=1}^{M} \|x_i\|_0 s.t. \|y_i - Ax_i\|_2 \leq \epsilon, 1 \leq i \leq M \tag{4}$$

This problem describes each given signal y_i as the sparsest representation x_i over the unknown dictionary A, and aims to jointly find the proper represenations and the dictionary. In the proposed algorithm MOD algorithm is used for training the dictionary. We preferred MOD over K-SVD because of the gain in computational speed. We applied both MOD and K-SVD on synthetic data consisting 1000 signals for training dictionary of size 64x1000, K-SVD took 6 minutes and MOD took 20 seconds for training. In this step MOD algorithm is employed on low-resolution image patches $\{p_l^k\}_k$ for training low resolution dictionary using the equation (5).

$$A_l, q^{(k)} = {}^{argmin}_{A_l, q^k} \sum_k \|p_l^k - A_l q^k\|^2 s.t. \|q^k\| \leq L \tag{5}$$

High resolution dictionary A_h, is formed to recover high resolution patches by approximating them as $\{p_h^k = A_h q^k\}$. The dictionary A_h is constructed such that this approximation is as exact as possible, thus the dictionary is defined as the one which minimizes the mean approximation error, i.e.,

$$A_h = {}^{argmin}_{A_h} \sum_k \|p_h^k - A_h q^k\|_2^2 = {}^{argmin}_{A_h} \|P_h - A_h Q\|_F^2, \tag{6}$$

Where the matrix P_h is constructed with high resolution patches $\{p_h^k\}_k$ as its columns, and Q contains $\{q^k\}$ as its columns. The solution to the above problem is given by following pseudo-inverse expression [10] (given that Q has full low rank).

$$A_h = P_h Q^+ = P_h Q^T (QQ^T)^{-1} \tag{7}$$

The construction of high resolution and low resolution dictionaries A_h, A_l completes training phase of super resolution algorithm.

4.2 Reconstruction Phase

In this phase our task is to super resolve the given low resolution image z_l using the following steps.

i. Obtaining high resolution estimate of z_l: The low resolution image is scaled by a factor of D using Hermit interpolation resulting in $y_l \in R_N$. High frequency components are extracted from y_l using DCT, then patches $\{p_l^k\}_k$ are obtained from the extracted image. OMP algorithm is applied on these patches and dictionary A_l, allocating L atoms to represent each patch for finding sparse representation vectors $\{q^k\}_k$. The sparse representation vectors $\{q^k\}_k$ are multiplied with high resolution dictionary A_l and the approximated high resolution patches $\{A_h q^k\}_k = \{p_h^k\}_k$ are obtained. The high resolution estimate $\hat{y_h}$ is obtained from $\{p_h^k\}_k$ by solving following minimization problem with respect to $\hat{y_h}$

$$\hat{y_h} = {}^{argmin}_{A_h} \sum_k \|R_k(\hat{y_h} - y_l) - p_h^k\|_2^2 \tag{8}$$

Where R_k is patch extraction operator. The above problem states that patches from the image $\hat{y_h} - y_l$ should be as close as possible to approximated patches, this problem has a least square solution given by

$$\hat{y_h} = y_l + [\sum_k R_k^T R_k]^{-1} \sum_k R_k^T p_h^k \tag{9}$$

The term $R_k^T p_h^k$ in the above equation positions the high resolution patch in the kth location and the term $R_k^T R_k$ is diagonal matrix that weighs each pixel in the high resolution outcome. Above equation can be simply explained as putting p_h^k in proper locations and adding y_l to them for getting high resolution estimate.

ii. Regularizing high resolution estimate: SVD is obtained for the high resolution estimate F. The least few eigen values of the obtained SVD are retained and used to reconstruct the image G which essentially contains rich edge information. In the next step we obtain the edge information by subtracting G from F, $F_{edge} = F - G$. This information is used as a prior to regularize high resolution estimate by minimizing the following cost function.

$$C = \sum_{i,j} \|z_{l(i,j)} - F_{(i,j)}\|^2 - \|F_{edge(i,j)} - f_{edge(i,j)}\|^2 \tag{10}$$

Where f_{edge} is edge information of interpolated image. $z_{l,(i,j)}$ is the low resolution test image and F_{edge} is edge information of high resolution estimate.We minimize the above cost function using gradient descent optimization. Regularized image is the final super resolution image.

5 Experimental Results

We constructed a training set with 30 high resolution images and the corresponding low resolution images by down sampling the high resolution images by a factor of 2 . Features are extracted from the training set using DCT. Around 7860 patches are obtained from the extracted feature images. MOD algorithm is applied on the low resolution patches for constructing A_l of size 64x6000. Using equation(6) we obtained A_h of size 64x6000. We applied our algorithm on several test images of similar classes used by Roman .et.al[10] and Yang .et.al[9]. The results were compared with the proposed approach and the results obtained by Roman .et.al[10] and Yang .et.al[9]. The results are tabulated in Table1. We compared our results with standard interpolation techniques[2] in terms of PSNR, RMSE and SSIM. The results were tabulated in Table1, Table2, Table3. Fig.1 gives visual comparis on of the results of the proposed methods with other listed methods. The entire code for the algorithm is written in Matlab and simulated.

Table 1. PSNR comparison

Image No.	PSNR			
	Mitchell interpolation	Yang et. al.	Roman et. al.	Proposed algorithm
1	33.20	33.10	33.50	39.89
2	33.10	33.10	33.50	35.72
3	24.50	24.80	25.00	29.66
4	24.76	24.80	25.00	31.51
5	28.00	28.20	28.40	32.27

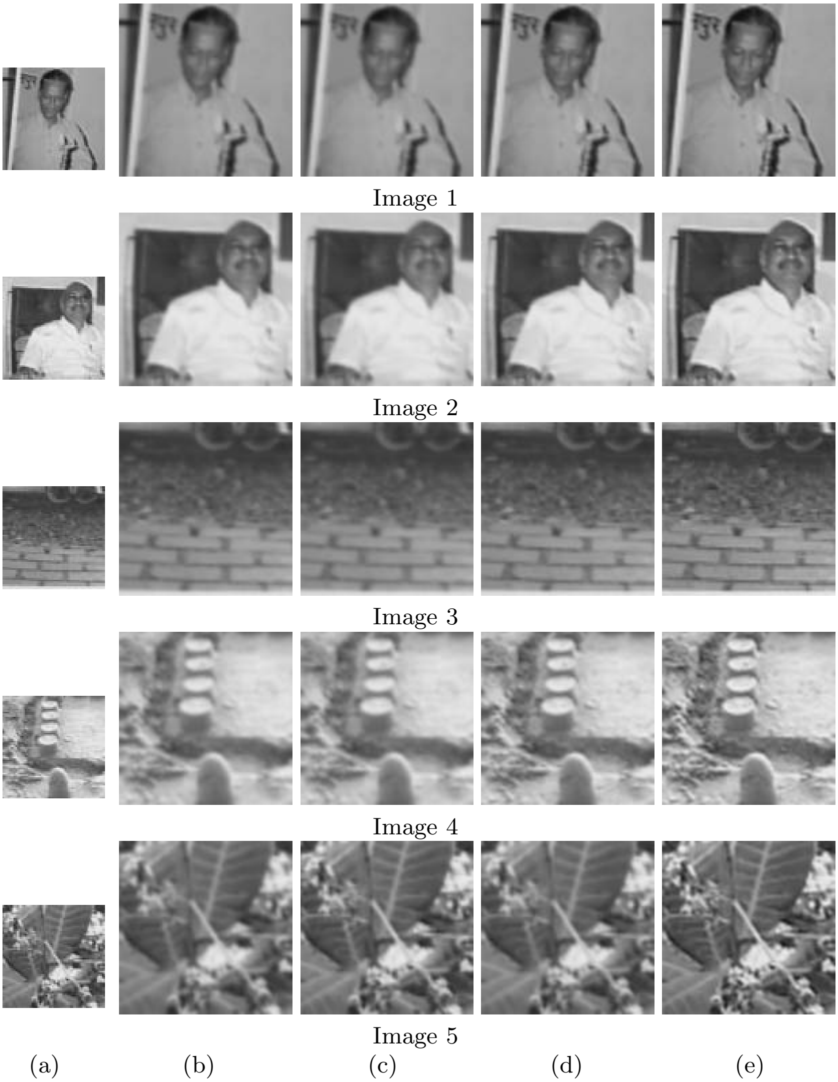


Fig. 1. (a) LR Image, (b) Mitchell Interpolation, (c) B-spline Interpolation (d) Bell Interpolation, (e) Proposed Method

Table 2. RMSE comparison

Image No.	RMSE			
	Mitchell interpolation	B-spline interpolation	Bell interpolation	proposed algorithm
1	12.2838	12.7950	11.6912	11.211
2	12.1617	12.6742	11.5603	11.06
3	9.9298	10.2821	9.6024	9.4011
4	14.7493	15.2244	14.2904	14.10
5	17.2603	18.0352	16.3310	14.987

Table 3. SSIM comparison

Image No.	SSIM			
	Mitchell interpolation	B-spline interpolation	Bell interpolation	Proposed algorithm
1	0.8750	0.8613	0.8911	0.9132
2	0.8435	0.8267	0.8608	0.8987
3	0.7808	0.7569	0.8092	0.8651
4	0.7395	0.7171	0.7663	0.8694
5	0.7547	0.7250	0.7896	0.8575

6 Conclusion

In the proposed method we have successfully simulated the concept of single image super resolution using sparse images obtained from trained dictionaries of high resolution and low resolution image patch pairs. We compared our results with the results obtained by Yang et. al. [9], Roman et. al.[10]. Our simulated results indicate the effectiveness of use of sparse representations of the learnt image and Singular values as priors. We modified the method used by Roman et.al. [10] by employing MOD algorithm for dictionary training instead of K-SVD algorithm, and DCT was used for feature extraction. DCT is simpler and effective because it eliminates patch concatenation and helps in dimensionality reduction. We used very less number of training patches as compared to Roman et. al. [10]. and got better results. The perfor mance gain compared to Roman et. al. [10] is due to the regularization of the cost function using singular values as priors. The efficiency of algorithm can be further improved by using K-COD algorithm instead of MOD algorithm and optimizing thresholds for DCT and extracting patches with overlapping.

Acknowledgement. The authors wish to thank Ron Rubinstein research scholar, Technion Israel Institute of Technology for all his technical support.

References

1. Liu, H.Y., Zhang, Y.S.: Study on the methods of super resolution image reconstruction. In: The International archives of photogrammetry, remote sensing and spatial information sciences, Beijing, vol. XXXVII.part B2 (2008)

2. Hou, H.S., Andrews, H.C.: Cubic splines for image interpolation and digital filtering. IEEE transactions on ASSP 26(6) (December 1978)
3. Kim, S.P., Bose, N.K., Valenzuela, H.A.: Recursive reconstruction of high resolution image from noisy under sampled multi frames. IEEE Transactions on ASSP 38(6) (June 1990)
4. Freeman, W., Jones, T., Pasztor, E.: Example based super resolution. IEEE Computer graphics and Applications 22(2), 56–65 (2002)
5. Elad, M., Feuer, A.: Restoration of single super resolution image from several blurred, noisy, and undersampled measured images. IEEE Transactions on Image Processing 6(12) (December 1997)
6. Hardie, R.C., Barnard, K.J., Armstrong, E.A.: Joint MAP registration and high resolution image estimation using a sequence under sampled images. In: IEEE TIP (1997)
7. Farisu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. In: IEEE TIP (2004)
8. Tipping, M.E., Bishop, C.M.: Bayesian image super resolution. In: Proc. Neural Information Processing Systems (2003)
9. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super resolution via sparse representation. In: Proc. IEEE Transactions on Image Processing (2009)
10. Zeyde, R., Elad, M., Protter, M.: On Single Image Scale-Up using sparse representations. Curves and Surfaces, Avignon-France (2010)
11. Engan, K., Asae, S.O., Husoy, J.H.: Multi-frame compression: Theory and design. EURASIP Signal Processing 80(10), 2121–2140 (2000)
12. Elad, M.: Sparse and redundant representations: from theory to applications in signal and image processing, 1st edn. Springer publications, Heidelberg (2010)
13. Pickup, L.C., Roberts, S.J., Zisserman, A.: A sampled texture prior for image super resolution. Proc. Neural Information Processing Systems (2003)
14. Rudraraju, K., Joshi, M.V.: Nonhomogeneous AR model based prior for multiresolution fusion. In: Proc. IEEE International Geosciences and Remote Sensing Symposium (2009)
15. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive-field properties by learning sparse code for natural images. Nature 381(13), 607–609 (1996)
16. Lewicki, M.S., Olshausen, B.A.: A probabilistic frame work for the adoptation and comparision of image codes. J. Opt. Soc. Amer. Opt. Image Sci. 16(7), 1587–1601 (1999)

Improved Gaussian Mixture Model for the Task of Object Tracking

Ronan Sicre[1,2] and Henri Nicolas[1]

[1] LaBRI, University of Bordeaux, 351 Cours de la libration, 33405 Talence, France
sicre-nicolas@labri.fr
[2] MIRANE S.A.S., 16 rue du 8 mai 1945, 33150 Cenon, France

Abstract. This paper presents various motion detection methods: temporal averaging (TA), Bayes decision rules (BDR), Gaussian mixture model (GMM), and improved Gaussian mixture model (iGMM). This last model is improved by adapting the number of selected Gaussian, detecting and removing shadows, handling stopped object by locally modifying the updating process. Then we compare these methods on specific cases, such as lighting changes and stopped objects. We further present four tracking methods. Finally, we test the two motion detection methods offering the best results on an object tracking task, in a traffic monitoring context, to evaluate these methods on outdoor sequences.

Keywords: Computer vision, Motion detection, Object tracking.

1 Introduction

Object tracking is an important task of the computer vision field [8]. There are two main steps in object tracking: interest moving object detection and tracking these objects from frame to frame. Then analysis can determine the objects' behaviors in various context, such as traffic monitoring [4], marketing application [6], etc. This paper focuses on the detection of moving objects. We use motion detection that aims at separating moving areas, or foreground, from the background.

The main challenge is to detect the relevant changes and filter out the irrelevant one. Another challenge is the textural similarity between foreground objects and background. Complex scenes bring along other issues such as partial or complete occlusions caused by the background, or several people aligned with the camera.

In this paper, we first review some previous work. Then we present four motion detection techniques: temporal averaging, Bayes decision rules, Gaussian mixture model and the proposed improved Gaussian mixture model. We further present four tracking method that are used for the evaluation and conclude.

2 Previous Work

The aim of motion detection is to distinguish the background from the moving objects. The motion detection techniques can be separated in two categories,

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 389–396, 2011.

depending if they are using a background model or not. Without a background model, the algorithms are based on temporal differencing or calculation of optical flow. These methods can be fast, but the results are noisy. When a model is available, we can classify them depending on the method used. The model can be pixel based, local, or global.

Pixel based models assign to each pixel of the image a value or an intensity function that gives the appearance of the background. We only use the measurement taken on the specific pixel. The model can be as simple as an image of the background without objects or more complex by using Gaussian distribution [7].

Local based models use the neighborhood of a pixel instead of the pixel itself to calculate the similarity measurement. Specifically, these methods calculate if a block of pixels belongs or not to the background.

Global based models use the entire image at each moment to build a model of the entire background. The k-means algorithm and eigen background are two examples of global models.

In our study, we choose a pixel based model that offers a good compromise between precision, quality, and speed.

3 Motion Detection

This section presents four motion detection algorithms: temporal averaging, Bayes decision rules, Gaussian mixture model, and improved Gaussian mixture model.

3.1 Temporal Averaging

To model the background, we calculate the mean over several images while no object is in the scene. In fact, we approximate this mean using an infinite impulse response filter. Once the model is generated, we calculate the difference between the current image and the model. If there is a high difference, pixels are considered as being part of the foreground, i.e. corresponding to a moving region. Otherwise, pixels belong to the background.

As we see on figure 1 first row, shadows generate false positive with such a model. Thus, we converted the images into YUV color space. Only the luminance channel, Y, is sensible to shadows, see figure 1 first row.

However using this technique, small illumination variations on highly textured areas can lead to false positives, see figure 1 second row.

It is interesting to note that the background model is further improved in order to detect people stopping in the scene. Using the temporal averaging technique, if a person enters the scene and stops moving, the person slowly disappears and becomes part of the background.

Assuming the background remains relatively stationary while being occluded by an object. We save the model values before the occlusion occurs and keep it identical while occluded by a moving object.

To conclude, since this method does not handle properly light changes, a multi-modal background model can probably offer better results.

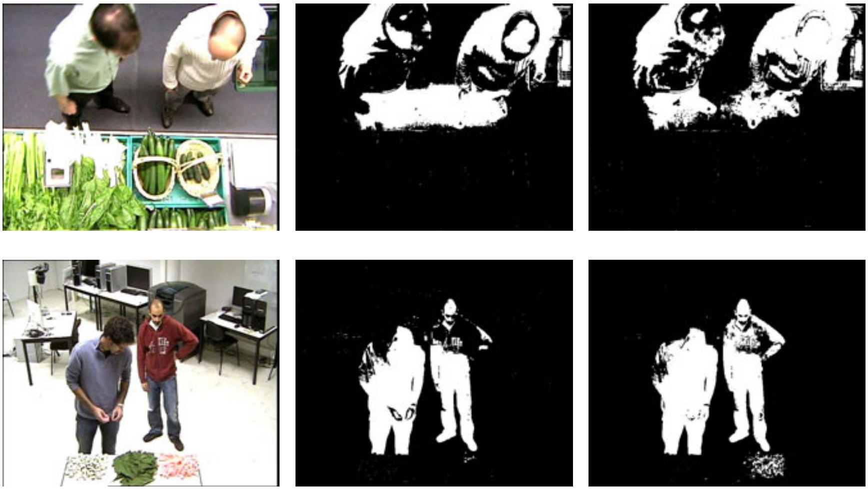


Fig. 1. Results for the temporal averaging technique using RGB color space, in the middle column, and YUV color space, in the right column

3.2 Bayes Decision Rules

This section shortly presents an elaborated method to model the background [3]. The first idea is that background pixels may have multiple states in complex environment: stationary or moving. Therefore, these different parts of the background should be analyzed with different features.

Specifically, this method aims at formulating Bayes decision rules in order to classify background and foreground using selected feature vectors. Stationary background objects are described with color features, while moving background objects are described with color co-occurrence features. Then, foreground objects are extracted from the fusion of the classification results from both stationary and moving pixels. More details on this method can be found in [3].

3.3 Proposed Improved Gaussian Mixture Background

We use a method based on the Gaussian mixture model (GMM) first introduced in [7]. The GMM is composed of a mixture of weighted Gaussian densities, which allows the color distribution of a given pixel to be multi-modal. Such a model is robust against illumination changes.

Weight ω, mean μ, and covariance $\varSigma$ are the parameters of the GMM that are updated dynamically over time. The following equation defines the probability P of occurrence of a color at the pixel coordinate s, at time t, in the image sequence I.

$$P(I(s,t)) = \sum_{i=1}^{k} \omega_{i,s,t} N(I(s,t), \mu_{i,s,t}, \varSigma_{i,s,t}) \tag{1}$$

Where $N(I(s,t), \mu_{i,s,t}, \Sigma_{i,s,t})$ is the i-th Gaussian model and $\omega_{i,s,t}$ its weight. The covariance matrix $\Sigma_{i,s,t}$ is assumed to be diagonal, with $\sigma^2_{i,s,t}$ as its diagonal elements. k is the number of Gaussian distributions.

For each pixel, $I(s,t)$, the first step is to calculate the closest Gaussian. If the pixel value is within T_σ deviation of the Gaussian mean, then parameters of the matched component are updated. Otherwise, a new Gaussian with a mean $I(s,t)$, a large initial variance, and a small initial weight ω_0 is created to replace the existing Gaussian with the lower weight. Once Gaussians are updated, weights are normalized and distributions are ordered based on the value $\omega_{i,s,t}/\sigma_{i,s,t}$.

Adaptive number of selected Gaussian. As proposed by [9], we improve the GMM by adapting the number of selected Gaussian distributions. The GMM selects only the most reliable distributions. Then, we modify the calculation of the distributions weights. The weight is decreased when the distribution is not observed for a certain amount of time.

$$\omega_{k,t} = \omega_{k,t-1} + \alpha(M_{k,t} - \omega_{k,t-1}) - \alpha c_T \tag{2}$$

Where α is the learning rate and $M_{k,t}$ is equal to 1 for the matched distribution and 0 for the others. c_T represents the prior evidence.

Pixels that are matched with any of these selected distributions are labeled as foreground. Otherwise, pixels belong to the background. We note that the model is updated at every frame.

Shadows detection. A second improvement to the GMM is to reduce the sensibility to shadows. Thus, we use a shadow detection algorithm, based on [1]. Shadows detection requires a model that can separate chromatic and brightness components. We use a model that is compatible with the mixture model. We compare foreground pixels against current background model. If the differences in chromatic and brightness are within some thresholds, pixels are considered as shadows. We calculate the brightness distortion a and color distortion c as follow:

$$a = argmin(I(s,t) - zE)^2 \;\; and \;\; c = \|I(s,t) - aE\| \tag{3}$$

Where E is a position vector at the RGB mean of the pixel background and $I(s,t)$ is the pixel value at position s and time t.

A foreground pixel is considered as a shadow if a is within T_σ standard deviations and $\tau < c < 1$. Where τ is the brightness threshold.

Handling stopped objects. Finally, we modify the updating process to better handle objects stopping in the scene. In fact, various applications require the detection of such meaningful objects. At a point of sale for example, people use to stop to look at products or prices. Concerning traffic monitoring, a stopped vehicle has to be detected for a potential traffic jam. Finally, detecting meeting in a video-surveillance in based on the detection of stopped people.

With the current model, stopped people starts disappearing, because they become part of the background. We modify the updating process for the distributions parameters, i.e. we do not refresh areas that are considered as belonging to a tracked object. Object tracking is presented in the following section.

We introduce $F_{s,t}$ that is a binary image representing these tracked objects. $F_{s,t}$ is a filtered foreground image where regions that were tracked for several frames, or objects, are displayed. Pixels covered by an object have value 1 while the others have value 0. We modify the distribution parameters updating equations:

$$\omega_{i,t} = \omega_{i,t-1} + (1 - F_{s,t})(\alpha(M_{i,t} - \omega_{i,t-1}) - \alpha c_T) \tag{4}$$

$$\mu_t = \mu_{t-1} + (1 - F_{s,t})\rho(I(s,t) - \mu_{t-1}) \tag{5}$$

$$\sigma_t^2 = \sigma_{t-1}^2 + (1 - F_{s,t})\rho((I(s,t) - \mu_t)^T(I(s,t) - \mu_t) - \sigma_{t-1}^2) \tag{6}$$

Where $\rho = \alpha\eta(I(s,t)|\mu_k, \sigma_k)$, $M_{i,t}$ is 1 for the matched density and 0 for the others, α is the learning rate, and c_T is the complexity reduction prior [9].

Once shadows are detected and erased, morphological filters are finally applied on this result to reduce noises, fill holes, and improve regions shape.

4 Object Tracking

This section shortly presents four existing tracking methods that are used to evaluate the motion detection techniques presented in the previous section.

4.1 Connected Component

This method is directly based on the motion detection output. For each frame, each detected regions is added to a region list. For each region in the list, a Kalman filter is used to estimate the location and size of the region in the next frame. Once the estimate is calculated for each region of the previous frame, we look for the region, in the current frame, that is the closest to the estimate. Closest regions are matched to the corresponding regions and are used to update the region location and size.

We note that this method only uses the location and size of each region and is therefore very fast. Then, the longer an object is matched, the better the estimate is.

4.2 Mean-Shift

The mean-shift algorithm is based on color information and is composed of 3 main phases:

- Calculation of an initial color histogram that identifies the tracked object.

- Apply this histogram on various locations around the estimated new location at the next frame.

- Find the closest histogram, i.e. the most probable match for the object.

We note that the distance calculation, between two histograms, is invariant to the scale of the target.

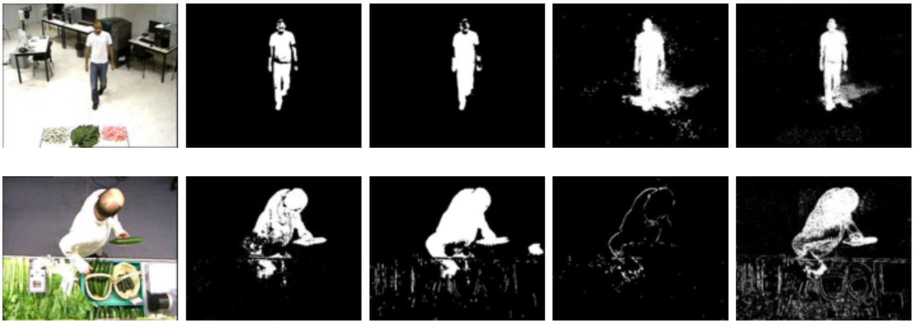


Fig. 2. Results for TA, BDR, GMM, and iGMM, in a shopping setting, from the left to the right. On the second row, the person is stopped.

4.3 Particle Filtering

Particle filter is a technique for implementing recursive Bayesian filter by Monte Carlo sampling [2]. The idea is to represent the posterior density by a set of random particles with associated weights. Particle filter computes estimates based on these samples and weights.

Particle filter main advantage is that it can handle non-linear, and non-Gaussian tracking problems.

The particle filter algorithm is composed of four main steps. First, the tracking task must be represented by a state space model. Then, Bayesian filter generates a posterior estimate. Monte Carlo simulates random samples. Finally, weights are calculated using importance sampling method.

4.4 Combined Connected Components and Particle Filtering

This last method is based on connected component technique. When objects are tracked and the estimated new locations of several objects correspond to a collision case. The system uses the particle filtering method presented above to continue tracking the objects as long as they are in a "collision state".

5 Results

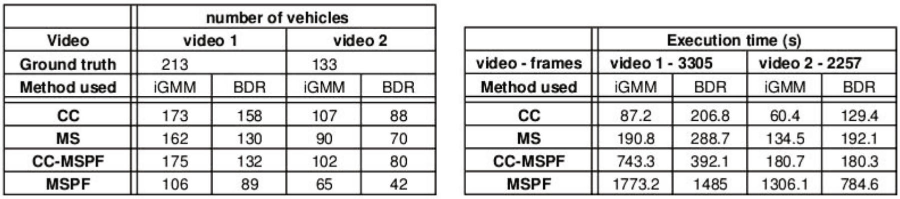
Motion detection: Tests are achieved on all the motion detection methods.

As we see on figure 1 first row, TA handles shadows. However, light changes are generating noises in highly textured areas, see figure 1 second row.

The original GMM is really sensible to shadows, see 2 first row. Moreover, this method do not handle stopped objects, see figure 2 second row.

BDR offers very good visual results, see figure 2 and 5. However, this methods remain sensible to noises in very complex scenes, see figure 5.

iGMM handles the limitations of GMM and offers very good visual results, see figure 2 and 5. We note that grey pixels correspond to detected shadows. This method is more robust than BDR on very complex scenes, as we see on figure 5.



	number of vehicles			
Video	video 1		video 2	
Ground truth	213		133	
Method used	iGMM	BDR	iGMM	BDR
CC	173	158	107	88
MS	162	130	90	70
CC-MSPF	175	132	102	80
MSPF	106	89	65	42

	Execution time (s)			
video - frames	video 1 - 3305		video 2 - 2257	
Method used	iGMM	BDR	iGMM	BDR
CC	87.2	206.8	60.4	129.4
MS	190.8	288.7	134.5	192.1
CC-MSPF	743.3	392.1	180.7	180.3
MSPF	1773.2	1485	1306.1	784.6

Fig. 3. Tables relating the number of tracked vehicles and the execution time

Fig. 4. Screen-shots of the two traffic monitoring videos

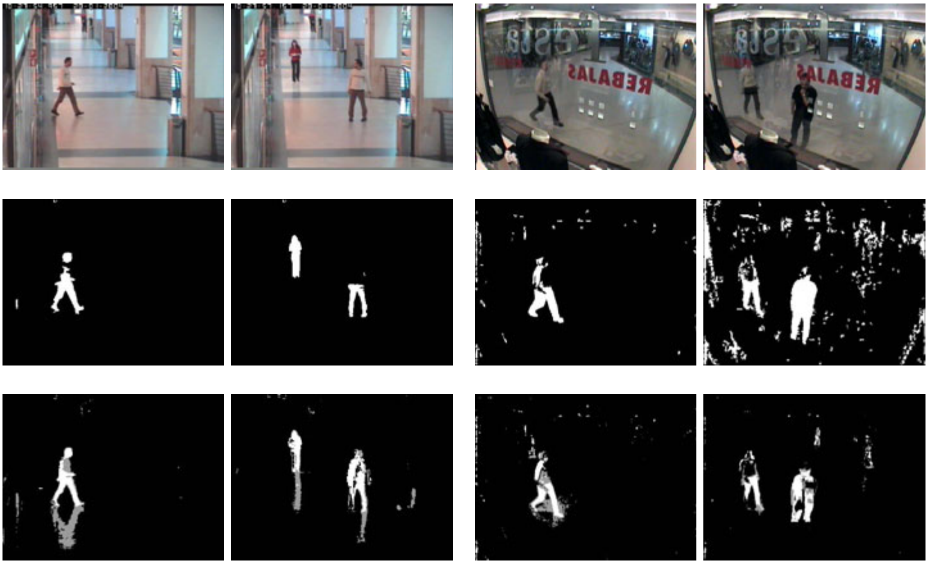


Fig. 5. Results for BDR, in the middle row, and iGMM, on the lowest row, in two complex scenes: Caviar 2004 on the left and PETS 2002 on the right [5]

We note that figure 2 second row presents some noises on the lower part of the image for most of the methods. These noises are due to camera motion.

Traffic monitoring: We compare the two best motion detection methods, BDR and iGMM, by applying four tracking algorithms in a specific task. We want to count vehicles for traffic monitoring purposes[1], see figure 5. We run the trackers and count all objects that are tracked for more than 30 frames, see table 1.

We note that iGMM offers better results than BDR on all tracking methods.

Computation time: As we see on table 2, we calculated the computation expenses of the various systems. We note that iGMM is faster on all methods except for particle filtering. We further calculate that the processing of a 704x576 image, on a Pentium M, 1.73 GHz, takes 0.05 to 0.08 second for iGMM and 0.3 to 0.4 second for BDR.

6 Conclusion

This paper presents various motion detection and object tracking methods. We compare our improved Gaussian mixture model with the other methods. Although Bayes decision rules offers accurate results, our model is more robust on complex scenes. Moreover, both methods are tested on outdoor sequences and our method offers better results. Finally our method is fast and allows real-time object tracking.

References

1. KaewTraKulPong, P., Bowden, R.: An improved adaptive background mixture model for real-time tracking with shadow detection. In: EWAVBSS (2001)
2. Korhonen, T., et al.: Particle filtering in high clutter environment. In: FINSIG (2005)
3. Li, L., et al.: Foreground object detection from videos containing complex background. ACM M, 2–10 (2003)
4. Morris, B., Trivedi, M.: A survey of vision-based trajectory learning and analysis for surveillance. IEEE Trans. Circuits Syst. Video Tech. 18(8) (2008)
5. Performance Evaluation of Tracking and Surveillance: pets2010.net/ (2010)
6. Sicre, R., Nicolas, H.: Human Behaviour Analysis and Event Recognition at a Point of Sale. In: PSIVT (2010)
7. Strauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR (1999)
8. Yilmaz, A., et al.: Object tracking: a survey ACM Computing Surveys (2006)
9. Zivkovic, Z., Heijden, F.V.D.: Efficient adaptive density estimation per image pixel for the task of background subtraction. Patt. Reco. Letters 27(7) (2006)

[1] ADACIS sarl and CETE sud-ouest provided the traffic monitoring sequences

Driver's Fatigue and Drowsiness Detection to Reduce Traffic Accidents on Road

Nawal Alioua[1], Aouatif Amine[1,2], Mohammed Rziza[1], and Driss Aboutajdine[1]

[1] LRIT, associated unit to CNRST, Faculty of Sciences,
Mohammed V-Agdal University, Rabat, Morocco
nawal.alioua@yahoo.fr, {rziza,aboutaj}@fsr.ac.ma
[2] ENSA, Ibn Tofail University, Kenitra, Morocco
amine_aouatif@univ-ibntofail.ac.ma

Abstract. This paper proposes a robust and nonintrusive system for monitoring driver's fatigue and drowsiness in real time. The proposed scheme begins by extracting the face from the video frame using the Support Vector Machine (SVM) face detector. Then a new approach for eye and mouth state analysis -based on Circular Hough Transform (CHT)- is applied on eyes and mouth extracted regions. Our drowsiness analysis method aims to detect micro-sleep periods by identifying the iris using a novel method to characterize driver's eye state. Fatigue analysis method based on yawning detection is also very important to prevent the driver before drowsiness. In order to identify yawning, we detect wide open mouth using the same proposed method of eye state analysis. The system was tested with different sequences recorded in various conditions and with different subjects. Some experimental results about the performance of the system are presented.

Keywords: Driver's hypovigilance, Fatigue detection, Drowsiness detection, Circular Hough Transform.

1 Introduction

Driver's hypovigilance is an important cause of traffic accidents, since it produces about 20% of these accidents [1]. The hypovigilance reduces capacity to react, judge and analyze information and it is often caused by fatigue and/or drowsiness. However, fatigue and drowsiness are different. The first one refers to a cumulative process producing difficulty to pay attention while the second one refers to the inability to stay awake. Therefore, it is important to monitor driver's vigilance level and issue an alarm when he is not paying attention.

Monitoring driver's responses, sensing physiological characteristics, driver operations, or vehicle responses, give a good indications of vigilance state. Driver operations and vehicle responses are studied by monitoring some vehicle parameters (steering wheel movement, vehicle speed, etc.). These are nonintrusive methods, but they are limited by vehicle type and driver condition. Monitoring driver's responses request sending periodically a response which is annoying.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 397–404, 2011.

The techniques based on sensing physiological phenomena are the most accurate. These techniques are implemented in two ways: measuring changes in physiological signals (brain waves, heart rate, eye blinking) and measuring physical changes (head orientation, eyes state, mouth state). The first one is not realistic since sensing electrodes must be attached on the driver's body which is distracting. The second one is nonintrusive and more suitable for real world driving conditions since it uses video cameras. Many research has been done on this last technique for extracting and analyzing facial features, especially the state of the eyes [2], head motion [3], or mouth motion [4]. The eye state is assumed to give a good indication of drowsiness level characterized by micro-sleep which is a short period (2-6 s) during which the driver rapidly closes the eyes and sleep. Many researchers use also the Percent of Eyelid Closure (PERCLOS) as a drowsiness indicator [1]. Others use the presence of the iris to determine if the eye is open [5]. To extract information about driver fatigue level, some authors consider that a high yawning frequency is a strong fatigue indicator [6]. In this work, we propose a system that permits to detect driver's vigilance level using two criteria. The first criterion allows detecting drowsiness by extracting micro-sleeps using the presence of the iris in the eyes. The second criterion detects fatigue using yawning frequency. Both methods are based on Circular Hough Transform (CHT) [7]. The organization of this paper is as follows. Section 2 explains the steps of proposed system. In Section 3, experimental results are exposed. Finally, conclusion is presented.

2 Proposed System

The aim of this study is to develop a closed-eye and opened-mouth detection algorithm aimed respectively at driver drowsiness and fatigue assessment. The present study is an extended work of [8]. The implemented methods are tested on real captured video sequences using low cost webcam. The proposed system performs some steps before determining driver's vigilance level. Firstly, the face is extracted from the video frames. Then, the localization of the eyes and the mouth is performed. Finally, we apply the proposed methods for detecting fatigue and drowsiness based on CHT.

2.1 Face Detection

The face is extracted from the frame to reduce the search region and therefore reduce the computational cost required for the subsequent step. We have used an existing method, based on SVM technique [9], developed by Kienzle [10].

2.2 Eyes and Mouth Localization

The reduced region where the eyes are situated is obtained for the same purpose as in the previous step. This reduction eliminates the confusion of the eyes with the mouth or the nose. We begin by applying gradient filter to highlight the edge. After that, horizontal projection is computed to detect upper and lower eye region boundaries. Next, we compute vertical projection on resulting image, which procures the right and the left limits of the face and separates the eyes. The same idea is adopted for mouth localization to reduce search region.

2.3 Driver's Drowsiness Analysis

This step detects micro-sleep periods in real time and issues immediately an alarm to avert the drowsy driver. To do this, we apply CHT on eye region images to identify iris. The eye is considered open if an iris is found. The CHT extracts circles from edge images. So, the obtained results depend on applied edge detector. Some classic edge detectors (Sobel, Prewitt, Roberts, Laplacian of Gaussian (LoG) and Canny) were tested to extract iris edge from eye region images. Unfortunately, the obtained edges did not provide the desired form, i.e. a kind of circular form referring to the iris. In order to solve this problem, we propose a new iris edge detector more suitable to the eye's morphology.

Iris Edge detector. Our iris edge detector respects the eye's morphology. If we observe an open eye, we see three main components. These are the *pupil* which is the little black circle in the center of eye surrounded by the *iris*, the circle distinguished by the eye color. The white outer area represents the *sclera.* This distinguished eye structure permits iris edge extraction from significant intensity variations between iris and sclera. Our iris edge detector considers only pixels x with grayscale intensity lower than an optimal threshold which must be chosen to handle only with pixels appertaining to iris. For each pixel x, a neighbourhood containing n pixels at left and right of x is specified. The difference between x and its n right and left neighbours is then computed.

- Left (*resp.* Right) edge: if n or $n-1$ left (*resp.* right) neighbours of x provide a difference with x higher than the high threshold and also if n or $n-1$ right (*resp.* left) neighbours of x provide a difference with x lower than the low threshold, we deduct that x is a left (*resp.* right) edge pixel of the iris and we put it at 1.
-Interpretation: In the case where the pixel appertains to the left edge, its left (*resp.* right) neighbours pixel's intensity is very high (*resp.* similar). Inversely, when the pixel appertains to the right edge, the right (*resp.* left) neighbours pixel's intensity is very different (*resp.* similar). So, the high threshold should distinguish the large difference between iris and sclera pixel's intensity. On the other side, the low threshold should respect the similarity between iris pixels. Fig.1 shows some examples of iris edge detection obtained by the proposed method compared to different classic edge detectors results. As can be seen, the classic edge detectors cannot provide good iris edge detection. For example, some edge components having circular form are detected in closed eye by classic edge detectors, while our edge detector did not identify such component.

Drowsiness detection based on eye state analysis. Once the appropriate iris edge detector is found, we apply the CHT on this edge to obtain the iris radius from which we decide if the eye is open. In the following, we present the CHT algorithm steps. At each iteration, three edge pixels are randomly chosen. If these pixels are not collinear and if the distance between each two pixels coordinates is higher than a fixed threshold, we compute the radius and the center coordinates of the candidate circle defined by these three pixels. If candidate circle parameters are between two specific thresholds for each parameter, they are assigned to an accumulator. After that, we compute the distance between the center and all edge pixels. If this distance is lower than a threshold, we increment the counter of the candidate circle pixels. If this counter is higher than a threshold, we consider that the candidate circle can represents the iris and we keep the other pixels as a new edge and we repeat the previous steps. The algorithm stops when the edge contains few pixels or when maximal iterations are reached. Since we need to detect the circle representing the iris, we select the one having the highest radius after the end of the algorithm. Driver's drowsiness is characterized by micro-sleep periods. So, we need to find the sleep intervals of at least 2 seconds. To reduce computational time, we analyze firstly the left eye state. If this eye is open, we proceed to fatigue detection (see section 2.4). If not, we analyze the right eye. If it is open, we switch to fatigue detection. Else, we increment the consecutive closed eye counter. We issue an alarm to avert the drowsy driver when the eyes remain closed for a certain period of time.

2.4 Driver's Fatigue Analysis

Driver's fatigue is characterized by a high yawning frequency. Because drowsiness occurs sometimes after fatigue, we decide to add this step to prevent the driver before micro-sleep. For this, we analyze the mouth state in order to find yawning. We assumes that yawning is a wide open mouth which lasts from 2 to 10 seconds. To detect yawning, we apply CHT on mouth edge images. As we have already seen in section 2.3, the well known edge detectors do not provide the desired results. So, we also propose an edge detector respecting the wide open mouth structure and similar to the previous one.

Wide open mouth edge detector. The proposed edge detector is based on the mouth structure. If the mouth is widely open, a large dark area having a circular form appears. If we replace in 2.3 the term "left" with the term "top" and "right" with "bottom", we obtain the wide open mouth edge detector. In the other terms, the pixels appertaining to the bottom edge have very different bottom neighbours and similar top neighbours and vice versa for the top edge pixels. Fig. 1 shows the wide open mouth edge detection obtained by the proposed method compared to several classic edge detectors results.

Fatigue detection based on mouth state analysis. The same algorithm (see section 2.3) is used to detect wide open mouth based on CHT. So, we

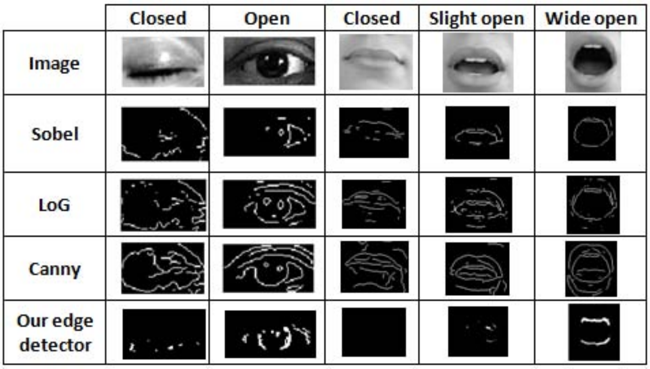


Fig. 1. Proposed and classic edge detectors for eye and mouth regions

select the circle having the highest radius after the end of the CHT algorithm. The considered driver's fatigue indicator is yawning which is an interval of wide open mouth of at least 2 seconds. To reduce computational time, we analyze the mouth state if only one of the two eyes is open. If the mouth is widely open, the counter of consecutive wide open mouths is incremented. When this counter exceeds the specified threshold, the yawning counter is incremented. Once this second counter is higher than a fixed threshold, the system indicated that the driver is suffering from fatigue, issuing warning signals. We also chose to analyze the mouth state if only one eye is open because a driver who is yawning having closed eyes does not pay attention to the road and he is consider drowsy.

Fig.2 illustrates our proposed system for Driver's fatigue and drowsiness detection.

3 Experimental Results

We conduct several tests on 18 real video sequences of 14 different subjects in various lighting conditions to validate the proposed system. All sequences are acquired by a low cost webcam connected to a laptop by USB port, providing images of 640x480 at 30 frames per second (fps). To meet the real-time constraints, we reduce the considered number of fps from 30 to 2. In the experiments, automatic detection of face, eyes and mouth has been integrated but not evaluated in this work. The main purpose of these integrations is to take them into account in assessing the runtime system. All experiments are done on laptop having Intel Core 2 Duo Processor. Results are presented as statistical measures such True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) and total number of samples (T). We also use Correct Classification Rate (CCR) and Kappa Statistic κ. Note that κ is a measure of non-random agreement degree between observations of the same categorical variable interpreted by Table 1. In Fig.3, 6 video sequences are considered to show examples of TP, FP, TN and FN of iris detection and wide open mouth detection. Table 2 presents the corresponding statistical measures. The first column of each statistic is the measure of

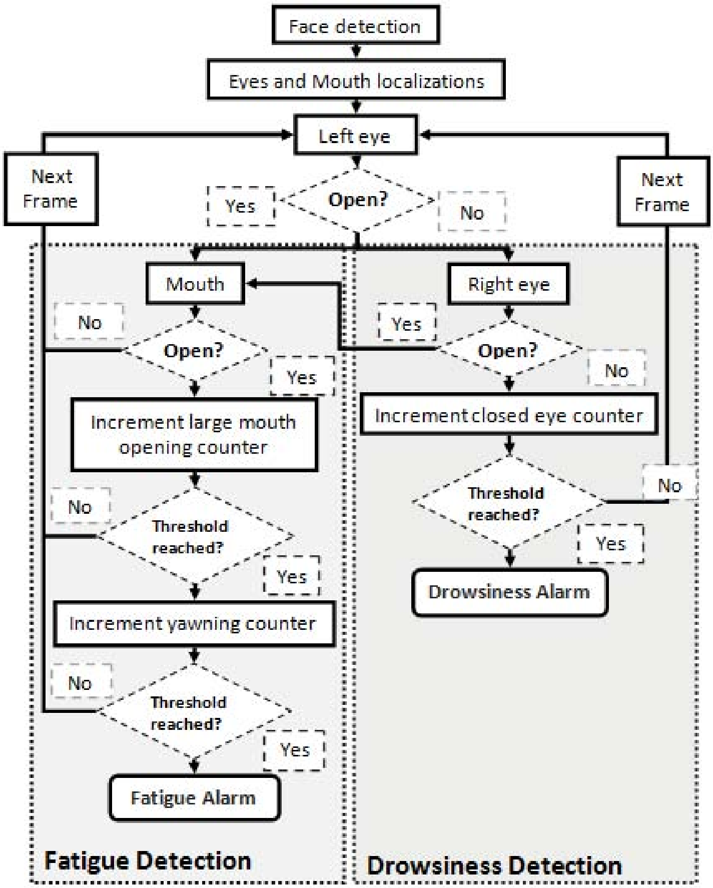


Fig. 2. Driver's fatigue and drowsiness detection system

iris detection and the second one is the measure of wide open mouth detection. The two last columns represent time in seconds. The *vid.D* refers to video duration and the *exec.T* refers to execution time of the whole system. From Table 2, the average (Avr.) of CCR is 94% and the average of κ is 86%. According to κ interpretation, the proposed system procures an almost perfect agreement between the observers. This means that both iris and wide open mouth detections permit assignation of the right classes in the most cases. After comparing the two lasts columns, we deduce that the system respect the real time constraints since execution time and video duration are almost the same. Thus we deduct that the proposed system can be used to provide a good and real-time estimation of the driver's vigilance state.

The last experiment exposes a comparison between our system and other existing systems of driver's hypovigilance detection. The system depicted in [5] is also based on CHT and uses 173 images of ORL database for evaluation, this system provides success accuracy rate of 90.3%. The second system presented

Table 1. Kappa statistic interpretation

Kappa Statistic	Interpretation
> 0.81	Almost perfect agreement
> 0.61 and < 0.8	Strong agreement
> 0.2 and < 0.6	Moderate agreement
> 0.0 and < 0.2	Poor agreement
< 0	Disagreement

Table 2. Statistical measures of fatigue and Drowsiness detections

V.	**TP**		**TN**		**FP**		**FN**		**T**		**CCR**		**Kappa**		**vid.D**	**exec.T**
1	103	32	44	70	0	0	3	0	150	102	0.98	1	0.95	1	61	57
2	108	32	80	70	0	1	11	5	199	108	0.94	0.94	0.87	0.86	72	66
3	68	6	30	58	0	0	2	3	100	67	0.98	0.95	0.95	0.74	41	46
4	82	24	24	54	0	0	2	5	108	83	0.98	0.93	0.94	0.83	47	54
5	57	12	29	45	2	0	14	2	102	59	0.84	0.96	0.66	0.88	40	46
6	69	71	31	0	4	0	11	0	115	71	0.87	1	0.71	1	44	48
									Avr.		0.94		0.86			

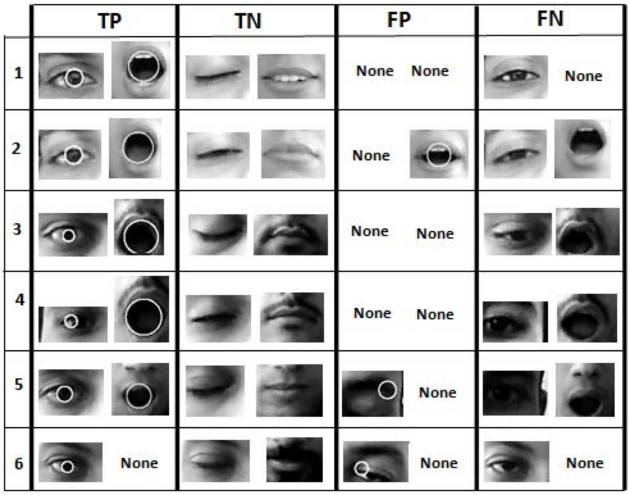


Fig. 3. Results of fatigue and drowsiness detections

in [2] uses 70 images taken with an infra-red camera and obtains a success rate of 90%. While the last system [6] detects fatigue using yawning analysis: the authors proposed a method to locate and track driver's mouth using cascade of classifiers. SVM is then used to train the mouth and yawning images. The tests are effectuated on real images providing a CCR of 81% for yawning detection. We deduct that our proposed system provides a high success rate comparing to the mentioned existing systems.

4 Conclusion

This paper presents a new approach of micro-sleep detection algorithm for drowsiness assessment and yawning detection algorithm for fatigue assessment, based on CHT. The whole driver's drowsiness and fatigue detection system uses three steps: face detection using SVM, eye/mouth region localization, closed eyes detection and wide open mouth detection. In the last step, we apply the CHT on our proposed edge detectors. With 94% accuracy and 86% of Kappa Statistic value, it is obvious that our driver's drowsiness and fatigue detection system is robust. As future works, we plan on one hand, to use multiple cameras in order to detect irises in various head orientations. On the other hand, we will introduce gaze tracking to detect driver's inattention, and to give us idea of where the driver is looking: at the dash display, at the roadside signs, or road ahead.

References

1. Bergasa, L., Nuevo, J., Sotelo, M., Vazquez, M.: Real-time system for monitoring driver vigilance. In: IEEE Intelligent Vehicle Symposium, pp. 78–83 (2004)
2. Hrishikesh, B., Mahajan, S., Bhagwat, A., Badiger, T., Bhutkar, D., Dhabe, S., Manikrao, L.: Design of drodeasys (drowsy detection and alarming system). Advances in computational algorithms and data analysis, 75–79 (2009)
3. Smith, P., Shah, M., Da Vitoria Lobo, N.: Monitoring head/eye motion for driver alertness with one camera. In: Proceedings of the International Conference on Pattern Recognition, pp. 636–642 (2000)
4. Mohanty, M., Mishra, A., Routray, A.: A non-rigid motion estimation algorithm for yawn detection in human drivers. International Journal of Computational Vision and Robotics 1, 89–109 (2009)
5. Tripathi, D.P., Rath, N.P.: A novel approach to solve drowsy driver problem by using eye-localization technique using CHT. International Journal of Recent Trends in Engineering (2009)
6. Saradadevi, M., Bajaj, P.: Driver fatigue detection using mouth and yawning analysis. IJCSNS International Journal of Computer Science and Network Security 6 (2008)
7. Duda, R.O., Hart, P.E.: Use of the hough transformation to detect lines and curves in picture. Commun. ACM, 11–15 (1972)
8. Alioua, N., Amine, A., Rziza, M., Aboutajdine, D.: Eye state analysis using iris detection to extract driver's micro-sleep periods. In: International Conference on Computer Vision Theory and Applications VISAPP (2011)
9. Burge, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 121–167 (1998)
10. Kienzle, W., Franz, M., Bakir, G., Scholkopf, B.: Face detection – efficient and rank deficient. Advances in Neural Information Processing Systems, 673–680 (2005)

Image Synthesis Based on Manifold Learning

Andrés Marino Álvarez-Meza[1], Juliana Valencia-Aguirre[1],
Genaro Daza-Santacoloma[2],
Carlos Daniel Acosta-Medina[1], and Germán Castellanos-Domínguez[1]

[1] Signal Processing and Recognition Group - Universidad Nacional de Colombia, Manizales, Colombia
{amalvarezme,jvalenciaag,cacostame,cgcastellanosd}@unal.edu.co,
[2] Faculty of Electronic Engineering, Universidad Antonio Nariño, Bogotá, Colombia
gdazas@unal.edu.co

Abstract. A new methodology for image synthesis based on manifold learning is proposed. We employ a local analysis of the observations in a low-dimensional space computed by Locally Linear Embedding, and then we synthesize unknown images solving an inverse problem, which normally is ill-posed. We use some regularization procedures in order to prevent unstable solutions. Moreover, the Least Squares-Support Vector Regression (LS-SVR) method is used to estimate new samples in the embedding space. Furthermore, we also present a new methodology for multiple parameter choice in LS-SVR based on Generalized Cross-Validation. Our methodology is compared to a high-dimensional data interpolation method, and a similar approach that uses low-dimensional space representations to improve the input data analysis. We test the synthesis algorithm on databases that allow us to confirm visually the quality of the results. According to the experiments our method presents the lowest average relative errors with stable synthesis results.

1 Introduction

In many real-life problems, we need to infer some observations that are not available in the dataset, which is usually known as data synthesis, as for example, being given a set of images from a rotating object to analyze from an unseen angle. This learning problem can be viewed as an approximation of an unknown function that maps from a parameter space to the sample space (e.g. image space). Therefore, a new sample is synthesized by learning an unknown function and computing it for a given reference. For this purpose, it can be used as an interpolation algorithm such as those based on radial basis functions, neural networks, statistical learning theory, splines, among others [1,2]. Particularly, splines are commonly used in fields as computer-aided design and computer graphics, because of their simplicity, accuracy of evaluation, and their capacity to approximate complex shapes [3]. Learning an unknown function directly from the image space demands a high-computational cost, because each dimension of the input data (pixel information) is analyzed. In addition these methods are sensitive to little changes in the original features (such as noise), computing unstable results.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 405–412, 2011.

In order to properly analyze and interpret high-dimensional data, several feature extraction techniques have been used. Nonlinear mapping algorithms based on manifold learning assume that the data lie on a manifold and aim to compute a low-dimensional space representation of the input data, discovering a small number of factors that suitably model the studied phenomenon. Particularly, LLE has been used to deal with high-dimensional data, even for large datasets [4], due to the fact that the optimization problem has an analytical solution avoiding local minima, and few free parameters need to be fixed by the user [5]. Thus, in [4] a low-dimensional representation of the available data is computed by LLE and the high-dimensional reconstructions of unseen observations are inferred using the two schemes. The first one employs a local analysis in the low-dimensional space to synthesize an unseen sample by solving an inverse problem. Nonetheless, the solution is unstable as it does not take into account an appropriate regularization process. The other scheme uses a regression method based on statistical learning theory, but it demands a high computational cost solving a regression problem for each variable in the image space.

In this sense, we propose a new methodology for image synthesis based on LLE. We analyze the samples in a low-dimensional space to identify the data underlying structure, and then we synthesize new samples as a local combination of their nearest neighbors, using a regularization procedure that ensures stable results. Our method employs Least Squares - Support Vector Regression (LS-SVR) to estimate unseen samples in low-dimensional space. We aim to discover the underlying structure of the data using LS-SVR rather than the traditional interpolation techniques, statistical methods or neural networks, in order to avoid the high number of free parameters to be fixed, and the overfitting. We also propose a method for automatic selection of the LS-SVR free parameters based on the Generalized Cross-Validation - (GCV). Thus, our algorithm of synthesis diminishes the relative synthesis errors, even for noisy conditions, and it employs automatic parameter tuning. The methodology is experimentally verified on three datasets that allow to visually confirm whether the high-dimensional data synthesis was correctly calculated. Also, the synthesis quality is measured using the average relative error among the targets and the predicted samples.

This paper is organized as follow. In Section 2, we present a brief description about LLE and LS-SVR, and we explain the proposed methodology for image synthesis based on manifold learning. In Section 3, we describe the experimental conditions and show the results obtained. Finally, in Sections 4 and 5 we discuss and conclude about the obtained results of this work.

2 Data Synthesis Based on Manifold Learning

2.1 Locally Linear Embedding

Let $\mathbf{X}_{n\times p}$ be the input data matrix with sample vectors $\{\mathbf{x}_i\}_{i=1}^{n}$ in $\Re^p$. The LLE algorithm maps $\mathbf{X}$ to a single global coordinate system represented by the matrix $\mathbf{Y}_{n\times m}$, with output sample vectors $\{\mathbf{y}_i\}_{i=1}^{n}$ in $\Re^m$, $(m < p)$. This algorithm makes a mapping that preserves the neighborhood relationships, assuming that the data are placed on a nonlinear manifold [6]. The algorithm has three principal steps. First, the k nearest neighbors per point as measured by Euclidean distance are found. Second, each point is represented as a weighted linear combination of its neighbors, minimizing $\varepsilon(\mathbf{W}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \sum_{j=1}^{n} w_{ij}\mathbf{x}_j\|^2$, subject to $w_{ij} = 0$, if $\mathbf{x}_j$ is not k-neighbor of $\mathbf{x}_i$, and $\sum_{j=1}^{n} w_{ij} = 1$. In third step, the low-dimensional embedding is calculated by minimizing $\Phi(\mathbf{Y}) = \sum_{i=1}^{n} \|\mathbf{y}_i - \sum_{j=1}^{n} w_{ij}\mathbf{y}_j\|^2$, subject to $\sum_{i=1}^{n} \mathbf{y}_i = \mathbf{0}$, and $\sum_{i=1}^{n} \mathbf{y}_i^\top \mathbf{y}_i/n = \mathbf{I}$.

2.2 Least Squares - Support Vector Regression with Multiple Parameter Choice

Given an output data matrix $\mathbf{Y}_{n\times m}$ and a reference vector $\mathbf{z}_{n\times 1}$, it is possible to learn the function $f(\mathbf{z}) = \mathbf{s}$ by means of statistical learning theory [1], where $\mathbf{s}_{n\times 1}$ is an univariate output vector that represents one coordinate of $\mathbf{Y}$. We construct m functions to interpolate each coordinate of $\mathbf{Y}$. It is important to note that m is a small number avoiding a high computational cost. We analyze the Least Squares version of Support Vector Machines regression oriented (LS-SVR) [7], which computes the function $f(\mathbf{z}) = \mathbf{v}^\top \varphi(\mathbf{z}) + b$, where $\mathbf{v}_{n\times 1}$ is a vector of weights, $b \in \Re$, and $\varphi(\cdot)_{n\times 1} : \Re \rightarrow \Re^n$. Let $\mathbf{K}_\sigma$ a gaussian kernel matrix, the LS-SVR solution can be obtained using the Lagrange theorem, thence, $\mathbf{1}b + \tilde{\mathbf{A}}\mathbf{a} = \mathbf{s}$, where $\tilde{\mathbf{A}} = \mathbf{K}_\sigma + C^{-1}\mathbf{I}$, $C \in \Re^+$ is a trade-off parameter between the training error and the system generalization, and $\mathbf{a}_{n\times 1}$ are the Lagrange multipliers with $\mathbf{1}^\top \mathbf{a} = 0$.

In LS-SVR training stage it is required to find suitable values for C and the kernel parameters (band-width σ for the gaussian kernel). In this sense, we use the GCV method [8], which finds a parameter γ in problems with the form $\mathbf{u}_\gamma^\varepsilon = \mathbf{A}_\gamma^\dagger \mathbf{r}^\varepsilon$, where $\mathbf{A}_\gamma^\dagger = (\mathbf{A} + \gamma\mathbf{I})^{-1}$ is an approximation of the generalized inverse of the matrix $\mathbf{A}$, $\mathbf{r}^\varepsilon$ is a disturbed sample, and $\mathbf{u}_\gamma^\varepsilon$ is the solution. Looking for a balance between the regularization error and the disturbance in the solution, we propose to compute the optimum values for C and σ by minimizing

$$\vartheta(\sigma, C) = \frac{\|\mathbf{K}_\sigma \mathbf{a}_{\sigma,C} - \mathbf{b}_{\sigma,C}\|^2}{tr\left(\mathbf{I} - \mathbf{K}_\sigma\left(\mathbf{K}_\sigma + C^{-1}\mathbf{I}\right)^{-1}\right)^2}, \tag{1}$$

where $\mathbf{b}_{\sigma,C} = \mathbf{s} - \mathbf{1}b$, and $\mathbf{a}_{\sigma,C} = \tilde{\mathbf{A}}^{-1}\mathbf{b}_{\sigma,C}$. Then, for a reference z, $f(\mathrm{z})$ can be rewriting as $f(\mathrm{z}) = \sum_{i=1}^{n} \mathbf{a}^\top \mathbf{K}_\sigma(\mathrm{z}_i, \mathrm{z}) + b$.

2.3 Regularized High-Dimensional Data Synthesis - RHDDS

Using the LS-SVR learning function with automatic parameter selection (minimizing (1)), a new reference z_{new} can be mapped to $\mathbf{Y}$ computing the projection $\mathbf{y}_{\text{new}} = f(\mathrm{z}_{\text{new}})$. We aim to take advantage of the small number of dimensions of the space $\mathbf{Y}$, in order to reduce the computational cost for the synthesis procedure. In this sense, we synthesize unknown samples only considering the local properties of the embedding space, thus, our synthesis is less sensitive to little changes of the original features. Therefore, we represent $\mathbf{y}_{\text{new}}$ as a combination of its k nearest neighbors $\boldsymbol{\eta}$, and then we minimize

$$\varepsilon_y\left(\mathbf{w}_y\right) = \|\mathbf{y}_{\text{new}} - \sum\nolimits_{j=1}^{k} w_{y_j}\boldsymbol{\eta}_j\|^2 \quad \text{s.t} \quad \sum\nolimits_{j=1}^{k} w_{y_j} = 1. \tag{2}$$

So, the Gram matrix $\mathbf{G}$ is calculated, with $\{G_{jl}\}_{j,l=1}^{k} = \langle\left(\mathbf{y}_{\text{new}} - \boldsymbol{\eta}_j\right), \left(\mathbf{y}_{\text{new}} - \boldsymbol{\eta}_l\right)\rangle$. Then, rewriting (2) as $\varepsilon_y = {\mathbf{w}_y}^{\top}\mathbf{G}\mathbf{w}_y$, the $\mathbf{w}_y$ vector that minimize ε_y can be computed using the Lagrange theorem as $\mathbf{w}_y = \left(\lambda/2\right)\mathbf{G}^{-1}\mathbf{1}$, with $\lambda = 2/\left(\mathbf{1}^{\top}\mathbf{G}^{-1}\mathbf{1}\right)$. However, in this case $\mathbf{G}$ is singular (or close), being necessary a regularization process to find an appropriate solution. In [4] a similar problem is analyzed, and they suggest to find $\mathbf{w}_y$ as

$$\mathbf{w}_y = \frac{\lambda}{2}\mathbf{G}^{-1}\mathbf{1} = c\,\mathbf{G}^{-1}\mathbf{1}, \tag{3}$$

where c is set to 1, and $\mathbf{w}_y$ is scaled to sum 1. Nevertheless, note that the proposed solution does not consider the ill-posed condition of $\mathbf{G}$, which may lead to wrong results. Consequently, we analyze two methodologies to improve the inverse problem solution considering the regularized version of $\mathbf{G}$. First, in [6] is proposed to calculate $\mathbf{G}$ as $G_{jl} \leftarrow G_{jl} + \alpha_1$, where

$$\alpha_1 = \delta_{jl}\left(\mathit{\Delta}^2/k\right)\operatorname{tr}\left(\mathbf{G}\right), \tag{4}$$

being $\mathit{\Delta}^2 \ll 1$ (usually $\mathit{\Delta} = 0.1$). It is important to note that the optimal value for this parameter can vary over a wide range and depends on the particular application. The second methodology, presented in [5], is an automatic regularization process for $\mathbf{G}$, which allows to obtain consistent solutions even for data variations caused by noise or sample randomness. For this purpose, an error optimization problem is formulated as $\varepsilon_{y_{reg}} = \mathbf{w}_y^{\top}\mathbf{G}\mathbf{w}_y + \alpha_2^2\mathbf{w}_y^{\top}\mathbf{w}_y$, subject to $\sum_{j=1}^{k} w_{y_j} = 1$. Using the Lagrange multipliers

$$\begin{bmatrix} \mathbf{w}_y \\ -\lambda \end{bmatrix} = \begin{bmatrix} \frac{1}{\mathbf{1}^{\top}\left(2\left(\mathbf{G}+\alpha_2^2\mathbf{I}\right)\right)^{-1}\mathbf{1}}\left(2\left(\mathbf{G}+\alpha_2^2\mathbf{I}\right)\right)^{-1}\mathbf{1} \\ -\frac{1}{\mathbf{1}^{\top}\left(2\left(\mathbf{G}+\alpha_2^2\mathbf{I}\right)\right)^{-1}\mathbf{1}} \end{bmatrix}, \tag{5}$$

as a result, the desired regularization parameter α_2 minimizes the function $g\left(\alpha_2\right) = \|\mathbf{w}_y\|^2 + \lambda^2$, and $\mathbf{G}$ is calculated as $\mathbf{G} = \mathbf{G} + \alpha_2^2\mathbf{I}$.

Finally, given the vector $\mathbf{w}_y$, the high-dimensional projection $\mathbf{x}_{\text{new}}$ of $\mathbf{y}_{\text{new}}$ can be computed as $\mathbf{x}_{\text{new}} = \mathbf{w}_y\boldsymbol{\varPsi}$, where $\boldsymbol{\varPsi}$ is the projection of $\boldsymbol{\eta}$ in the high-dimensional space. The proposed methodology for data synthesis can be summarized as in Figure 1.

3 Experiments

The proposed methodology is tested on three real-world datasets. Moreover, the synthesis quality is measured using the average relative error (ARE) between the targets and the synthesized samples.

$$\text{ARE}\left(\hat{\mathbf{X}}\right) = 100\frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}\frac{\|\hat{\mathbf{x}}_i - \mathbf{x}_i\|^2}{\|\mathbf{x}_i\|^2}[\%], \tag{6}$$

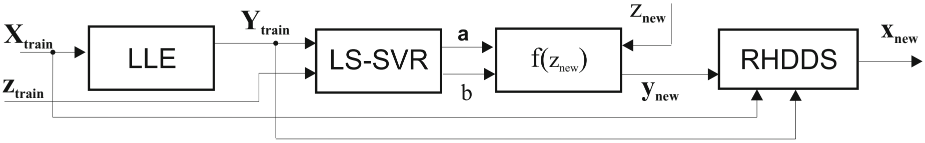


Fig. 1. Block diagram for image synthesis based on manifold learning

where $\hat{\mathbf{x}}$ is a synthesized sample, $\mathbf{x}$ is the original observation, and n_{test} is the size of the test set. We employ a 10-fold cross validation analysis to determinate the algorithms performance. Besides, the performance of the methodology is compared using both of the proposed regularization procedures against the state of the art found in [4], and a traditional direct interpolation method in the high-dimensional space based on cubic splines [2,3]. Hence, we obtained four possible synthesis results: $\hat{\mathbf{X}}_{\alpha_1}$ and $\hat{\mathbf{X}}_{\alpha_2}$ using the regularization equations (4) and (5) (our approach), $\hat{\mathbf{X}}_c$ employing the scaling equation (3) (similar approach), and $\hat{\mathbf{X}}_s$ by means of cubic splines (direct interpolation method).

The databases are: CMU MoBo (Gait) [9], COIL-100 (Maneki Neko) [10], and CMU/VASC (Hand) [11]. The first database holds 25 individuals walking a treadmill. For concrete testing, we used the slow walking sequence of one person captured from a side view (camera $vr03\_7$). The images in JPG format are resized to 80×61 and mapped to a gray scale space. We reduced the number of frames to obtain one gait cycle. The second dataset contains 72 RGB-color images of 100 rotating objects in PNG format. We use the gray scale images subsampled to 64×64 pixels for the Maneki Neko. Finally, the third database contains 481 gray-scale pictures in PNG format related with a hand holding a rice bowl. We subsampled the images to 72×77 pixels. Moreover, we test perturbing all the three datasets with Gaussian noise $\sim N(0, 0.03)$, in order to verify the synthesis robustness. The number of nearest neighbors k for LLE is chosen as in [12], which computes an specific number of neighbors for each input point. It uses graph theory and geodesic distance to analyze the density and linearity of each manifold patch. The dimension m of the embedding space is fixed according to [13], considering a local analysis of variance in the high-dimensional space to conserve the expected signal to noise ratio in the embedding space. Further, the reference vector $\mathbf{z}$ for Gait and Hand is set as $\mathbf{z} = [1, ..., n]$ (image position in the sequence), and for Maneki Neko as $\mathbf{z} = [0, 5, ..., 355]$ (rotation angle).

4 Results and Discussion

According to the ARE for the synthesis results shown in Table 1, it is possible to notice that our approach exhibits a better performance than the other methods found in the state of the art. Moreover, the proposed synthesis methodology using the automatic regularization technique (5) shows better results ($\hat{\mathbf{X}}_{\alpha_2}$) than when the empirical regularization (4) is set ($\hat{\mathbf{X}}_{\alpha_1}$), because the former can adapt its values depending on the particular conditions of the samples (noise

Table 1. Average Relative Error Results (target: ARE = 0)

Dataset	ARE($\hat{\mathbf{X}}_{\alpha_2}$)[%]	ARE($\hat{\mathbf{X}}_{\alpha_1}$)[%]	ARE($\hat{\mathbf{X}}_{c}$)[%]	ARE($\hat{\mathbf{X}}_{s}$)[%]
Gait	$\mathbf{0.58 \pm 0.14}$	0.92 ± 0.58	$9.05e7 \pm 2.14e6$	5.35 ± 9.46
Noisy Gait	$\mathbf{2.42 \pm 0.32}$	13.61 ± 8.63	$6.72e7 \pm 2.12e8$	151.07 ± 347.70
Maneki Neko	$\mathbf{2.35 \pm 0.97}$	2.43 ± 1.65	$1.20e11 \pm 3.79e11$	57.43 ± 157.02
Noisy Maneki Neko	$\mathbf{7.16 \pm 1.57}$	12.75 ± 3.78	$1.19e15 \pm 3.73e15$	66.99 ± 50.15
Hand	$\mathbf{0.75 \pm 0.14}$	0.95 ± 0.87	$4.30e13 \pm 6.64e14$	$\mathbf{0.44 \pm 0.80}$
Noisy Hand	$\mathbf{4.41 \pm 0.51}$	24.10 ± 7.68	$1.42e17 \pm 3.67e17$	92.99 ± 112.70

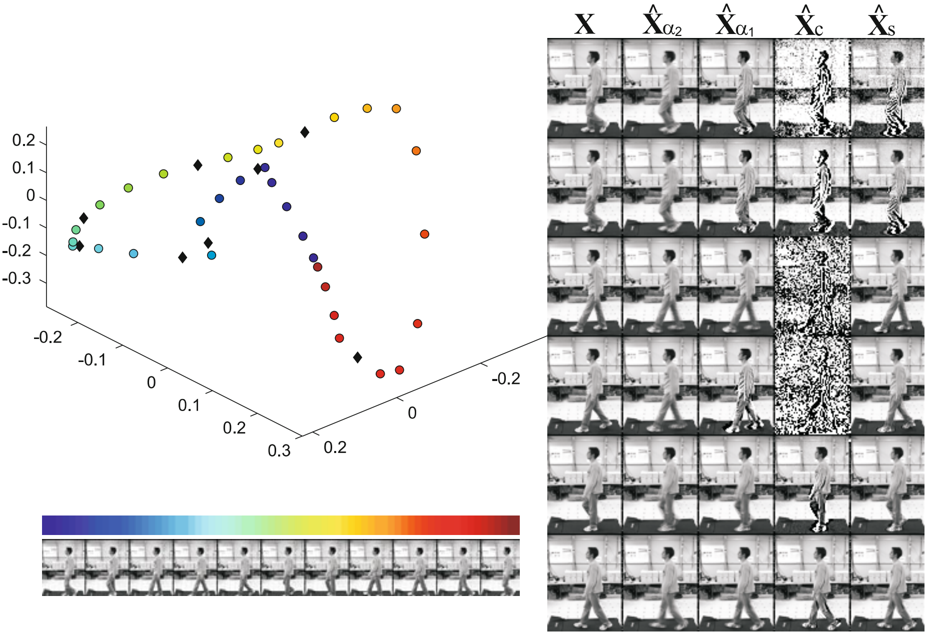


Fig. 2. Gait Results ($n = 42, p = 4880, m = 3$). *Left*: training images and embedding space, *color points*: training samples projected by LLE, *black diamonds*: interpolated samples using LS-SVR. *Right*: Examples of target and synthesized images.

and randomness), the latter method is high-sensitive to noise perturbations of the dataset. These results are visually confirmed by Figures 2, 3, and 4. The synthesis results obtained with cubic splines ($\hat{\mathbf{X}}_s$) are unappropriated, because the behavior of the high-dimensional data interpolation methods is sensitive to little variations in the features of the data. For this reason, even though some visual results are according to the expectancies, in most of the cases the algorithm does not identify the features variability, specially when the available data is perturbed with noise. This can be corroborated by the high average relative error and the high standard deviation obtained by $\hat{\mathbf{X}}_s$. Also, the synthesized samples $\hat{\mathbf{X}}_c$ are not accurate, because the methodology proposed in [4] only takes

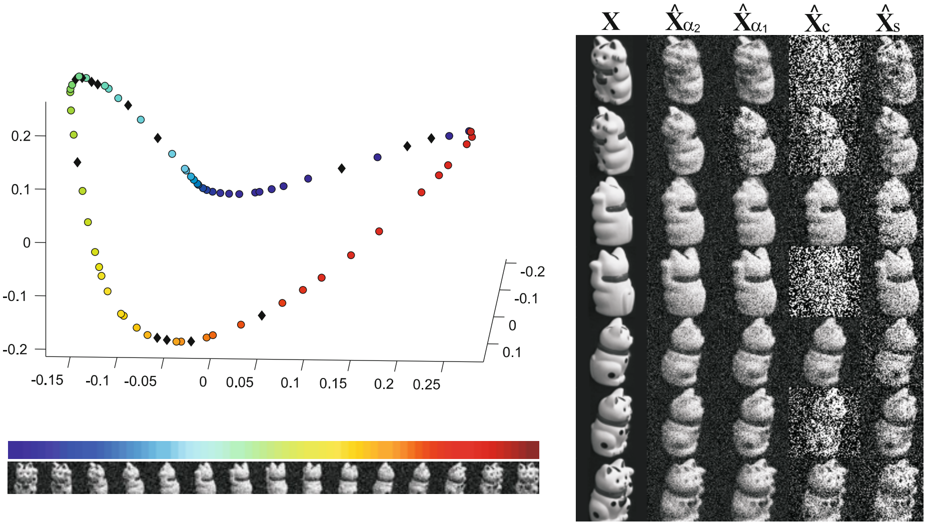


Fig. 3. Noisy Maneki Neko Results ($n = 72, p = 4096, m = 3$). *Left*: training images and embedding space, *color points*: training samples projected by LLE, *black diamonds*: interpolated samples using LS-SVR. *Right*: Examples of target and synthesized images.

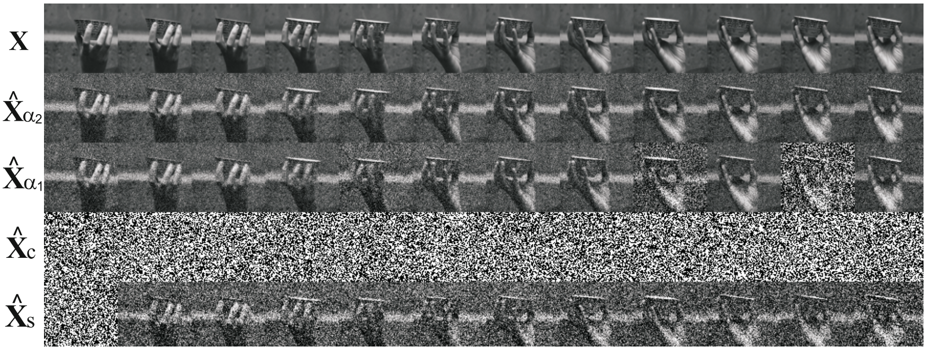


Fig. 4. Examples of target and synthesized images for Noisy Hand database ($n = 481, p = 5544, m = 6$)

into account a scale parameter to improve the inverse problem solution, and it does not consider the ill-posed condition of $\mathbf{G}$. Thence, some reconstructed images exhibit poor quality and high average relative errors. Finally, according to the embedding spaces shown in Figures 2, 3, and the ARE presented in Table 1, the LS-SVR interpolation together with the proposed approach for multiple parameter selection show be an efficient algorithm.

5 Conclusion

In this paper a methodology for image synthesis based on manifold learning and LS-SVR was presented. Our approach improves the synthesis performance by diminishing the perturbation caused by the feature variability among objects. Moreover, it takes into account the ill-condition characteristic of the inverse problem for image reconstruction, incorporating a regularization process. Additionally, an algorithm for multiple parameter choice in LS-SVR was proposed, which does not need prior knowledge to properly identify the underlying data structure of a given manifold. According to the attained results our work outperformed a previous similar approach [4], and cubic spline interpolation [3]. Particularly, the proposed methodology was tested on noisy datasets, showing the best performance in all cases. Otherwise, the methods found in the state of the art were quite sensitive to perturbations on the intensity of the pixels, being incapable of producing suitable synthesized images. Thus, our work seems to be adequate for real-time applications, which will be interesting for future studies.

Acknowledgments. This research was carried out under grants provided by ARTICA, a master thesis support project funded by U. Nacional de Colombia, and the project "*Análisis de deficiencias de patrones básicos de movimiento empleando técnicas de extracción no lineal de características*" funded by U. de Caldas and U. Nacional de Colombia.

References

1. Scholkopf, B., Smola, A.J.: Learning with Kernels. The MIT Press, Cambridge (2002)
2. de Boor, C.: A Practical Guide to Splines. Springer, Heidelberg (2005)
3. Shih, F., Fu, C., Zhang, K.: Multi-view face identification and pose estimation using b-spline interpolation. Information Sciences 169, 189–204 (2005)
4. Zhang, C., Wang, J., Zhao, N., Zhang, D.: Reconstruction and analysis of multipose face images based on nonlinear dimensionality reduction. Patter Recognition 37(3), 325–336 (2004)
5. Daza-Santacoloma, G., Acosta-Medina, C.D., Castellanos-Dominguez, G.: Regularization parameter choice in locally linear embedding. Neurocomputing 73, 1595–1605 (2010)
6. Saul, L., Roweis, S.: Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Machine Learning Research 4, 119–155 (2003)
7. Suykens, J.A.K., Gestel, V.T., Brabanter, J.D., Moor, B.D., Vandewalle, J.: Least squares support vector machines. World Scientific, Singapore (2002)
8. Hansen, C., Nagy, J., Oleary, D.: Deblurring Images: Matrices, Spectra, and Filtering. Society for Industrial and Applied Mathematics, Philadelphia (2006)
9. Gross, R., Shi, J.: The CMU motion of body database. Carnegie Mellon University, Tech. Rep. (2001)
10. Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library: Coil-100. Columbia University, Tech. Rep. (1996)
11. Wang, C.: CMU/VASC database. Carnegie Mellon University, Tech. Rep. (2006)
12. Álvarez-Meza, A., Valencia-Aguirre, J., Daza-Santacoloma, G., Castellanos-Domínguez, G.: Global and local choice of the number of nearest neighbors in locally linear embedding (minor revision). Patter Recognition Letters (2010)
13. de Ridder, D., Kouropteva, O., Okun, O., Pietikäinen, M., Duin, R.P.W.: Supervised locally linear embedding. International Conference on Artificial Neural Networks (2003)

Hierarchical Foreground Detection in Dynamic Background

Guoliang Lu[1], Mineichi Kudo[2], and Jun Toyama[3]

Graduate School of Information Science and Technology
Hokkaido University, Sapporo, 060-0814, Japan
{luguoliang,mine,jun}@main.ist.hokudai.ac.jp

Abstract. Foreground detection in dynamic background is one of challenging problems in many vision-based applications. In this paper, we propose a hierarchical foreground detection algorithm in the HSL color space. With the proposed algorithm, the experimental precision in five testing sequences reached to 56.46%, which was the best among compared four methods.

Keywords: foreground detection, dynamic background, HSL color space.

1 Introduction

Robust detection of foreground of interest is a fundamental and important module in many vision-based applications such as video surveillance, tracking, traffic monitoring. Since no prior information is available about potential foreground, in most of applications foreground is detected by subtracting the newly observed frame from the reference background modeled from a preceding sequence of background images. Such a simple subtraction algorithm works well for stationary background. However, one also has to deal with non-stationary background (dynamic background) such as water ripples, swaying trees, fluttering flags and so on [1,2,3]. Certainly, there are other practical problems [4] to be considered such as light switch, bootstrapping and foreground aperture. In this paper, however, we focus on foreground detection in dynamic background.

Background subtraction methods are pixel-based originally. Several techniques such as a mixture of Gaussians [5] and Kalman filter [6] are employed to model the background on the pixel basis. However, such a pixel-based background subtraction is sensitive to the change of background. In other words, pixel-based methods are in principle applicable to stationary background only. To deal with time-varying background, block-based background subtraction is proposed to exploit information of neighbors of each pixel [3,7]. Compared with pixel-based subtraction, block-based background subtraction is insensitive to dynamically changing background and is also more beneficial at computation cost. The detection performance is satisfactory for applications of coarse-level in dynamic background. For obtaining a better performance even in fine-level, hierarchical methods have been proposed in recent years [2,8]. For example, Chen et al.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 413–420, 2011.

proposed an efficient hierarchical method by combining pixel-based and block-based approaches into a single framework [2].On the other hand, recent studies use other information than RGB color or gray-level values to improve the robustness of background model, such as texture [7], depth information [9].

On the basis of these studies, in this paper we propose a robust hierarchical method for foreground detection in dynamic background using HSL color values, as shown in Fig.1. Firstly the newly observed frame is examined on the basis of blocks by using color and texture information in coarse-level and then pixels in detected foreground blocks are classified into foreground or background on the basis of clustering using hue and lighting values in fine-level. Contributions of this work are: 1) We propose a hierarchical method to detect foreground in dynamic background where an image is represented by three components of HSL color space; 2) In the coarse-level process, color and texture information is used for robust block-wise foreground detection; and 3) In the fine-level process, hue and lighting information is used for precise pixel-wise foreground detection.

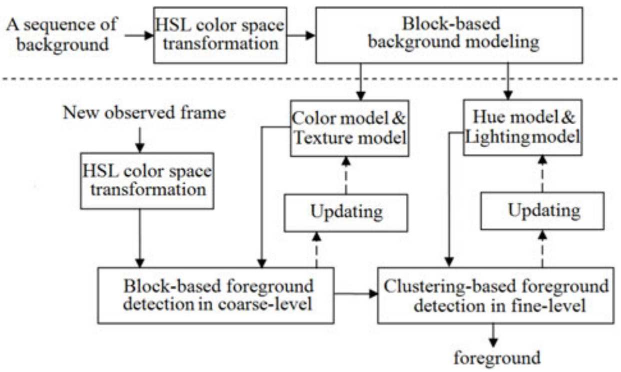

A sequence of background
HSL color space transformation
Block-based background modeling
New observed frame
Color model & Texture model
Hue model & Lighting model
HSL color space transformation
Updating
Updating
Block-based foreground detection in coarse-level
Clustering-based foreground detection in fine-level
foreground

Fig. 1. Structure of our proposed approach

Organization of this paper is given as: Section 1 reviews the background subtraction based foreground detection. Our proposed approach is presented in Sections 2 and 3. In Section 4, the experimental results are shown with those of some state-of-the-art methods. Finally, discussion and conclusion are given in Sections 5 and 6.

2 Block-Based Foreground Detection in Coarse-Level

In the coarse-level, foreground detection is made on the block basis. Each block of a newly observed frame is compared with the block of reference background at the same position that is modeled from a preceding sequence of background images. In the background modeling, color and texture information is widely employed [3,7,9]. As a texture descriptor, Local Binary Pattern (LBP) has been proven its satisfactory performance for texture recognition in gray level [7]. However it is not sufficient to represent a dynamic image block for which another descriptor of LBP is proposed recently [2,10]. In our background modeling, we reconstruct contrast histogram [10] and combine it with a color descriptor.

2.1 Color Description of an Image Block

Unlike three components of RGB color space, the components of hue (H), saturation (S) and lighting (L) of HSL color space are said to be independent to each other. Furthermore, the values of H and S are stably obtained in different conditions of lighting [11]. We, therefore, use HSL color space and denote the color information of a pixel by (h, s, l), more concretely, that of the ith pixel in the cth block B_c is reprecented by (h_i, s_i, l_i). Then we give a color description of block B_c by a 4-tuple $C(B_c) = [\mu_c^H, \sigma_c^H, \mu_c^S, \sigma_c^S]^T$, where μ_c^H is the mean value of H components of block B_c and σ_c^H is the standard derivation; μ_c^S and σ_c^S are those in S components.

2.2 Texture Description of an Image Block

The component L of HSL color space contains rich texture information. Unlike an extension of contrast histogram proposed by Chen et al. [2], in which the contrast histogram is constructed by cross-contrast values in RGB color space, we construct the contrast histogram in L component of HSL color space.

Contrast value c_i of the ith pixel in B_c of $N \times N$ pixels is calculated [10] as:

$$c_i = l_i - \mu_c^l \tag{1}$$

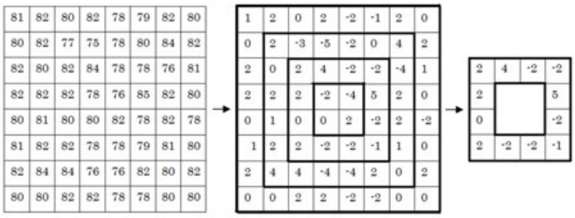
where μ_c^l is the mean value of L components of B_c.

Then, this block is divided into square bands with maximum distance q of one to $\lfloor \frac{N}{2} \rfloor$ shown in Fig.2, and the positive contrast histogram value CH_q^+ and negative contrast histogram value CH_q^- are calculated in each square band:

$$CH_q^+ = \frac{\sum\{c_i | c_i \in q,\ c_i > 0)\}}{m_q^+},\ CH_q^- = \frac{\sum\{c_i | c_i \in q,\ c_i < 0)\}}{m_q^-} \tag{2}$$

where $q = 1, 2, ..., \lfloor \frac{N}{2} \rfloor$ and m_q^+ (m_q^-) is the number of pixels with positive (negative) contrast value in square band q.

Finally, the text description $T(B_c)$ of block B_c is obtained by integrating both of positive and negative contrast histogram values in each of square bands, that is, $T(B_c) = \{CH_1^+(B_c), CH_1^-(B_c), ..., CH_{\lfloor \frac{N}{2} \rfloor}^+(B_c), CH_{\lfloor \frac{N}{2} \rfloor}^-(B_c)\}^T$. Note that our contrast histogram descriptor is almost invariance to rotation, which is valuable nature as a descriptor of dynamic background. Besides, this descriptor has a smaller number of features than those of the descriptor of Chen [2]: $T(B_c)$ has eight features for a block of 8×8 pixels, but the number of features necessary for the descriptor of [2] is 48.



81	82	80	82	78	79	82	80
80	82	77	75	78	80	84	82
82	80	82	84	78	78	76	81
82	82	82	78	76	85	82	80
80	81	80	80	82	78	82	78
81	82	82	78	78	79	81	80
82	84	84	76	76	82	80	82
80	80	82	82	78	78	80	80

1	2	0	2	-2	-1	2	0
0	2	-3	-5	-2	0	4	2
2	0	2	4	-2	-2	-4	1
2	2	2	-2	-4	5	2	0
0	1	0	0	2	-2	2	-2
1	2	2	-2	-2	-1	1	0
2	4	4	-4	-4	2	0	2
0	0	2	2	-2	-2	0	0

2	4	-2	-2
2			5
0			-2
2	-2	-2	-1

Fig. 2. From left to right: an image block of 8×8 pixels with L values; contrast values in 5 square bands enclosed by bold lines; values of positive and negative contrast of the 2th square band producing $CH_2^+ = \frac{2+4+5+2+2}{5} = 3$ and $CH_2^- = \frac{-2-2-2-1-2-2}{6} = -1.833$.

2.3 Background Modeling and Updating

On the basis of the color description $C(B_c)$ and the texture description $T(B_c)$, the preceding sequence of this block B_c is used for building a 'Color' model $M_c(B_c)$ and a 'Texture' model $M_t(B_c)$. As a model, we employ a mixture of Gaussians (MoG) with K components [5]. Note that an MoG is independently prepared in each block. At time $t+1$, the probability of feature X_{t+1} ($X_{t+1} = C(B_c)$ or $T(B_c)$) is considered as:

$$P(X_{t+1}) = \sum_{i=1}^{K} \omega_{i,t} \cdot N(X_{t+1}; \mu_{i,t}, \Sigma_{i,t}) \tag{3}$$

where $N(X; \mu, \Sigma) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(X-\mu)^T \Sigma^{-1} (X-\mu)\}$, with $\sum_{i=1}^{K} \omega_i = 1$, $p = 4$ (for $C(B_c)$) and $p = 8$ (for $T(B_c)$). In addition, K is set to 3 and Σ is assumed to be the form of $\Sigma = \sigma^2 I$ for computational reasons [5].

Given a newly observed frame at block B_c, when the block is classified into background, $M_c(B_c)$ and $M_t(B_c)$ are updated with X_{t+1}, respectively, as follows:

$$\omega_{i,t+1} = (1-\alpha)\omega_{i,t} + \alpha M_{i,t} \tag{4}$$

where $M_{i,t}$ is 1 for the component of $M_c(B_c)$ or $M_t(B_c)$ with the highest value of likelihood to the block B_c (0 for the remaining), and $\{\mu, \sigma\}$ of the component are updated as:

$$\mu_{i,t+1} = (1-\rho)\mu_{i,t} + \rho X_{t+1} \tag{5}$$

$$\sigma_{i,t+1}^2 = (1-\rho)\sigma_{i,t}^2 + \rho(X_{t+1} - \mu_{i,t+1})^T (X_{t+1} - \mu_{i,t+1}) \tag{6}$$

where α is a learning rate and $\rho = \alpha N(X_{t+1}; \mu_{i,t}, \Sigma_{i,t})$. With these updating formulae, we keep $M_c(B_c)$ and $M_t(B_c)$ as the latest. Furthermore, weights $\{\omega_{i,t+1}\}$ of K Gaussians in $M_c(B_c)$ or $M_t(B_c)$ are normalized to ensure $\sum_{i=1}^{K} \omega_{i,t+1} = 1$.

2.4 Foreground Detection in Coarse-Level

In detection of foreground in coarse-level, for a newly observed block B_c, color description $C(B_c)$ and texture description $T(B_c)$ are extracted and then compared with background model $M_c(B_c)$ and $M_t(B_c)$ to measure the likelihood as background, respectively. If at least one component gives a higher value of likelihood than a threshold θ in $M_c(B_c)$ and $M_t(B_c)$, that is, when there exists i such that $\sqrt{(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)} \leq \theta$, the block is judged as background. Here, θ is defined as three times the standard deviations in one component.

3 Foreground Detection in Fine-Level

It is known that if low saturation values are dominant in an image then many pixels are hardly distinguished even if they have a wide range of hue values. On the contrary, if an image has high saturation values, then even a small difference of hue values is detectable. Thus a color can be approximated by the gray value (intensity level) or the hue value [12]. In the fine-level process of our approach, pixels in the blocks judged as foreground in the preceding coarse-level are classified according to their hue and lighting values and the final decision is made by integrating respective classification results in hue and lighting components.

3.1 Background Modeling by Using Hue and Lighting Values

In fine-level, a 'Hue' model $M_h(B_c)$ of block B_c is learned as an MoG with K components from the hue values of all pixels seen in the preceding sequence of B_c. Similarly, a 'Lighting' model $M_l(B_c)$ is learned. K is set to 3 in our experiments.

3.2 Pixels Classification in Fine-Level

In our approach, pixel-based classification is carried out as follows:

a) Classification by hue value. For the block B_c judged as foreground in preceding coarse-level, all pixels in it are collected and are clustered in hue values by k-means algorithm. Then each cluster is compared with the model $M_h(B_c)$. If the proximity is close enough to a component Gaussian of $M_h(B_c)$, all pixels belonging to this cluster are assigned to 'background' and the Gaussian is updated by formulae (4)-(6), otherwise the cluster is judged as 'foreground'. The proximity is judged in binary if the centre of a cluster lies within three times the standard deviations of a component Gaussian of $M_h(B_c)$ or not.

b) Classification by lighting value. Similarly, pixels are classified according to model $M_l(B_c)$ in the lighting values.

c) Final decision. If a pixel is classified to 'foreground' in the hue value or in the lighting value or both, the final decision of the pixel is 'foreground'.

4 Experiment

4.1 Testing Sequences and Evaluation Measurement

Our approach focuses on modeling the dynamic background with random spatial movements of background elements. As datasets meeting this hypothesis, we used five sequences {*Fountains*, *Waving leaves*, *Ocean waves*, *Waving river*,

$Moving\ escalator\}$ [1,2]. We used the same training and testing frames as Ref. [2] in our experiments. Some previous studies used a pair of false negative and false positive to evaluate the performance of algorithms, but Jaccard index is chosen in this paper as a similarity $S(A, B)$ between the binary detected foreground A and the corresponding ground truth B on the pixel basis:

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

4.2 Experimental Results

The size of a block is an importance factor in our approach. A larger size of block gives more robust classification against noise, but a smaller size gives a better precision and resolution. In our experiments, we tested block size of 8 × 8, 10 × 10, 14 × 14 and 20 × 20, respectively, in the five testing sequences.

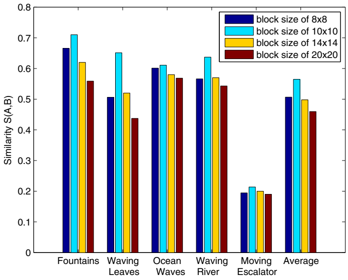


Fig. 3. Detection performance with blocks of different sizes

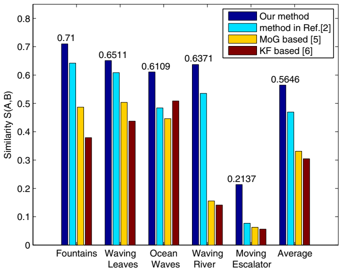


Fig. 4. Comparative results of four methods in five testing sequences

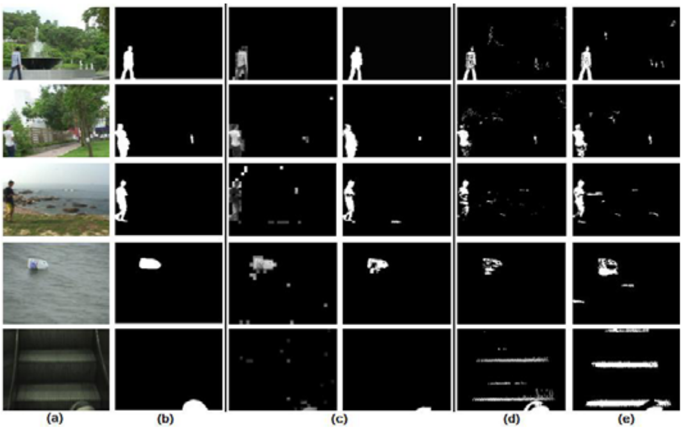


Fig. 5. (a)Example frames, from top to bottom: Fountains, Waving leaves, Ocean waves, Waving river, Moving escalator. (b)The ground truth. (c) Results by proposed approach in coarse and fine level. (d) MoG based [5]. (e)KF based [6].

Experimental results shown in Fig.3 reveal that the proposed approach achieved the best detection performance with block size of 10 × 10 in all of testing sequences.

Furthermore, we compared our approach, with block size of 10×10, with three state-of-the-art methods: the method proposed in Ref.[2], MoG based foreground detection [5] and Kalman Filter (KF) based foreground detection [6]. The experimental results show that better performance can be obtained by the proposed approach from statistical evaluation shown in Fig.4[1] and visualization in Fig.5.

5 Discussion

Experimental results in the five testing sequences with dynamic background indicate that the proposed approach is superior to the other three methods. However, compared with the ground truth given by a human inspector, there are still partial corruptions in the true foreground and noisy/false foregrounds in the true background even in the proposed method. Such partial corruptions are typically seen in two situations: one is in the uniform region (including shadow) between foreground and background, and, the other is in the region near to edge, as shown in Fig.6. In the first situation, color or texture or both show similar values between foreground and background,so foreground is possible to be classified to background by our approach in coarse-level; besides, in the later situation, if a newly observed block includes only few part of foreground, this block is often considered as background, because of the weak evidences.

In addition, for the testing sequence $\{MovingEscalator\}$, satisfactory performance was not obtained by any of four methods, the possible reason of which is the lack of training frames (only 50) except for the above two situations.

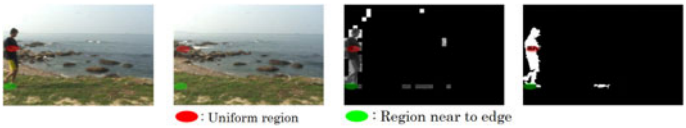


Fig. 6. Partial corruptions in two situations

6 Conclusion

In this paper, we have proposed a hierarchical approach using HSL color values for foreground detection in dynamic background. In coarse-level, foreground blocks are detected by color and texture information and then pixels in blocks judge as foreground in coarse-level are classified on the basis of clustering in fine-level. In the proposed approach, an image is represented multiply and respectively in three components of HSL color space, by which better performance was obtained in dynamic background compared with three typical methods.

[1] Results of foreground detection by method of Chen and MoG-based foreground detection are taken from Ref.[2].

One drawback of the proposed approach is that it shows a weak discrimination power in uniform regions, which may cause some partial corruptions or noisy/false foregrounds in the final detected foreground. In the future, we will study in the following two aspects: 1) To improve the performance in uniform regions; 2) By post-processing, to increase the quality of detected foreground such as morphologic processing and region growing.

Acknowledgment. This work was supported by JSPS Grant-in-Aid for Scientific Research (C) 10213216.

References

1. Zhong, J., Sclaroff, S.: Segmenting Foreground Objects from a Dynamic Textured Background via a Robust Kalman Filter. In: ICCV 2003, vol. 1, pp. 44–50 (2003)
2. Chen, Y.T., Chen, C.S., Huang, C.R., Hung, Y.P.: Efficient hierarchical method for background subtraction. Pattern Recognition 40(10), 2706–2715 (2007)
3. Matsuyama, T., Ohya, T., Habe, H.: Background subtraction for non-stationary scenes. In: ACCV 2000, pp. 662–667 (2000)
4. Toyama, K., Krumm, J., Brumitt, B., Meyers, B., Wallflower: Principles and practice of background maintenance. In: ICCV 1999, vol. 1, pp. 255–261 (1999)
5. Stauffer, C., Grimson, W.E.L.: Adaptive Background mixture Models for real-time Tracking. In: CVPR 1999, pp. 246–252 (1999)
6. Ridder, C., Munkelt, O., Kirchner, H.: Adaptive Background Estimation and Foreground Detection Using Kalman-Filtering. In: ICRAM 1995, pp. 193–199 (1995)
7. Heikkila, M., Pietikainen, M., Heikkila, J.: A texture-based method for detecting moving objects. In: BMVC 2004, vol. 1, pp. 187–196 (2004)
8. Javed, O., Shafique, K., Shah, M.: A hierarchical approach to robust background subtraction using color and gradient information. In: MVC 2002, pp. 22–27 (2002)
9. Harville, M., Gordon, G., Woodfill, J.: Foreground segmentation using adaptive mixture models in color and depth. In: DREV 2001, pp. 3–11 (2001)
10. Huang, C.R., Chen, C.S., Chung, P.C.: Contrast context histogram?An efficient discriminating local descriptor for object recognition and image matching. Pattern Recognition 41, 3071–3077 (2008)
11. http://en.wikipedia.org/wiki/HSL_and_ HSV
12. Sural, S., Qian, G., Pramanik, S.: Segmentation and histogram generation using the HSV color space for image retrieval. In: ICIP 2002, vol. 2, pp. 589–592 (2002)

Image Super-Resolution Based Wavelet Framework with Gradient Prior

Yan Xu[1,2], Xueming Li[1], and Chingyi Suen[2]

[1] Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Centre for Pattern Recognition and Machine Intelligence, Concordia University, Montreal, Quebec H3G 1M8, Canada
yanxu.yx2008@gmail.com

Abstract. A novel super-resolution approach is presented. It is based on the local Lipschitz regularity of wavelet transform along scales to predict the new detailed coefficients and their gradients from the horizontal, vertical and diagonal directions after extrapolation. They form inputs of a synthesis wavelet filter to perform the undecimated inverse wavelet transform without registration error, to obtain the output image and its gradient map respectively. Finally, the gradient descent algorithm is applied to the output image combined with the newly generated gradient map. Experiments show that our method improves in both the objective evaluation of peak signal-to-noise ratio (PSNR) with the greatest improvement of 1.32 dB and the average of 0.56 dB, and the subjective evaluation in the edge pixels and even in the texture regions, compared to the "bicubic" interpolation algorithm.

Keywords: super-resolution, wavelet framework, gradient prior, local Lipschitz regularity.

1 Introduction

Super Resolution (SR) is a software or hardware technique of generating one or more High Resolution (HR) images from one or more Low Resolution (LR) images. This means that HR images have more detailed information with the help of resolution enhancement, and contain more pixels as a result of storing more contents, compared to the original LR images. The fundamental idea is to exchange the time resolution (or the time bandwidth) with the spatial resolution. In the past ten years, many surveys [1] had emerged related to this research field.

In general, SR reconstruction algorithms can be divided into three groups: the interpolation-based method, the reconstruction-based method [2] and the learning-based (or statistical-based) method [3]. There are two major problems to be solved in the process of reconstruction, namely the registration and the restoration, and some specific tasks need to be performed, such as the removal of blurring, noise and artifacts (e.g. blocking, ringing, jaggy and aliasing).

Using the property of multi-resolution analysis in the wavelet domain, it is

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 421–428, 2011.

possible to predict detailed coefficients of next higher fine scales from that of former lower coarse scales. That is based on the fact that local extrema of the wavelet transform propagate across scales, which can be used for extrapolation of higher frequency scales [4]. Several approaches in the wavelet-based SR algorithm have been proposed from various points of view: Hidden Markov Tree (HMT), Neural Network (NN), local approach (such as local Lipschitz regularity across scales [5], local linear embedding in the contourlet domain [6] and edge information [7]).

To design a good image super-resolution algorithm, it is essential to apply good prior or constraint on the HR image because of the ill-posedness of image super-resolution. Sun [8] presented another prior, belonging to generic smoothness prior, which is one branch of two widely used priors, and the other is edge smoothness prior [9].

Our method belongs to the interpolation-based approach for only using one LR image. It focuses on the restoration operation, avoiding the registration error in the procedure of registration. It also increases more details along the edge pixels by means of supplementing modulus maximum and gradient prior, within the framework of the undecimated wavelet transform and the inverse wavelet transform.

The reminder of this paper is organized as follows: Section 2 and Section 3 introduce basic concepts of the local Lipschitz regularity and gradient prior. Section 4 describes our proposed method based on the wavelet framework with gradient prior. Experiments and results are in Section 5. Section 6 concludes and discusses the future of work.

2 Local Lipschitz Regularity

The Lipschitz regularity gives an indication of the differentiability of a function, but it is more precise [5]. The Fourier transform can only provide the global Lipschitz regularity of functions, while compactly supported wavelet frames and bases can evaluate their local Lipschitz regularity. Mallat [5] proposed two theorems to confirm that one can estimate the Lipschitz regularity of a function over intervals and even at a point.

Let S be the set of index pairs (k, l) such that for some $\epsilon > 0$, an interval $(x_0 - \epsilon, x_0 + \epsilon) \subset \mathit{support}(\psi_{k,l})$. A signal has local Holder exponent α in the neighborhood $(x_0 - \epsilon, x_0 + \epsilon)$ if there exists a finite constant C such that wavelet transform coefficients $\omega_{k,l} =< f, \psi_{k,l} >$ satisfy

$$\max_{(k,l)\in S} |\omega_{k,l}| \leq C \cdot 2^{-k(\alpha+\frac{1}{2})} \tag{1}$$

where $\psi_{k,l}$ is a mother wavelet with all scales $k \in R^{+}$ and all translations $l \in R$ [10]. This inequality requires only those underlying wavelet basis functions are more regular than analyzed functions, independent of the particular decomposition filters $H(z)$ and $G(z)$.

Many researchers benefited a lot from employing that theory in diverse applications. The paper [11] proposed a wavelet-based interpolation method to estimate the regularity of edges. They firstly did the undecimated dyadic wavelet transform, assessed C and α in (1) locally for each edge by a linear least-squares fit, copied the feature magnitude of the known subband into that of the new subband, and adjusted the magnitude along the edge pixels to satisfy the bound in (1). The original image and the new subband were up-sampled by a factor of 2, and were added to a single wavelet synthesis stage to produce an output HR image. However, their interpolation method is only available for a factor of 2, and the high frequency subbands are mixed together.

3 Gradient Prior

The meaningful discontinuities in intensity values are detected by using first- and second-order derivatives, and the first-order derivative of choice in image processing is the gradient [12]. A basic property of the gradient vector is that it points to the direction of the maximum rate of change at coordinates (x, y), usually corresponding to the edge pixels, thus we just regard the gradient of edge pixels.

The paper [8] considered the gradient profile as the reconstruction constraint in super-resolution, which had the regularity in their observations of 1000 natural images, described it completely by shape and sharpness parameters, and computed the ratio of gradient profile to forecast the gradient of a HR image. In their reconstruction, the energy function was established by enforcing both the reconstruction constraint $E_i(I_h|I_l)$ in the image domain and the gradient constraint $E_g(\nabla I_h|\nabla I_h^W)$ in the gradient domain:

$$E(I_h|I_l, \nabla I_h^W) = E_i(I_h|I_l)+\beta E_g(\nabla I_h|\nabla I_h^W) = |(I_h * G) \downarrow - I_l|^2 + \beta|\nabla I_h - \nabla I_h^W|^2 \tag{2}$$

where G is a Gaussian filter for spatial filtering, $*$ is the convolution operator, and $\downarrow$ is the down-sampling operation. ∇I_h is the gradient of I_h, and ∇I_h^W is the new gradient of I_h followed by predicting the gradient of detailed coefficients and performing the inverse wavelet transform.

The energy function (2) can be minimized by the gradient descent algorithm:

$$I_h^{t+1} = I_h^t - \tau \cdot \frac{\partial E(I_h)}{\partial I_h} \tag{3}$$

$$\text{where } \frac{\partial E(I_h)}{\partial I_h} = ((I_h * G) \downarrow - I_l) \uparrow * G - \beta \cdot (\nabla^2 I_h - \nabla^2 I_h^W) \tag{4}$$

4 Proposed Method

The wavelet transform and the gradient magnitude are both linear operations, while wavelet coefficients obey the decay of the wavelet transform across scales, therefore the gradient magnitude also abides by the similar criterion. It is possible to predict their gradient magnitude at higher resolution levels from three directions as new detailed coefficients. Combining the wavelet framework with gradient prior, we propose a novel approach. Its procedure is shown in Figure 1:

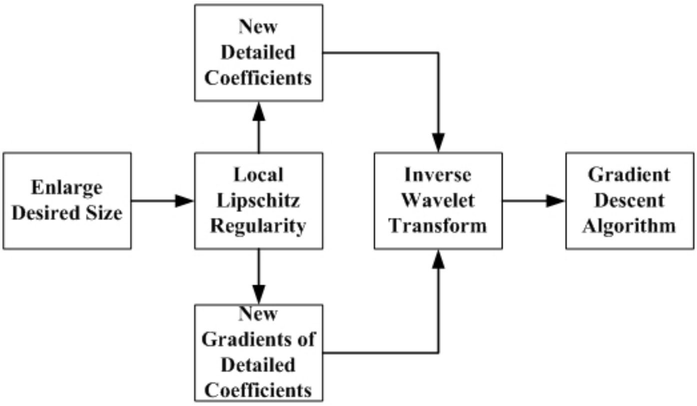

New Detailed Coefficients
Enlarge Desired Size
Local Lipschitz Regularity
Inverse Wavelet Transform
Gradient Descent Algorithm
New Gradients of Detailed Coefficients

Fig. 1. Block diagram of our proposed algorithm

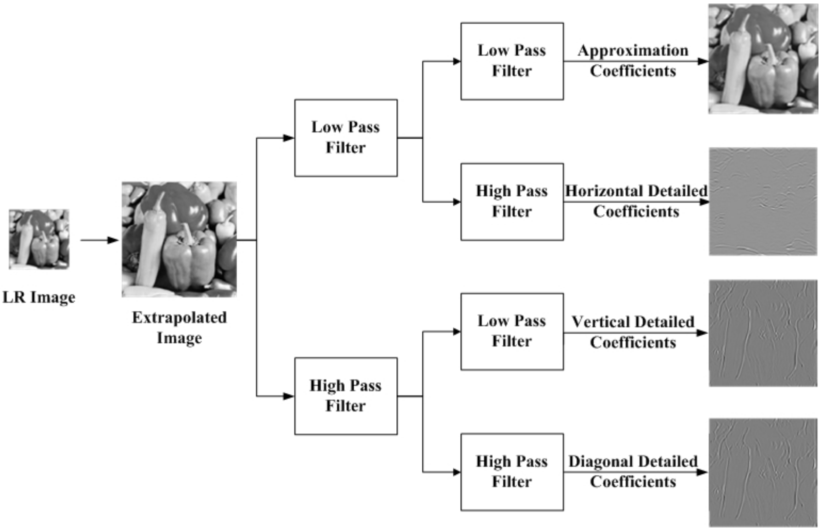


Fig. 2. Undecimated discrete wavelet decomposition of an image

Step 1: Extrapolate the input LR image I_l into the desired image size I_u with the "bicubic" interpolation, transform it into the wavelet domain to obtain one approximation coefficient app_1 and three detailed coefficients deh_1, dev_1 and ded_1 from three directions, as shown in Figure 2, then take app_1 as the input of the next decomposition level, and repeat this process until the final decomposition level is reached.

Step 2: Calculate the gradient of I_u and detailed coefficients deh_m, dev_m, ded_m at each decomposition level, to obtain I_{ug}, $dehg_m$, $devg_m$, and $dedg_m$, where the decomposition level $m = 1, \ldots, 4$.

Step 3: Use the Canny detector to produce the edge image I_{uedg} of I_u, and label the neighboring edge pixels as the same group with the similar tag in a predefined distance (e.g. 5 pixels).

Step 4: Measure local parameters C_w, α_w and C_g, α_g for each labeled edge group in the wavelet domain and in the gradient domain separately according to (1).

Step 5: Take half the value of detailed coefficients deh_1, dev_1, ded_1 and their corresponding gradient $dehg_1$, $devg_1$, $dedg_1$ at the first decomposition level as the original value of new detailed coefficients $dehnew$, $devnew$, $dednew$ and their new gradients $dehgnew$, $devgnew$, $dedgnew$ respectively. According to (1), change their values along the edge pixels with parameters calculated from Step 4 to get new detailed coefficients $dehnew'$, $devnew'$, $dednew'$ and their new gradient $dehgnew'$, $devgnew'$, $dedgnew'$.

Step 6: Regard I_u as one input of two low-pass wavelet filters, and $dehnew'$, $devnew'$, $dednew'$ from Step 5 are as other inputs of a high-pass and low-pass, a low-pass and high-pass, and a high-pass and high-pass wavelet filters respectively, which is the inverse process of Figure 2 to obtain a new HR image I_h. The generation of its new gradient is in the same way, and the difference is that the gradient of I_{ug} from Step 2 and $dehgnew'$, $devgnew'$, $dedgnew'$ from Step 5 are as inputs to get a new gradient I_{hg} of I_h.

Step 7: Iterate (3) hundreds of times to calculate the final HR image I'_h by the gradient descent algorithm.

5 Experiments and Results

In our experiment, the Cohen-Daubechies-Feauveau (CDF) 9/7 bi-orthogonal symmetric wavelet filter is exploited, which is used for lossy compression in the JPEG 2000 compression standard. Four decomposition levels are sufficient, because the more levels an image can be decomposed into, the less information

Table 1. The value of PSNR of different sizes with different factors

Name	LR Size	Factor*(2^1)		Factor(2^2)		Factor(2^3)	
		Bicubic	Proposed	Bicubic	Proposed	Bicubic	Proposed
mandrill	256 × 256	20.70	21.61	/	/	/	/
peppers		28.24	29.56	/	/	/	/
sails		24.95	25.94	/	/	/	/
mandrill	128 × 128	21.23	21.87	17.84	18.15	/	/
peppers		25.87	26.79	22.52	22.81	/	/
sails		22.73	23.04	20.46	20.79	/	/
mandrill	64 × 64	22.05	22.08	18.55	18.91	16.46	16.70
peppers		22.76	23.81	19.95	20.55	18.26	18.74
sails		21.71	22.40	18.71	19.05	17.62	17.90

[*] Note: The data in the table for each column with the same factor correspond to the "bicubic" interpolation algorithm and our proposed method respectively. The symbol "/" expresses no PSNR values because of lack of reference images.

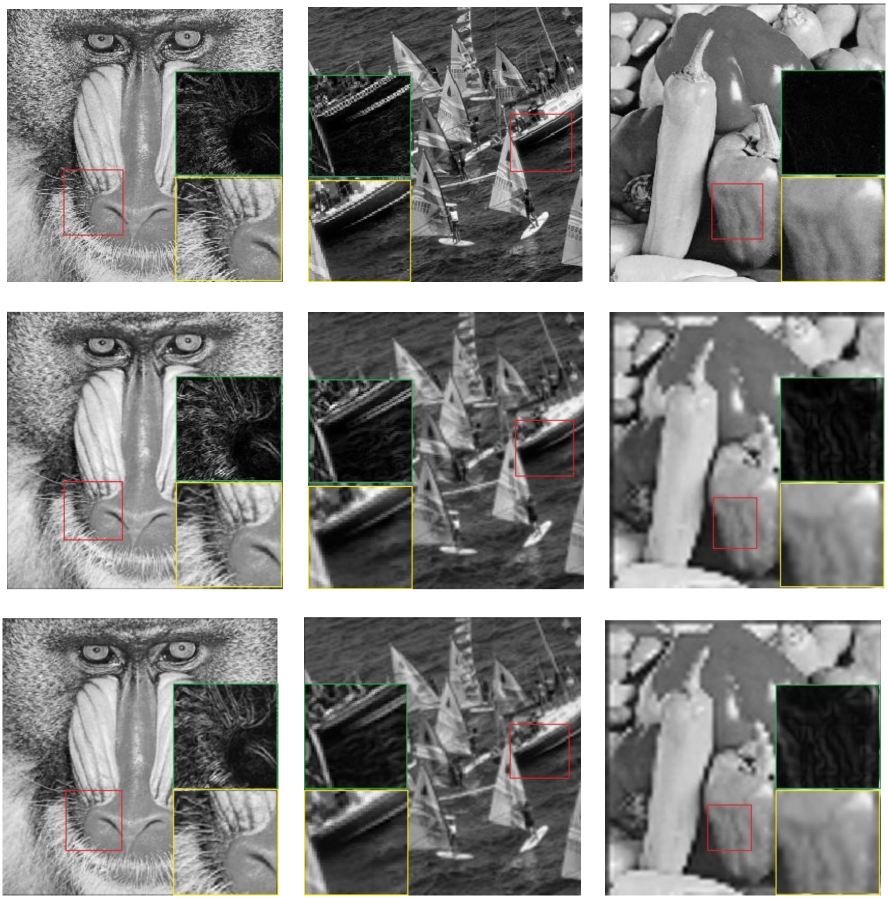


Fig. 3. Super-resolution results with up-sampling factors of 2, 4 and 8 for the first, second and third columns from left to right. There are three images in each column from top to bottom, the reference image, the "bicubic" interpolation image and our resulting image using the proposed algorithm separately. Each image contains three color blocks, the red one (left or right) for the position of the original image, the yellow one (below) for the magnification part, and the green one (above) for its gradient map.

it can hold. I_l can be produced by the original reference 512 × 512 image I_r. I_r goes through a Gaussian low pass filter with the variance of 0.8, down-sampled by a factor of 2, and this process is repeated until I_l is scaled from 64 × 64 to 128 × 128. Much research has been done on super-resolution of color images, but most of them just simply apply their methods to the grayscale channel, and then up-sample the color channel to fuse them together. Furthermore, the human vision system (HVS) is more sensitive to the brightness information. Thus, we just select the grayscale image.

We compute the PSNR to measure super-resolution results from the objective perspective, which are illustrated in Table 1. It can be seen from Table 1, the image “peppers” includes more edges, and thus it has the best restoration of details. While the images “mandrill” and “sails” contain more textures, whose PSNR is lower than the “peppers” image, especially the “mandrill” image. With the increase of the scaling factor, the value of PSNR for each type of image is gradually decreased. While with the increase of the LR size, the magnification image is remarkably better because more information is included. The differences between two approaches in the table can be simply calculated. The greatest improvement of PSNR is 1.32 dB and the average is 0.56 dB, compared to the “bicubic” interpolation.

The subjective result can be found from Figure 3. It is easy to discover obvious improvements along the edge pixels, hairs near the nose in the image “mandrill”, ripples under the boat in the image “sails”, and textures of the pepper surface in the image “peppers”. Compared to the “bicubic” interpolation approach, our algorithm generates more detailed information especially along the edge pixels, even in some texture regions. It is evident for the gradient map to appear more intensity changes. The reason why the edge of “peppers” image is a sawtooth curve is that when the scaling factor is 8, the strong blocking effect is exhibited as the result of the inherent drawback of the “bicubic” interpolation, which is as the initial value of our algorithm.

6 Conclusion

A novel method of image super-resolution has been proposed. It is based on the wavelet transform and inverse wavelet transform framework, combined with gradient prior. It considers the detailed information in the horizontal, vertical and diagonal directions, and adds the predicted detailed coefficients and gradient priors along the edge pixels using the local Lipschitz regularity, regardless of the procedure of registration. Furthermore, our method can be used not only for a certain specific image, but suitable for diverse images in many situations because the information used comes from the image itself by the wavelet decomposition framework. It is reasonable for our approach to improve the detailed information. Experiments prove that it keeps good objective and subjective effects.

The future of work can be focused on the following aspects: (a) to find out the best edge pixels including more details and the best initial values for newly generated coefficients and gradients; (b) to extend interesting pixels not only limited to the edge pixels; (c) to adjust the energy function of the resulting HR image; and (d) to make full use of the generated gradient to increase the amount of information.

References

1. VJiji, C., Chaudhuri, S., Chatterjee, P.: Single frame image super-resolution: should we process locally or globally? Multidimensional Systems and Signal Processing 18, 123–152 (2007)

2. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. IEEE Trans. on PAMI 24(9), 1167–1183 (2002)
3. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neigbor embedding. In: Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 275–282 (2004)
4. Chang, S.G., Cvetkovic, Z., Vetterli, M.: Resolution enhancement of images using wavelet transform extrema extrapolation. In: Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 4, pp. 2379–2382 (1995)
5. Mallat, S., Hwang, W.L.: Singularity detection and processing with wavelets. IEEE Trans. Information Theory 38(2), 617–643 (1992)
6. Do, M.N., Vetterli, M.: The contourlet transform: an efficient directional multi resolution image representation. IEEE Trans. on Image Processing 14(12), 2091–2106 (2005)
7. Velisavljevic, V.: Edge-preservation resolution enhancement with oriented wavelets. In: Proc. of the IEEE International Conference on Image Processing (ICIP), pp. 1252–1255 (2008)
8. Sun, J., Sun, J., Xu, Z.B., Shum, H.Y.: Gradient profile prior and its applications in image super-resolution and enhancement. IEEE Trans. on Image Processing 20(6), 1529–1542 (2011)
9. Dai, S.Y., Han, M., Xu, W., Wu, Y., Gong, Y.H., Katsaggelos, A.K.: Softcuts: a soft edge smoothness prior for color image super-resolution. IEEE Trans. on Image Processing 18(5), 969–981 (2009)
10. Daubechies, I.: Ten lectures on wavelets. SIAM, Philadelphia (1992)
11. Carey, W.K., Chuang, D.B., Hemami, S.S.: Regularity-preserving image interpolation. IEEE Trans. on Image Processing 8(9), 1293–1297 (1999)
12. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital image processing using MATLAB. Publishing House of Electronics Industry, Beijing (2004)

Are Performance Differences of Interest Operators Statistically Significant?

Nadia Kanwal[1,2], Shoaib Ehsan[1], and Adrian F. Clark[1]

[1] VASE Laboratory, Computer Science & Electronic Engineering University of Essex,
Colchester CO4 3SQ, UK
[2] Lahore College for Women University, Pakistan
{nkanwa,sehsan,alien}@essex.ac.uk

Abstract. The differences in performance of a range of interest operators are examined in a null hypothesis framework using McNemar's test on a widely-used database of images, to ascertain whether these apparent differences are statistically significant. It is found that some performance differences are indeed statistically significant, though most of them are at a fairly low level of confidence, *i.e.* with about a 1-in-20 chance that the results could be due to features of the evaluation database. A new evaluation measure i.e. accurate homography estimation is used to characterize the performance of feature extraction algorithms.Results suggest that operators employing longer descriptors are more reliable.

Keywords: Feature Extraction, Homography, McNemar's Test.

1 Introduction

Recent years have seen interest point operators becoming a focus of attention in computer vision. As new operators are developed, publications have generally compared their performances with existing ones using the database[1]; and there have been some dedicated comparison papers that also use this same dataset[1] [1,2]. The purpose of this paper is to establish whether such differences in performance are indeed significant. The general approach adopted is null hypothesis testing, widely used for establishing the effectiveness of drug treatments, for example [3]. As well as answering the question posed in the title, a useful research result in its own right, this paper attempts to present null hypothesis testing in an approachable way, one that encourages others to use it in establishing whether their own algorithms perform significantly differently from prior art. Another motivation behind this work is to provide a non-parametric evaluation framework for vision algorithms with special focus on local feature extraction algorithms. This type of investigation is considered necessary due to the complexity of the image data and variety of results presented in literature [1,4].

The remainder of this paper is structured as follows. Section 2 reviews briefly some related work in this filed. Section 3 presents the evaluation framework based

[1] http://www.robots.ox.ac.uk/~vgg/research/affine/

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 429–436, 2011.

around null hypothesis testing. Sections 4 and 5 attempt to answer the question posed in the title by performing null hypothesis testing. Finally, section 6 draws conclusions and indicates further work.

2 Related Work

Interest point extraction from an image is a fundamental step in many vision applications such as tracking, navigation, panorama stitching, and mosaicking [2]. It is subdivided into two main stages: firstly, the detection of image interest points ('features' or 'keypoints') and computation of distinctive descriptors for detected interest points.

We define an interest point extractor as a system that includes detection and description of a feature point in an image. Well-known detectors that have received recent interest in the literature include Harris Affine detector [5], Hessian Affine detector [5], Laplacian of Gaussian detector, Difference of Gaussian detector and Fast Hessian based detector [4]. Image feature description, schemes like that in the Scale Invariant Feature Transform (SIFT) [6], Gradient Location Orientation Histogram (GLOH) [1], and Speeded Up Robust Feature (SURF) [4] represent the state-of-the-art.

There are several well-known procedures developed to evaluate and characterize feature detectors and descriptors separately in the literature [1,7,8]. Unfortunately, the importance of evaluating interest point extraction as a system (*i.e.,* detection *and* description) has been overlooked. Evaluations measures such as repeatability [7,8], a theoretic measure which does not reflect the actual performance of detectors [2], and ROC (Reciever Operating Curves) and precision–recall curves have all been used to evaluate interest point descriptors. In [1], which evaluated descriptors, performance measures were calculated for only for correct matches between images, overlooking the need for the descriptor to identify correctly true and false negative matches. In [9], the evaluation of feature descriptors for correctly-matched video frames was presented using the same precision–recall as in [1], hence also failing to assess false negative performance. The authors contend that assessing performance on false negatives cannot be ignored in identifying the true performance of an algorithm.

Most of the comparative studies for local feature descriptors ranked Scale Invariant Feature Transform (SIFT) [6] descriptor and SIFT-based descriptors such as Gradient Localization Orientation Histogram (GLOH) [1] on top. The more recently-developed technique known as Speeded Up Robust Features (SURF) claimed to produce better results [4,9] using the criteria given in [1]. However, the reliability of these rankings is questionable because, of the evlaution criteria used. The dataset used in [1] is composed of several sets of real images with different geometric and photometric transformations and is publicly available. In the following, to avoid confusion, the term 'dataset' is taken to mean one of the component datasets, while 'database' refers to all of them together. We have used exactly the same datasets as previously-published evaluations of these detectors [1,4,7,8,9]. The results presented in this paper demonstrate that these previous evaluations need to be treated with some caution.

3 Null Hypothesis Testing

3.1 McNemar's Test

In null hypothesis testing as applied to the assessment of interest operators, one starts by assuming that there is no difference between their performances and then assessing whether the evidence obtained from testing does or does not support the hypothesis. As in this case it is possible to run each interest detector on the same set of images, one can build up a kind of 'truth table' for a pair of algorithms:

	Algorithm A Failed	Algorithm A Succeeded
Algorithm B Failed	N_{ff}	N_{sf}
Algorithm B Succeeded	N_{fs}	N_{ss}

where N_{sf} is the number of tests on which algorithm A succeeded and algorithm B failed, and so on. With this, an appropriate statistic to use is McNemar's test, a form of χ^2 test for matched paired data:

$$Z = \frac{(|N_{sf} - N_{fs}| - 1)}{\sqrt{N_{sf} + N_{fs}}} \tag{1}$$

where the -1 is a continuity correction. If $N_{sf} + N_{fs} \gtrsim 20$, then Z should be reliable [10]. If Algorithms A and B give similar results, then $Z \approx 0$; as their results diverge, Z increases. It is interesting to note that this expression involves cases where one algorithm succeeds and the other fails, whereas performance evaluation in vision largely focuses on where algorithms succeed. Confidence limits can be associated with the Z value as shown in Table 1; $Z > 1.96$, for example, corresponds to two standard deviations from the mean of a Gaussian distribution, in which case the results from the algorithms are expected to differ by chance only one time in 20. In the physical sciences, it is common to use a more exacting 1-in-5,000 criterion [11]. Values for two-tailed and one-tailed predictions are shown in Table 1 as either may be needed, depending on the hypothesis used: If we are assessing whether the performances of two algorithms differ, a two-tailed test should be used; but if we are determining whether one algorithm is better than another, a one-tailed test is needed.

Table 1. Converting Z scores onto confidence limits

Z value	Degree of confidence	Degree of confidence
	Two-tailed prediction	One-tailed prediction
1.645	90%	95%
1.960	95%	97.5%
2.326	98%	99%
2.576	99%	99.5%

3.2 Comparing Many Algorithms

McNemar's test compares only two algorithms. When several algorithms are to be compared, as in this paper, one needs to consider that each pairwise comparison involves a selection from a population of algorithms — the use of multiple comparisons tends to increase the family-wise error rate. In such cases, there are several corrections that one could use, the best-known of which is due to Bonferroni [12]: for A algorithms, if the significance level for the whole family of tests to be at most α, then each individual test needs to be performed with a significance level of α/A. The Bonferroni correction is actually a first-order Taylor series expansion of the more general Šidák correction, $1-(1-\alpha)^{1/A}$.

However, there are concerns over the use of *any* correction [13]. For example, these corrections control only the probability of false positives and come at the cost of increasing the probability of false negatives. In this paper, we have avoided performing such corrections.

3.3 The Independence of Tests

An important requirement for any kind of statistical testing is that each test is independent of the others. As the interest detectors reviewed in section 2 may identify features anywhere in an image, performance cannot be measured just on where features are found. Instead, the approach that is normally taken is to match feature points between pairs of images and, from those, to determine the homography (transformation) between the images. The closeness of the homography to the correct value determines the effectiveness of the detector. For the data described in section 2, 'true' homographies, calculated from verified feature matches, are provided. Using homographies ameliorates the effect of finding features at multiple scales, though their lack of dependence may become problematic if a correspondence is calculated from few matches.

4 Experimental Procedure and Evaluation Framework

The algorithms selected for evaluation are all combinations of a detector and a descriptor. SIFT comes with its own DoG detector and 128-bin descriptor [6]; SURF defined its own Hessian based detector but comes with two variations in terms of descriptor lenght, with 64 or 128 bins [4]. Conversely, GLOH is a descriptor that can be used with any detector, and its authors also developed Harris-affine, Harris-Laplace, Hessian-affine and Hessian-Laplace detectors [2] for use with their GLOH descriptor. In this work, GLOH is combined with all of these detectors in turn. All of the experiments employ the database described in [1].

Using these algorithms, corresponding matched feature points between a pair of images are determined as their nearest neighbors in terms of descriptor distance. From the set of matched features, a homography matrix, the *estimated homography matrix* is calculated using RANSAC (RANdon SAmple Consensus). The test is based on finding the accuracy of the estimated homography matrix as compared to the original homography representing actual transformation between the pair of images. Though it is true that the spatial distribution of matches affects the accuracy of the resulting homography, but gives scatterness (a requirement for tracking and navigation applications) of features in an image and our approach encapsulates both coverage and repeatability into a single performance measure. In performing these experiments, the various tuning parameters of the algorithms have been left at the values suggested by their authors. The experiments have been performed on all image pairs belonging to all eight datasets within the database. The criterion adopted is *a homography matrix is well approximated if the transformation it represents is close to the actual transformation.* Therefore, instead of asking how well the two images matched we test the homography, estimated from truly matched points. For this some100 points (x, y) are generated randomly and their projections are calculated using the 'ground truth' homography (x_g, y_g) and the estimated homography (x_i, y_i). If Euclidean distance between (x_g, y_g) and (x_i, y_i)) is less than two pixels, then the estimated homography represents a true transformation between image pair; and pass the test; otherwise it fails. The scores of estimated homographies being 'pass' or 'fail' are counted, so that McNemar's Z-score is calculated for a sample size of 500 points per dataset.

Table 2. McNemar's test result for Graffiti dataset

	SIFT pass	SIFT fail
GLOH-HarAff pass	224	124
GLOH-HarAff fail	28	124

For example, the Z-score calculated for data given in Table 2 for the Graffiti dataset between SIFT and GLOH-HarAff is 7.70, showing that the performances of the algorithms are significantly different and that there is 99% confidence that the performance of GLOH-HarrisAffine is better than SIFT.

5 Results

Z-scores for all eight datasets are recorded in Table 3, where every block represents pair wise comparison of all algorithms for one dataset. These Z-scores show that SIFT and SURF-128 are more effective than the other algorithms. Converting the Z-scores into confidence limits, one can be 99% confident that SURF-128 with Fast-Hessian detector will extract more stable features than SIFT for matching textured images under transformations such as rotation, viewpoint change and blurring. Conversely, SIFT performs better if the images have homogeneous regions with similar transformations. The high Z-score of SURF for the Bark, Trees, Graffiti and Wall datasets proves that the use of Haar wavelet response for feature description is more distinctive than the simple gradient direction based description as in SIFT. For features in homogeneous regions, a description of gradient direction is sufficient for matching. From the Z-scores

Table 3. Z-scores between feature extraction algorithms for datasets. Arrowhead direction points towards the better operator in each pair-wise comparison.

	SURF-64	SURF-128	GLOH-HarLap	GLOH-HarAff	GLOH-HesLap	GLOH-HesAff
SIFT	←0.9	↑4.6	← 5.9	←6.2	←1.3	←1.4
SURF-64		↑4.6	←4.1	←5.9	←1.2	←0.8
SURF-128			←7.1	←8.8	←5.8	←3.2
GLOH-HarLap				←4.2	↑2.6	↑3.3
GLOH-HarAff					↑4.5	↑4.8
Bark Dataset						
SIFT	←4.7	←4.8	←0.6	←4.4	←2.9	←4.1
SURF-64		0	↑3.4	←0.1	↑2.1	↑1.6
SURF-128			↑2.1	←0.7	↑2.1	↑2.3
GLOH-HarLap				←4.5	←2.5	←3.5
GLOH-HarAff					←0.5	←1.0
Boat Dataset						
SIFT	←1.7	←6.2	←10.6	←5.8	←2.7	←1.7
SURF-64		←1.1	←9.9	←3.2	0	←1.1
SURF-128			←6.3	←1.0	↑4.1	↑2.9
GLOH-HarLap				↑7.7	↑10.3	↑9.8
GLOH-HarAff					↑3.7	←1.9
Bikes Dataset						
SIFT	←2.4	↑3.3	← 4.2	←5.5	←7.4	←8.9
SURF-64		↑2.1	←2.5	←5	←6.4	←8.3
SURF-128			←5.9	←6.7	←6.8	←9.4
GLOH-HarLap				←1.4	←3.6	←7.2
GLOH-HarAff					←4.4	←7.7
Trees Dataset						
SIFT	←5.4	↑2.7	↑ 1.6	↑7.7	←4.3	↑8.9
SURF-64		↑6.9	↑6.7	↑8.6	0	←1.1
SURF-128			←0.5	↑5	←7.2	↑9.1
GLOH-HarLap				↑5.4	←6.7	↑2.5
GLOH-HarAff					←8.6	↑2.4
Graffiti Dataset						
SIFT	↑4.4	↑6.7	← 4.9	↑7.1	←1.9	↑2.7
SURF-64		↑4.2	←6.1	↑4	←2.6	←0.5
SURF-128			←4.7	←0.1	←3.1	←2.9
GLOH-HarLap				↑5.2	↑5.2	↑4.6
GLOH-HarAff					←1.1	←5.6
Wall Dataset						
SIFT	0	0	←9.9	←3.6	0	←1.2
SURF-64		0	←9.9	←4	0	0
SURF-128			←9.9	←4.2	0	←1.2
GLOH-HarLap				↑7	↑9.9	↑9.6
GLOH-HarAff					↑4.7	↑2.9
Leuven Dataset						
SIFT	0	←3.6	0	0	0	0
SURF-64		↑2.2	←0	←0	←0	←0
SURF-128			↑2.7	↑3.5	↑2.7	↑3.2
GLOH-HarLap				0	0	0
GLOH-HarAff				0	0	0
UBC Dataset						

Table 4. Ranking of feature extraction algorithms based on McNemar's test results

	Blurring	View-point change	Zoom & Rotation	Change in Illumination	JPEG Compression
SIFT	++++	++	++++	++++	++++
SURF-64	+++	++	++	++++	++++
SURF-128	+++	+++	++	++++	++
GLOH+Harlap	+	++	+++	+	++++
GLOH+HarAff	++	++++	++	++++	++++
GLOH+HesLap	+	+++	+++	+++	++++
GLOH+HesAff	+	++++	+++	++++	++++

in the comparison of SURF-64 and SURF-128, it is interesting to see that descriptor length plays an important role in a descriptor's matching performance. Where SURF-128 performed much better than SURF-64 when there is significant amount of transformation in textured images. The ranking of interest point operators given in Table 4 reflects the statistical difference in the performances of these algorithms. This ordering differs from previous performance evaluation studies [1,4], reinforcing the need to use sufficiently large datasets and statistically valid methods for evaluation.

As McNemar's test performs pairwise comparisons between algorithms, it is possible to compare the effectiveness of alternative detectors with the same descriptors, such as GLOH with Harris-Affine and GLOH with Hessian-Affine. This comparison yields a Z-score of 2.4, showing that Hessian-Affine's features are more robust than those of Harris-Affine when there is an affine transformation in images (Graffiti dataset). It is particularly interesting that a previous comparative study [2] shows contradictory results, a discrepancy that is still being explored.

6 Conclusions and Further Work

A null hypothesis framework employing McNemar's test has been used to investigate performance differences of feature extraction algorithms. The results clearly indicate the advantage of SIFT and SURF features (both with 128-bin descriptors), being more distinctive and robust for matching under different image transformations. Hessian or Harris detectors excels only when there is a significant viewpoint change in images, and therefore should prove most useful when used in conjunction with a good feature detector such as SIFT or SURF.

The *characterization*, as opposed to *evaluation*, of algorithm performance based on statistically-valid evaluation is an important bonus of this approach. Although McNemar's test is valuable in identifying performance differences between algorithms and thus has a place in evaluation studies, the fact that it identifies cases where one algorithm succeeds while another fails also has an important place in algorithm development: when this is so, the algorithms are operating in significantly different way and a hybrid algorithm that incorporates elements of both intelligently is likely to achieve a much better performance than either in isolation. Our future work aims to explore this potential.

References

1. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1615–1630 (2005)
2. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3), 177–280 (2008)
3. Saag, M.S., Powderly, W.G., Cloud, G.A., Robinson, P., Grieco, M.H., Sharkey, P.K., Thompson, S.E., Sugar, A.M., Tuazon, C.U., Fisher, J.F., et al.: Comparison of amphotericin B with fluconazole in the treatment of acute AIDS-associated cryptococcal meningitis. New England Journal of Medicine 326(2), 83–89 (1992)
4. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
5. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60(1), 63–86 (2004)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
7. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. International Journal of computer vision 37(2), 151–172 (2000)
8. Ehsan, S., Kanwal, N., Clark, A.F., McDonald-Maier, K.D.: Improved repeatability measures for evaluating performance of feature detectors. Electronics Letters 46(14), 998–1000 (2010)
9. Valgren, C., Lilienthal, A.: SIFT, SURF and seasons: Long-term outdoor localization using local features. In: Proceedings of the European Conference on Mobile Robots (ECMR), pp. 253–258 (2007)
10. Clark, A.F., Clark, C.: Performance Characterization in Computer Vision A Tutorial (1999)
11. Crease, R.P.: Discovery with statistics. Physics World 23(8), 19 (2010)
12. Abdi, H.: Bonferroni and Šidák corrections for multiple comparisons. Sage, Thousand Oaks, CA (2007)
13. Perneger, T.V.: What's wrong with bonferroni adjustments. British Medical Journal 316, 1236–1238 (1998)

Accurate and Practical Calibration of a Depth and Color Camera Pair

Daniel Herrera C., Juho Kannala, and Janne Heikkilä

Machine Vision Group
University of Oulu
{dherrera,jkannala,jth}@ee.oulu.fi

Abstract. We present an algorithm that simultaneously calibrates a color camera, a depth camera, and the relative pose between them. The method is designed to have three key features that no other available algorithm currently has: accurate, practical, applicable to a wide range of sensors. The method requires only a planar surface to be imaged from various poses. The calibration does not use color or depth discontinuities in the depth image which makes it flexible and robust to noise. We perform experiments with particular depth sensor and achieve the same accuracy as the propietary calibration procedure of the manufacturer.

Keywords: calibration, depth camera, camera pair.

1 Introduction

Obtaining depth and color information simultaneously from a scene is both highly desirable and challenging. Depth and color are needed in applications ranging from scene reconstruction to image based rendering. Capturing both simultaneously requires using two or more sensors. A basic device for scene reconstruction is a depth and color camera pair. Such a camera pair consists of a color camera rigidly attached to a depth sensor (e.g. time-of-flight (ToF) camera, laser range scanner, structured light scanner).

In order to reconstruct a scene from the camera pair measurements the system must be calibrated. This includes internal calibration of each camera as well as relative pose calibration between the cameras. Color camera calibration has been studied extensively [1,2] and different calibration methods have been developed for different depth sensors. However, independent calibration of the cameras may not yield the optimal system parameters, and a comprehensive calibration of the system as a whole could improve individual camera calibration as it allows to use all the available information.

1.1 Previous Work

A standard approach is to calibrate the cameras independently and then calibrate only the relative pose between them [3,4,5]. This may not be the optimal solution as measurements from one camera can improve the calibration of the

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 437–445, 2011.

other camera. Moreover, the independent calibration of a depth camera can require a high precision 3D calibration object that can be avoided using joint calibration.

Fuchs and Hirzinger [6] propose a multi-spline model for ToF cameras. Their model has a very high number of parameters and it requires a robotic arm to know the exact pose of the camera. Lichti [7] proposes a calibration method for an individual laser range scanner using only a planar calibration object. It performs a comprehensive calibration of all parameters. However, it relies on the varying response of the scanner to different surface colors to locate corner features on the image.

Zhu et al. [8] describe a method for fusing depth from stereo cameras and ToF cameras. Their calibration uses the triangulation from the stereo cameras as ground truth. This ignores the possible errors in stereo triangulation and measurement uncertainties. The different cameras are thus calibrated independently and the parameters obtained may not be optimal.

1.2 Motivation

As a motivation for our work, we propose three requirements that an optimal calibration algorithm must have. To the best of our knowledge, no available calibration algorithm for a depth and color camera pair fulfills all three criteria.

Accurate: The method should provide the best combination of intrinsic and extrinsic parameters that minimizes the reprojection error for both cameras over all calibration images. This may seem like an obvious principle but we stress it because partial calibrations, where each camera is calibrated independently and the relative pose is estimated separately, may not achieve the best reprojection error.

Practical: The method should be practical to use with readily available materials. A high precision 3D calibration object is not easy/cheap to obtain and a robotic arm or a high precision mechanical setup to record the exact pose of the camera pair is usually not practical, whereas a planar surface is usually readily available.

Widely applicable: To be applicable to a wide range of depth sensors, one cannot assume that color discontinuities are visible on the depth image. Moreover, some depth sensors, like the one used for our experiments, may not provide accurate measurements at sharp depth discontinuities. Thus, neither color nor depth discontinuities are suitable features for depth camera calibration. The method should use features based on depth measurements that are most reliable for a wide range of cameras.

2 The Depth and Color Camera Pair

Our setup consists of one color camera and one depth sensor rigidly attached to each other. Our implementation and experiments use the Kinect sensor from Microsoft, which consists of a projector-camera pair as the depth sensor that measures per pixel disparity. The Kinect sensor has gained much popularity in the scientific and the entertainment community lately. The complete model includes $20 + 6N$ parameters where N is the number of calibration images. The details of the model are described below.

2.1 Color Camera Intrinsics

We use a similar intrinsic model as Heikkilä and Silven [1] which consists of a pinhole model with radial and tangential distortion correction. The projection of a point from color camera coordinates $\mathbf{x}_c = [x_c, y_c, z_c]^\top$ to color image coordinates $\mathbf{p}_c = [u_c, v_c]^\top$ is obtained through the following equations. The point is first normalized by $\mathbf{x}_n = [x_n, y_n]^\top = [x_c/z_c, y_c/z_c]^\top$. Distortion is then performed:

$$\mathbf{x}_g = \begin{bmatrix} 2k_3 x_n y_n + k_4(r^2 + 2x_n^2) \\ k_3(r^2 + 2y_n^2) + 2k_4 x_n y_n \end{bmatrix} \tag{1}$$

$$\mathbf{x}_k = (1 + k_1 r^2 + k_2 r^4)\mathbf{x}_n + \mathbf{x}_g \tag{2}$$

where $r^2 = x_n^2 + y_n^2$ and k is a vector containing the four distortion coefficients.

Finally the image coordinates are obtained:

$$\begin{bmatrix} u_c \\ v_c \end{bmatrix} = \begin{bmatrix} f_{cx} & 0 \\ 0 & f_{cy} \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} u_{c0} \\ v_{c0} \end{bmatrix} \tag{3}$$

The complete color model is described by $\mathcal{L}_c = \{f_{cx}, f_{cy}, u_{c0}, v_{c0}, k_1, k_2, k_3, k_4\}$.

2.2 Depth Camera Intrinsics

In our experiments we used the increasingly popular Kinect sensor as a depth camera [9]. However, the method allows any kind of depth sensor to be used by replacing this intrinsic model. The Kinect consists of an infrared projector that produces a constant pattern and a camera that measures the disparity between the observed pattern and a pre-recorded image at a known constant depth. The output consists of an image of scaled disparity values.

The transformation between depth camera coordinates $\mathbf{x}_d = [x_d, y_d, z_d]^\top$ and depth image coordinate $\mathbf{p}_d = [u_d, v_d]$ follows the same model used for the color camera. The distortion correction did not improve the reprojection error and the distortion coefficients were estimated with very high uncertainty. Therefore we do not use distortion correction for the depth image.

The relation between the disparity value d and the depth z_d is modeled using the equation:

$$z_d = \frac{1}{\alpha(d - \beta)} \tag{4}$$

where α and β are part of the depth camera intrinsic parameters to be calibrated. The model for the depth camera is described by $\mathcal{L}_d = \{f_{dx}, f_{dy}, u_{d0}, v_{d0}, \alpha, \beta\}$.

2.3 Extrinsics and Relative Pose

Figure 1 shows the different reference frames present in a scene. Points from one reference frame can be transformed to another using a rigid transformation denoted by $\mathcal{T} = \{\mathbf{R}, \mathbf{t}\}$, where $\mathbf{R}$ is a rotation and $\mathbf{t}$ a translation. For example, the transformation of a point $\mathbf{x}_w$ from world coordinates $\{W\}$ to color camera coordinates $\{C\}$ follows $\mathbf{x}_c = \mathbf{R}_c\mathbf{x}_w + \mathbf{t}_c$. Reference $\{V\}$ is anchored to the corner of the calibration plane and is only used for initialization. The relative pose $\mathcal{T}_r$ is constant, while each image has its own pose $\mathcal{T}_c$, resulting in $6 + 6N$ pose parameters.

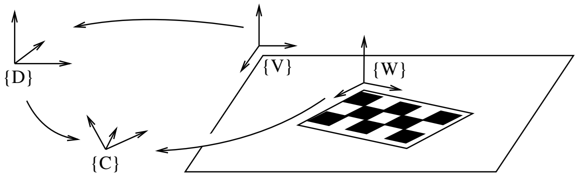


Fig. 1. Reference frames and transformations present on a scene. $\{C\}$ and $\{D\}$ are the color and depth cameras' reference frames respectively. $\{V\}$ is the reference frame anchored to the calibration plane and $\{W\}$ is the world reference frame anchored to the calibration pattern.

3 Calibration Method

We use a planar checkerboard pattern for calibration which can be constructed from any readily available planar surface (e.g. a flat table, a wall). The checkerboard corners provide suitable constraints for the color images, while the planarity of the points provides constraints on the depth image. The pixels at the borders of the calibration object can be ignored and thus depth discontinuities are not needed. Figure 2 shows a sample image pair used for calibration. Figure 3 shows the steps of the calibration and its inputs. An initial estimation for the calibration parameters is obtained by independently calibrating each camera. The depth intrinsic parameters $\mathcal{L}_d$ and the relative pose $\mathcal{T}_r$ are then refined using a non-linear optimization. Finally, all parameters are refined simultaneously.

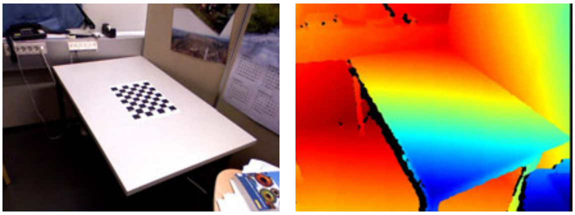


Fig. 2. Sample calibration images. Note the inaccuracies at the table's edge.

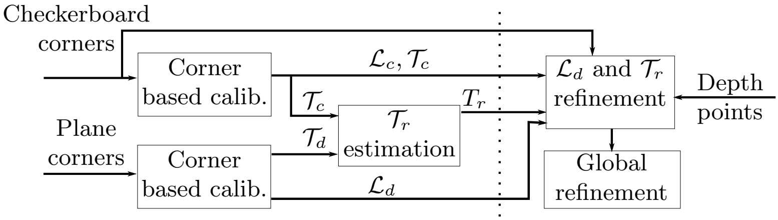


Fig. 3. Block diagram of the calibration algorithm. **Left of dashed line:** initialization. **Right of dashed line:** non-linear minimization.

3.1 Corner Based Calibration

The calibration of a color camera is a well studied problem, we use Zhang's method [2,10] to initialize the camera parameters. Briefly, the steps are the following. The checkerboard corners are extracted from the intensity image. A homography is then computed for each image using the known corner positions in world coordinates $\{W\}$ and the measured positions in the image. Each homography then imposes constraints on the camera parameters which are then solved with a linear system of equations. The distortion coefficients are initially set to zero.

The same method is used to initialize the depth camera parameters. However, because the checkerboard is not visible in the depth image, the user selects the four corners of the calibration plane (the whole table in figure 2). These corners are very noisy and are only used here to obtain an initial guess. The homography is thus computed between $\{V\}$ and $\{D\}$. This initializes the focal lengths, principal point, and the transformation $\mathcal{T}_d$. Using these initial parameters we obtain a guess for the depth of each selected corner. With this depth and the inverse of the measured disparity an overdetermined system of linear equations is built using (4), which gives an initial guess for the depth parameters (α and β).

3.2 Relative Pose Estimation

The independent calibrations give an estimation of the transformations $\mathcal{T}_c$ and $\mathcal{T}_d$. However, the reference frames $\{W\}$ and $\{V\}$ are not aligned. By design we know that they are coplanar. We can use this information by extracting the plane equation in each reference frame and using it as a constraint. We define a plane using the equation $\mathbf{n}^\top \mathbf{x} - \delta = 0$ where $\mathbf{n}$ is the unit normal and δ is the distance to the origin.

If we divide a rotation matrix into its colums $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3]$ and know that the parameters of the plane in both frames are $\mathbf{n} = [0, 0, 1]^\top$ and $\delta = 0$, the plane parameters in camera coordinates are:

$$\mathbf{n} = \mathbf{r}_3 \quad \text{and} \quad \delta = \mathbf{r}_3^\top \mathbf{t} \tag{5}$$

where we use $\mathbf{R}_c$ and $\mathbf{t}_c$ for the color camera and $\mathbf{R}_d$ and $\mathbf{t}_d$ for the depth camera.

As mentioned by Unnikrishnan and Hebert [4] the relative pose can be obtained in closed form from several images. The plane parameters for each image are concatenated in matrices of the form: $\mathbf{M}_c = [\mathbf{n}_{c1}, \mathbf{n}_{c2}, ..., \mathbf{n}_{cn}]$, $\mathbf{b}_c = [\delta_{c1}, \delta_{c2}, ..., \delta_{cn}]$, and likewise for the depth camera to form $\mathbf{M}_d$ and $\mathbf{b}_d$. The relative transformation is then:

$$\mathbf{R}'_r = \mathbf{M}_d \mathbf{M}_c^\top \quad \text{and} \quad \mathbf{t}_r = (\mathbf{M}_c \mathbf{M}_c^\top)^{-1} \mathbf{M}_c (\mathbf{b}_c - \mathbf{b}_d)^\top \tag{6}$$

Due to noise $\mathbf{R}'_r$ may not be orthonormal. We obtain a valid rotation matrix through SVD using: $\mathbf{R}_r = UV^\top$ where $USV^\top$ is the SVD of $\mathbf{R}'_r$.

3.3 Non-linear Minimization

The calibration method aims to minimize the weighted sum of squares of the measurement reprojection errors. The error for the color camera is the Euclidean distance between the measured corner position $\mathbf{p}_c$ and its reprojected position $\mathbf{p}'_c$. Whereas for the depth camera it is the difference between the measured disparity d and the predicted disparity d' obtained by inverting (4). Because the errors have different units, they are weighted using the inverse of the corresponding measurement variance, σ_c^2 and σ_d^2. The resulting cost function is:

$$c = \sigma_c^{-2} \sum \left[(u_c - u'_c)^2 + (v_c - v'_c)^2 \right] + \sigma_d^{-2} \sum (d - d') \tag{7}$$

Note that (7) is highly non-linear. The Levenberg-Marquardt algorithm is used to minimize (7) with respect to the calibration parameters. The initialization gives a very rough guess of the depth camera parameters and relative pose, whereas the color camera parameters have fairly good initial values. To account for this, the non-linear minimization is split in two phases. The first phase uses fixed parameters for the color camera $\mathcal{L}_c$ and external pose $\mathcal{T}_c$, and optimizes the depth camera parameters $\mathcal{L}_d$ and the relative pose $\mathcal{T}_r$. A second minimization is performed over all the parameters to obtain an optimal estimation.

3.4 Variance Estimation

An initial estimate of the color measurement variance σ_c^2 is estimated from the residuals after the first independent calibration. An estimate of the disparity variance σ_d^2 is obtained from the disparity residuals after the first non-linear minimization. It is noted that, because $\mathcal{L}_c$ and $\mathcal{T}_c$ are fixed, the color residuals do not need to be computed and σ_d^2 plays no role in this minimization. The second minimization stage, when all parameters are refined, is then run iteratively using the previously obtained residual variances as the measurement variances for the next step until they converge.

4 Results

We tested our calibration method with an off-the-shelf Kinect device. The device consists of a color camera, an infrared camera and an infrared projector arranged horizontally. The electronics of the device compute a depth map for the infrared image based on the observed pattern from the projector. We ignore the infrared image and use only the depth information and treat it as a generic depth and color camera pair. We used a dataset of 55 images, 35 were used for calibration and 20 for validation. Both sets cover similar depth ranges (0.5m to 2m) and a wide range of poses. For the validation set, (7) was minimized only over the external pose $\mathcal{T}_c$ to find the best pose for the previously obtained calibration.

4.1 Parameters and Residuals

The obtained calibration parameters and their uncertainties are presented in Table 1. Figure 4 presents histograms of the residuals for the validation set. The formulation of our cost function (7) allows us to use the uncertainty analysis presented by Hartley and Zisserman [11]. They show that the covariance of the estimated parameters Σ_P can be obtained directly from the Jacobian of the cost function J and the covariance of the measurements Σ_X using:

$$\Sigma_P = \left(J^\top \Sigma_X J\right)^{-1} \tag{8}$$

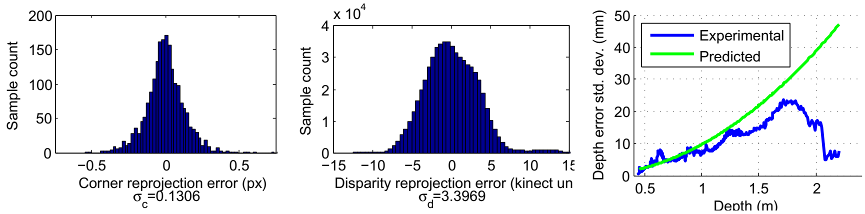
4.2 Depth Uncertainty

The disparity errors are well modeled by a gaussian distribution. Using (4) and the estimated disparity variance, we obtained numerically the expected

Table 1. Obtained calibration parameters. Error estimates correspond to three times their standard deviation.

Color internals							
f_{cx}	f_{cy}	u_{c0}	v_{c0}	k_1	k_2	k_3	k_4
532.90	531.39	318.57	262.08	0.2447	-0.5744	0.0029	0.0065
±0.06	±0.05	±0.07	±0.07	±0.0004	±0.0017	±0.0001	±0.0001

Depth internals						Relative pose (rad, mm)			
f_{dx}	f_{dy}	u_{d0}	v_{d0}	α	β	θ_r	t_{rx}	t_{ry}	t_{rz}
593.36	582.74	322.69	231.48	-0.00285	1091.0	0.024	-21.4	0.7	1.0
±1.81	±2.48	±1.34	±1.59	±0.00001	±1.0	±0.003	±1.5	±1.5	±1.9



(a) Color residuals **(b)** Depth residuals **(c)** Depth uncertainty

Fig. 4. Obtained error residuals and depth uncertainty

variance in depth for each disparity value. Separate statistics are computed for each depth present in the validation set to obtain an experimental depth variance. Both curves are shown in Figure 4c. The experimental curve shows the expected increase in variance as the depth increases. The final drop in variance is due to low sample count at the end of the range.

4.3 Comparison with Manufacturer Calibration

The manufacturer of the Kinect sensor, PrimeSense, has a proprietary camera model and calibration procedure. They provide an API to convert the disparity image to a point cloud in world coordinates. To validate our calibration against the one from the manufacturer, we took an image from a slanted planar surface that covers a range of depths. The disparity image was reprojected to world coordinates using our model and the manufacturer's API. A plane was fitted to each point cloud and the distance of the points to the plane was computed. The manufacturer's reprojection had a standard deviation of 3.10mm from the plane, while ours was 3.00mm. This proves that our calibration of the depth camera has comparable accuracy to that of the manufacturer.

4.4 Colorized Point Cloud

The fully calibrated system can be used to obtain a colored point cloud in metric coordinates. For illustration purposes, Figure 5 shows an example scene and a reprojection from a different view point.

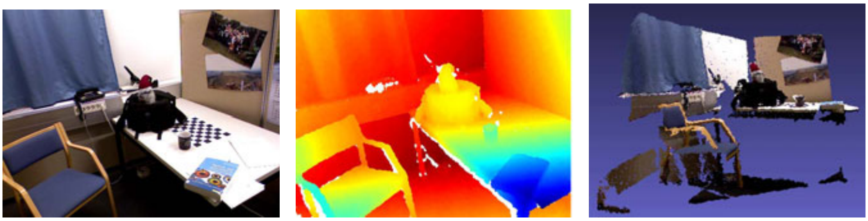


Fig. 5. Sample scene. Color image, depth map, and change of view point.

5 Conclusions

The results show that our algorithm performed adequately for the chosen camera pair. In addition, we believe that our algorithm is flexible enough to be used with other types of depth sensors by replacing the intrinsics model of the depth camera. The constraints used can be applied to any type of depth sensor. Future work can include the calibration of a ToF and color camera pair.

We have presented a calibration algorithm for a depth and color camera pair that is optimal in the sense of the postulated principles. The algorithm takes into account color and depth features simultaneously to improve calibration of the camera pair system as a whole. It requires only a planar surface and a simple checkerboard pattern. Moreover, the method is flexible to be used with different types of depth sensors. Finally, our method showed comparable accuracy to the one provided by the manufacturer of a particular depth camera.

Acknowledgements. This project has been funded by the Academy of Finland's project #127702.

References

1. Heikkilä, J., Silven, O.: A Four-step Camera Calibration Procedure with Implicit Image Correction. In: CVPR, p. 1106. IEEE, Los Alamitos (1997)
2. Zhang, Z.: Flexible camera calibration by viewing a plane from unknown orientations. In: ICCV, pp. 666–673 (1999)
3. Zhang, Q., Pless, R.: Extrinsic calibration of a camera and laser range finder (improves camera calibration). In: IROS, vol. 3, pp. 2301–2306 (2004)
4. Unnikrishnan, R., Hebert, M.: Fast extrinsic calibration of a laser rangefinder to a camera. Robotics Institute, Pittsburgh, Tech. Rep. CMU-RI-TR-05-09 (2005)
5. Scaramuzza, D., Harati, A., Siegwart, R.: Extrinsic self calibration of a camera and a 3D laser range finder from natural scenes. In: IROS, pp. 4164–4169 (2007)
6. Fuchs, S., Hirzinger, G.: Extrinsic and depth calibration of tof-cameras. In: CVPR, pp. 1–6 (2008)
7. Lichti, D.: Self-calibration of a 3D range camera. In: ISPRS, vol. 37(3) (2008)
8. Zhu, J., Wang, L., Yang, R., Davis, J.: Fusion of time-of-flight depth and stereo for high accuracy depth maps. In: CVPR, pp. 1–8 (June 2008)
9. Gesture keyboarding, United State Patent US 2010/0 199 228 A1 (August 5, 2010)
10. Bouguet, J.: Camera calibration toolbox for matlab (March 2010), http://www.vision.caltech.edu/bouguetj/calib_doc/
11. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004)

Contourlet-Based Texture Retrieval Using a Mixture of Generalized Gaussian Distributions

Mohand Saïd Allili and Nadia Baaziz

Université du Québec en Outaouais,
Département d'informatique et d'ingénierie,
Gatineau, Québec, J8X 3X7, Canada
mohandsaid.allili, nadia.baaziz@uqo.ca

Abstract. We address the texture retrieval problem using contourlet-based statistical representation. We propose a new contourlet distribution modelling using finite mixtures of generalized Gaussian distributions (MoGG). The MoGG allows to capture a wide range of contourlet histogram shapes, which provides better description and discrimination of texture than using single probability density functions (pdfs). We propose a model similarity measure based on Kullback-Leibler divergence (KLD) approximation using Monte-Carlo sampling methods. We show that our approach using a redundant contourlet transform yields better texture discrimination and retrieval results than using other methods of statistical-based wavelet/contourlet modelling.

Keywords: Contourlet transform, mixture of generalized Gaussians (MoGG), texture discrimination & retrieval.

1 Introduction

Multiscale transforms, such as the discrete wavelet transform (DWT) and the contourlet transform (CT) [8], decompose an image into a set of sub-images exhibiting details and structure at multiple scales and orientations. Each sub-image corresponds to a frequency subband; a localized partition of the image spectrum. These multiscale representations are very useful tools for texture analysis [9]. This is motivated by three main facts: a) the human visual system adequacy to spatial-frequency representation of signals, b) the inherent nature of texture patterns in terms of presence of geometrical structures, relationship between texture elements and variation in scales and orientations, and c) the psychological research on human texture perception which suggests that two textures are often difficult to discriminate when they produce similar distributions of responses from a bank of linear filters [11,12].

The CT has been introduced to overcome the wavelet inefficiency for describing directional selectivity. Taking into account the relevance of multiple scales and orientations in textures, the CT carries an important impact on the quality of texture analysis and feature extraction. Recently, some studies have successfully investigated the use of contourlets to capture image features for

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 446–454, 2011.

texture retrieval. The reported works adopted either energy measures through spatial-frequency subbands [16,19] or statistical modelling parameters of contourlet subbands [10,14,15]. Po et al. [14] conducted qualitative and quantitative investigation on the statistical modelling of CT coefficients in natural images. New properties are revealed such as: a) CT coefficients strongly depend on their spatial, multi-scale and directional neighborhood, especially for highly textured images, and b) conditioned on their neighborhood, CT coefficients are "approximately" Gaussian. Based on these properties, each CT subband is modeled using a hidden Markov tree model with mixtures of Gaussians (MoG) that capture all neighborhood dependencies. For similarity measurement between two images, the Kullback-Liebler distance (KLD) is suggested. The approach demonstrated better retrieval rates than using wavelets for searching texture images exhibiting high directionality. A variant of this method restricts the MoG modelling to intra-band dependencies in luminance images [10], where estimation of model parameters is based on Markov dependencies among neighboring coefficients according to the directional orientation in each contourlet subband.

In [15], a generalized Gaussian density (GGD) is used to model the marginal distribution of contourlet coefficients in each subband. Estimated model parameters (namely the scale and the shape) are taken as components of the feature vector. Minimum value of KLD between two feature vectors is employed to find similar images. In comparison to the Laplacian or the Gaussian distribution, the GGD has an additional free parameter that controls the shape of the distribution and provides it more flexibility to fit different *platykurtic* or *leptokurtic* histogram shapes [2,5]. In [6,7], the authors demonstrated the superiority of using the wavelet-based GGD modelling over energy-based methods for texture retrieval. However, the main assumption of these works is that a single GGD can capture the shape of the wavelet distribution in each wavelet subband. When examining several examples of texture images, one might clearly notice that for a wide range of natural texture images, wavelet distribution as well as CT distribution is overly heavy-tailed and the representation with a single GGD will lack accuracy. Based on this observation, [1] proposed a model based on finite mixtures of GGDs, which has proven to be more powerful than using the GGD to capture the variety of wavelet histogram shapes.

In this paper, we propose to use a mixture of GGDs (denoted by MoGG) to model the contourlets distribution of texture images. The MoGG offers more flexibility than using a single GGD to fit various shapes of data. Furthermore, it provides the ability to combine low-pass and high-pass contourlet coefficients for image description. We demonstrate that the combination of redundant contourlet image representation and MoGG modelling offers a powerful tool for texture discrimination and retrieval. We prove experimentally that this combination outperforms the wavelet-based retrieval approach. Furthermore, we propose a similarity measurement between images based on Monte-Carlo sampling to approximate the KLD. Experiments on the well known Vistex dataset [13] demonstrate the superiority of the proposed approach for texture retrieval over recent state-of-the-art methods while maintaining comparable computational time.

The remainder of this paper is organized as follows: Section 2 describes relevant properties of the redundant contourlet transform. Section 3 details the MoGG model and its use in representing contourlet distribution. Section 4 presents our application to texture retrieval. Section 5 presents some experimental results.

2 The Redundant Contourlet Transform (RCT)

Directional multiscale image representation is achieved efficiently by means of contourlet transform and its variants [8]. Compared to the DWT, the strength of this true two-dimensional transform is its enhanced directional selectivity. It is well known that fine contours are efficiently reconstructed from their compact representation as few contourlet coefficients localized in the right directional subbands [8]. These facts provide enough motivation to use contourlets in extracting significant image features for texture retrieval. In general, image processing on a multi-scale representation where redundancy is not a big issue can take benefit from over-sampled image decompositions to capture more accurate image characteristics through scale and orientation. This is the reason why the redundant contourlet transform (RCT) has been introduced [4]. The RCT variant aims at reducing aliasing problems and shift sensitivity.

As for the contourlet transform, the RCT decomposition scheme is divided in two parts: a multiscale decomposition based on a Laplacian pyramid and a directional filter bank (DFB) using two-dimensional filtering and critical down-sampling. However, in RCT all down-sampling operations are discarded from the Laplacian stage and a set of symmetric low-pass filters having adequate frequency selectivity and pseudo-Gaussian properties are employed (see Fig. 1). Filter impulse responses $h_b(n)$ are given in Eq. (1) where b is a factor influencing the frequency bandwidth:

$$h_b(n) = e^{-2\frac{n}{b}} - e^{-2}\left(e^{-2\left(\frac{n-b}{b}\right)^2} e^{-2\left(\frac{n+b}{b}\right)^2}\right) \qquad (1)$$

Using L filters (with b=2, 4, 8, 16) results into a redundant Laplacian pyramid (RLP) having $L+1$ equal-size sub-images: one coarse image approximation and L band-pass sub-images. Then, a DFB with $D = 4$ orientations and 1:4 critical down-sampling is applied on each of the L RLP subbands to obtain $4L$ equal-size directional subbands $(C_{ld}, l = 1, ..., L; d = 1, ..., D)$ in addition to a 1:4 down-sampled image approximation C_L as shown in Fig. 1.

3 MoGG Modelling of RCT

The general Gaussian distribution for a univariate random variable $\mathbf{x} \in \mathbb{R}$ is defined in its general form as follows (see ref. [5]):

$$p(x|\mu, \sigma, \beta) = \frac{\beta\sqrt{\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}}}{2\sigma\Gamma\left(1/\beta\right)}\exp\left(-A(\beta)\left|\frac{x-\mu}{\sigma}\right|^{\beta}\right), \qquad (2)$$

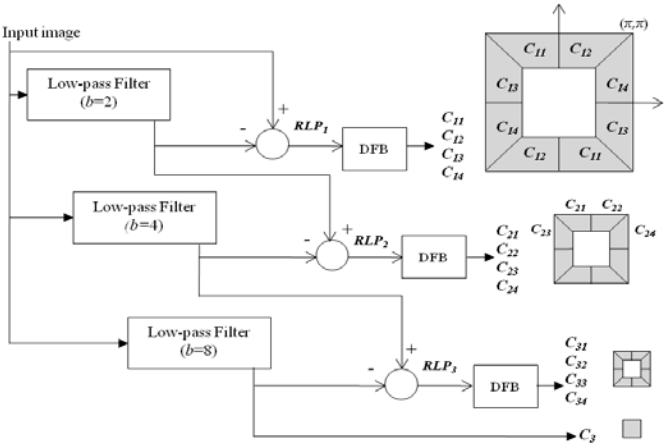


Fig. 1. Three-level RCT decomposition scheme ($L = 3$) and corresponding frequency partition. All down-sampling operations are discarded from the Laplacian stage.

where $A(\beta) = \left[\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}\right]^{\frac{\beta}{2}}$, $\Gamma(\cdot)$ denotes the gamma function, and μ and σ are the distribution mean and standard deviation parameters. The parameter $\beta \geq 1$ controls the kurtosis of the pdf and determines whether the distribution is peaked or flat: the larger the value of β, the flatter the pdf; and the smaller β is, the more peaked the pdf is around its mean. As $\beta \rightarrow \infty$, the distribution becomes uniform; whereas, when $\beta \rightarrow 0$, the distribution becomes a delta function at μ [2,5]. For multi-modal data, we propose to use a mixture of GGDs (MoGG) as proposed in [2]. Given a MoGG with K components, the marginal distribution of the random variable $\mathbf{x}$ is given by

$$p(x|\boldsymbol{\theta}) = \sum_{i=1}^{K} \pi_i p(x|\mu_i, \sigma_i, \beta_i), \tag{3}$$

where $0 < \pi_i \leq 1$ and $\sum_{i=1}^{K} \pi_i = 1$. Here, $\boldsymbol{\theta}$ designates the set of model parameters $\{\pi_i, \mu_i, \sigma_i, \beta_i, i = 1, \ldots, K\}$. The model selection and parameter estimation of the MoGG is achieved in an unsupervised fashion using the minimum message length (MML) principle [18]. Given a sample of data $\mathcal{X} = \{x_1, x_2, ..., x_N\}$, the MML provides a natural trade-off between model complexity (i.e., number of mixture components) and goodness-of-fit to $\mathcal{X}$. With a MoGG with K components, the estimation of the GGD parameters is given as follows [2]:

$$\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} p(\theta_k|x_i) \tag{4}$$

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} p(\theta_k|x_i)|x_i - \mu_k|^{\beta_k - 2} x_i}{\sum_{i=1}^{n} p(\theta_k|x_i)|x_i - \mu_k|^{\beta_k - 2}} \tag{5}$$

$$\hat{\sigma}_k = \left[\frac{\beta_k A(\beta_k)\sum_{i=1}^{n} p(\theta_k|x_i)|x_i - \mu_k|^{\beta_k}}{\sum_{i=1}^{n} p(\theta_k|x_i)}\right]^{1/\beta_k} \tag{6}$$

Fig. 2 shows an example of second-level RCT subbands of image 'Fabric.02' of [13]. Fig. 3 shows the MoGG modelling of the approximation image and the subband C_{23}. Clearly, the histogram of C_{23} is sharply peaked, heavily tailed and not perfectly symmetric at the same time, making the modelling using one GGD inaccurate, as shown in Fig. 3.c. Using a MoGG (see Fig. 3.e) yields better approximation in this case where the coefficients histogram is perfectly fitted. Fig. 3.a. and 3.b compare the accuracy of using MoG versus MoGG to fit the histogram of the approximation image. Clearly MoGG has a better fitting than the MoG due to the flexility of the GGD to adapt heavy-tailed and sharply peaked histogram modes.

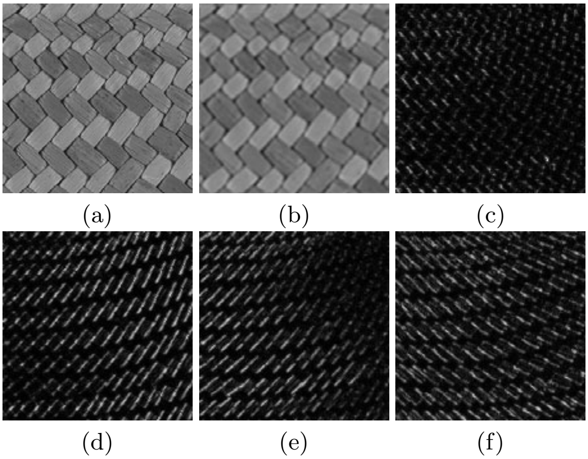


Fig. 2. Illustration of the RCT: (a) original image, (b) approximation image, and RCT subbands (c) C_{21}, (d) C_{22}, (e) C_{23}, and (f) C_{24}

4 MoGG Similarity Measurement for Texture Retrieval

Do et al. [7] developed a closed-form KLD between univariate centered GGDs. However, when dealing with mixture of GGDs with arbitrary number of components, a closed-form solution for the KLD is intractable. To remedy this issue, we resort to approximating the KLD using Monte-Carlo sampling techniques. Given two MoGG models $f(x) = \sum_{i=1}^{K} \pi_i p(x|\theta_i)$ and $g(x) = \sum_{j=1}^{M} \omega_j p(x|\theta_j)$ representing the distribution of contourlet coefficients in the same subband in two images, the KLD distance between the two models is given by:

$$D(f||g) = \int \sum_{i=1}^{K} \pi_i p(x|\theta_i) \log\left(\frac{\sum_{i=1}^{K} \pi_i p(x|\theta_i)}{\sum_{j=1}^{M} \omega_j p(x|\theta_j)}\right) dx, \tag{7}$$

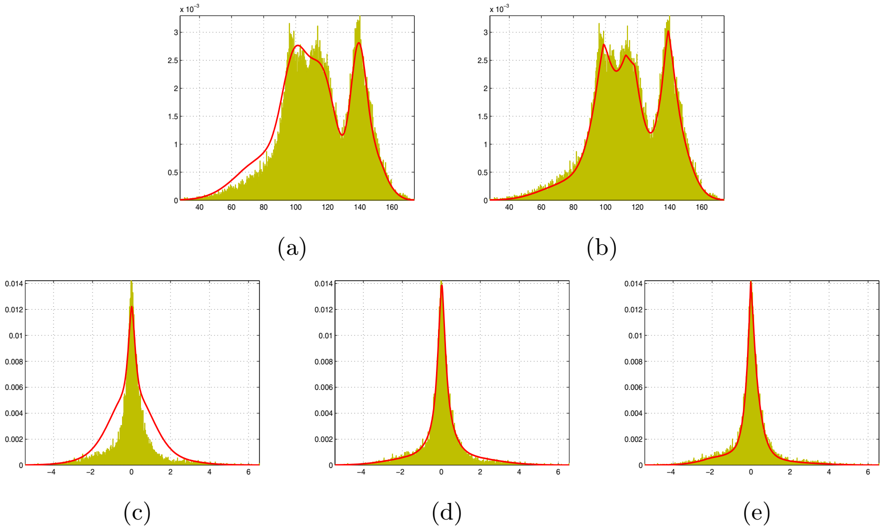


Fig. 3. Approximation of the histograms of the RCT subbands. The first row shows the approximation of a low-pass band using: (a) a mixture of 8 Gaussians and (b) a mixture of 8 GGDs. The second row shows the approximation of a detail subband C_{23} using: (c) one GGD, (d) a mixture of two GGDs and (e) a mixture of three GGDs.

The goal of sampling is to generate a sufficiently large sample $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ from the distribution f, in order to approximate the KLD using Monte-Carlo integration:

$$D_{mc}(f||g) = \frac{1}{n}\sum_{i=1}^{n} \log\frac{f(x_i)}{g(x_i)}, \tag{8}$$

which converges to $D(f||g)$ when $n \to \infty$. We tested several sampling techniques to find the most accurate approximation of KLD in (7), namely: rejection, importance and MCMC Metropolis-Hasting sampling techniques [17]. We found rejection sampling as the most accurate in our case [3]. In what follows, we use this sampling as our reference for developing our texture distance measurement.

Given a pyramidal decomposition of two images I_1 and I_2, we obtain for each image $4L$ high-pass subbands: C_{l1}, C_{l2}, C_{l3} and C_{l4} and one low-pass subband C_L. We calculate the similarity of I_1 and I_2 using the following function:

$$S(I_1, I_2) = \frac{1}{2}\sum_{l=1}^{L}\sum_{d=1}^{4}\left(D(f_1^{ld}||f_2^{ld})\right) + D(f_1^{L}||f_2^{L}), \tag{9}$$

where f_1^{ld} and f_2^{ld} stand for the subbands of level l and direction d for the images I_1 and I_2, respectively. The f_1^L and f_2^L correspond to the approximation (low-pass) subbands for the images I_1 and I_2, respectively. As mentioned previously, to calculate $S(I_1, I_2)$, we use rejection sampling since it provides the best approximation for the KLD.

5 Experiments

To measure the performance of the proposed approach, we conducted experiments using images from the VisTex dataset [13]. The dataset contains images of size 512 × 512. We selected 47 images from various texture categories and extracted 16 sub-images of size 256 × 256 in each image to construct a database of 752 sub-images. To eliminate the effect of range variation in the luminance of the sub-images, and thus reduce bias in the retrieval phase, we normalized the luminance of each sub-image as follows. Let μ_M and σ_M be the medians of the means and the standard deviations of the 16 sub-images that originate from the same image. Each sub-image I_i of the 16 is normalized using the following formula:

$$I_i = \frac{(I_i - \mu_i)}{\sigma_i} \times \sigma_M + \mu_M, \tag{10}$$

where μ_i and σ_i are the mean and the standard deviation of I_i. We measured the retrieval performance of the compared methods in terms of the retrieval rate (%), which refers to the number of relevant images found among the top-N retrieved images. In what follows, all the presented retrieval results have been obtained by averaging the retrieval rates corresponding to 47 randomly selected queries (a random query per each image category). We compared our approach for texture retrieval, denoted by RCT-MoGG, with recent state-of-the-art methods, namely:

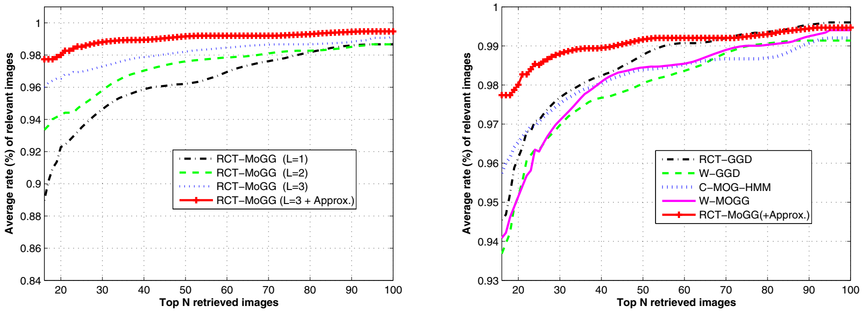
1. W-GGD: Wavelet-based texture retrieval using single GGD and KLD [7].
2. W-MoGG: Wavelet-based texture retrieval using MoGG and KLD [1].
3. C-MoG-HMM: Contourlet-based texture retrieval using mixtures of Gaussians and hidden Markov model trees [14].
4. RCT-GGD: RCT-based texture retrieval using single GGD and KLD.

Tab. 1 presents the average retrieval rates (%) in the top 16 images obtained by the compared methods. We can observe that methods using single GGDs for both wavelets and contourlets lead to the least retrieval rates. Our experiments agree with [1] in that W-MoGG gives better performance than W-GGD. The best method is RCT-MoGG(+Approx.) which corresponds to the MoGG modelling of RCT coefficients including the approximation subband. The method C-MoG-HMM gives slightly better results in comparison to RCT-GGD, but it is less accurate than RCT-MoGG. Fig. 4.a shows the retrieval rates with different combinations of scale levels. We can see that each further scale increases the performance of the retrieval rate. Excellent results, with retrieval rates above 98%, are obtained using 3 scale levels plus the approximation subband.

Finally, Fig. 4.b shows a comparison of RCT-MoGG to the other methods. We used 3 levels of pyramidal decomposition for each method. We can clearly see the superiority of the RCT-MoGG to the compared methods. From our observations, this performance comes from two main factors: 1) The use of MoGG to model the redundant contourlet coefficients which gives highly accurate histogram fitting, and 2) the inclusion of the approximation subband for retrieval which adds

Table 1. Average retrieval rate (%) in the top 16 images vs. the number of decomposition levels. Six methods are compared

Number of levels	W-GGD	W-MoGG	RCT-GGD	C-MoG-HMM	RCT-MoGG	RCT-MoGG (+ Approx.)
$L = 1$	86.11	88.83	83.45	91.48	88.96	97.74
$L = 2$	91.20	92.42	92.69	93.75	93.35	97.74
$L = 3$	92.60	93.88	94.55	95.14	96.01	97.74



(a) (b)

Fig. 4. (a) Retrieval performance of the RCT-MoGG using different numbers of RCT scale levels, (b) Comparison of our RCT-MoGG approach to the other methods

substantial information about the image appearance. In terms of computational time, we observed that W-GGD and RCT-GGD are more efficient than C-MoG-HMM, and RCT-MoGG is more efficient than C-MoG-HMM.

6 Conclusions

A new statistical-based texture characterization is proposed using finite MoGG modelling of redundant contourlets. The proposed approach allows leveraging the power of low-pass and high-pass contourlet coefficients for texture description and characterization. We successfully applied our approach for texture discrimination and retrieval. Compared to recent state-of-the-art methods, our approach has shown better results with retrieval rates above 98%. In the future, other contourlet decomposition filters will be investigated as well as applications of our approach for natural image classification and retrieval.

Acknowledgment. This work has been completed thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC) support.

References

1. Allili, M.S.: Wavelet-Based Texture Retrieval Using a Mixture of Generalized Gaussian Distributions. In: Int'l Conf. Pattern Recognition, pp. 3143–3146 (2010)
2. Allili, M.S., Bouguila, N., Ziou, D.: Finite General Gaussian Mixture Modelling and Application to Image and Video Foreground Segmentation. J. of Electronic Imaging 17, 013005 (2008)
3. Allili, M.S.: Similarity Measurements Between Finite Mixtures of Generalized Gaussian Distributions. Technical report (2011)
4. Baaziz, N.: Adaptive Watermarking Schemes Based on a Redundant Contourlet Transform. In: IEEE Int'l Conf. on Image Process., pp. I-221 –224 (2005)
5. Box, G.-E.P., Tiao, G.C.: Bayesian Inference in Stat. Analysis. Wiley classis, Chichester (1992)
6. Choy, S.-K., Tong, C.S.: Supervised Texture Classification Using Characteristic Generalized Gaussian Density. J. of Math. Imaging and Vision 29(1), 37–47 (2007)
7. Do, M.N., Vetterli, M.: Wavelet-Based Texture Retrieval Using Generalized Gaussian Density and KLD. IEEE Trans. on Image Process. 11(2), 146–158 (2002)
8. Do, M.N., et al.: The Contourlet Transform: An Efficient Directional Multiresolution Image Representation. IEEE Trans. on Image Process. 14(12), 2091–2106 (2005)
9. Fan, G., Xia, X.G.: Wavelet-Based Texture Analysis and Synthesis Using Hidden Markov Models. IEEE Trans. on Circuits and Systems-I 50(1), 106–120 (2003)
10. He, Z., Bystrom, M.: Color Texture Retrieval Through Contourlet-Based Hidden Markov Model. In: IEEE Int'l Conf. on Image Process., pp. 513–516 (2005)
11. Heeger, D.J., Bergen, J.R.: Pyramid-Based Texture Analysis/Synthesis. In: IEEE Int'l Conf. on Image Process., vol. 3, pp. 648–651 (1995)
12. Mirmehdi, M., et al.: Handbook of Texture Analysis. Imperial College Press, London (2008)
13. Vision Texture, http://vismod.media.mit.edu/
14. Po, D.D.-Y., Do, M.N.: Directional Multiscale Modelling of Images Using the Contourlet Transform. IEEE Trans. on Image Process. 15(6), 1610–1620 (2006)
15. Qu, H., et al.: Texture Image Retrieval Based on Contourlet Coefficient Modelling with Generalized Gaussian Distribution. In: Kang, L., Liu, Y., Zeng, S. (eds.) ISICA 2007. LNCS, vol. 4683, pp. 493–502. Springer, Heidelberg (2007)
16. Rao, S., Srinivas Kumar, S., Chatterji, B.N.: Content-Based Image Retrieval Using Contourlet Transform. ICGST-GVIP Journal 7(3) (2007)
17. Robert, C., Casella, G.: Monte Carlo Statistical Methods. Springer, Heidelberg (2010)
18. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. Information Science and Statistics. Springer, Heidelberg (2005)
19. Yifan, Z., Liangzheng, X.: Contourlet-Based Feature Extraction on Texture Images. In: Int'l Conf. on Computer Science and Software Engineering, pp. 221–224 (2008)

Evaluation of Histogram-Based Similarity Functions for Different Color Spaces

Andreas Zweng, Thomas Rittler, and Martin Kampel

Computer Vision Lab, Vienna University of Technology,
{zweng,kampel}@caa.tuwien.ac.at, thomas.rittler@tuwien.ac.at

Abstract. In this paper we evaluate similarity functions for histograms such as chi-square and Bhattacharyya distance for different color spaces such as RGB or L*a*b*. Our main contribution is to show the performance of these histogram-based similarity functions combined with several color spaces. The evaluation is done on image sequences of the PETS 2009 dataset, where a sequence of frames is used to compute the histograms of three different persons in the scene. One of the most popular applications where similarity functions can be used is tracking. Data association is done in multiple stages where the first stage is the computation of the similarity of objects between two consecutive frames. Our evaluation concentrates on this first stage, where we use histograms as data type to compare the objects with each other. In this paper we present a comprehensive evaluation on a dataset of segmented persons with all combinations of the used similarity functions and color spaces.

Keywords: similarity functions, tracking, color spaces, evaluation.

1 Introduction

Color histograms are the most common technique for extracting color features in computer vision due to theirs computational low cost. Furthermore, similarity functions are used to compute the similarity or the difference between two objects. In the field of tracking, color histograms and similarity measures take an important part in mean shift-based [3] and particle filter-based [4] tracking algorithms. In [5] similarity is computed in order to track detected persons. They use several stages of data association to increase the robustness of the tracker. In the first stage they combine single detections of consecutive frames, which has so called tracklets as output. A tracklet contains several temporal detections which are then combined in a second stage. In [6] a similar approach is used, where three different stages are used to track objects. In a low-level stage simple features are used to compute the similarity between persons. Features like position, size and appearance are used to compute a similarity only between two consecutive frames. In further stages the tracklets are then analyzed using the hungarian algorithm for example. Our motivation was to evaluate all combinations of well known similarity functions and color spaces in order to find the best combination which can then be used in different applications such as tracking.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 455–462, 2011.

This paper is organized as follows: In the next two sections we will provide a brief overview of the used color spaces and similarity functions. The evaluation results will be presented and discussed in section 4. The paper is concluded in section 5.

2 Color Spaces

In order to describe colors reproducibly, several color spaces have been evolved to meet different requirements. We have evaluated six well-established color spaces in the area of computer vision like tracking or image retrieval. In this section we shortly describe the area of usage of color spaces and their properties.

The RGB color space is an additive color space based on the three primary colors: red, green and blue. Colors are produced by adding components, with white having all colors present and black being the absence of any color. Accordingly, this implies that the RGB components are highly correlated with one another. The RGB color space is most commonly used for active devices such as computer monitors, digital cameras and television.

The HSV color space is similar to the human perception of color and considered to be better suited for human interaction. Colors are described in terms of the three components: hue, saturation and value. Hue refers to the pure spectrum color, saturation is the colorfulness and value is the brightness.

In the YCbCr color space is divided into luminance (Y) and chrominance (Cb, Cr). Cb and Cr denote the blue-yellow and the red-green color difference, respectively. This representation uses the characteristics of the human eye to be particularly sensitive to green light. The YCbCr color space is most frequently used in digital video and photography systems. Since the human visual system is less sensitive to changes in chrominance than to changes in luminance, the separation of Y and chroma components enables a higher compression rate and a reduction of bandwidth and storage space using chroma subsampling [7].

The I1I2I3 color space is a linear transformation of the RGB color space introduced in [8]. I1 represents intensity and I2 and I3 correspond to chromatic information, respectively. Since these components are less correlated, the I1I2I3 color space is more efficient in terms of region segmentation using color features.

XYZ is a standardized color space based on the CIE tristimulus values: X, Y and Z. These values were derived from the responsivity of the cone cells in the retina of the eye that are responsible for color vision. The projection to the unit plane X+Y+Z=1 defines the CIE chromaticity diagram, discarding the luminance Y. The XYZ color space comprises all human-visible colors.

The L*a*b* color space was derived from the XYZ color space to achieve perceptual uniformity. Similar to YCbCr, it consists of one lightness dimension (L) and two chrominance dimensions (a*, b*) based on the color opponent process. Additionally, L*a*b* is a equidistant and device independent color space, covering all percievable colors.

3 Similarity Functions

Similarity functions compute a similarity measure between datasets, in our case, two histograms. A taxonomy on histogram distance measures has been presented in [9]. Subsequently, let H_1 and H_2 denote the histograms of two images I_1 and I_2 respectively, each containing n bins with $i = 1 \ldots n$. Furthermore, it needs to be marked that the similarity functions differ in the range of their outcomes, which is elucidated, additionally to their main attributes, in this section.

Histogram intersection (HI) has been introduced in computer vision by [10], which is defined as:

$$d_{intersection}(H_1, H_2) = \sum_i \min(H_1(i), H_2(i)) \tag{1}$$

If both histograms are normalized by the total number of image pixels, HI is a similarity measure such that $0 \leq d_{intersection}(H_1, H_2) \leq 1$, where 1 indicates a perfect match and 0 is a complete mismatch. The resultant value of the intersection is the proportion of pixels from image I_1 that have corresponding pixels of the same color in image I_2.

The Bhattacharyya distance is a similarity measure between two probability distributions based on the Bhattacharyya coefficient [11][12]. The relation between Bhattacharyya and the χ^2 measure is presented in [13].

The chi-square (χ^2) measure is a commonly used statistical measure to calculate the similarity between frequency distributions based on the Pearsons's chi-square test statistic.A total mismatch of the chi-square similarity measure is unbounded, i.e., the value of its maximum depends on the size of the histogram. In contrast, 0 indicates a perfect match. Furthermore, a singularity problem occurs when comparing frequencies of the distributions that are both zero. An upper bound of 1.0 has been found empirically as the best choice for evaluation.

The *Pearson Product-Moment Correlation Coefficient* is a widely used measure of linear dependence between two random variables or the linear dependence of a bivariate random variable. It is defined as the quotient between their covariance and the product of their respective standard deviations. Perfect positive and negative linear relationship between X and Y yields to, respectively, 1 and -1, whereas a value of 0 means that X and Y are not linearly correlated. Experiments showed, that the output range lies between -0.1 and 1.0. Since negative values are unlikely we assigned the value 0 to all negative outputs for evaluation to achieve the range [0, 1].

The earth mover distance computes the effort for shifting a histogram into another histogram for similarity computation [2]. We have not evaluated the EMD because the computational cost of the algorithm is too high for a real-time computation in video sequences (with image sizes of e.g.: 640x480 pixel).

4 Evaluation

In the sequence `S2.L1.12-34.View001` of the PETS 2009 dataset we have choosen 50 consecutive frames and segmented three different persons (see Fig. 1). The

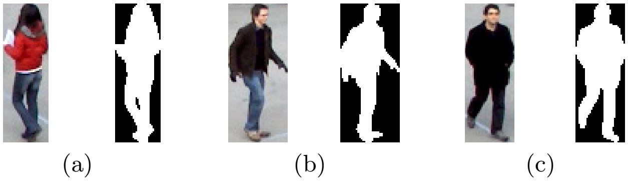

(a) (b) (c)

Fig. 1. Example of subject images and corresponding binary masks (PETS 2009)

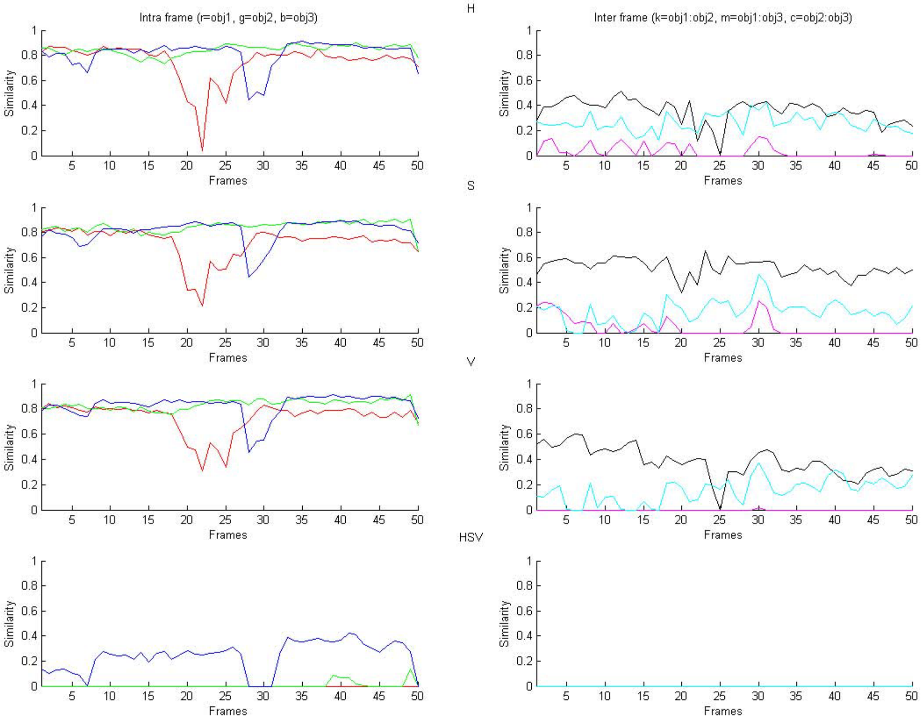


Fig. 2. Detailed similarities for the HSV space using χ^2

persons are manually pre-segmented in order to increase the precision of the evaluation of the similarity functions. After the color space transformation of the subject images, the color histograms using the binary masks are computed per frame. The histograms are constructed for each color channel independently as well as for all three channels combined (3D-histogram), to preserve their spatial linkage. Finally, after histogram normalization, the several similarity functions are applied and their outcomes are adapted to the range $[0, 1]$, where a value of 0 indicates complete dissimilarity.

The left column in Figures 2 and 3 represents the Intra-object similarity and the right column the Inter-object similarity, where rows depict the individual color channels and 3D. The subjects a, b, c are color coded as red, green, blue as

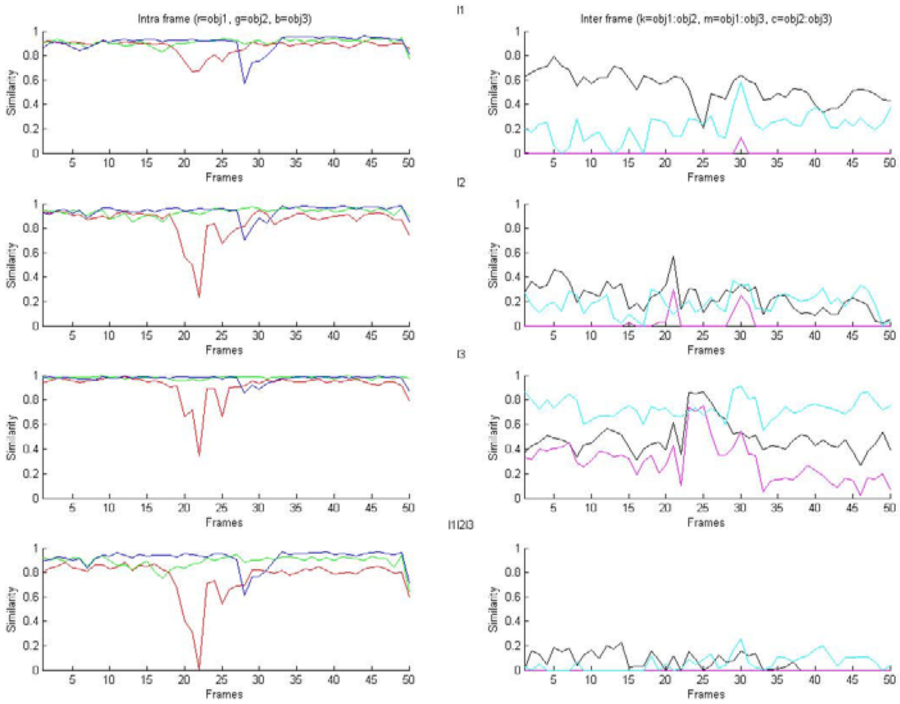
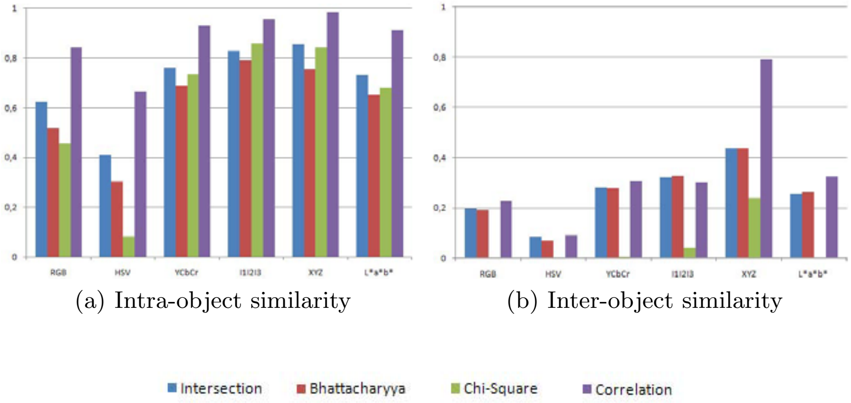


Fig. 3. Detailed similarities for the I1I2I3 color space using χ^2

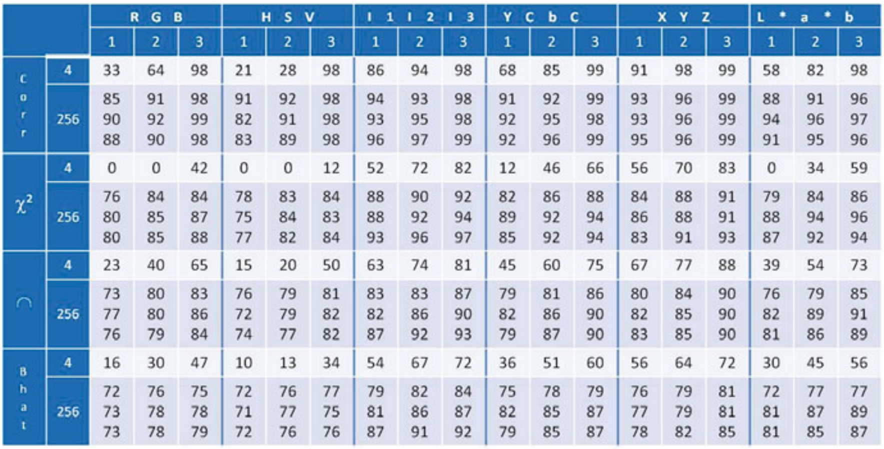


(a) Intra-object similarity

(b) Inter-object similarity

Fig. 4. Similarities combining all three channels (3D-histograms)

well as their combinations thereof 1:2, 1:3, 2:3 as black, magenta, cyan, respectively. Similarities of one person, named *Intra-object similarity*, are calculated on consecutice frames of this subject. On the other side, similarities (here: dissimilarities) between different objects, named *Inter-object similarity*, are computed



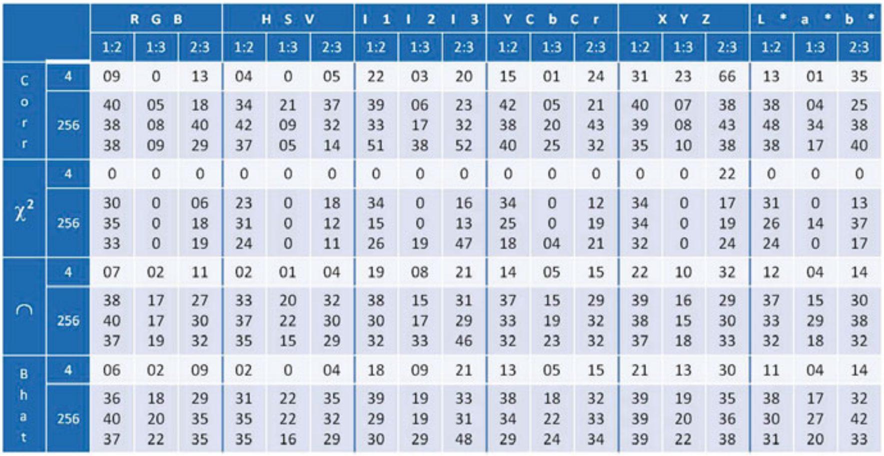
		RGB			HSV			I1I2I3			YCbC			XYZ			L*a*b		
		1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Corr	4	33	64	98	21	28	98	86	94	98	68	85	99	91	98	99	58	82	98
	256	85	91	98	91	92	98	94	93	98	91	92	99	93	96	99	88	91	96
		90	92	99	82	91	98	93	95	98	92	95	98	93	96	99	94	96	97
		88	90	98	83	89	98	96	97	99	92	96	99	95	96	99	91	95	96
χ^2	4	0	0	42	0	0	12	52	72	82	12	46	66	56	70	83	0	34	59
	256	76	84	84	78	83	84	88	90	92	82	86	88	84	88	91	79	84	86
		80	85	87	75	84	83	88	92	94	89	92	94	86	88	91	88	94	96
		80	85	88	77	82	84	93	96	97	85	92	94	83	91	93	87	92	94
$\cap$	4	23	40	65	15	20	50	63	74	81	45	60	75	67	77	88	39	54	73
	256	73	80	83	76	79	81	83	83	87	79	81	86	80	84	90	76	79	85
		77	80	86	72	79	82	82	86	90	82	86	90	82	85	90	82	89	91
		76	79	84	74	77	82	87	92	93	79	87	90	83	85	90	81	86	89
Bhat	4	16	30	47	10	13	34	54	67	72	36	51	60	56	64	72	30	45	56
	256	72	76	75	72	76	77	79	82	84	75	78	79	76	79	81	72	77	77
		73	78	78	71	77	75	81	86	87	82	85	87	77	79	81	81	87	89
		73	78	79	72	76	76	87	91	92	79	85	87	78	82	85	81	85	87

Fig. 5. Mean Intra-object similarities of all frames per subject (0,100)

on the subject images of the same frame. Moreover, for the 3D-histograms the color space is divided into subcubes, where the edge length of one cube is equal to the segment size, i.e., a transformation of the three color dimensions to 1D and the number of bins is defined by $(\frac{\text{number of bins}}{\text{segment size}})^3$. In Figure 2 - Figure 7 we set the bin count to 256 and the segment size to 8. The number of histogram bins influences the tradeoff between robustness and discriminability to the extent that a lower quantity of bins increases the *Intra-object similarity* as well as *Inter-object similarity*, since it decreases the number of different color features.

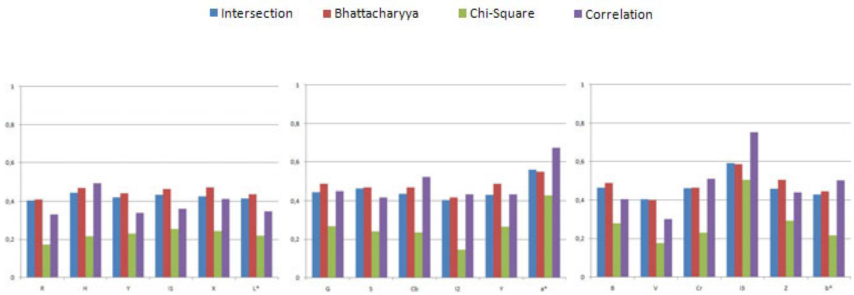
Experiments show that clamping the values of the unbounded χ^2 - function to the range [0,1] delivers reasonable results for normalized histograms. In a similar way negative values of the correlation are transformed. In Figure 2 and Figure 3 the results of the bounded χ^2 - similarity measure for the I1I2I3 and HSV color space are presented. The hugh dips for subject 1 in frame 22 and subject 3 in frame 28 are due to occlusions in the image sequence. Since greater differences are more punished by the χ^2 - function, these dips are less marked in the remaining similarity measures. Due to the properties of the individual HSV dimensions, combining all three channels is ineffectual in general.

In Figure 4 the average similarities for 3D-histograms are visualized. Here, correlation yields the highest scores for *Intra-object similarity* as well as for *Inter-object similarity* except from the I1I2I3 color space, where the color dimensions are almost uncorrelated. As a result, the I1I2I3 color space delivers the best *Intra-/Inter-object similarity* ratio on the stated similarity measures followed by YCbCr. Additionally, it can be especially observed that the XYZ color space is inapplicable for color histograms due to its definition. Figure 7 endorses that the I3 color dimensions can be neglected for similarity measures, since it contains less information as stated in [8]. *Intra-object similarity* for single channel histograms yields values between 0.75 and 0.95.



		R G B			H S V			I 1 I 2 I 3			Y C b C r			X Y Z			L * a * b *		
		1:2	1:3	2:3	1:2	1:3	2:3	1:2	1:3	2:3	1:2	1:3	2:3	1:2	1:3	2:3	1:2	1:3	2:3
Corr	4	09	0	13	04	0	05	22	03	20	15	01	24	31	23	66	13	01	35
	256	40 38 38	05 08 09	18 40 29	34 42 37	21 09 05	37 32 14	39 33 51	06 17 38	23 32 52	42 38 40	05 20 25	21 43 32	40 39 35	07 08 10	38 43 38	38 48 38	04 34 17	25 38 40
χ^2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0
	256	30 35 33	0 0 0	06 18 19	23 31 24	0 0 0	18 12 11	34 15 26	0 0 19	16 13 47	34 25 18	0 0 04	12 19 21	34 34 32	0 0 0	17 19 24	31 26 24	0 14 0	13 37 17
$\cap$	4	07	02	11	02	01	04	19	08	21	14	05	15	22	10	32	12	04	14
	256	38 40 37	17 17 19	27 30 32	33 37 35	20 22 15	32 30 29	38 30 32	15 17 33	31 29 46	37 33 32	15 19 23	29 32 32	39 38 37	16 15 18	29 30 33	37 33 32	15 29 18	30 38 32
Bhat	4	06	02	09	02	0	04	18	09	21	13	05	15	21	13	30	11	04	14
	256	36 40 37	18 20 22	29 35 35	31 35 35	22 22 16	35 32 29	39 29 30	19 19 29	33 31 48	38 34 29	18 22 24	32 33 34	39 39 39	19 20 22	35 36 38	38 30 31	17 27 20	32 42 33

Fig. 6. Mean Inter-object similarities of all frames per subject combination (0,100)



(b) First Channel (c) Second Channel (d) Third Channel

Fig. 7. Inter-object similarities seperated into individual channels

An overview of different results of our evaluation is shown in Figures 5 and 6. Rows represent the different similarity functions, where row '256' indicates single channel histograms using 256 bins and '4' depicts 3D histograms segmented in blocks of size 4, i.e. $(256/4)^3$ bins, respectively; columns represent the different color spaces and subjects combinations, where '1:2' shows the similarity between subject 1 and subject 2 for example. In Figure 6 higher scores indicate better results, where in Figure 5 lower scores indicate better results.

5 Conclusion

In this paper we examined the effect of several histogram-based similarity functions for different color spaces. Our analysis show that the best result for high

Intra-object similarity and low *Inter-object similarity* is achieved by a bounded version of the χ^2 - similarity measure in the I1I2I3 and the YCbCr color space considering all three color dimensions. For single channel histogram comparision I2 of the I1I2I3 color space provides a decent choice. Further, histogram intersection provides similar results to the Bhattacharyya distance at computational lower cost. Our results can be used for applications which are related to tracking. People detection such as the human detector using the histograms of oriented gradients feature [1] or a connected component computation are required to be computed before tracking. A tracker then tries to assign the single detections in consecutive frames where similarity functions are used. Using our evaluation the best combination using the mentioned color spaces and similarity functions is known.

References

1. 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 (2005)
2. Levina, E., Bickel, P.: The Earth Mover's distance is the Mallows distance: some insights from statistics. In: Proceedings of the eighth IEEE International Conference on Computer Vision (ICCV 2001), vol. 2, pp. 251–256 (2001)
3. Changjun, W., Li, Z.: Mean shift based orientation and location tracking of targets. In: 6th International Conference on Natural Computation (ICNC), pp. 3593–3596 (2010)
4. Jiang, F., Cheng, Y., Li, H.: Research of particle filter tracking algorithm based on structural similarity. In: 3rd International Congress on Image and Signal Processing (CISP), pp. 368–371 (2010)
5. Li, Y., Huang, C., Nevatia, R.: Learning to associate: HybridBoosted multi-target tracker for crowded scene. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 2953–2960 (2009)
6. Li, Y., Huang, C., Nevatia, R.: Robust Object Tracking by Hierarchical Association of Detection Responses. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 788–801. Springer, Heidelberg (2008)
7. Poynton, C.A.: A Technical Introduction to Digital Video, p. 175. John Wiley & Sons, Inc., Chichester (1996)
8. Ohta, Y.I., Kanade, T., Sakai, T.: Color information for region segmentation. Computer Graphics and Image Processing 13, 222–241 (1980)
9. Cha, S.: Taxonomy of nominal type histogram distance measures. In: Proceedings of the American Conference on Applied Mathematics, pp. 325–330 (2008)
10. Swain, M.J., Ballard, D.H.: Indexing via colour histograms. In: Proceedings of the Third International Conference on Computer Vision, pp. 390–393 (1990)
11. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. In: Bulletin of the Calcutta Mathematical Society, pp. 99–109 (1943)
12. Comaniciu, D., Ramesh, V., Meer, P.: Real-Time Tracking of Non-Rigid Objects using Mean Shift. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR 2000), Vol. 2, pp. 142–149 (2000)
13. Aherne, F., Thacker, N., Rockett, P.: The Bhattacharyya metric as an absolute similarity measure for frequency coded data. Kybernetika 32(4), 1–7 (1997)

Color Contribution to Part-Based Person Detection in Different Types of Scenarios

Muhammad Anwer Rao, David Vázquez, and Antonio M. López

Computer Vision Center and Computer Science Dpt.,
Universitat Autònoma de Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
{muhammad,david.vazquez,antonio}@cvc.uab.es, www.cvc.uab.es/adas

Abstract. Camera-based person detection is of paramount interest due to its potential applications. The task is difficult because the great variety of backgrounds (scenarios, illumination) in which persons are present, as well as their intra-class variability (pose, clothe, occlusion). In fact, the class *person* is one of the included in the popular PASCAL visual object classes (VOC) challenge. A breakthrough for this challenge, regarding *person* detection, is due to Felzenszwalb *et al.* These authors proposed a part-based detector that relies on histograms of oriented gradients (HOG) and latent support vector machines (LatSVM) to learn a model of the whole human body and its constitutive parts, as well as their relative position. Since the approach of Felzenszwalb *et al.* appeared new variants have been proposed, usually giving rise to more complex models. In this paper, we focus on an issue that has not attracted sufficient interest up to now. In particular, we refer to the fact that HOG is usually computed from RGB color space, but other possibilities exist and deserve the corresponding investigation. In this paper we challenge RGB space with the opponent color space (OPP), which is inspired in the human vision system. We will compute the HOG on top of OPP, then we train and test the part-based human classifier by Felzenszwalb *et al.* using PASCAL VOC challenge protocols and *person* database. Our experiments demonstrate that OPP outperforms RGB. We also investigate possible differences among types of scenarios: indoor, urban and countryside. Interestingly, our experiments suggest that the benefits of OPP with respect to RGB mainly come for indoor and countryside scenarios, those in which the human visual system was *designed* by evolution.

1 Introduction

Camera-based person detection is of great interest for applications in the fields of content management, video-surveillance and driver assistance. Person detection is difficult because the great variety of backgrounds (scenarios, illumination) in which persons are present, as well as their intra-class variability (pose, clothe, occlusion). Currently, discriminative part-based approaches [1,2], that heavily rely on dynamic part detection, constitute the state of the art for detecting persons.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 463–470, 2011.

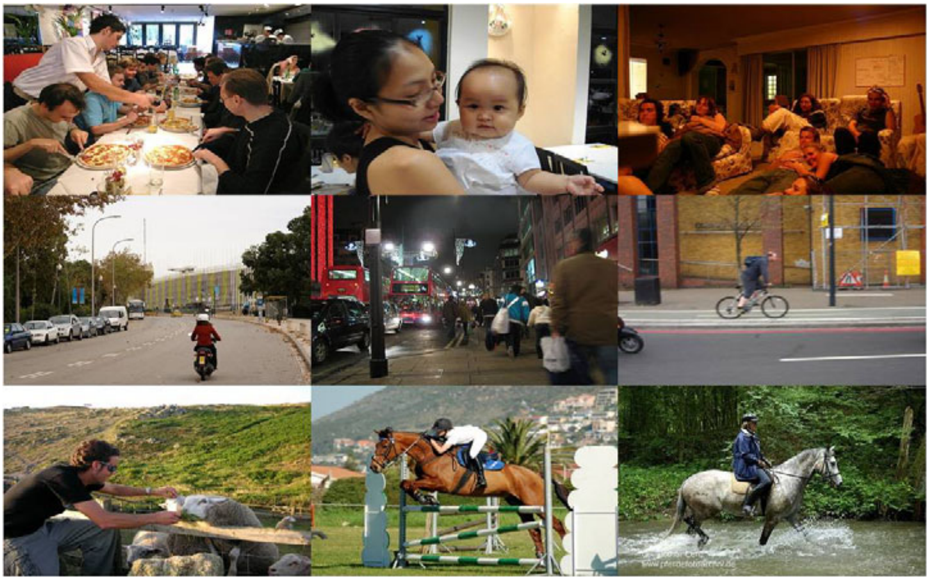


Fig. 1. Annotation enrichment for PASCAL VOC 2007 dataset. First, second and third rows show images that we have annotated as *indoor*, *urban* and *countryside*, resp.

The part-based human detectors generally use the histograms of oriented gradients (HOG) introduced in [3] by Dalal *et al.* as low-level features for building person models. HOG features are computed on top of RGB color space. On the other hand, in the context of image categorization [4] it has been demonstrated the usefulness of the so-called *opponent color space* (OPP) when working with the so-called SIFT descriptor [5]. Since HOG are SIFT-inspired, we think it is worth to test the use of opponent colors for person detection, *i.e.*, replacing the RGB color space by the OPP one in the part-based person detection method described in [2]. Moreover, we are interested in assessing if person detection performance can be affected by the type of scenario where it is performed. In other words, we want to perform a scenario-based comparison between the OPP and RGB color spaces, when *pugged-in* for HOG-part-based person detection.

As scenarios we have chosen three relevant types: indoor, countryside and urban. In order to conduct our experiments we use the class *person* included in the popular PASCAL visual object classes (VOC) challenge [6]. We have enriched the annotation with the *indoor, countryside* and *urban* labels, both for training and testing data (Fig. 1). As we will see, our experiments suggest that the benefits of OPP with respect to RGB mainly come for indoor and countryside scenarios, those in which the human visual system was *designed* by evolution.

The rest of the paper is organized as follows. In section 2 we summarize our proposal of using opponent colors with part-based person detection. Section 3 details the conducted experiments, while in Sect. 4 we discuss the obtained results. Finally, Sect. 5 draws the conclusions and future work.

2 Part-Based Person Detector Based on Opponent Colors

The part-based paradigm, introduced by Fischler and Elshlager dates back to 1973 [7]. It provides an elegant way of representing an object category and is particularly efficient for object localization. This model has been built and extended in many direction according to different problems in the computer vision field. Here, we will briefly overview the main principles of part-based methods.

In part-based models, the focus remains on modelling an object as having a number of parts arranged in a deformable configuration. Each part captures the appearance of the object at local level and there is some flexibility in object-parts placement to account for global deformations. The best configuration of such a model is framed on an image as an energy minimization problem which measures the score for each part and deformation score for each pair of connected parts. Part-based models can be separated into many categories depending upon the connection structure to represent the parts: constellation model, star-shaped, tree-shaped, bag of features, etc. Recently, [1,2] has adopted the star-structured part-based model, which has shown to provide excellent results on human detection [6]. The appearance of an object is represented by histograms of oriented gradients (HOG) features in a 31-dimensional feature vector. HOG of part filters are captured at twice the resolution of the root (full-body) filter to model appearance at multiple scales. Here we follow the implementation associated to [2], whose code has been kindly put publicly available by the authors.

In this implementation, and many others derived from it, HOG features are computed on top of RGB color space. Or more precisely, on top of the *max-gradient* operation on RGB color space (*i.e.*, $\max\{\nabla R, \nabla G, \nabla B\}$). This way of computing HOG derives from the original work by Dalal *et al.* [3] where HOG features were defined in the context of a holistic person detector.

However, in the context of image categorization [4] it has been demonstrated the usefulness of the so-called *opponent color space* (OPP) when working with the so-called SIFT descriptor [5]. Since HOG are SIFT-inspired, we think it is worth to test the use of opponent colors for person detection, *i.e.*, replacing the RGB color space by the OPP one in the part-based person detection method in [2]. Accordingly, we briefly summarize OPP in the rest of the section.

Opponent process theory postulates that yellow-blue and red-green information is represented by two parallel channels in the visual system that combine cone signals differently. It is now accepted that at an early stage in the red-green opponent pathway, signals from L and M cones are opposed and, in the yellow-blue pathway, signals from S cones oppose a combined signal from L and M cones [8]. In addition, there is a third luminance or achromatic mechanisms in which retinal ganglion cells receive L- and M- cone input. Thus, L, M and S belong to a first layer of the retina whereas luminance and opponent colors belong to a second layer of it, forming the basis of chromatic input to the primary visual cortex. Note also that this mechanism is not random since human color vision evolved for increasing the probability of subsistence [9].

Seeing the RGB space used for codifying color in digital images as the LMS color space of the first layer of human retina, we can also compute an opponent colors (OPP) space as follows [4]:

$$\begin{aligned}\text{red-green}: O_1 &= (R-G)/\sqrt{2}\ ,\\ \text{yellow-blue}: O_2 &= ((R+G)-2B)/\sqrt{6}\ ,\\ \text{luminance}: O_3 &= (R+G+B)/\sqrt{3}\ ,\end{aligned} \tag{1}$$

3 Experiments

In this paper we want to address the following specific questions in the context of part-based person detection: (**Q1.**) *if our detector must work in specific scenarios, is it better to use OPP or RGB?*. This is useful to know it for specific systems that must work in specific locations (e.g., intruder detection, pedestrian detection, etc.) rather than as general computer vision systems. (**Q2.**) *if we don't know a prior the scenario in which our detector must work, is it better to use OPP or RGB?*. This question is more related to general systems that must work in a broad spectrum of environments (e.g., automatically detecting people for focusing before a camera shot).

In order to answer **Q1** we will run experiments where person classifiers, based on RGB and OPP, are trained and tested in specific scenarios. We have selected three different and relevant scenarios: indoor, countryside and urban (Fig. 1). In particular we will run the part-based method summarized in Sect. 2, with the only difference of the input color space used before computing the HOG descriptors: we run equivalent experiments for RGB and OPP. We will use the *person* class of the PASCAL VOC detection challenge of 2007. The reason for using the data from 2007 is that it was the last time that testing annotations were provided. We need such annotations to enrich them with the different scenarios we have mentioned (Fig. 1). After doing such enrichment for training and testing data, we obtain the numbers of training windows and testing images per scenario summarized in Tab. 1.

In order to answer **Q2** we run experiments analogous to the scenario-based ones, but without actually distinguishing the scenario. In other words, we perform the type of experiments that PASCAL VOC challenge participants do, for the cases of RGB and OPP color spaces. Additionally, we not only present the

Table 1. Training and testing numbers per scenario: person windows (+); images without persons (-); initial background windows (-) after sampling 200 one per image without persons; number of images for testing as well as persons to be detected

	Training			Testing	
	Windows (+)	Images (-)	Initial Windows (-)	Images	Windows (+)
Indoor	(45.5%) 4268	(36.0%) 516	(36.0%) 103200	(41.0%) 2031	(49.1%) 2252
Countryside	(18.8%) 1762	(29.0%) 414	(29.0%) 82800	(29.5%) 1463	(22.2%) 1004
Urban	(35.7%) 3350	(35.0%) 501	(35.0%) 100200	(29.5%) 1458	(28.1%) 1272
Overall	9380	1431	286200	4952	4528

overall result on the full testing dataset but also the results of applying the overall classifiers (*i.e.*, the ones trained without taking into account the scenarios) to each considered scenario separately.

It is worth to mention that Felzenszwalb *et al.* method computes the HOG over the *max-gradient* as we have seen in Sect. 2, however, we compute separate HOG features for each OPP channel. Thus, our features are of a dimension three times higher than the usually used for HOG computation. Nevertheless, for a fair comparison we did similar experiments using the separate R, G and B channels in an analogous use to the three OPP channels. The results were basically analogous to the use of *max-gradient* for RGB, thus, the conclusions of this paper do not change. Accordingly, here on we will only focus on the usual procedure found in the literature, *i.e.*, computing the *max-gradient* for RGB. Note that while RGB channels are highly correlated ones, OPP ones are not.

For the training of any classifier we apply the bootstrapping method to collect hard negatives. We follow the scheme provided by the publicly available software of Felzenszwalb *et al.*, which collects all possible hard negatives until filling 3GB of working memory. In practice, this means to perform about ten bootstrappings.

In order to evaluate the obtained results, we follow the PASCAL VOC 2007 protocol, which is based on *precision-recall* (PR) curves and the associated *average precision* (AP). Please, refer to [6] for more details about such protocol.

In summary, the experiments to be done are:

– *Indoor, countryside* and *urban* classifiers: they are learnt from indoor images and applied to indoor images. The same for countryside and urban ones.
– *Overall* classifier: it is learnt from all the images but tested in different ways: on all the test images; only in the test images classified as *indoor*; only *countryside*; only *urban.*

These experiments must be run for OPP and RGB color spaces. Thus, we get a total of 14 PR curves and corresponding APs. Figure 2 shows all the PR curves in a meaningful way and Tab. 2 presents the corresponding APs. Additionally, we also applied each scenario-specifically-trained classifier to the other scenarios (not trained). We do not plot the corresponding PR curves for the sake of simplicity but we include the respective APs in Tab. 2.

4 Discussion

Results summarized in Fig. 2 and Tab. 2 allow to answer the questions **Q1** and **Q2** stated in Sect. 3.

Table 2 shows that AP in indoor scenarios is 1.7 points higher for OPP than for RGB when using only such type of scenarios for training. In the case of countryside the difference is even higher, 3.1 points. However, in urban scenarios RGB performs 0.6 points better.

A closer look to the PR curves (Fig. 2) for indoor, urban and countryside scenarios gives more detailed insight. In the case of indoor scenarios we appreciate that for the specifically trained and tested classifiers the difference between OPP

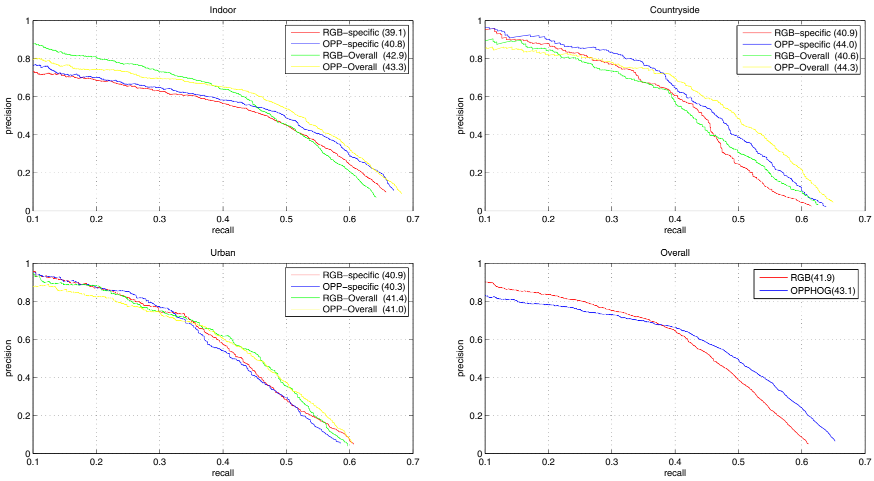


Fig. 2. Precision-recall (PR) curves obtained from the different experiments are shown: using RGB and OPP color spaces, for the indoor, countryside, urban and overall classifiers. The average precision (AP) of each PR curve is the number shown inside the respective parenthesis. The PRs of the specific classifiers are plotted together with the PRs of the overall classifiers applied only in the corresponding specific scenarios.

Table 2. Average precision (AP) in % of the different trained and tested classifiers. Indoor/Countryside/Urban/Overall in the first column refer to the training step, while Indoor/Countryside/Urban in the second row refer to testing. Bold numbers indicate the higher APs comparing the counterpart RGB and OPP results. For the overall classifiers we not only include the overall APs, but also the APs corresponding to apply such classifiers only to specific scenarios during testing.

	RGB			OPP		
	Indoor	Countryside	Urban	Indoor	Countryside	Urban
Indoor	39.1	21.8	21.2	**40.8**	23.4	22.8
Countryside	22.0	40.9	31.1	24.9	**44.0**	33.4
Urban	29.9	34.9	**40.9**	33.3	39.8	40.3
Overall	41.9			**43.1**		
	42.9	40.6	**41.4**	**43.3**	**44.3**	41.0

and RGB is higher for higher recall. This fact is not captured by the AP computation method used in PASCAL VOC 2007 detection challenge. Note, that detection systems are usually interested in having higher recall. In countryside scenarios we observe an analogous situation, but with higher differences. In the case of urban scenarios we see that the specifically trained classifiers are pretty similar along the whole PR plot.

From these observations we conclude that the answer to question **Q1** is: *for indoor and countryside scenarios OPP color space performs better than RGB, while for urban scenarios it seems that there is not a clear preference for mid-to-high recalls.* The major benefit of OPP is for countryside scenarios. Interestingly, OPP color space is the result of human evolution inside primitive indoor and countryside environments, not urban ones, where humans were targets of interest among others. Primitive indoor scenarios are of different background than modern ones. However, countryside colors remain constant. Of course, we don't argue here that our experiments are supporting psychological/evolutive claims about the human vision system, we only want to point out here what in our modest opinion is an interesting fact.

Regarding question **Q2**, Tab. 2 shows that when jointly using all human windows and backgrounds for training, the AP is 1.2 points higher for OPP than for RGB. Again, by a closer look to PR curves (Fig. 2) for the overall case, we observe that the major benefit of OPP comes for recalls over 40%, *e.g.*, for a recall of the 50% we obtain about ten points more of precision with OPP. We can also assess the performance of these overall classifiers focused on our specific scenarios. We observe then that for the indoor ones, for recalls below the 40% RGB is giving higher precision, however, over such recall the situation changes. The AP is 0.4 points higher for OPP than for RGB. The case of countryside scenarios is analogous but here the OPP starts to offer better precision before, approximately for recalls higher than the 22%. The AP is 3.7 points higher for OPP. In urban scenarios precision is higher with RGB than with OPP for recalls lower than approximately the 30%, however, over such recall OPP and RGB behave pretty similar. The AP is 0.4 points higher for RGB.

From these observations we conclude that the answer to question **Q2** is: *combining data coming from different scenarios during training helps to potentially obtain benefits from OPP over RGB, however, the final benefits will only be obtained if the classifier is used in indoor and countryside scenarios.* Note that the best scenario for OPP, i.e., countryside according to our experiments, is the less represented in the training of overall classifiers (Tab. 1). During testing, countryside and urban scenarios are, basically, equally represented, but indoor scenarios gain in testing presence (Tab. 1), which probably is the reason for OPP offering an overall improvement over RGB (countryside cases help AP for OPP while urban cases help RGB).

In summary, using OPP for human detection is worth out of urban scenarios, specially for countryside. Examining Tab. 2 one could be also tempted to conclude that overall detectors outperform the specifically trained ones, however, we think that this can be only an effect of the number of examples and counter-examples during training. What is clear (and expected), however, is that classifiers trained only in one type of scenario perform poorly in the other types of scenarios.

5 Conclusion

In this paper we have investigated the effect of using the opponent color space, which is based on the human vision system, for person detection. We have taken as baseline person detector the HOG and part-based method proposed by Felzenszwalb *et al.*. Then, by following the protocols of the PASCAL VOC challenge of 2007, applied to the *person* class, we have collected experimental results that state that opponent color space is a better choice for computing HOG in indoor and, specially, countryside environments. In urban scenarios, there is no clear benefit. Interestingly, indoor and countryside scenarios, those in which the human visual system was *designed* by evolution. The combination of opponent color scape and Felzenszwalb *et al.* method as well as the scenario-based study are new up to the best of our knowledge.

As future work we plan to combine scenario-specific trained classifiers with image classifiers so that, given a new image of unknown type, we can compute the type of scenario to which it belongs (or a probability for each type) and apply a selection methodology (or a fusion scheme) in order to obtain the best benefit of the learned classifiers.

Acknowledgments. This work has been supported by the Spanish Government projects TRA 2010-21371-C03-01 and Consolider Ingenio 2010: MIPRCV (CSD200700018).

References

1. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, AK, USA (2008)
2. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with dscriminatively trained part based models. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(9), 1627–1645 (2010)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conf. on Computer Vision and Pattern Recognition, San Diego, CA, USA (2005)
4. van de Sande, K., Gevers, T., Snoek, C.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
5. Lowe, D.: Object recognition from local scale-invariant features. In: Int. Conf. on Computer Vision, Kerkyra, Greece (1999)
6. Everingham, M., Gool, L.V., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. Journal on Computer Vision 88(2), 303–338 (2010)
7. Fischler, M., Ekschlager, R.: The representation and matching of pictorial structures. IEEE Transactions on Computers 100(22), 67–92 (1973)
8. Krauskopf, J., Williams, D.R., Heeley, D.W.: Cardinal directions of color space. Vision Research 22(9), 1123–1132 (1982)
9. Mollon, J.D.: "tho she kneel'd in that place where they grew.." the uses and origins of primate colour vision. Journal of Experimental Biology 146(1), 21–38 (1989)

Content Adaptive Image Matching by Color-Entropy Segmentation and Inpainting

Yuanchang Sun and Jack Xin

Math Department, Univ of California Irvine, Irvine, CA 92697, USA
yuanchas@uci.edu, jxin@math.uci.edu

Abstract. Image matching is a fundamental problem in computer vision. One of the well-known techniques is SIFT (scale-invariant feature transform). SIFT searches for and extracts robust features in hierarchical image scale spaces for object identification. However it often lacks efficiency as it identifies many insignificant features such as tree leaves and grass tips in a natural building image. We introduce a content adaptive image matching approach by preprocessing the image with a color-entropy based segmentation and harmonic inpainting. Natural structures such as tree leaves have both high entropy and distinguished color, and so such a combined measure can be both discriminative and robust. The harmonic inpainting smoothly interpolates the image functions over the tree regions and so blurrs and reduces the features and their unnecessary matching there. Numerical experiments on building images show around 10% improvement of the SIFT matching rate with 20% to 30% saving of the total CPU time.

Keywords: Content Adaptivity, Color-Entropy Segmentation, Inpainting, Image Matching.

1 Introduction

Image matching is a fundamental aspect of numerous tasks in computer vision, such as object and scene recognition, 3D structure reconstruction from multiple planar images, stereo correspondence, and motion tracking. Two key ingredients are critical to matching accuracy and efficiency. One is the selection of image features that are distinctive between different image classes yet invariant to unimportant attributes of the images (e.g. scaling and rotation), and partially invariant to change in illumination and camera viewpoint. The other is the design of a reliable matching criterion that quantifies feature similarity. A successful technique is Lowe's scale-invariant feature transform (SIFT, [5]). SIFT extracts from an image a collection of feature vectors (feature point descriptors), each of which is invariant to image translation, scaling, and rotation, partially invariant to illumination changes and robust to local geometric distortion. Such invariant features have proven useful in the context of image registration and object recognition. SIFT assigns location, scale, and orientation to each feature point. The location is defined as extrema of Difference of Gaussian functions

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 471–478, 2011.

applied in scale-space to a series of smoothed and re-sampled images. Due to the large numbers of extrema, there could be up to thousands of detected feature points in an image, see Fig. 1. Yet many of the SIFT feature points are often found to be insignificant as they may well fall on uninteresting part of an image. As in Fig. 1, a building is typically surrounded by a natural landscape consisting of trees, lawns or bushes which can attract a large number of SIFT feature points. In the image of the left panel of Fig. 4, there are 4692 SIFT feature points. Almost half of them are on trees and lawns. Clearly, the feature points on the building are more important and reliable (less sensitive to weather and season) for building identification.

In this paper, we propose a content adaptive measure to exclude the uninteresting feature points, so subsequent image matching focuses more on the man-made corner-like features on the buildings. The adaptive measure combines color and entropy information of the image pixels. We shall locate the tree like natural regions by a combined measure of color and entropy, and apply image inpainting technique to blur and smooth out features in the tree regions. SIFT then acts on the inpainted image where regions of no interest have been weakened. Matching of feature points follows that of [5]. In brief, a descriptor (a multi-dimensional vector) is assigned to each feature point. A match p' in image A for a feature point p in image B is found by identifying the nearest neighbor of p from the feature points in image A. The nearest neighbor is defined as the feature point with minimum Euclidean distance in terms of the associated descriptor vectors. In [6], Lowe proposed a more robust matching criterion which we shall discuss later.

The paper is organized as follows. In section 2, we discuss the identification of tree like natural structures by a combined color and entropy measure. In section 3, we present an image inpainting method based on solving the Laplace equation on irregular domains to blur out the trees regions. In section 4, we discuss feature points matching criterion, and present experimental matching results on building images. The conclusion and remarks on future work are in section 5.

2 Color and Entropy Based Segmentation

Let us consider to match a building image to a similar image of the same scene. The images may contain background of trees, bushes, lawns, and sky. The goal is to locate the trees (or other vegetation) in an automatic fashion. A straightforward way is to use the colors. It is known that L*a*b color model is a better choice for images of natural scenes than RGB color space. The L*a*b* color space is derived from the CIE XYZ tristimulus values [4]. The L*a*b* space consists of a luminosity layer 'L*', chromaticity-layer 'a*' indicating where color falls along the red-green axis, and chromaticity-layer 'b*' indicating where the color falls along the blue-yellow axis. All of the color information is in the 'a*' and 'b*' layers. L*a*b color is designed to approximate human vision. It aspires to perceptual uniformity, and its L* component closely matches human perception of lightness. Therefore we propose to segment colors in the L*a*b* color space with K-means clustering method. We first convert image from RGB color

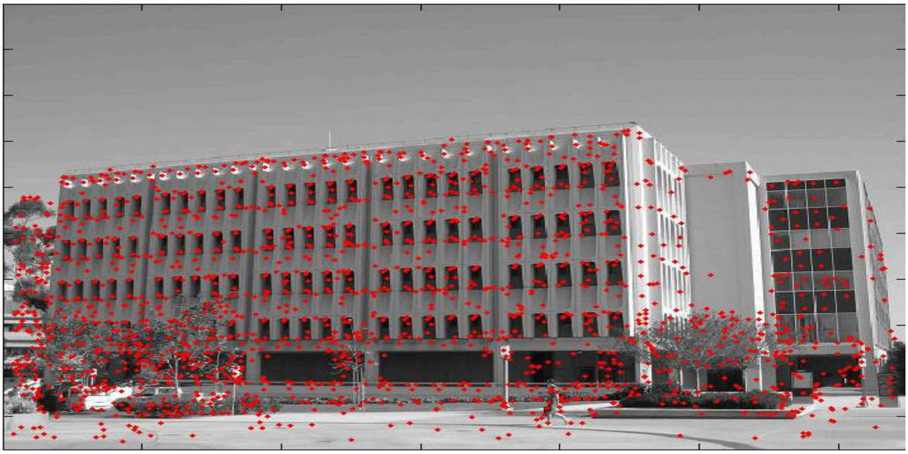


Fig. 1. There are 2280 SIFT feature points in the image, many of them fall on tree, lawn or bush areas

space to L*a*b* color space. Then we classify the colors in ‘a*b*’ space using K-means clustering. For example, three colors are used to segment the image into 3 images consisting of trees, building and sky. The segmented tree/lawn region is shown in the right panel of Fig. 2. Colors alone suffice for the segmentation in this example. In general, identification of tree like objects solely based on color is not enough. Some man-made structures could have nearly the same color as trees. Additional information is necessary. We observe that tree leaves and branches appear as rough or fractal, and so the entropy of image patches is helpful for their separation from uniform, and smoothly varying object surfaces. To compute the entropy of the gray-level images, let $V_{\max}$ be the maximum value in an image patch, the patch entropy is defined as

$$E = -\sum_{i=0}^{V_{\max}} h_i \log(h_i) , \tag{2.1}$$

where $h_i = n_i/N$ is the $i-$th histogram count n_i divided by the total number of pixels (N) in the image patch. Image patches in the tree regions in general have larger entropy than those in the uniform or regularly shaped objects such as buildings and sky. Measure of entropy proves to be rotation-invariant, and can serve as an indication of tree regions. However, other objects such as walls of building in the image may also have high entropy. Fig. 3 shows the white high entropy regions in an image, it can be seen that some parts of the building have large entropy too. This example suggests that a combination of color and entropy produces a better segmentation criterion. The idea is to first segment the image via colors, then compute the entropy of each segmented region. The part with high entropy will be recognized as belonging to trees. Now with the tree regions detected, we smooth and blur the image function in the tree regions by the Laplacian inpainting technique.

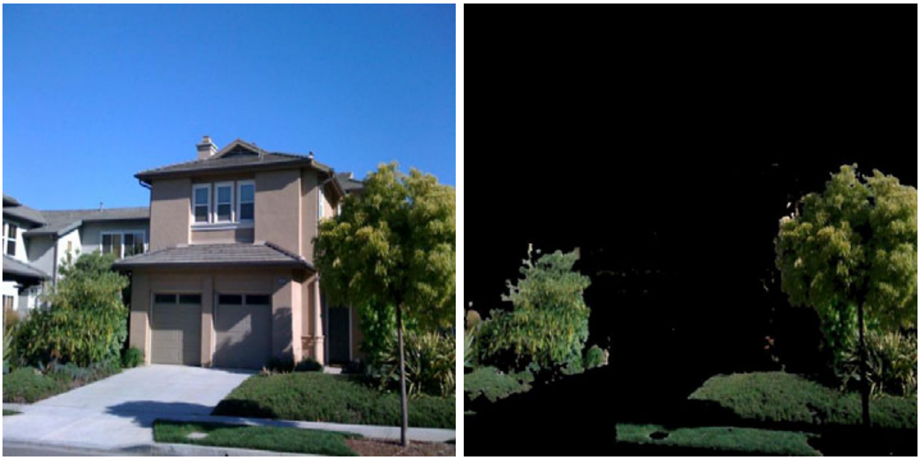


Fig. 2. The left panel is the original image. The right panel is the color segmented image consisting of trees and lawn.

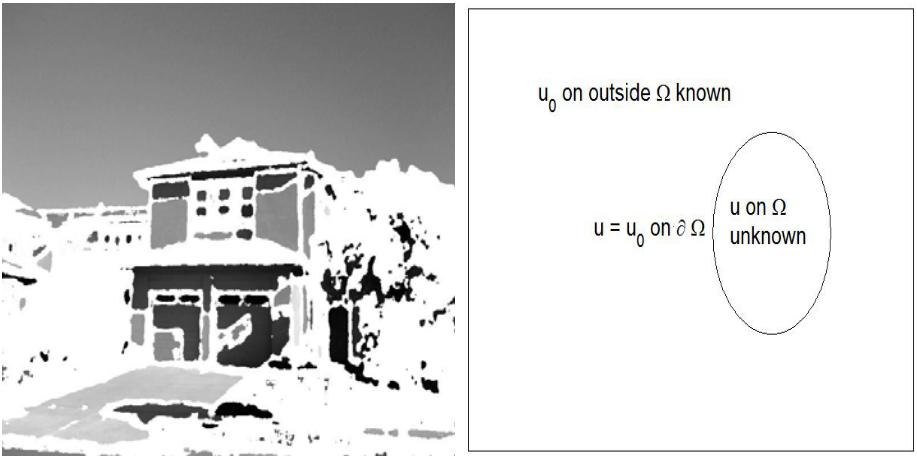


Fig. 3. Left panel: the white regions have high entropy. Right panel: the inpainting set-up. The domain *Ω* may have highly irregular boundaries in real-world images.

3 Harmonic In-Painting

Inpainting aims to restore a damaged or corrupted image. The goal of inpainting algorithm varies from making the restored parts of the image consistent with the rest of the image, to making it as close as possible to the original image. We consider the former, and hope to interpolate the values of the image function in the tree regions smoothly by outside values. This way, the image functions are blurred in the tree regions and so feature points will be sharply reduced there.

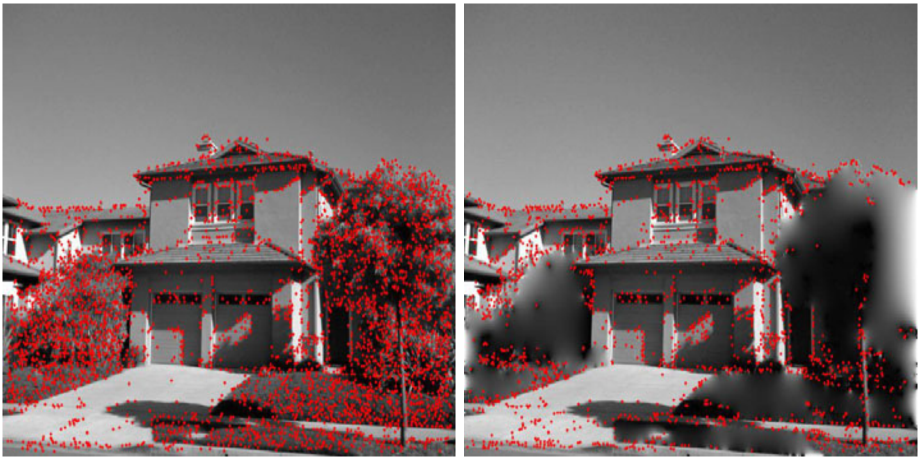


Fig. 4. SIFT features detected from the original image (left) and the preprocessed image (right). There are 4692 SIFT feature points detected from the original image. In contrast, there are 2097 SIFT feature points in the pre-processed image.

As a result, feature points are more focused on buildings or other man-made objects in the inpainted image. Since the original work of Bertalmio et al. [1], automatic inpainting algorithms have appeared in recent years. Inpainting is an interpolation problem: Given a rectangular image u_0 known outside a hole Ω (see Fig. 3), we want to find an image u, an inpainting of u_0, that matches u_0 outside the hole and has meaningful content inside the hole Ω. Broadly speaking, probabilistic [3] and variational [1,2] approaches are available for inpainting. We shall use the latter which has two representative prototypes: the harmonic and total variation (TV) models [2]. They amount to minimizing the square and absolute value of the gradient of the image, or solving the variational problems:

$$\text{Harmonic}: \min_{u\in\mathcal{A}_H}\int_\Omega |\nabla u|^2 dx \tag{3.1}$$

and TV : $\min_{u\in\mathcal{A}_{TV}}\int_\Omega |\nabla u| dx$. The admissibility sets are $\mathcal{A}_H = \{u \in W^{1,2}(\Omega) : u = u_0 \text{ on } \partial\Omega\}$ and $\mathcal{A}_{TV} = \{u \in BV(\Omega; R) : u = u_0 \text{ on } \Omega\}$. Here $W^{1,2}$ is Sobolev space of square integrable functions and square integrable gradients, BV is the function space of bounded variations. The minimizers can be found by solving the Euler-Lagrange partial differential equation: $\Delta u = 0$ on Ω, $u = u_0$ on $\partial\Omega$, for (3.1) and $\mathrm{div}\left(\frac{\nabla u}{|\nabla u|}\right) = 0$ on Ω, $u = u_0$ on $\partial\Omega$ for the TV case. Harmonic inpainting gives a desired smoother inpainted image than TV inpainting which preserves the intensity jumps or edges better. Harmonic inpainting is also much faster to compute. A five point stencil finite differencing discretizes the Laplace equation to a linear algebra problem $Au = b$, where b stores the known values of the image function at the boundary of Ω, A is a banded matrix with irregular off-diagonal entries reflecting the irregular $\partial\Omega$. The harmonic inpainting blurs the background regions (trees and grass). The right panel of Fig. 4 shows the effects of our proposed content adaptive pre-processing, after which more than half of the un-interesting SIFT feature points are removed for more efficient SIFT building matching. We remark that though the simple strategy of turning the tree regions blank also removes the unimportant features, it can introduce new feature points and cause mis-matches at the boundaries of tree regions which tend to be irregular. Our numerical tests suggest that inpainting is a better solution.

4 Feature Matching and Test Results

4.1 Matching

Matching of feature points is carried out as in SIFT which is an Euclidean-distance nearest neighbor approach in the descriptor space. A pair of feature points is considered matched if the distance ratio between the nearest neighbor distance and a second nearest neighbor distance is below τ,

$$\frac{d(p, p_{1st})}{d(p, p_{2nd})} < \tau \ ,$$

where $p \in \mathbb{R}^n$ is the descriptor to be matched and d_{1st} and d_{2nd} are the nearest and second nearest descriptors respectively, with d denoting the Euclidean distance. The $\tau = 0.8$ is suggested in [6]. This measure performs well because correct matches need to have the nearest neighbor significantly closer than the nearest incorrect match to achieve reliable matching. For false matches, there will likely be a number of other false matches within similar distances due to the high dimensionality of the descriptor space. We can think of the second nearest match as providing an estimate of the density of false matches within this portion of the feature space and at the same time identifying specific instances of feature ambiguity.

4.2 Examples of Image Matching

Here we show two experiments with images containing buildings and trees. Each of the tested images contains 800×600 pixels. In the first experiment, we match two near views of the same scene in Fig. 4. In Fig. 5, the left plot shows the matching of the SIFT feature points of the two original images, the right plot is the SIFT matching of reduced features in the pre-processed images. Computation is performed on a Dell laptop with 6G RAM and 1.6 GHz i7 CPU. The cpu time for the left plot is 47.16 seconds, and it is 39.09 seconds for the right plot. The the cpu time reported here and below included preprocessing. The matching rate for the left plot is 53.80 % (333 correct matches, 286 false matches), it is 67.08 % for the right plot (218 correct matches, 107 false matches). In the second experiment, we compare SIFT feature matching on two images with and without preprocessing, where one image is a near view of a farm house, and the other

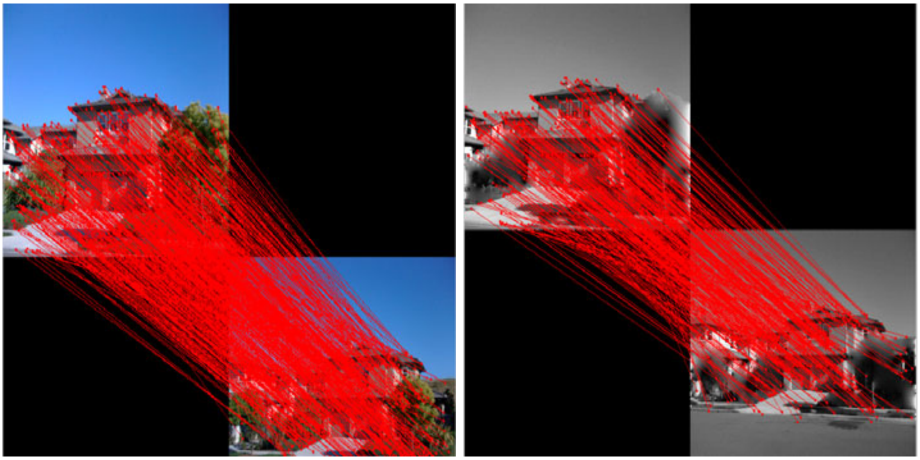


Fig. 5. Image matching of SIFT features with and without preprocessing. Left panel is the matching between two original images without segmentation and inpainting (matching rate 53.80 %). Right panel is the matching between two pre-processed images with trees regions removed (matching rate 67.08 %) and 17.11 % total cpu saving.

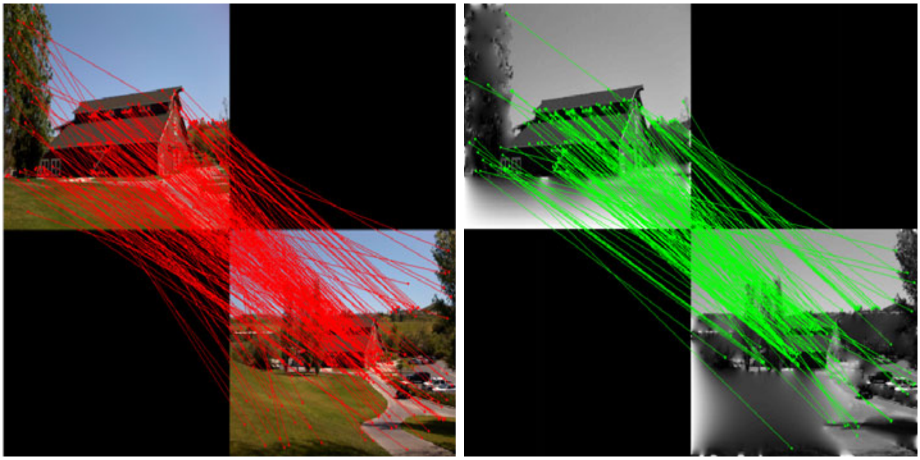


Fig. 6. Near (upper left) and far (lower right) views of a farm house in each panel. The left panel shows matching of SIFT features on the original images, with matching rate 28.73 % (77 correct matches, 191 false matches). The right panel shows SIFT feature matching on preprocessed images, with matching rate 37.17 % (71 correct matches, 120 false matches) and 29.51% total cpu saving.

is a far view. The far view contains more objects in the environment, such as additional trees occluding part of the building, and cars in the parking lot. Near (upper left) and far (lower right) views of a farm house are shown in Fig. 6. The left panel shows matching of SIFT features on the original images, with matching rate 28.73 % (77 correct matches, 191 false matches), and cpu time is 57.30 seconds. The right panel shows SIFT feature matching on the preprocessed images, with matching rate 37.17 % (71 correct matches, 120 false matches), and cpu time is 40.39 seconds. The improvement is consistent across 20 other images of landscaped buildings from street scenes in residential and commercial areas as long as the tree/bush/grass regions occupy roughly the same percentage of pixels in the images.

5 Concluding Remarks

We introduced a novel and efficient color-entropy segmentation for content adaptive image matching of natural building images. Experimental results showed a robust increase in matching rate by approximately 10% and decrease in cpu time (from 20 % to 30%) with pre-processing time included. A future line of work will study how to avoid modifying the images. Instead, one may design a weighting function to adapt the matching score. The weighting function is negatively correlated with the entropy-color measure and is essentially supported away from the tree regions. We also plan to study content adaptive preprocessing for more recent versions of corner detectors [7], and perform more extensive experiments on landscaped building images preferably on a public database.

The work was partially supported by NSF grant DMS-0911277.

References

1. Bertalmío, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpaiting. In: Proc. of SIGGRAPH 2000, New Orleans, USA, pp. 417–424 (July 2000)
2. Chan, T., Shen, J.: Mathematical Models for Local Nontexture Inpaintings. SIAM J. Appl. Math. 62, 1019–1043 (2002)
3. Geman, G., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell. 6, 721–741 (1984)
4. L*a*b color space: http://en.wikipedia.org/wiki/Lab_color_space
5. Lowe, D.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1150–157 (1999)
6. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
7. Rosten, E., Porter, P., Drummond, T.: Faster and Better: A Machine Learning Approach to Corner Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 105–119 (2010)

Face Image Enhancement Taking into Account Lighting Behavior on a Face

Masato Tsukada[1], Chisato Funayama[1], Masatoshi Arai[2], and Charles Dubout

[1] Information and Media Processing Laboratories, NEC Coproration, Kawasaki, Japan
[2] Advanced Technology Solution Division, NEC Informatic Systems, Ltd., Kawasaki, Japan
[3] Idiap Research Institute, Martigny, Switzerland
m-tsukada@cj.jp.nec.com, c-funayama@ce.jp.nec.com,
masa-a@pb.jp.nec.com, charles.dubout@idiap.ch

Abstract. This paper presents a face image enhancement taking into account lighting behavior with a low computational cost. We already proposed the enhancement method using a 3D face model. It is however difficult to implement it in color imaging appliances due to the high computational cost. The newly proposed method decomposes color information of a face into three components, i.e., specularites, shadows and albedo by using a light reflection model. Spectral reflectance is recovered from the albedo, and improved by bringing it close to a predefined reference. By modifying only the reflectance, the synthesized images appear naturally enhanced, maintaining the original expression of human faces that is influenced by specularities and shadows. The proposed method reduces its processing time to one seventy-fifth of that of the method using a 3D face model. Experiments demonstrate that the proposed method is effective for imaging appliances in terms of computational cost and image quality.

Keywords: Image quality enhancement, Skin color, Albedo, Surface reflectance, Color constancy.

1 Introduction

With recent, rapid popularization of color imaging appliances such as digital still camera, camera phone, etc., digital still images and digital movies have been easily obtained and appreciated in our daily life. A number of articles on image quality of color imaging appliances are flooded in magazines and the internet. Accordingly, it is obvious that image quality is one of the most significant factors for color imaging appliances. Especially human face is one of the most attractive objects for us. Human beings are particularly sensible to the appearance of human face, making its enhancement both desirable and difficult.

A way to solve this problem is to augment the image with additional information about the scene and objects in the image, so that the estimation of those parameters can be done automatically, and so that only the intended parameters are modified. Conventional techniques for enhancing a face image assume that there is a preferred skin tone, to which they can shift the skin colors in an image [1] [2]. This

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 479–488, 2011.

assumption, however, may not be adequate, as those values depend heavily on the particular the lighting condition of the scene and imaging conditions. In this paper, we make the assumption that there exists a preferred reflectance for faces in an image, independent of lighting and imaging conditions.

Several methods based on physical models have been proposed for skin color synthesis [3] [4] [5], but we are not aware of any targeting face image enhancement. We proposed a face image enhancement [6] that works by canceling the effects of obstructive lighting on a face using 3D face estimation. The method first decomposes color information on a face into three components, i.e., specularites, shadows and albedo by using 3D face estimation. The method produces good results applying a color correction to the only albedo which includes real skin color. Since the 3D face estimation needs high computational cost, however, it is difficult to realize it as software implementation in color imaging appliances.

In this paper, we propose a face image enhancement taking into account lighting behavior on a face with low computational cost. By using a light reflection model on a face instead of 3D face estimation, the method reduces the processing time to one seventy-fifth of that of the method using 3D face estimation. In experiments, the result by the method is compared with those by conventional methods. The processing time of the proposed method is estimated when it runs on a typical computational environment for color imaging appliances such as a camera phone. We show that the proposed method is effective for a color imaging appliances according to the balance of the calculation cost and the image quality improvement.

2 Physical Parameters Estimation

The appearance of a face under different lighting conditions can vary significantly, even though the spectral reflectance of the skin stays constant. However, as shown both by Basri and Jacobs [7] and Ramamoorthi and Hanrahan [8], if one neglects the effects of cast shadows and near-field illumination, the irradiance of a face is then a function of the surface normal $\boldsymbol{n}$ only and can be well approximated analytically in terms of spherical harmonic coefficients. These assumptions are reasonable since human heads are mostly convex and the distance to the light is usually much greater than the size of the face. They derived an analytic formula for the irradiance, showing that it can be treated as a convolution of the incident illumination with the Lambertian reflectance function (a clamped cosine). A key result of their works is that Lambertian reflection acts as a low-pass filter, so that the radiance lies close to a comparatively small number of subspaces. The eigenvectors of the subspaces are simply quadratic polynomials of the Cartesian components, and are illustrated in Fig. 1. Positive values of spheres in Fig.1 are in light gray, negative values in dark and zero is set to the gray of the background.

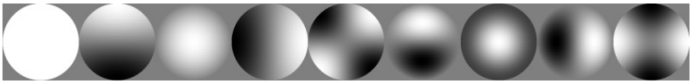


Fig. 1. Spherical harmonic basis vectors

It is thus possible to closely model the reflected radiance of a solid diffuse object under any distant illumination with a small number of basis vectors. In the case of a textured object, the irradiance $\boldsymbol{E_k}$ is simply scaled by the albedo $\rho_k(\boldsymbol{x})$ which depends on the position $\boldsymbol{x}$ of an image and gives the reflected radiance $\boldsymbol{V_k}$, directly related to image intensity, where the subscript k shows a color channel such as R, G and B.

$$V_k(x,n) = \rho_k(x)E_k(n) \tag{1}$$

As the method [6] takes only a single 2D image, 3D face estimation models [9] [10] were introduced to recover the surface normals $\boldsymbol{n}$ at an each pixel of a face in an input image. Augmenting the image with 3D information enables us to decompose each pixel's intensity into albedo, specularities and shading terms. This improves the effectiveness of the skin reflectance recovery, as it allows the estimation to be performed on the specularities and shading free skin albedo which is the real target to be modified.

Under the assumption that skin albedo is constant at low frequency on a human face, the weighted coefficients of the spherical harmonics basis vectors can be solved using a least square procedure [10]. The coefficients will be scaled by the constant skin albedo, which thus must be estimated to obtain the true irradiance. Once the irradiance $\boldsymbol{E_k}$ has been recovered, the albedo can be calculated by dividing the image intensities $\boldsymbol{V}_k$ by the irradiance $\boldsymbol{E_k}$ to remove shading components.

An additional improvement comes from the fact that it is also easy to estimate specularities. Image pixel intensities of value greater than the recovered reflected radiance $\boldsymbol{V}$ are simply clamped, and the residual parts are taken as specularities, i.e.:

$$\delta_k(x) = \max(\sigma_k(x) - V_k(x,n),0) \tag{2}$$

$$\rho_k(x) = \frac{\sigma_k(x) - \delta_k(x)}{E_k(n)}, \tag{3}$$

where $\sigma_k(x)$ is the intensity of k channel at the position $\boldsymbol{x}$ in an input image, $\rho_k(\boldsymbol{x})$ is the albedo, $\delta_k(\boldsymbol{x})$ is the estimated specular component. The processing of 3D face estimation in the method, however, needs a comparatively high computational cost.

2.1 Simple Estimation of Specualities and Albedo

Introducing a simple light reflection model for human skin is one of the solutions to improve the computational cost of our previous method using 3D face estimation. We focus on the dichromatic light reflection model [11] which is proposed for a dielectric material. Under the dichromatic light reflection model, the appearance color, which is reflected from a point on a dielectric material, is represented as a mixture of the light (surface reflection or specular component) reflected at the surface of the material and the light (body reflection or diffuse component) reflected from the material body. The specular component depends on viewing geometry and is maximum at the viewpoint where the incident angle of the incident light is equal to the angle of the reflection. On the other hand, the diffuse component is comparatively invariant from any angle of viewpoint. The reflection characteristics of the two components are different in the spatial-frequency of light reflection. That is, specular component has high spatial-frequency and diffuse component has low one.

As for a light reflection for human skin, specular component and diffuse component can be visually distinguished from the lights reflected from a face in an image. The difference of the spatial-frequency characteristic for the two components can be used to remove the specular components from a face region in an image by introducing the dichromatic light reflection model as a simple light reflection model of human skin shown in Fig.2.

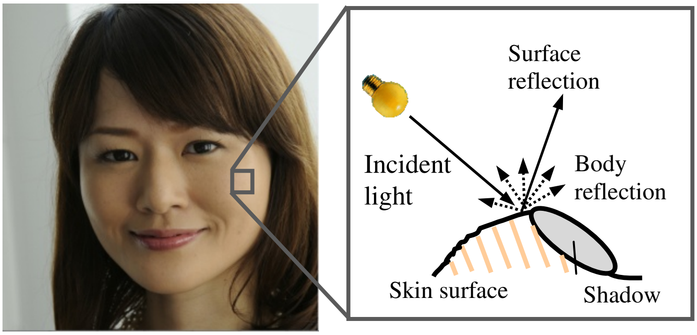


Fig. 2. Simple light reflection model of human skin

Based on the simple light reflection model of human skin, the specular components $\delta_k(\boldsymbol{x})$ on a face in an input image can be estimated by (4).

$$\delta_k(x) = \max(\sigma_k(x) - VL_k(x), 0) \tag{4}$$

where VL_k $(\boldsymbol{x})$ is the pixel intensity at the position $\boldsymbol{x}$ of the low frequency image of the input image. The albedo can be also obtained by the normalization of the luminance on a face region. The calculation means the removal of shading components.

$$\rho_k(x) = \frac{\sigma_k(x) - \delta_k(x)}{Y(x)}, \tag{5}$$

where $Y(\boldsymbol{x})$ is the luminance at the position $\boldsymbol{x}$ in the low frequency image.

2.2 Surface Reflectance Estimation

As detailed in [12], supposing that we can ignore the surface characteristics, lighting, and viewing geometry by using a relative spectral power distribution $E(\lambda)$ of the illumination in a scene, instead of physical irradiance measures, the color response $\sigma_k(\mathbf{x})$ of a color channel k with the sensitivity $R_k(\lambda)$ is:

$$\sigma_k(x) = \int_{vs} S(x, \lambda) E(\lambda) R_k(\lambda) d\lambda \tag{6}$$

where $S(\boldsymbol{x}, \lambda)$ is the spectral reflectance of an object at position $\boldsymbol{x}$ in an input image and *vs* indicates the visible spectrum. As shown in [13], it is usually enough to represent the functions $R(\lambda)$, $S(\boldsymbol{x}, \lambda)$ and $E(\lambda)$ by samples taken at 10 nm intervals over the visible spectrum, i.e., the spectral range of 400 to 700 nm. Using linear algebra notations, surface reflectance $S(\boldsymbol{x}, \lambda)$, illumination $E(\lambda)$, and sensor sensitivities $R_k(\lambda)$ can thus respectively be expressed as the 31 × 1 vectors $\boldsymbol{s}$, $\boldsymbol{e}$, $\boldsymbol{r_k}$ and (6) can be simply written:

$$\sigma_k = s^{\mathrm{T}} diag\,(e) r_k \tag{7}$$

where $^{\mathrm{T}}$ indicates the transpose and *diag* is an operator that turns a vector into a diagonal matrix. Since our goal is to enhance images taken by a standard digital color camera, which process colors so as to be viewable by the human visual system, the CIE 1931 color matching functions, i.e., $\bar{x}(\lambda)$, $\bar{y}(\lambda)$ and $\bar{z}(\lambda)$ can be used instead of the sensitivity $R_k(\lambda)$, and then the color space of input images are appropriately converted to CIEXYZ color space.

Having first estimated the power spectral distribution of the illuminant by using color constancy theory [14], it is easy to recover the surface reflectance vector $\boldsymbol{s}$ at each pixel of a face in the input image. We first determined basis vectors for skin reflectance by principal component analysis (PCA) over a data set consisting of surface reflectance of 4407 Japanese faces from the standard object colour spectra database SOCS [15]. We then solved the linear system obtained from (7) at each pixel, setting:

$$s = c_0 + \textstyle\sum_{i=1}^{3} s_i c_i = c_0 + \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} s_1 & s_2 \,.\, s_3 \end{bmatrix}^{\mathrm{T}} \tag{8}$$

Since we have three tristimulus values, i.e., CIEXYZ of the albedo, only the first three weighted coefficients $\mathrm{s_i}$ (i = 1, 2, 3) corresponding to the first three eigenvectors $\mathbf{c}_\mathrm{i}$ of the basis for surface reflectance of Japanese faces can be recovered, where $\mathbf{c}_0$ is the mean surface reflectance vector of 4407 Japanese faces, which was subtracted before performing the PCA analysis. Three basis vectors are enough to get a good approximation of the real skin reflectance, as human skin reflectance function is fairly smooth. Our PCA analysis reveals that the first three eigenvectors already account for 85% of the distribution of the data set we used. The linear system to be solved is:

$$\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 \end{bmatrix}^{\mathrm{T}} diag(e)(c_0 + \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} \begin{bmatrix} s_1 \\ {,}s_2 \\ s_3 \end{bmatrix}) \tag{9}$$

where σ_1 , σ_2 and σ_3 are the tristimulus values CIEXYZ of the albedo obtained by (5).

Defining the 3 × 31 matrix M = $[\mathbf{r}_1\ \mathbf{r}_2\ \mathbf{r}_3]^{\mathrm{T}} diag(\mathbf{e})$, converting the surface reflectance of skin to color responses, it is equivalent to:

$$\begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} = (M \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix})^{-1} (\begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix} - \dot{M} c_0) \tag{10}$$

The surface reflectance of at the position $\boldsymbol{x}$ in a face an input image can thus be estimated very efficiently by a matrix multiplication and vector subtraction.

3 Skin Color Enhancement Processing

The surface reflectance of every pixel in a face region in an input image can now be recovered by using (10) and the albedo $\boldsymbol{\rho}_k$ ($\mathbf{x}$) obtained by (5). The perceived the image quality of the face in the input image can be improved by matching the mean surface reflectance $\boldsymbol{s}_{avg}(\lambda)$ of the face region in an input image to a preferred reference $\boldsymbol{s}_{ref}(\lambda)$ which is taken as the estimated skin reflectance of a face in a target photograph.

First, the function $\boldsymbol{f}$ is determined, matching the mean surface reflectance $\boldsymbol{s}_{avg}(\lambda)$ to the preferred reference $\boldsymbol{s}_{ref}(\lambda)$:

$$s_{ref}(\lambda) = f(\lambda) \cdot s_{avg}(\lambda) \Leftrightarrow f(\lambda) = \frac{s_{ref}(\lambda)}{s_{avg}(\lambda)} \tag{11}$$

Using algebraic notation, the function $\boldsymbol{f}$ can be represented by a linear transformation $\boldsymbol{F}$, such that $\boldsymbol{s}_{ref} = \boldsymbol{F}\, \boldsymbol{s}_{avg}$ with:

$$F = diag\left(\begin{bmatrix} \frac{s_{ref,1}}{s_{avg,1}} & \cdots & \frac{s_{ref,31}}{s_{avg,31}} \end{bmatrix}\right) \tag{12}$$

The estimated surface reflectance of each pixel in the face region in the image is then multiplied by the function $\boldsymbol{f}$ (or the matrix $\boldsymbol{F}$ in algebraic notation), giving the enhanced surface reflectance which can be converted back to color stimuli to get the enhanced image.

Once the power spectral distribution of the illuminant is estimated, the albedo can be enhanced by a linear transformation which enables us to expect a high speed transformation. The whole process can be summarized as:

$$\begin{bmatrix} \rho_1' \\ \rho_2' \\ \rho_3' \end{bmatrix} = MF\left(c_0 + \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix} \left(M\begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}\right)^{-1} \left(\begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \end{bmatrix} - Mc_0 \right) \right) \tag{13}$$

where $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$, $\boldsymbol{\rho}_3$ are the tristimulus values XYZ of the albedo at a position in a face region in an input image, and $\boldsymbol{\rho}'_1$, $\boldsymbol{\rho}'_2$, $\boldsymbol{\rho}'_3$ are the enhanced one.

An additional improvement comes from smoothing the appearance of the face in the image by scaling down the specularities image including high frequency components. Specularities come primarily from the skin surface lipid film (SSLF) [16]. Reducing its intensity corresponds thus roughly to reducing the amount of sebum and sweat on the face. Such specularities reveal the skin's imperfections, and are thus undesirable to most people.

4 Experiments

We developed a software program implementing the proposed method to confirm the feasibility of the embed function in a color imaging appliance such as a camera phone. The software program was coded in conformity with ANSI C, without using any native SIMD instruction prepared for a CPU. The program uses fixed-point number representations to get the high speed processing.

A processing flow enhancing a face image of the program is summarized in Fig. 3. The diffuse component image is created from an input image. The specular component image is derived by subtracting the diffuse component image from the input image. The diffuse component image divided by the luminance image makes the albedo image. The albedo image is enhanced in the manner of skin color enhancement described in Section III. Finally the enhanced image is synthesized the enhanced albedo image, the specular component image and the luminance image which is related to the shadows on the face, referring to the skin color mask image created from the input image. The program applies the enhancement processing to the only region of skin color by citing the skin mask image.

We estimated the processing time of the program when it runs on ARM11 865MHz, which is often used as the application CPU of a color imaging appliance, by using ARM Developer Suite (ADS) 1.2. Clock cycles needed for the processing of the program were counted and compared with the clock cycles of the reference program which the relationship between the clock cycles and the processing time on the CPU is known. The estimated processing time is 390ms for an image with a face region consisting of 1M pixels. We also estimated the processing time of the program using 3D face estimation under the same condition. The processing time is more than 30s. The processing time of the proposed method is reduced to about one seventy-fifth of the method using 3D face estimation.

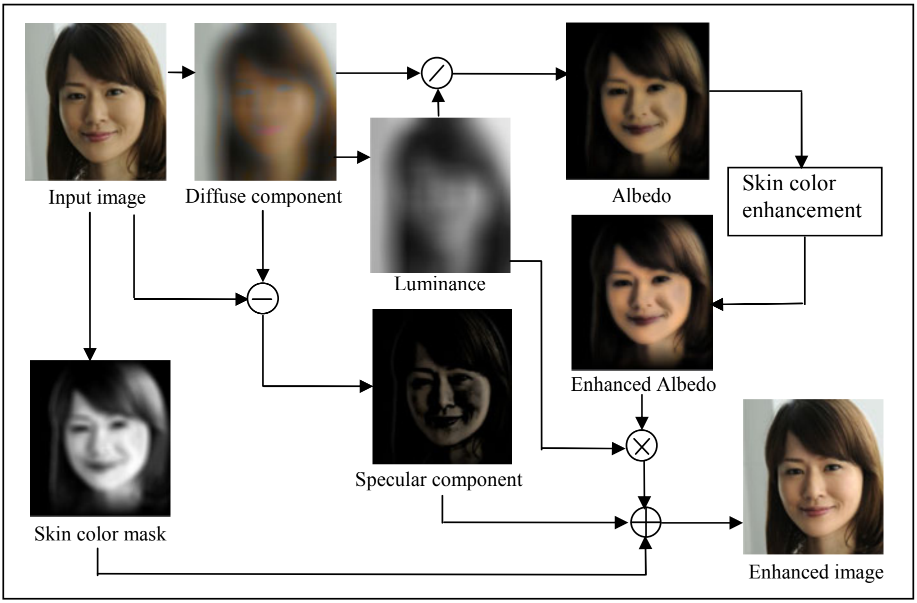
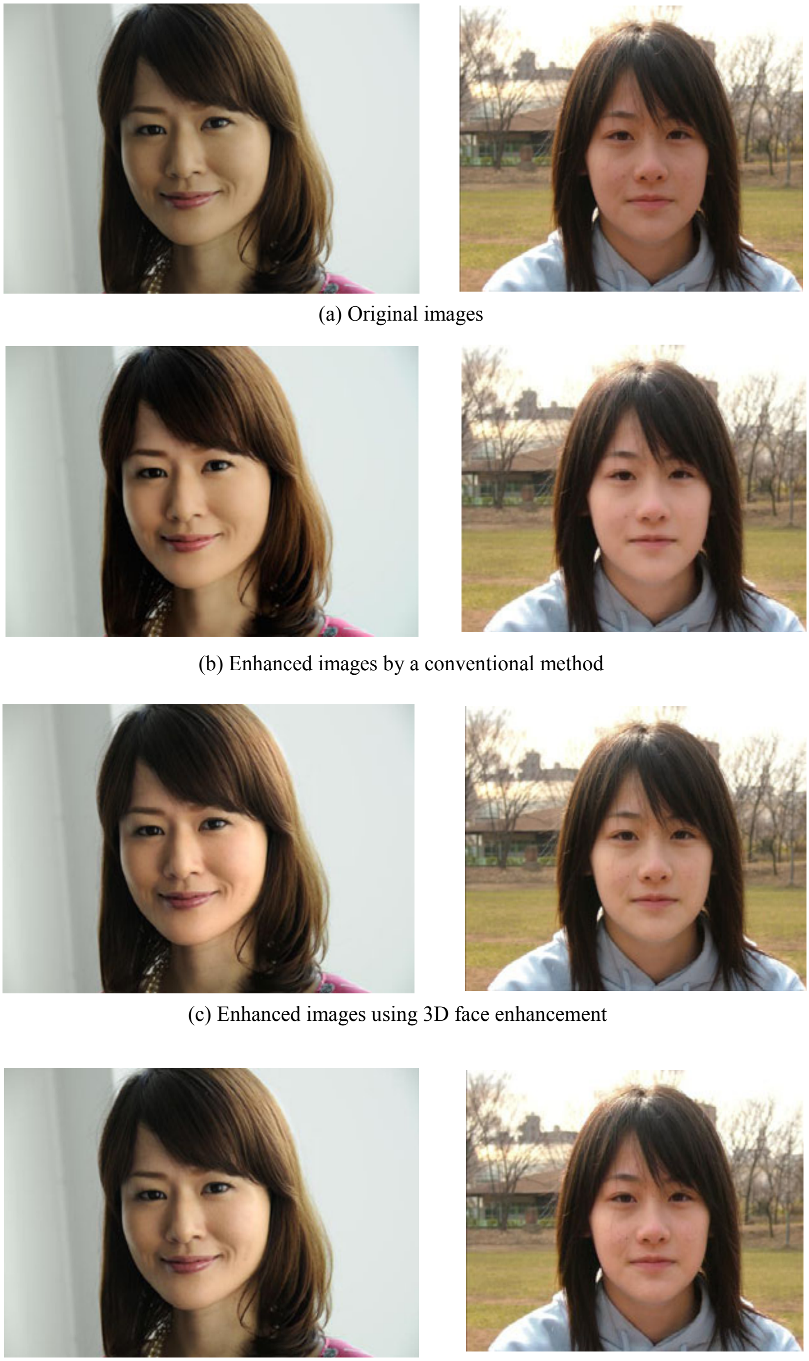


Fig. 3. Processing flow in the proposed method



(a) Original images

(b) Enhanced images by a conventional method

(c) Enhanced images using 3D face enhancement

(d) Enhanced images without 3D face estimation (Proposed method)

Fig. 4. Comparison of enhanced images and the original

In our previous work [6], the image quality of face images enhanced by the method using 3D face estimation, which is the original method of the proposed method, was closely evaluated by using the pair wise comparison method. In the subjective experiments, ten Japanese portrait images were evaluated by fourteen subjects (seven men and women) who are all Japanese with normal color vision. We could confirm that our enhance method by using 3D face estimation obtains significantly higher scores.

To evaluate the image quality of the proposed method, face images enhanced by the method were compared with the original one and images enhanced by a conventional method and the method using 3D face estimation. Sample images are shown in Fig. 4. The original images were taken by a digital single lens reflex camera. Images (b) were enhanced by commercial software as a conventional method. Images (c) and (d) were enhanced by the method using 3D face estimation and the proposed method respectively. Since unnatural smoothing applied to images (b) deletes fine textures on the faces and makes the face-looking flat, we can easily find out that these images are modified artificially. For most Japanese, the image quality of images (c) and (d) is more natural and preferable than the others, because the skin color is preferably reproduced and the original expression of the women influenced by specularities and shadows is represented naturally.

Although the degree of enhancement in images (d) is slightly weaker than that of images (c), images (d) can be obtained by one seventy-fifth of the computational cost to get images (c). It can be said that the proposed method is very effective for a color imaging appliances in terms of the calculation cost and the image quality.

5 Conclusion

This paper presented a method to enhance the perceived quality of a face image taking account into light behavior on a face with low computational cost. The proposed method decomposes the face color information in an image into three components (specularites, shadow and albedo) using a light reflection model on a skin surface instead of 3D face estimation which needs high computational cost. In the experiments, we confirmed that skin color is preferably reproduced and the original expression influenced by specularities and shadows is represented naturally in the proposed method. This produces good impression of face image quality. The processing time of the proposed method is reduced to one seventy-fifth of that of the method using 3D face estimation. The proposed method is very effective for a color imaging appliances in terms of the calculation cost and the image quality.

References

1. Tsukada, M., Funayama, C., Tajima, J.: Automatic color preference correction for color reproduction. In: Proc. SPIE, vol. 4300, pp. 216–223 (2000)
2. Kim, D.-H., Do, H.-C., Chien, S.-I.: Preferred skin color reproduction based on adaptive affine transform. IEEE Transactions on Consumer Electronics 51, 191–197 (2005)
3. Tsumura, N., Ojima, N., Sato, K., Shiraishi, M., Shimizu, H., Nabeshima, H., Akazaki, S., Hori, K., Miyake, Y.: Imagebased skin color and texture analysis/synthesis by extracting hemoglobin and melanin information in the skin. ACM Transactions on Graphics 22, 770–779 (2003)

4. Weyrich, T., Matusik, W., Pfister, H., Bickel, B., Donner, C., Tu, C., McAndless, J., Lee, J., Ngan, A., Jensen, H., Gross, M.: Analysis of human faces using a measurement based skin reflectance model. ACM Transactions on Graphics 25, 1013–1024 (2006)
5. Li, L., Ng, C.-S.: Rendering human skin using a multi-layer reflection model. Int. J. of Math Comput. Simul. 3, 44–53 (2009)
6. Dubout, C., Tsukada, M., Ishiyama, R., Funayama, C., Süsstrunk, S.: Fase image enhancement using 3D and spectral information. In: Proc. International Conference on Image Processing, pp. 697–700 (2009)
7. Basri, R., Jacobs, D.: Lambertian reflectance and linear subspaces. In: Proc. International Conference on Computer Vision, vol. 25, pp. 383–390 (2001)
8. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: Proc. SIGGRAPH, pp. 117–128 (2001)
9. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proc. SIGGRAPH, pp.187–194 (1999)
10. Ishiyama, R., Sakamoto, S.: Fast and accurate facial pose estimation by aligning a 3D appearance model. In: Proc. International Conference on Pattern Recognition, vol. 4, pp. 388–391 (2004)
11. Klinker, G.J., Shafer, S.A., Kanade, T.: A Physical Approach to Color Image Understanding. Int. J. of Computer Vision. 4(1), 7–38 (1990)
12. Peres, M., Süsstrunk, S. (eds.): The Focal Encyclopedia of Photography, 4th edn. Digital Imaging, Theory and Applications, History, and Science. Focal Press (2007)
13. Smith, B., Spiekermann, C., Sember, R.: Numerical methods for colorimetric calculations: Sampling density requirements. COLOR Research and Application 17(6), 394–401 (1992)
14. Tsukada, M., Ohta, Y.: Color reproduction based on memory color and its implementation. In: Proc. 2nd Int. Workshop on Image Media Quality and Its Applications (IMQA), pp.47–52 (2007)
15. ISO/TC130: Graphic technology – Standard object colour spectra database for colour reproduction evaluation (SOCS). ISO/TR 16066 (2003)
16. Wen, Z., Liu, Z., Huang, T.: Face relighting with radiance environment maps. In: Proc. IEEE Conf. on Computer Vision & Pattern Recognition, pp. 158–165 (2003)

Adaptive Matrices for Color Texture Classification

Kerstin Bunte, Ioannis Giotis, Nicolai Petkov, and Michael Biehl

Johann Bernoulli Institute for Mathematics and Computer Science,
University of Groningen, The Netherlands
k.bunte@rug.nl, i.e.giotis@rug.nl

Abstract. In this paper we introduce an integrative approach towards color texture classification learned by a supervised framework. Our approach is based on the Generalized Learning Vector Quantization (GLVQ), extended by an adaptive distance measure which is defined in the Fourier domain and 2D Gabor filters. We evaluate the proposed technique on a set of color texture images and compare results with those achieved by methods already existing in the literature. The features learned by GLVQ improve classification accuracy and they generalize much better for evaluation data previously unknown to the system.

Keywords: adaptive metric, Gabor filter, color texture analysis, classification, Learning Vector Quantization.

1 Introduction

Texture analysis and classification are topics of particular interest mainly due to their numerous possible applications, such as medical imaging, industrial quality control and remote sensing. A wide variety of methods for texture analysis has been already developed such as co-occurrence matrices [8], Markov random fields [24], autocorrelation methods [18,16], Gabor filtering [22,5,11,13,15,6] and wavelet decomposition [23]. However, these methods mostly concern intensity images and since color information is a vector quantity an adaptation to the color domain is not always straightforward. With regards to color texture the possible approaches can be distinguished in three categories [17]. The most popular among them is called the integrative approach [10,4,17,9] and it describes texture by combining color information with the spatial relationships of image regions within each color channel and between different color channels.

In this contribution we introduce a novel integrative approach towards color texture classification and recognition based on 2D Gabor filters through supervised learning. Given a set of labeled color images (RGB) for training and a bank of 2D Gabor filters the goal here is to learn a transformation of a color image to a single channel (intensity) image, such that the Gabor responses of the transformed images will yield the best possible classification. Most signal processing techniques are based on insights or empirical observations from neurophysiology or optical physics. The proposed, novel approach incorporates data-driven

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 489–497, 2011.

adaptation of the system, e.g. example based learning. Furthermore, the filters used in our approach can be substituted, depending on the data domain and the task at hand. As an example we explore in this paper the use of rotation and scale invariant descriptors based on Gabor filter responses [7]. We demonstrate that our novel approach yields very good generalization ability with respect to previously unknown data.

2 Review of the Generalized LVQ

In this section we introduce a methodology to learn discriminative transformations for images. Our adaptation is based on the Generalized Learning Vector Quantization (GLVQ) [19]. GLVQ is an extension which introduces a cost function to the original Learning Vector Quantization (LVQ) [12] formulation. LVQ is a supervised prototype-based classification method, easy to implement and interpret, which makes it popular for many applications. The training is based on data points $\boldsymbol{x}^i \in \mathbb{R}^D$ and their corresponding label information $y^i \in 1, \ldots, C$, where D denotes the dimension of the feature vectors and C the number of classes. The prototypes are characterized by their location in the feature space $\boldsymbol{w}^i \in \mathbb{R}^D$ and the respective class label $c(\boldsymbol{w}^i) \in 1, \ldots, C$. Given a dissimilarity measure $d(\boldsymbol{x}, \boldsymbol{w})$ (e.g. the Euclidean distance), any data point $\boldsymbol{x}$ is assigned to the class label $c(\boldsymbol{w}^i)$ of the closest prototype $\boldsymbol{w}^i$ with $d(\boldsymbol{x}, \boldsymbol{w}^i) \leqslant d(\boldsymbol{x}, \boldsymbol{w}^j)$ for all $j \neq i$. The training algorithm is guided by the minimization of a cost function

$$f_c(d, J, K) = \sum_i \frac{d(\boldsymbol{x}^i, \boldsymbol{w}^J) - d(\boldsymbol{x}^i, \boldsymbol{w}^K)}{d(\boldsymbol{x}^i, \boldsymbol{w}^J) + d(\boldsymbol{x}^i, \boldsymbol{w}^K)} \tag{1}$$

where the quantities $d(\boldsymbol{x}^i, \boldsymbol{w}^J)$ with $c(\boldsymbol{w}^J) = \boldsymbol{y}^i$ and $d(\boldsymbol{x}^i, \boldsymbol{w}^K)$ with $c(\boldsymbol{w}^K) \neq \boldsymbol{y}^i$ correspond to the distances of the feature vector $\boldsymbol{x}^i$ from the respective closest correct prototype $\boldsymbol{w}^J$ and the closest wrong prototype $\boldsymbol{w}^K$. The original algorithm follows a stochastic gradient descent for the optimization of the cost function Eq. (1). The gradients are evaluated with respect to the contribution of single instances $\boldsymbol{x}^i$, which are presented at random and sequentially during training. Further extensions like, for instance, the Generalized Matrix LVQ (GMLVQ) [20] employ an adaptive dissimilarity measure $d^{\Omega}(\boldsymbol{x}, \boldsymbol{w}) = (\boldsymbol{x} - \boldsymbol{w})^{\top} \Omega^{\top} \Omega (\boldsymbol{x} - \boldsymbol{w})$ which corresponds to a generalized Euclidean metric. GMLVQ and its modifications have proven beneficial in many applications, including classification, content based image retrieval and supervised dimension reduction [21,2,3]. In the following section we extend the original GLVQ formulation for color texture classification.

3 Adaptive Matrices for Texture Classification

Consider a data set consisting of color image patches of a priorly defined size $(p \times p)$ and a bank of Gabor kernels $\mathbf{G}$ with different scales and orientations. We use for both the image patches and the filter kernels their representation in the Fourier domain. After vectorizing we end up with complex data points $\boldsymbol{x}^i \in \mathbb{C}^D$ of dimension $D = p \cdot p \cdot 3$ carrying a label $y^i \in \{1, \ldots, C\}$ that belong to one of C classes and a filter bank $\mathbf{G}$, where $\mathbf{G}^l \in \mathbb{C}^M$ with $M = p \cdot p$ is the vectorized kernel of the l-th filter of the bank. The general form of the descriptor for a vectorized image patch $\boldsymbol{v}$ given the filter bank $\mathbf{G}$ and parameterized by local transformations Ω_k can be written as $f_{\Omega_k}(\boldsymbol{v}, \mathbf{G}) : \mathbb{C} \to \mathbb{C}$. Here k corresponds to the index of the prototype $\boldsymbol{w}^k$ or the index of its class label $c(\boldsymbol{w}^k)$ for class-wise transformations. For the proposed optimization procedure it is necessary, that f_{Ω_k} is differentiable. In this contribution f_{Ω_k} corresponds to the sum of the responses of all filter kernels in $\mathbf{G}$ to the vectorized image patch, thus defining the descriptor:

$$f_{\Omega_k}(\boldsymbol{v}, \mathbf{G}) : \boldsymbol{v} \to \boldsymbol{r}_k(\boldsymbol{v}) = \sum_l \boldsymbol{v} {\Omega_k}^\top \ast \mathbf{G}^l \ , \tag{2}$$

where $\ast$ denotes the convolution. The filter bank $\mathbf{G}$ may be chosen based on the user's preference, suitable to the data and the task at hand. The vector $\boldsymbol{v}$ is defined in the data domain $\mathbb{C}^D$ and $\Omega_k \in \mathbb{C}^{M \times D}$ is the local transformation, which maps the color values to scalar, "intensity" values used for filtering. The dissimilarity measure is defined by:

$$d_{\mathbf{G}}^{\Omega_k}(\boldsymbol{x}^i, \boldsymbol{w}^k) = \| \ |\boldsymbol{r}_k(\boldsymbol{x}^i)|^2 - |\boldsymbol{r}_k(\boldsymbol{w}^k)|^2 \ \|^2 \ , \tag{3}$$

and corresponds to the difference of descriptor magnitudes. This considers two patches containing the same texture pattern as similar, independent of the position where the pattern occurs within the patches.

We use the same cost function as in the original GLVQ algorithm Eq. (1). We follow a stochastic gradient descent procedure and present the samples $\boldsymbol{x}^i$ of the training set sequentially and update the parameters accordingly. We will refer to this algorithm as Color Image Analysis LVQ (CIA-LVQ) and to one sweep through the training set as one epoch E.

Explicit form of the learning rules: For the sake of completeness we present the explicit form of the learning rules of CIA-LVQ. The parameter updates read as follows:

$$\boldsymbol{w}^L = \boldsymbol{w}^L - \alpha \Delta \boldsymbol{w}^L, \ \Delta \boldsymbol{w}^L = \frac{\partial f_c(d_{\mathbf{G}}^{\Omega_J}, d_{\mathbf{G}}^{\Omega_K}, J, K)}{\partial \Re(\boldsymbol{w}^L)} + i \frac{\partial f_c(d_{\mathbf{G}}^{\Omega_J}, d_{\mathbf{G}}^{\Omega_K}, J, K)}{\partial \Im(\boldsymbol{w}^L)} \tag{4}$$

$$\Omega_L = \Omega_L - \epsilon \Delta \Omega_L, \ \Delta \Omega_L = \frac{\partial f_c(d_{\mathbf{G}}^{\Omega_J}, d_{\mathbf{G}}^{\Omega_K}, J, K)}{\partial \Re(\Omega_L)} + i \frac{\partial f_c(d_{\mathbf{G}}^{\Omega_J}, d_{\mathbf{G}}^{\Omega_K}, J, K)}{\partial \Im(\Omega_L)} \tag{5}$$

where $L \in \{J, K\}$ and α and ϵ are the learning rates for the prototypes and the matrix respectively. The derivatives with respect to the closest correct $\boldsymbol{w}^J$ and closest wrong prototype $\boldsymbol{w}^K$ together with their corresponding matrices Ω_J and Ω_K for the given training data point $\boldsymbol{x}^i$ read:

$$\Delta \boldsymbol{w}^L = -4 \cdot \gamma^L \left[\left(|\boldsymbol{r}_L(\boldsymbol{x}^i)|^2 - |\boldsymbol{r}_L(\boldsymbol{w}^L)|^2 \right) \cdot \boldsymbol{r}_L(\boldsymbol{w}^L)^* \ast \left(\sum_l \Omega_L \ast \mathbf{G}^l \right) \right]^* \tag{6}$$

$$\Delta \Omega_L = \gamma^L \Bigg(4 \left(|\boldsymbol{r}_L(\boldsymbol{x}^i)|^2 - |\boldsymbol{r}_L(\boldsymbol{w}^L)|^2 \right) \tag{7}$$

$$\cdot \left[\boldsymbol{r}_L(\boldsymbol{x}^i)^* \ast \left(\sum_l \boldsymbol{x}^i \ast \mathbf{G}^l \right) - \boldsymbol{r}_L(\boldsymbol{w}^L)^* \ast \left(\sum_l \boldsymbol{w}^L \ast \mathbf{G}^l \right) \right] \Bigg)^*, L \in \{J, K\}$$

with $\gamma^J = \frac{2 \cdot d_{\mathbf{G}}^{\Omega_K}(\boldsymbol{x}^i, \boldsymbol{w}^K)}{d_{\mathbf{G}}^{\Omega_J}(\boldsymbol{x}^i, \boldsymbol{w}^J) + d_{\mathbf{G}}^{\Omega_K}(\boldsymbol{x}^i, \boldsymbol{w}^K) \; ^2}$, $\gamma^K = \frac{-2 \cdot d_{\mathbf{G}}^{\Omega_J}(\boldsymbol{x}^i, \boldsymbol{w}^J)}{d_{\mathbf{G}}^{\Omega_J}(\boldsymbol{x}^i, \boldsymbol{w}^J) + d_{\mathbf{G}}^{\Omega_K}(\boldsymbol{x}^i, \boldsymbol{w}^K) \; ^2}$ and * denoting the complex conjugate. Note, that since we are working with complex values we have to take all derivatives with respect to the real and imaginary parts respectively.

In the next section we experiment with the algorithm and show its use in practice.

4 Experiments

In order to evaluate the usefulness of the proposed algorithm, we perform classification on patches of pictures taken from the VisTex database [1]. Our data consists of color images with size 128×128 pixels from the groups Bark, Brick, Tile, Fabric and Food. Although in texture classification literature each such image is often considered as a different class, here we distinguish into five different classes equivalent to the five aforementioned groups. Despite its increased difficulty, this classification task allows us to better demonstrate the ability of CIA-LVQ to describe general characteristics of real-world texture patterns.

At first we draw 15×15 patches randomly from each image shown in Fig. 1. The training set contains 150 patches per image, resulting in 3000 samples in total, while the test set holds 50 patches from each image. The test set may contain patches which partially overlap with those used for training. Therefore the images in Fig. 2 are used in order to create an evaluation set that was never seen in the training process. The evaluation set consists of 50 randomly drawn patches per image and is used to show the generalization ability of the approach.

A note is due here to the nature of the filter used. A 2D Gabor filter is defined as a Gaussian kernel function modulated by a sinusoidal plane wave. All filter kernels can be generated from one basic wavelet by dilation and rotation. In this experiment our filter bank consists of 12 Gabor filters of bandwidth equal to 1 at six orientations $\theta = 0$, 30, 60, 90, 120 and 150 degrees and two scales (wavelengths) varying by one octave: $\lambda = 7$ and $7\sqrt{2}$. These scales ensure that the Gabor function yields an adequate number of visible parallel excitatory and

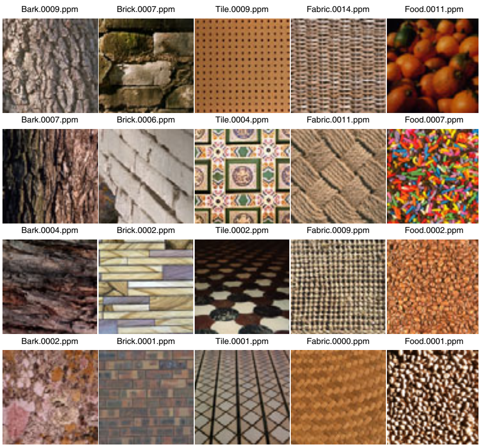


Fig. 1. Images, which are used to provide random patches for training and test

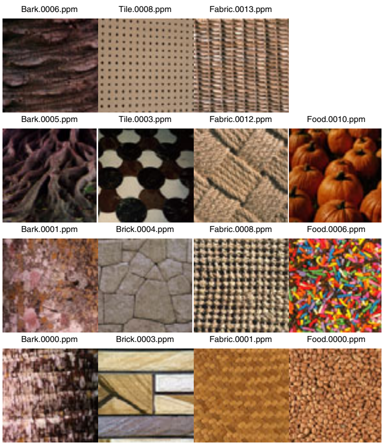


Fig. 2. Images, which are used to provide random patches for evaluation

inhibitory stripe zones. Dependent on the patch size different scales might be adequate. We set the phase offset $\phi = 0$ and the aspect ratio $\gamma = 1$ for all filters. In this way we create center-on symmetric filters with circular support. We run the localized version of CIA-LVQ with class-wise matrices Ω_c initialized with the identity matrix and 4 prototypes per class for E =300 epochs. The learning rates were chosen as $\alpha(t) = 0.002\,(0.005)^{t/E}, \epsilon(t) = 10^{-3}\left(10^{-2}\right)^{t/E}$ where t is the current epoch. Using more filters and more localized matrices Ω_j may cause overfitting effects. So it is advisable to increase the complexity of the system carefully. The training error is 10.6% and the error on the test set 28%.

We use the same data sets and the same filter bank to compare with the Opponent Color Features (OCF) [10] and the common approach of deriving textural information only from the luminance plane of images [4]. OCF attempts to extend the use of features extracted from Gabor filter responses to color texture classification motivated by mechanisms of human vision. The luminance approach is considered to often outperform combined color and texture features [14]. For the latter an RGB to gray (RGB2G) transformation is used, which builds intensity values by a weighted sum of the color components of every pixel: 0.2989·R+0.587·G+0.114·B. We again vectorize all patches s and in this case the image patch descriptor is given by $\boldsymbol{r}_2(\boldsymbol{s}) = \sum_l \boldsymbol{s} * \mathbf{G}^l$. For OCF we use a one nearest neighbor (1-NN) classification scheme with precisely the set of features and the dissimilarity measure suggested by the authors of [10], whereas for the RGB2G approach we use the 1-NN scheme with a dissimilarity measure similar to Eq. (3): $d_{\mathbf{G}}(\boldsymbol{x}^i, \boldsymbol{x}^j) = \|\ |\boldsymbol{r}_2(\boldsymbol{x}^i)|^2 - |\boldsymbol{r}_2(\boldsymbol{x}^j)|^2\ \|^2$. The NN scheme shows a test error of 14.5% based on the OCF and of 37.5% based on the RGB2G transformation, but most interesting is the comparison of the classification errors on the evaluation set. Here the NN scheme shows an error of 43.1% and 61.9% for OCF and RGB2G respectively, while the CIA-LVQ still has an error of 28.8%.

Table 1. Confusion matrices (eval. set)

CIA-LVQ:						
	1	2	3	4	5	
1	176	10	12	7	2	207
2	1	57	11	9	3	81
3	18	25	43	31	10	127
4	1	5	23	127	4	160
5	4	3	11	26	131	175
	200	100	100	200	150	750
class-wise accuracy of estimation in %						
	88.00	57.00	43.00	63.50	87.33	
OCF:						
	1	2	3	4	5	
1	94	13	12	32	23	174
2	46	62	23	28	19	178
3	30	14	62	15	6	127
4	25	10	3	125	18	181
5	5	1	0	0	84	90
	200	100	100	200	150	750
class-wise accuracy of estimation in %						
	47.00	62.00	62.00	62.50	56.00	
RGB2G: class-wise accuracy of estimation in %						
	26.00	45.00	51.00	41.50	36.67	

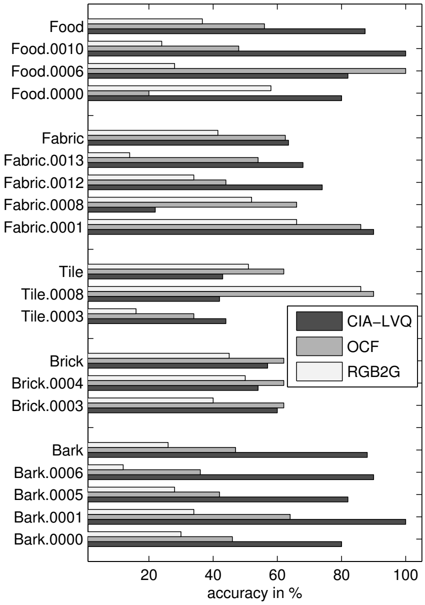


Fig. 3. Classwise and individual image accuracies

The LVQ scheme displays very good generalization, which is shown in Table 1 and Fig. 3. Note, that the accuracy rates among individual images of the same class can vary. Brick and Tile are the most difficult classes, because the texture is large, so it cannot be captured very well with such a small patch size, since a lot of patches might be drawn from non-textured regions. On the other side classes like Food and Bark with less diversity regarding textural structures can be learned quite well.

The prototypes, which classify the evaluation set are shown in Fig. 4. Additionally we show some example patches from the evaluation set, which are classified correctly together with their descriptors in Fig. 5 and some examples of wrongly classified patches in Fig. 6. Some obvious problems occur due to the random sampling and the very small patch size: we observe, that classes which vary a lot in the size of the actual structure (e.g. Brick and Tile) are more difficult to recognize than classes with small variations in the scale of texture (like Bark and Food). It is interesting to notice that random patches drawn from Food.0010.ppm are 100% correctly classified, even though no patch from this image was ever used to train the algorithm. The learned local transformation recognizes the channels leading to the orange color and increased their weights to distinguish this class from others.

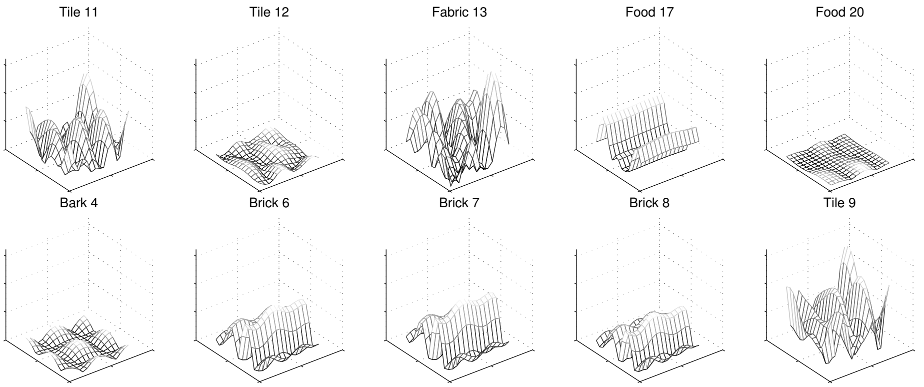


Fig. 4. Magnitude of the descriptors $|\boldsymbol{r}_L(\boldsymbol{w}^L)|$ of the prototypes which classify the evaluation set

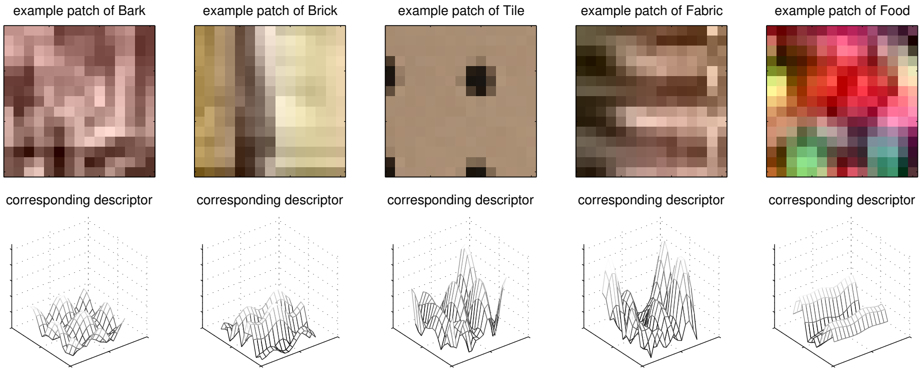


Fig. 5. Magnitude of the descriptors $|\boldsymbol{r}_L(\boldsymbol{w}^L)|$ of some correct classified example patches of the evaluation set

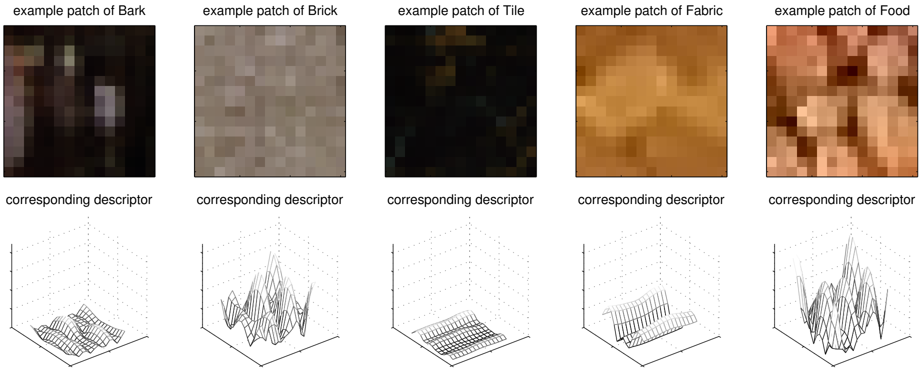


Fig. 6. Magnitude of the descriptors $|\boldsymbol{r}_L(\boldsymbol{w}^L)|$ of some wrongly classified example patches of the evaluation set

5 Conclusion and Outlook

In this contribution we proposed a prototype based framework for color texture analysis. Contrary to standard approaches which are either based on a single channel representation of the images through a fixed transformation or empirical observations for combining color and textural information, we offer the alternative of data driven learning of suitable, parameterized image descriptors. The ability of weighting different color channels automatically according to their importance for the classification task is the most important factor which distinguishes our approach. We have formulated a novel general principle: based on a differentiable convolution and a predefined filter bank the CIA-LVQ algorithm optimizes the classification. It is also of conceptual value that this adaptation of LVQ is suitable for learning in the complex numbers domain. In principle every adaptive metric method could be extended following our suggestion, but we consciously choose LVQ because of its easily interpretable results and the lower computational costs in comparison to other approaches. As an example we used Gabor filters to classify texture patterns in 15×15 patches randomly drawn from images of the VisTex database. The results show that the algorithm can learn typical texture patterns with very good generalization, even from relatively small patches and filter banks. Similarly to Gabor filters any other family of 2D filters commonly used to describe gray scale image information could be adapted and applied to color image analysis with this algorithm. A filter bank with differences of Gaussians for color edge detection is a possible example. Investigation of the performance of the system on other filters can be addressed in future. Furthermore, depending on the task it might be desirable that two patches in which the same texture occurs on different positions should not be interpreted as similar. In this case another similarity measure should be used: $\| \, |\boldsymbol{r}(\boldsymbol{x}^i) - \boldsymbol{r}(\boldsymbol{w}^L)| \, \|^2$, which is not based on the difference of magnitudes. This might be of advantage for example in the recognition of objects such as traffic signs, were a corner or an edge might have different meanings dependent on its position in the image.

References

1. Database VisTex of color textures from MIT, http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html
2. Bunte, K., Biehl, M., Petkov, N., Jonkman, M.F.: Learning effective color features for content based image retrieval in dermatology. Pattern Rec. (2011)
3. Bunte, K., Hammer, B., Schneider, P., Biehl, M.: Nonlinear discriminative data visualization. In: Verleysen, M. (ed.) Proc. of European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, pp. 65–70 (April 2009)
4. Drimbarean, A., Whelan, P.F.: Experiments in colour texture analysis. Pattern Recognition Letters 22(10), 1161–1167 (2001)
5. Fogel, I., Sagi, D.: Gabor filters as texture discriminator. Biological Cybernetics 61(2), 103–113 (1989)
6. Grigorescu, S., Petkov, N., Kruizinga, P.: Comparison of texture features based on gabor filters. IEEE Trans. on Image Processing 11(10), 1160–1167 (2002)

7. Han, J., Ma, K.-K.: Rotation-invariant and scale-invariant gabor features for texture image retrieval. Image and Vision Computing 25(9), 1474–1481 (2007)
8. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. IEEE Trans. on Systems, Man and Cybernetics 3(6), 610–621 (1973)
9. Hoang, M.A., Geusebroek, J.-M., Smeulders, A.W.: Color texture measurement and segmentation. Signal Processing 85(2), 265–275 (2005), SI on Content Based Image and Video Retrieval
10. Jain, A., Healey, G.: A multiscale representation including opponent color features for texture recognition. IEEE Trans. on Image Processing 7(1), 124–128 (1998)
11. Jain, A.K., Farrokhnia, F.: Unsupervised texture segmentation using gabor filters. Pattern Recognition 24(12), 1167–1186 (1991)
12. Kohonen, T.: Self-Organizing Maps, 3rd edn. Springer, Heidelberg (2001)
13. Kruizinga, P., Petkov, N.: A computational model of periodic-pattern-selective cells. In From Natural to Artificial Neural Computation. In: Sandoval, F., Mira, J. (eds.) IWANN 1995. LNCS, vol. 930, pp. 90–99. Springer, Heidelberg (1995)
14. Mäenpää, T., Pietikäinen, M.: Classification with color and texture: jointly or separately? Pattern Recognition 37(8), 1629–1640 (2004)
15. Manjunath, B., Ma, W.: Texture features for browsing and retrieval of image data. IEEE Trans. on Pattern Analysis and Machine Intelligence 18(8), 837–842 (1996)
16. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence 24, 971–987 (2002)
17. Palm, C.: Color texture classification by integrative co-occurrence matrices. Pattern Recognition 37(5), 965–976 (2004)
18. Pietikäinen, M., Ojala, T., Xu, Z.: Rotation-invariant texture classification using feature distributions. Pattern Recognition 33(1), 43–52 (2000)
19. Sato, A.S., Yamada, K.: Generalized learning vector quantization. In: Touretzky, D.S., Mozer, M.C., Hasselmo, M.E. (eds.) Advances in Neural Information Processing Systems, vol. 8, pp. 423–429. MIT Press, Cambridge (1996)
20. Schneider, P., Biehl, M., Hammer, B.: Adaptive relevance matrices in learning vector quantization. Neural Computation 21(12), 3532–3561 (2009)
21. Schneider, P., Bunte, K., Hammer, B., Biehl, M.: Regularization in matrix relevance learning. IEEE Transactions on Neural Networks 21(5), 831–840 (2010)
22. Turner, M.R.: Texture discrimination by gabor functions. Biological Cybernetics 55, 71–82 (1986), doi:10.1007/BF00341922
23. Wang, J.-W., Chen, C.-H., Chien, W.-M., Tsai, C.-M.: Texture classification using non-separable two-dimensional wavelets. Pattern Recogn. Lett. 19, 1225–1234 (1998)
24. Wang, L., Liu, J.: Texture classification using multiresolution markov random field models. Pattern Recognition Letters 20(2), 171–182 (1999)

Color Texture Classification Using Rao Distance between Multivariate Copula Based Models*

Ahmed Drissi El Maliani[1], Mohammed El Hassouni[2], Nour-Eddine Lasmar[3], Yannick Berthoumieu[3], and Driss Aboutajdine[1]

[1] LRIT, URAC 29, Mohammed V University, Agdal, Morocco
[2] DESTEC, FLSHR, Mohammed V University, Agdal, Morocco
[3] IMS- Groupe Signal- UMR 5218 CNRS, ENSEIRB, University Bordeaux, France

Abstract. This paper presents a new similarity measure based on Rao distance for color texture classification or retrieval. Textures are characterized by a joint model of complex wavelet coefficients. This model is based on a Gaussian Copula in order to consider the dependency between color components. Then, a closed form of Rao distance is computed to measure the difference between two Gaussian Copula based probabilty density functions on the corresponding manifold. Results in term of classification rates, show the effectiveness of the Rao geodesic distance when applied on the manifold of Gaussian Copula based probability distributions, in comparison with the Kullback-Leibler divergence.

Keywords: Color texture classification, Gaussian Copula, Rao distance.

1 Introduction

Texture classification is an important and challenging task in image analysis. Efficient texture classification is essentially based on a pertinent feature extraction and similarity measurement steps, especially when choosing the K-Nearest Neighbor approach. The feature extraction step consists of figuring out a set of attributes that best describe the texture, and best discriminate this latter from different textures. Many works, stressed that treating the texture in the wavelet domain allows accurate characterization, by modeling histograms of the wavelet subbands with appropriate models.

In this context, the Generalized Gaussian Density (GGD) proposed by Do and Vetterli [1] showed a good ability in modeling the heavely tailed and pickly pronounced behavior of the histograms. In the case of color textures, and more specifically when they are represented in the RGB color space, a big correlation exists between the color bands, thus considering univariate models leads to a considerable loss of information in the characterization. For this, joint models were proposed to describe the dependence across color bands. Verdoolaeg et al. [2], proposed a multivariate Generalized Gaussian distribution (MGGD) for multiscale color texture retrieval, modeling dependence across color components. A t-Student Copula based multivariate Weibull distribution was proposed

* This work was supported by franco-moroccan CNRS-CNRST project STIC 05/10.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 498–505, 2011.

by Kwitt et al. [3], describing dependence between complex wavelet coefficient magnitudes also in the context of color texture retrieval. The second step of a KNN based classification process, is the similarity measurement, which consists of measuring distance between textures according to the extracted features. In their pioneering work, Do and Vetterli [1] showed that the Kullback-Leibler (KL) divergence is suitable for comparing textures reposing on the estimated model parameters. Recently KL divergence has been used to measure similarity between Copula based distributions, in the context of texture retrieval [3]. A Monte-carlo approach was adopted to deal with the lack of a closed form of the KL divergence between Copula based models. However, the KL divergence cannot be considered as a metric on the space of probability distributions, since it is not symmetric and does not obey the triangle inequality.

To remedy to this problem, a Riemannian metric namely the Rao or Geodesic distance can be used instead of the commonly used Kullback-Leibler divergence. The Rao distance was first proposed by Rao [4] exploiting that the Fisher information for a set of probability density functions (pdfs) is a Riemannian metric on the corresponding manifold. Thus, the Rao distance is achieved reposing on the Fisher information metric and by resolving the corresponding geodesic equations. Rao distance was already used in different fields, such as segmentation and classification [5][6]. Closed forms of the Fisher-Rao metric were proposed for Extreme Value probability distributions namely the Gumbel, Cauchy-Fréchet, and Weibull [7]. In [2], the Rao distance was used as a similarity measure between zero-mean multivariate generalized Gaussian distributions (MGGD). A closed form of the Rao metric was computed in the case of a fixed shape parameter, while geodesic equations were solved numerically when assuming different shape parameters.

In this work, we propose a Rao metric as a similarity measure on the manifold of Gaussian Copula based multivariate probability distributions. This assumes that we know Rao distance for the marginal distributions, before computing the metric for the joint model. To test effectiveness of the Rao distance based similarity measure, we show results in comparison with the KL divergence based similarity measure in term of classification rates.

This paper is organized as follows. In the next section, we give a review of Copula theory and construct the multivariate models reposing on the Copula approach. In section 3, we derive the general form of the Rao distance on the manifold of Gaussian Copula based probability distributions. In section 4, we present results of texture classification using the Rao distance and the Kullback-Leibler divergence, before concluding in section 5.

2 Gaussian Copula Based Probability Distributions Manifold

2.1 Review of the Copula Theory

We draw on the Copula theory to incorporate the information of dependency between color components in the RGB color space. Copulas are an elegant tool for merging a set of marginal pdfs into a multivariate pdf with a particular dependence structure. A Copula is a multivariate cumulative distribution function defined on the d-dimensional unite cube $[0, 1]^d$ [8], with uniform one dimensional marginals. Given a d-dimensional vector $x = [x_1, ..., x_d]$ on the unit cube $[0, 1]$, with a cumulative distribution function F and marginal cumulative distribution functions (cdf) $F_1, ..., F_d$. The multivariate cdf is:

$$F(x_1, ..., x_d) = P(X_1 \leq x_1, ..., X_d \leq x_d) \tag{1}$$

Sklar theorem [9] shows that there exists a d-dimensional Copula C such that:

$$F(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d)) \tag{2}$$

Further, if C is continuous and differentiable, the Copula density is given by:

$$c(u_1, ..., u_d) = \frac{\partial^d C(u_1, ..., u_d)}{\partial u_1 ... \partial u_d} \tag{3}$$

The joint pdf is uniquely deduced from the margins and the Copula density as follows:

$$f(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d)) \prod_{i=1}^{d} f_i(x_i) \tag{4}$$

where $f_i, i = 1, ..., d$, represent the marginal densities. It appears that the Gaussian Copula is suitable to model linear dependence, which is the most popular in texture modeling. Gaussian Copula density is defined by:

$$c(u, \Sigma) = \frac{1}{|\Sigma|^{1/2}} \exp[-\frac{1}{2}\vartheta^T(\Sigma^{-1} - I)\vartheta] \tag{5}$$

with $\vartheta_i = \phi^{-1}(F_i(u_i))$, and ϕ represents the standard normal cumulative distribution function. Σ denotes the correlation matrix, and I denotes the d-dimensional matrix identity. From this we can derive the joint model as:

$$f(x, \theta) = \frac{1}{|\Sigma|^{1/2}} \exp[-\frac{1}{2}\vartheta^T(\Sigma^{-1} - I)\vartheta] \prod_{i=1}^{d} f_i(x_i) \tag{6}$$

where $\theta = (\eta, \Sigma)$ denotes the hyperparameters of the joint model, where $\eta = (\eta^{(1)}, \eta^{(2)}, ..., \eta^{(k)})$ represents a set of the marginal parameters and Σ denotes the covariance matrix of the Gaussian vector ϑ.

For estimating parameters of the Gaussian Copula based model we use the IFM (Inference From Margins) method [10]. In a first time, this consists of estimating the parameters of the marginals using the Maximum Likelihood (ML) procedure. ML estimators $\hat{\eta_i}$ are deduced as:

$$\hat{\eta_k} = argmax_{\eta_k} \sum_{i=1}^{n} \log(f_k(x_{ik}, \eta_k)) \tag{7}$$

Secondly, the log-likelihood function for the joint distribution is maximized using the margins estimators $\hat{\eta}$=($\hat{\eta}^{(1)}$, ..., $\hat{\eta}^{(k)}$):

$$\hat{\Sigma} = argmax_{\Sigma} \sum_{i=1}^{n} \log c(F_1(x_{1i}; \hat{\eta_1}), ..., F_d(x_{di}; \hat{\eta_d}); \Sigma) \tag{8}$$

2.2 Multivariate Weibull Distribution

From (6), the pdf of multivariate Weibull (Mwbl) distribution is defined as:

$$f_{Mwbl}(x,\theta) = \frac{1}{|\Sigma|^{1/2}} \exp[-\frac{1}{2}\vartheta^T(\Sigma^{-1} - I)\vartheta] \times (\frac{\tau}{\lambda})^d \prod_{i=1}^{d} x_i^{\tau-1} \exp - \sum_{i=1}^{d} (\frac{x_i}{\lambda})^\tau \quad (9)$$

with $\theta = (\tau, \lambda, \Sigma)$, τ represents the shape parameter, λ represents the scale parameter, and Σ denotes the covariance matrix.

2.3 Multivariate Gamma Distribution

The joint pdf of multivariate Gamma (Mgam) distribution is defined as:

$$f_{Mgam}(x,\theta) = \frac{1}{|\Sigma|^{1/2}} \exp[-\frac{1}{2}\vartheta^T(\Sigma^{-1} - I)\vartheta] \times (\frac{\beta^{-\alpha}}{\Gamma(\alpha)})^d \prod_{i=1}^{d} x_i^{\alpha-1} \exp - \sum_{i=1}^{d} (\frac{x_i}{\beta}) \quad (10)$$

with $\theta = (\alpha, \beta, \Sigma)$, α represents the shape parameter, β represents the scale parameter, and Σ denotes the covariance matrix.

2.4 Multivariate Laplacian Distribution

The pdf of multivariate Laplacian (Mlap) Distribution is defined as:

$$f_{Mlap}(x,\theta) = \frac{1}{|\Sigma|^{1/2}} \exp[-\frac{1}{2}\vartheta^T(\Sigma^{-1} - I)\vartheta] \times \frac{1}{b^d} exp - \sum_{i=1}^{d} \frac{(x_i - a)}{b} \quad (11)$$

with $\theta = (a, b, \Sigma)$, a represents the location parameter, b represents the scale parameter, and Σ denotes the covariance matrix.

3 Rao Geodesic Distance on the Manifold Of Gaussian Copula Based Distributions

Let us consider M_θ as the statistical manifold of Gaussian Copula based probability distributions, and $f(x;\theta)$ a pdf from this manifold, where θ a vector of parameters of f. Rao distance is a Riemannian metric defined by the fisher information matrix as:

$$ds^2 = \sum_{i,j=1}^{d} g_{ij}(\theta) d\theta_i d\theta_j \quad (12)$$

where g_{ij} represents the Fisher matrix elements:

$$g_{ij}(\theta) = E\Big[\frac{\partial}{\partial\theta^i} \log f(X;\theta) \frac{\partial}{\partial\theta^j} \log f(X;\theta)\Big] \quad (13)$$

$$= -E\Big[\frac{\partial^2}{\partial\theta^i\theta^j} \log f(X;\theta)\Big] \quad (14)$$

Hence, given two probability distributions $f(x;\theta_1)$ and $f(x;\theta_2)$ on M_θ, we can compute the Rao geodesic distance as:

$$L = \int_{\theta_1}^{\theta_2} ds = \int_0^1 \sqrt{\sum_{i,j} g_{ij}\dot{\theta}^i\dot{\theta}^j}\,dt \tag{15}$$

In M_θ, probability distributions are defined as:

$$f(x,\theta) = c_\Phi(u_1,...,u_d;\Sigma)\prod_{i=1}^{d} f_i(x_i;\eta_i) \tag{16}$$

So,

$$g_{ij}(\theta) = -E\Big[\frac{\partial^2}{\partial\theta^i\theta^j}\log c_\Phi(u_1,...,u_d;\Sigma)\prod_{i=1}^{d} f_i(x_i;\eta_i)\Big] \tag{17}$$

$$= -E\Big[\frac{\partial^2}{\partial\theta^i\theta^j}\log c_\Phi(u_1,...,u_d;\Sigma) + \log\prod_{i=1}^{d} f_i(x_i;\eta_i)\Big] \tag{18}$$

The vector $(u_1,...,u_d)$, is constructed by transforming the vector $(x_1,...,x_d)$ empirically with known parameters η_i via $u_i = F(x_i;\hat{\eta_i})$, thus we consider these observations without assumptions on the parametric form of the marginal distributions: $(u_1,...,u_d) = (F(x_i,\hat{\eta_1}),...,F(x_i,\hat{\eta_1}))$.

This approach has been already considered for estimating the parameters of the copula without estimating the marginals parameters as an alternative of the IFM method, and was named as CML (Canonical Maximum Likelihood) [14]. When this is supposed, we have:

$$\frac{\partial}{\partial\eta}c_\Phi(u;\Sigma) = 0 \tag{19}$$

so then, $g_{\eta\Sigma}(\theta) = g_{\Sigma\eta}(\theta) = 0$.

Thus

$$g_{\Sigma\Sigma}(\theta) = -E\Big[\frac{\partial^2}{\partial\Sigma\partial\Sigma}\log c_\Phi(u;\Sigma)\Big] \tag{20}$$

and

$$g_{\mu\nu}(\theta) = -E\Big[\frac{\partial^2}{\partial\mu\partial\nu}\log\prod_{i=1}^{d} f_i(x_i;\eta_i)\Big] \tag{21}$$

with $\mu,\nu \in \{\eta^{(1)},...,\eta^{(k)}\}$

$$g_{\mu\nu}(\theta) = -E\Big[\frac{\partial^2}{\partial\mu\partial\nu}\sum_{i=1}^{d}\log f_i(x_i;\eta_i)\Big] \tag{22}$$

$$= -E\Big[\sum_{i=1}^{d}\frac{\partial^2}{\partial\mu\partial\nu}\log f_i(x_i;\eta_i)\Big] \tag{23}$$

$$= \sum_{i=1}^{d} -E\Big[\frac{\partial^2}{\partial\mu\partial\nu}\log f_i(x_i;\eta_i)\Big] \tag{24}$$

Thus, from (12):

$$ds^2 = g_{\Sigma\Sigma} d\Sigma d\Sigma + \sum_{i=1}^{d} \sum_{\mu,\nu} g_{\mu\nu} \dot{\mu}\dot{\nu} dt \tag{25}$$

$$= ds^2_{Gauss} + \sum_{i=1}^{d} ds^2_{Margins} \tag{26}$$

Thus, the Rao distance between two probability density functions $f(x;\theta_1)$ and $f(x;\theta_2)$ on M_θ is defined from equation (15) as follows:

$$L(f(x;\theta_1)||f(x;\theta_2)) = L_{Gauss}(f(x;\Sigma_1)||f(x;\Sigma_2)) + \sum_{i=1}^{d} L_{Margins}(f(x;\eta_1)||f(x;\eta_2)) \tag{27}$$

and then, the Rao distance between two Mwbl pdfs is:

$$L(f_{Mwbl}(x;\theta_1)||f_{Mwbl}(x;\theta_1)) = [\frac{1}{2}\sum_{i=1}^{d}(\ln r^i)^2]^{1/2} +$$

$$\left(\frac{[\log(\tau_2/\tau_1) - s(\lambda_2-\lambda_1)^2/\lambda_1^2\lambda_2^2]^2 + q^2(\lambda_2-\lambda_1)^2/\lambda_1^2\lambda_2^2}{[\log(\tau_2/\tau_1) - s(\lambda_2-\lambda_1)^2/\lambda_1^2\lambda_2^2]^2 + q^2(\lambda_2+\lambda_1)^2/\lambda_1^2\lambda_2^2} \right)^{1/2} \tag{28}$$

where $r^i, i = 1, ..., d$ represents the eigenvalues of $\Sigma_1^{-1}\Sigma_2$, $s = 1 - \gamma$ and $q = \frac{\pi}{\sqrt{6}}$. The Rao distance between two Mlap pdfs is:

$$L(f_{Mlap}(x;\theta_1)||f_{Mlap}(x;\theta_1)) = [\frac{1}{2}\sum_{i=1}^{d}(\ln r^i)^2]^{1/2} +$$

$$\frac{1}{2}\log\frac{\{a_1-(K+R)\}/\{a_1-(K-R)\}}{\{a_2-(K+R)\}/\{a_2-(K-R)\}} \tag{29}$$

with, $K = \frac{1}{2}\frac{b_2^2-b_1^2}{a_2^2-a_1^2} + a_1 + a_2$, and $R^2 = \frac{b_1^2+b_2^2}{2} + \frac{(a_2-a_1)^2}{4} + \frac{(b_2^2-b_1^2)^2}{4(a_2-a_1)^2}$. Sometimes, the geodesic equations of the marginal probability distribution are difficult to solve as the case of the Gamma distribution. In such cases, we proceed by numerical approximations for the geodesic equations [13].

4 Experimental Results

Experiments are conducted on the MIT vision texture database (Vistex) [11]. We consider 30 texture classes as used in [12]. Each 512 × 512 texture image is splitted into sixteen subimages of size 128 × 128 resulting on a database of 480 samples. Then color bands of each subimage are decomposed via the Dual Tree Complex Wavelet Transform (DTCWT). DTCWT presents advantages over the classic DWT (Discret Wavelet Transform), in terms of translation invariance and directional selectivity, since DTCWT provides complex subbands in six orientations at each decomposition level. Every color band (R, G and B) is then

Table 1. Average classification rate for different number of training set

Q	2	3	4	5	6	7	8
Mwbl	86.75	86.90	91.72	92.14	94.25	95	95.96
Mgam	84.83	85.74	91.76	91.93	92.70	93.62	95.33
Mwbl+KL [3]	85	86.14	90.65	91.33	93.75	94.81	95.45
MGGD [2]	84.54	84.98	88.42	90.14	92.49	92.95	94.56

decomposed via DTCWT resulting on six subbands $r_i, g_i, b_i, i \in \{1, ..., 6\}$ for each color component (considering only subbands of the second decomposition level). Thus, the dataset to be modeled is constructed by arranging the absolute values of the subbands coefficients in an $n \times 18$ matrix, where n is the number of subband coefficients.

As said in the introduction we choose the K-Nearest Neighbor approach for the classification purpose. In the KNN approach, an instance is classified reposing on a similarity measure, and is accorded the label of the majority of its K-Nearest Neighbors.

We begin the classification process by dividing our dataset into training and testing sets. From each class of textures, we choose, randomly, Q samples to construct the training set. The rest of the samples are then considered as the testing set. This experiment is repeated 100 times before returning the average classification rate. In our experiments, we vary Q from 2 to 8 in order to test performances for different numbers of training samples.

Table 1 shows the average classification rates for two submanifolds of M_θ namely Mwbl and Mgam with Rao distance as a similarity measure in comparison with the MGGD model [2], also with the same similarity measure, and then with the approach proposed in [3], which uses the Monte-carlo based KL divergence with Multivariate Weibull distribution. We observe that for Mwbl and Mgam higher classification rates are achieved for $Q = 8$, with *95.96%* for Mwbl and *95.33%* for Mgam, which is better than the rates achieved when using MGGD (*94.56%* for $Q = 8$). This is due the flexibility of the Gaussian Copula based models in describing the information of dependence. Also, it can be seen that, in comparison with the KL divergence based approach, results with the Rao distance are slightly better. These little improvements in term of classification rates are valuable, since one can replace the use of the Kullback-Leibler divergence by the Rao distance, and then benefits the advantages of this latter over the KL divergence, as being a distance in the right sense of the word, respecting properties of symmetry and triangularity. Another advantage of the Rao distance, is that for multivariate distributions, and especially those based on copulas, the KL divergence is numerically computed using the Monte carlo method, known as being computationally expensive.

5 Conclusions

In this work, we have proposed to use the Rao distance as a similarity measure between Gaussian Copula based multivariate distributions. We derived a closed form for the Rao distance of these joint distributions when the Rao distance of the marginal distributions is known. Results in term of classification show the Rao distance can replace the use of the Kullback-Leibler divergence since rates are improved, and then benefit the advantages of the Rao distance.

References

1. Do, M., Vetterli, M.: Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. IEEE Transactions on Image Processing 11, 146–158 (2002)
2. Verdoolaeg, G., De Backer, S., Scheunders, P.: Multiscale colour texture retrieval using the geodesic distance between multivariate Generalized Gaussian models. In: Proceedings of the 15th IEEE Interational Conference On Image Processing (ICIP 2008), San Diego, California, USA, pp. 169–172 (2008)
3. Kwitt, R., Uhl, A.: A joint model of complex wavelet coefficients for texture retrieval. In: 16th IEEE International Conference on Image Processing, ICIP 2009, pp. 1877-1880 (2009)
4. Rao, C.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. 37, 81–89 (1945)
5. Pastore, J., Moler, E., Ballarin, V.: Segmentation of brain magnetic resonance images through morphological operators and geodesic distance. In: Digital Signal Processing, vol. 15(2) pp. 153-160 (2005)
6. Yong, Q., Jie, Y.: Modified kernel functions by geodesic distance. EURASIP Journal on Applied Signal Processing 16, 2515–2521 (2004)
7. Oller, J.M.: Information Metric for Extreme Value and Logistic Probability Distributions. The Indian Journal of Statistics, Series A 49, 17–23 (1987)
8. Nelsen, R.B.: An Introduction to Copulas, 2nd edn. Springer Series in Statistics. Springer, Heidelberg (2006)
9. Sklar, M.: Fonctions de répartition à n dimensions et leurs marges. Publications de l'institut de Statistique de l'Université de Paris. vol. 8, pp. 229-231 (1959)
10. Joe, H.: Multivariate Models and Dependence Concepts. Monographs on Statistics and Applied Probability. Chapman & Hall, Boca Raton (1997)
11. MIT vision and modeling group, http://vismod.media.mit.edu
12. Van De Wouwer, G., Scheunders, P., Van Dyck, D.: Statistical Texture Characterization From Discrete Wavelet Representations. IEEE transactions on image processing 8, 592–598 (1999)
13. Reverter, F., Oller, J.M.: Computing the Rao distance for Gamma distributions. Journal of Computational and Applied Mathematics 157, 155–167 (2003)
14. Durrleman, V., Nikeghbali, A., Roncalli, T.: Which copula is the right one. Groupe de Recherche Operationnelle, Credit Lyonnais (2000)

Texture Analysis Based on Saddle Points-Based BEMD and LBP

JianJia Pan and YuanYan Tang

Department of Computer Science, Hong Kong Baptist University,
Hong Kong SAR, China
{jjpan,yytang}@comp.hkbu.edu.hk

Abstract. In this paper, a new texture analysis method(EMDLBP) based on BEMD and LBP is proposed. Bidimensional empirical mode decomposition (BEMD) is a locally adaptive method and suitable for the analysis of nonlinear or nonstationary signals. The texture images can be decomposed to several BIMFs (Bidimensional intrinsic mode functions) by BEMD, which present some new characters of the images. In this paper, firstly, we added the saddle points as supporting points for interpolation to improve the original BEMD, and then the new BEMD method is used to decompose the image to components (BIMFs). After then, the Local Binary Pattern (LBP) method is used to detect the feature from the BIMFs. Experiments shown the texture image recognition rate based on our method is better than other LBP-based methods.

Keywords: texture analysis, empirical mode decomposition, LBP.

1 Introduction

Texture analysis is widely recognized as a difficult and challenging computer-vision problem. It provides many applications such as in remote sensing image, medical image diagnosis, document analysis, and target detection, etc.

Many methods have been used in texture analysis, among which, the local descriptors are widely used. The most popular local descriptors are the Gabor wavelet [Manjunath96] and local binary pattern (LBP) [Ojala02]. The Gabor representation has been shown to be optimal in the sense of minimizing the joint two-dimensional uncertainty in space and frequency [Manjunath96]. The Gabor wavelet has been widely used in image analysis, such as texture analysis and segmentation, face recognition. Another important local descriptor is local binary pattern (LBP), which has gained increasing attention due to its simplicity and excellent performance in various texture and face recognition. Many improvements of original LBP have been proposed [Guo10] [Zhao07] [Ahonen09]. LBP operator has been extended to use neighborhoods for different sizes [Ojala02]. Guo [Guo10] proposes an alternative hybrid scheme, globally rotation invariant matching with locally variant LBP texture features. Chen [Chen10] proposes the Weber Local Descriptor (WLD) inspired by Webers Law and LBP, which consists of two components: differential excitation and orientation.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 506–514, 2011.

Recently, Empirical mode decomposition (EMD), developed by Huang [Huang98], has been used for the texture analysis and face recognition [Nunes03]. EMD is a data driven processing algorithm which applies no predetermined filter. The EMD is based on the local characteristics scale of the data, which is able to perfectly analyze the nonlinear and nonstationary signals and presents the illumination robust features. EMD has been used to analyze the two-dimensional signals, for example, the images, which is known as bidimensional EMD (BEMD).

BEMD presents some better quality than Fourier, wavelet and other decomposition algorithms in extracting intrinsic components of textures because of its data driven property [Nunes03] [Huang03]. It is different from the wavelet-based multi-scale analysis that characterizes the scale of a signal event using pre-specified basis functions.

One research topic in BEMD is the extrema points detection method [Sharif08] [Nunes03], which supplies the supporting points for the interpolation. If the extrma detection loses some significant supporting points, the BIMFs' orthogonality will increase. In this paper we add the saddle points[Xu06] as the supporting points for the interpolation, and give the definition of saddle points. Combined with the neighbor local maxima and minima points [Pan10] , three types of points are treated in the same way and the interpolation in BEMD is improved.

Based on the new BEMD, in this paper, we propose using the Local Binary Patterns (LBP) as the texture descriptor to detect the characteristic of texture images. The BEMD decompose the original image to a new multi-scale components (BIMFs). In those new components, the LBP can be better work than in the original images. Experiments shows the texture image recognition rate based on our method is better than other LBP-based descriptors.

2 Review of BEMD

Empirical mode decomposition (EMD) is first proposed by Huang [Huang98] for the processing of non-stationary functions. This tool decomposes signal into components called Intrinsic Mode Functions (IMFs) satisfying the following two conditions:

(a)The numbers of extrema and zero-crossings must be either equal or differ at most by one;

(b)At any point, the mean value of the envelope defined by the local maxima and the envelope by the local minima is zero.

Huang [Huang98]has also proposed an algorithm called 'sifting' to extract IMFs from the original signal $f(t)$ as follows:

$$f(t) = \sum_{i=1}^{N} I_i(t) + r_N(t) \tag{1}$$

Where $I_i(t)$, $i = 1, , N$ are IMFs and $r_N$is the residue.

The bidimensional EMD (BEMD) process is conceptually the same as the one dimension EMD, except that the curve fitting of the maxima and minima envelope now becomes a surface fitting exercise and the identification of the local extrema is performed in space to take into account for the connectivity of the points.

The main process of the BEMD can be described as:

(a)Locate the maximum and minimum points in the image $I(i,j)$;

(b)Interpolate the surface among the all maxima (resp. minima) to build the envelope $X_{max,k}(i,j)$ and $X_{min,k}(i,j)$;

(c)Compute the mean envelope

$$X_k(i,j) = (X_{max,k}(i,j) + X_{min,k}(i,j))/2 \tag{2}$$

(d)Update the $I_k(i,j) = I_{k-1}(i,j) - X_k(i,j)$;

(e)Check the stopping criterion

$$SD = \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{(I_k(i,j) - I_{k-1}(i,j))^2}{(I_{k-1}(i,j))^2} \tag{3}$$

if SD is larger than a threshold ε, repeat the steps (a)-(e) with $I_k(i,j)$ as the input, other wise, $I_k(i,j)$ is an IMF $d_k(i,j)$;

(f)Update the residual $I_k(i,j) = I_{k-1}(i,j) - d_k(i,j)$;

(g)Input the $I_k(i,j)$ to steps(a)-(e) until it can not be decomposed, and the last residual is $I_k(i,j) = r(i,j)$.

After the BEMD, the decomposition of the image can be rewritten as following form:

$$I(i,j) = \sum_{k=1}^{K} d_k(i,j) + r(i,j) \tag{4}$$

The $d_k(i,j)$ is the BIMFs (bidimensional intrinsic mode functions) of the images, and $r(i,j)$ is the residual function.

3 BEMD Based on Saddle Points

In our previous work [Pan10], we have proposed some approaches for solving problems in BEMD. The local extrema point was detected based on its neighbor and the extended parts were rebuilt based on self-similarity. In this paper we add the saddle points [Xu06] as the supporting points for the interpolation.

The first step in BEMD is to detect the local extrema points. In one dimensional EMD, the extrema points are local maxima or minima points. In two dimensional, except the local maxima or minima points, there will be a new condition. One point may be a maxima point in one dimensional but a minima point in other dimensional. By use of the neighbor local extrema detection [Pan10], these points may not be detected. The saddle points are simultaneously a local maximum and local minimum point evaluated in different directions, and they also give important supporting features about the local variation of the original function [Xu06].

Definition 1. a two dimensional function or a three dimensional curve $u(x,y)$, in the points (x_0, y_0) can be Taylor expansion:

$$u(x,y) = u(x_0,y_0) + \frac{\partial u}{\partial x}\triangle x + \frac{\partial u}{\partial y}\triangle y + \frac{1}{2}[\frac{\partial^2 u}{\partial x^2}(\varDelta x)^2 + 2\frac{\partial^2 u}{\partial x \partial y}\varDelta x \varDelta y + \frac{\partial^2 u}{\partial y^2}(\varDelta y)^2] + \cdots \tag{5}$$

Where $\varDelta x = x - x_0$, $\varDelta y = y - y_0$,
if $\frac{\partial u}{\partial x}|_0 = 0$, $\frac{\partial u}{\partial y}|_0 = 0$, we have

$$u(x,y) - u(x_0,y_0) = \frac{1}{2}[u_{xx}(\varDelta x + \frac{u_{xy}}{u_{xx}}\varDelta y)^2 + (\triangle y)^2(u_{yy} - \frac{u_{xy}^2}{u_{xx}})] + \cdots \tag{6}$$

the extrema point is *saddle point*, if $u_{xy}^2 > u_{xx}u_{yy}$.

Definition 2. $f(i,j)$ is a maximum (or. minimum) if it is larger (or. lower) than the value of f at the nearest neighbors of (i,j).

Let the window size for local extrema determination be $(2w+1) \times (2w+1)$, Then

$$x_{mn} = \left\{ \begin{array}{l} x_{max} \text{ if } x_{mn} > x_{ij}, \\ x_{min} \text{ if } x_{mn} < x_{ij}. \end{array} \right\} \tag{7}$$

Where $x_{ij} = \{x|(m-w) : i : (m+w), (n-w) : j : (n+w), i \neq m, j \neq n\}$

From the experimental, we find that 3 × 3 window results is an optimum extrema map for given images.

These three type points, saddle points, neighbor local maxima and neighbor local minima points, are treated in the same way. All these points are detected from the original image and supply the supporting points for the BEMD's interpolation.

4 Texture Descriptor Based on BEMD and LBP

To analyze and classify the texture images, we propose using the LBP descriptor to extract the local features from the decomposed BIMFs. Local Binary Patterns (LBP) is introduced as a powerful local descriptor for microstructure of images [Ojala02]. In this section, we firstly give an introduction of the original LBP, and then present the new feature detection method based on LBP and BEMD.

4.1 Local Binary Patterns (LBP)

The LBP operator is originally developed for texture description. The operator assigns a label to every pixel of an image by thresholding the 3×3-neighborhood of each pixel with the center pixel value and considering the result as a binary number. Then the histogram of the labels can be used as a texture descriptor [Ojala02].

The form of the resulting 8-bit LBP code can be defined as follows [Ojala02]:

$$LBP(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n \tag{8}$$

where i_c corresponds to the gray value of the center pixel (x_c, y_c) into the gray values of the 8 neighborhood pixels, and function $s(x)$ is defined as:

$$s(x) = \begin{Bmatrix} 1 \text{ if } x \geq 0, \\ 0 \text{ if } x < 0. \end{Bmatrix} \tag{9}$$

From the above processing, the LBP presents that it will be not affected by any monotonic gray-scale transformation which preserves the pixel intensity order in a local neighborhood. Each bit of the LBP code has the same significance level and that two successive bit values may have a totally different meaning.

To deal with textures at different scales, the LBP operator is later extended to use neighborhoods for different sizes [Ojala02]. The local neighborhood is extended to as a set of sampling points evenly spaced on a circle centered at the pixel to be labeled allows any radius and number of sampling points[Zhao07]. If a sampling point is not in the center of a pixel, it will be rebuilt by bilinear interpolation. The notation (P, R) is defined as the pixel neighborhood which means P sampling points on a circle of radius of R. Figure 1 shows an example of circı

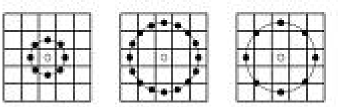


Fig. 1. The circular (8,1) (16,2) (8,2) neighborhoods

4.2 LBP Based on BEMD

In our method, the texture images are firstly decomposed by BEMD into several components BIMFs, and then, the LBP is used to extract the local features of BIMFs.

BEMD is based on the local characteristics scale of the data and the filtering occurs in time space rather than in frequency space; therefore, it is able to perfectly analyze and present the nonlinear and nonstationarity signals. The corresponding high-frequency components are more robust to the illumination changes [Zhang05]. Moreover, the corresponding BIMFs by the BEMD based methods are able to capture more representative features of the original signal, especially more singular information in high frequency ones. At the same time, because of the invariance of the LBP features, the LBP can be suit for the considerable gray-scale variations in images and no normalization of input images is needed. LBP is a nonparametric method, which means that no prior knowledge about the distributions of images is needed.

We use the following procedure for detect texture features:

Firstly, the original image I is decomposed into its BIMFs d_k:

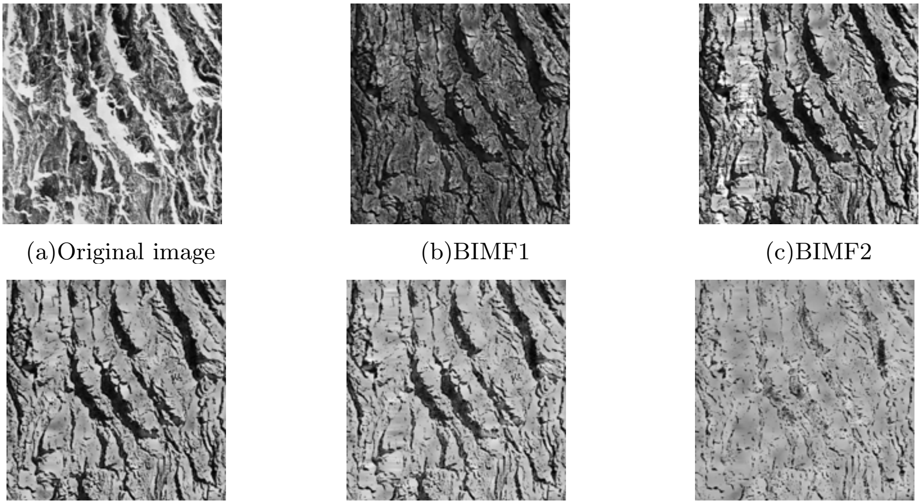
$$I(i, j) = \sum_{k=1}^{K} d_k(i, j) + r(i, j) \tag{10}$$

Secondly, as we can find from the texture images' BIMFs (Figure 2), the first and the second BIMF remain the main detail of the original image, and the last BIMFs represent the information in large scales. The characteristic points in BIMF3(BIMF4, BIMF5) are in a larger scales, using the same size of LBP to detect the LBP code of these BIMFs will be ineffective. In [Ojala02], the LBP operator was extended to use neighborhoods for different sizes.

In our experiment, the 1^{st} and 2^{nd} LBP code of BIMF are detected by the $LBP_{8,1}$, the last BIMFs are detected by $LBP_{16,2}$.

Lastly, features are detected from BIMFs by the LBP:

$$EMDLBP = [LBP_{8,1}^{BIMF1} \quad LBP_{8,1}^{BIMF2} \quad LBP_{16,2}^{BIMF3} \ldots LBP_{16,2}^{BIMFn}] \tag{11}$$



(a)Original image (b)BIMF1 (c)BIMF2

(d)BIMF3 (e)BIMF4 (f)BIMF5

Fig. 2. BIMFs of the texture image

5 Experimental Result

The proposed algorithm for texture analysis is tested and compared with other methods on two database: Brodatz database and KTH-TIPS2-a database.

The Brodatz database is a widely used database for texture recognition. It consists of 111 images. Following the same processing in previous approaches [Lazebnik05], each class image is partitioned into nine non-overlapping fragments, for a total of 999 images. The classifier is based on the nearest-neighbor classifier. For Brodatz database, the proposed feature is compared with LBP [Ojala02], LBPV [Guo10], RIFT [Lazebnik05].

Table1 shows recognition results for outputs of proposed EMDLBP method and other methods. The original LBP [Ojala02], LBPV [Guo10] and RIFT [Lazebnik05] methods are used in the same database. By comparing the classification rates, we can find that the proposed EMDLBP method gives better performance than LBP, LBPV and RIFT method. BEMD decompose the texture

Table 1. Classification accuracy of proposed EMDLBP in the Brodatz database, compared with LBP, LBPV, RIFT

Classification Accuracy				
LBP	LBPV(8,1)	LBPV(8,2)	RIFT	EMDLBP
0.7568	0.8258	0.7883	0.7628	0.8468

Table 2. Classification accuracy of proposed EMDLBP and other methods in the KTH-TIPS2-a textures database

Classification Accuracy						
Parameters	LBP^{u2}	LBP-HF	WLD	LBPV	FBL-LBP	EMDLBP
(8,1)	0.525	0.525	0.564	0.552	0.619	0.642
(16,2)	0.508	0.533	0.585	0.568	0.624	0.648
(8,1)+(16,2)	0.538	0.542	0.647	0.575	0.631	0.729

image into BIMFs, and the different sizes of LBP are suit for BIMFs' multiscale decomposition. The experimental results show that the multiscale LBP can better work in the BIMFs than in the original images.

The KTH-TIPS2-a database is another database widely used for texture classification and material categorization, which contains four physical, planar samples of each of 11 materials under varying illumination, pose, and scale[Chen10]. The KTH-TIPS2-a texture data set contains 11 texture classes with 4572 images.

The NN classifier is trained with one sample (i.e. 9×12 images) per material category. The remaining $3 \times 9 \times 12$ images are used for testing. This is repeated with 1000 random combinations as training and testing data and the result are reported as the average value.

Classification accuracy of the EMDLBP and other LBP-based methods are listed in Table2. Note that the results from LBP-HF [Ahonen09], FBL-LBP [Yimo10] in Table2 are quoted directly from the original papers. In Table2, the (8,1)+(16,2) for our proposed EMDLBP is the feature of combined (8,1) and (16,2) as described in Equation(11). The result of (8,1) and (16,2) is just simply used the $LBP_{8,1}$ and $LBP_{16,2}$ in the all BIMFs. It can be found that the EMDLBP has better performance than other methods in all cases. Specially, the multi-scale (8,1)+(16,2) can achieve higher performance than all other methods included the single (8,1) or (16,2) in BIMFs. As we mentioned in Part 4, the BIMFs are multi-scale components of the original images, using the different sizes of LBP to detect the LBP code of different BIMFs is more effective.

6 Conclusion

Texture analysis is widely recognized as a difficult and challenging computer-vision problem. In this paper, a new global-local feature (EMDLBP) is proposed for texture analysis. To improve the local descriptor LBP, a new decomposition method BEMD is used to supply new multi-scale components(BIMFs). In these new multi-scale components, the LBP descriptor can achieve better performance than original images.

Firstly, BEMD method based on saddle points is proposed, which can reduce the orthogonality index of BIMFs and improve the BEMD. And then the BEMD decompose images to different BIMFs, which present different scale information of the original image. After then, the LBP method is used to detect the local information of BIMFs. Experiments show the texture image recognition rate based on our method is better than other LBP-based methods.

References

[Huang98] Huang, N.E., Shen, Z., Long, S.R., Wu, M.C.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Royal Society. Lond 454, 903–1005 (1998)

[Nunes03] Nunes, J.C., Bouaoune, Y., Delechelle, E., Niang, O., Bunel, P.: Image analysis by bidimensional empirical mode decomposition. Image and Vision Computing 21(12), 1019–1026 (2003)

[Huang03] Huang, N.E., Wu, M.L.C., Long, S.R.: A confidence limit for the EMD and Hilerbet spectral analysis. Royal Society A 459, 2317–2345 (2003)

[Sharif08] Bhuiyan, S.M.A., Adhami, R.R., Khan, J.F.: Fast and Adaptive Bidimensional Empirical Mode Decomposition Using Order-Statistics Filter Based Envelope Estimation. Journal on Advances in Signal Processing (2008)

[Nunes05] Nunes, J.C., Guyot, S., Deechelle, E.: Texture analysis based on local analysis of the Bidimensional Empirical Mode Decomposition. Machine Vision and Applications 16(3), 177–188 (2005)

[Ojala02] Ojala, T., Pietikainen, M., Maenpaa, T.T.: Multiresolution gray-scale and rotation invariant texture classification with Local Binary Pattern. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)

[Guo10] Guo, Z., Zhang, L., Zhang, D.: Rotation invariant texture classification using LBP variance(LBPV) with global matching. Pattern Recognition 43(3), 706–719 (2010)

[Zhao07] 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)

[Lazebnik05] Lazebnik, S., Schmid, C., Ponce, J.: A Sparse Texture Representation Using Local Affine Regions. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1265–1278 (2005)

[Ahonen04] Ahonen, T., Hadid, A., Pietikainen, 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)

[Ahonen09] Ahonen, T., Matas, J., He, C., Pietikainen, 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)

[Chen10] Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X., Gao, W.: WLD: A Robust Local Image Descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1705–1720 (2010)

[Zhang05] Zhang, Z.B., Ma, S.L., Wu, D.Y.: The application of neural network and wavelet in human face illumination compensation. Proc. Advances in Neural Networks, 828–835 (2005)

[Pan10] Pan, J., Zhang, D., Tang, Y.: A fractal-based BEMD method for image texture analysis. In: IEEE International Conference on Systems, Man, and Cybernetics, Turkey, pp. 3817–3820 (October 2010)

[Xu06] Xu, Y., Liu, B., Liu, J., Rimenschneider, S.: Two-dimensional empirical mode decomposition by finite elements. In: Proceedings of Royal Society, pp. 3081–3090 (2006)

[Manjunath96] Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. IEEE Transactions on Pattern Analysis and Machine Intelligence 18(8), 837–842 (1996)

[Yimo10] Guo, Y., Zhao, G., Pietikainen, M., Xu, Z.: Descriptor Learning Based on Fisher Separation Criterion for Texture Classifcation. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 185–198. Springer, Heidelberg (2011)

A Robust Approach to Detect Tampering by Exploring Correlation Patterns

Lu Li[1], Jianru Xue[1], Xiaofeng Wang[1,2], and Lihua Tian[1]

[1] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[2] School of science, Xi'an University of Technology, China
{lilu.acrobat,jrxuester,xfwangxf,lihua.tian}@gmail.com,

Abstract. Exposing digital forgeries by detecting local correlation patterns of images has become an important kind of approach among many others to establish the integrity of digital visual content. However, this kind of method is sensitive to JPEG compression, since compression attenuates the characteristics of local correlation pattern introduced by color filter array (CFA) interpolation. Rather than concentrating on the differences between image textures, we calculate the posterior probability map of CFA interpolation with compression related Gaussian model. Thus our approach will automatically adapt to compression. Experimental results on 1000 tampered images show validity and efficiency of the proposed method.

Keywords: digital forgery, color filter array, posterior probability map.

1 Introduction

A wide distribution of digital cameras, in combination with sophisticated image editing softwares, makes altering digital images ubiquitous. The fact that seeing is not believing puts an urgent demand for the techniques of image forensic detection. It also motivates recent intense research on digital image forensic tools[1] that can assess the authenticity of digital images without access to the source image or source devices.

Fortunately, although digital forgeries may leave no visual clues of having been tampered with, they may alter the underlying patterns of pixels that make up a digital image. These patterns that many image forensic tools depend on can be summarized into three categories: (1) pixel-based patterns, which detect anomalies introduced at the pixel level by manipulations such as cloning, splicing, etc[2][3], (2) correlation patterns, which are induced by imaging sensors[4][5], imaging formulation[6][7], and formats for storage[8][9], and (3) explicitly object model based patterns, which detect anomalies in the 3D interactions among physical objects, light, and the camera[10][11].

Among these aforementioned underlying patterns, a particular kind of forensic method belonging to the second category, which relies on characteristics of color filter array (CFA) interpolation pattern between neighboring pixels have been widely used in many recent image forensic tools[12][13]. Nearly every digital

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 515–522, 2011.

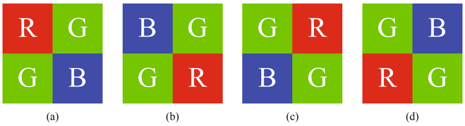


Fig. 1. CFA Configurations

forgery starts out as a photo taken by a digital camera. To obtain a full-color image, the vast majority of sensors employ CFA, such that each sensor element only captures light of a certain wavelength. The remaining color information has then to be estimated from the surrounding pixels of the raw image. This interpolation introduces specific correlation patterns between the pixels of a color image. The main idea of this kind of methods is that when creating a digital forgery these correlation patterns may be destroyed or altered. However, these methods are sensitive to lossy compression, which attenuates the trace of CFA correlation patterns.

In this paper, we propose a compression robust approach to detect tampering in images by detecting CFA correlation patterns, rather than specific CFA interpolation. The main idea is that the damage of these CFA correlation patterns due to lossy JPEG compression can be regarded as additive gaussian noise over the image. Thus the value of pixels can be described with statistics based Gaussian models, which can automatically adapt to different compression level. The proposed method consists of three subsequent key steps: 1) calculating residual map, 2) estimating the correlation pattern, and 3) detecting the periodicity. In the first two steps, we describe the correlation patterns with posterior probability map. Specifically, based on the residual distribution, we develop a technique which can automatically adjust the Gaussian models to make it more suitable for the compression level. In the final step to detect periodicity, we use the two dimensional discrete Fourier Transform(2D-DFT). The resulted method not only provides a new way for these CFA pattern based image forensic tools to extend their usage, but also shows validity and efficiency over thousands of tampered and computer generated images.

2 Detecting Traces of CFA

CFA plays an important role in digital color image formulation. To obtain a full-color image, the vast majority of sensors employ a CFA. Typically the CFA contains three color filters: red, green, and blue as shown in Fig. 1. CFA interpolation requires re-sampling the original image onto a new sampling lattice, introducing specific periodic correlations between neighboring pixels. This underlying pattern forms basis for the widely used image forensic tools proposed in[7], and image forensic can be detected by checking for the existence or consistency of CFA artifacts in digital camera images.

CFA interpolation introduces spatially periodic inter-pixel correlations into the image as a natural fingerprint of tamper detection. The complete forgeries detection method includes three steps. First, we calculate residual map by applying a highpass filter to the image. Then we choose the green channel and compute the posterior probability map of CFA interpolation. Finally, we detect the periodicity of the posterior probability map. According to a simple threshold, we can determine whether the test image is tampered or not.

2.1 Calculation of Residual Map

In the approach, we assume that CFA interpolation takes the form of bilinear interpolation. This simplifying assumption has been proved to be robust to different CFA interpolation algorithms in [14]. To make the approach robust to compression and in-camera processing, the residual map is calculated by applying a highpass filter H to the image's green channel, which contains more original information than others. The highpass filter H is defined as follows:

$$H = \frac{1}{4}\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \tag{1}$$

The local correlation patterns of CFA emerges in the residual map as a periodic distribution.

2.2 Estimation of Posterior Probability Map

For a full color image, its pixels fall into two classes: interpolated pixels that are obtained by interpolation of its neighboring pixels, and raw pixels that are obtained from the imaging sensor directly. We denote these two classes as M_1 and M_2, respectively. For pixels belonging to M_1, they satisfy that:

$$f(x,y) = \sum_{m,n=-N}^{N} u(m,n)f(x+m,y+n) + \delta_i(x,y) \tag{2}$$

where $f(x,y)$ is the pixel in green channel, and u is the bilinear filter. $\delta_i(x,y)$ denotes the residual of approximation for interpolated pixels. $\delta_i(x,y)$ is an independent and identically distributed error that follows a Gaussian distribution with zero mean and variance σ_i^2. Similarly, for pixels belonging to M_2, they satisfy that:

$$f(x,y) = \sum_{m,n=-N}^{N} u(m,n)f(x+m,y+n) + \delta_o(x,y) \tag{3}$$

where $\delta_o(x,y)$ denotes the residual of approximation for uninterpolated pixels, which follows a Gaussian distribution with zero mean and variance σ_o^2. According to the analysis in [12], σ_o is much bigger than σ_i. Because of nonlinear operations, the compression will induce quantization error, δ_Q, into the image. It can be regarded as additive gaussian noise. The residual of approximation in M_1 is defined as:

$$n_i = \delta_i(x,y) + \delta_Q(x,y) \tag{4}$$

The residual of approximation in M_2 is defined as:

$$n_o = \delta_o(x,y) + \delta_Q(x,y) \tag{5}$$

n_i is drawn from a normal distribution with variance $(\sigma_i^2 + \sigma_Q^2)$. And n_o is drawn from a normal distribution with variance $(\sigma_i^2 + \sigma_Q^2)$. Let σ_{M1} and σ_{M2} separately denotes the variance of n_i and n_o.

Based on Bayes' rule, the posterior probability of each pixel belonging to class M_1 is then defined as:

$$P\{f(x,y \in M_1 \mid f(x,y)\} = \frac{P\{f(x,y) \mid f(x,y) \in M_1\}P\{f(x,y) \in M_1\}}{\sum\limits_{i=1}^{2} P\{f(x,y) \mid f(x,y) \in M_i\}P\{fx,y \in M_i\}} \tag{6}$$

where priors $P\{f(x,y) \in M_1\}$ and $P\{f(x,y) \in M_2\}$ are assumed to be 1/2. The probability of a pixel belonging to class M_1, $P\{f(x,y) \mid f(x,y) \in M_1\}$, is defined as:

$$P\{f(x,y) \mid f(x,y) \in M_1\} = \frac{1}{\sigma_{M1}\sqrt{2\pi}} e^{\frac{(f(x,y) - \sum_{m=-N}^{N}\sum_{n=-N}^{N} u_{m,n} f(x+m,y+n))^2}{-2\sigma_{M1}^2}} \tag{7}$$

In addition, the probability of a pixel belonging to class M_2 is defined in the similar way as:

$$P\{f(x,y) \mid f(x,y) \in M_2\} = \frac{1}{\sigma_{M2}\sqrt{2\pi}} e^{\frac{(f(x,y) - \sum_{m=-N}^{N}\sum_{n=-N}^{N} u_{m,n} f(x+m,y+n))^2}{-2\sigma_{M2}^2}} \tag{8}$$

Now the only unknown parameters are the variances σ_{M1}^2 and σ_{M2}^2. In green channel, the CFA interpolation is with a finite kinds of configurations. The variance is separately computed in all kinds of CFA configurations. The smallest one is taken as the variance of $P\{f(x,y) \mid f(x,y) \in M_1\}$. And σ_{M2}^2 is subsequently identified.

The posterior probability map shows the periodic interpolation of the image in an intuitive way. If an image is CFA interpolated, the period of the posterior probability map will be 2 in both horizontal and vertical directions.

2.3 Detection of Periodicity

Generally, CFA has four kinds of configurations, as shown in Fig. 1. In the green channel, the samples captured by the camera are arrayed as rhombuses in the lattice. This is regarded as an inherent feature of CFA. In practice, we detect the periodicity of the posterior probability map $p(x,y)$ calculated in Sec. 2.2. By applying 2D-DFT to $p(x,y)$, $\left|P(e^{j\omega_1}, e^{j\omega_2})\right|$ is obtained. The high peak at the frequency $(\omega_1, \omega_2) = (\pi, \pi)$, indicates that the period of $p(x,y)$ is 2 in both horizontal and vertical directions. To overcome the influence of noise, we define feature R to measure the trace of CFA interpolation as:

$$R = \frac{|P_{\pi,\pi}|}{|P_{mid}|} \tag{9}$$

where $P_{\pi,\pi}$ is the element of the Fourier series at the frequency $(\omega_1, \omega_2) = (\pi, \pi)$, and P_{mid} is the element whose amplitude is the median of all the amplitudes. The higher value of R indicate more obvious CFA pattern.

3 Experimental Results

3.1 Robustness and Sensitivity

The proposed method is evaluated using an uncompressed image data set, and 1000 images are selected randomly as test images. Then they are compressed with different JPEG quality factors from 20 to 100, and are manipulated with different image editing tools, including blurring, down-sampling, up-sampling and rotation. These tampered images form our final test images.

We first test whether the method is robust to different interpolation algorithms. These 1000 images are first interpolated in five kinds interpolation algorithms including bilinear, bicubic, median, smooth hue and gradient based interpolation. The final results is shown in Fig. 2. The threshold of R is determined according to the image compression level. In uncompressed images the threshold of R can be determined as 32 with the accuracy 99.69%. Though the nonlinear operation have an adverse influence on the detection, the results indicate that the proposed method is not only robust to JPEG compression but also to difference interpolation method.

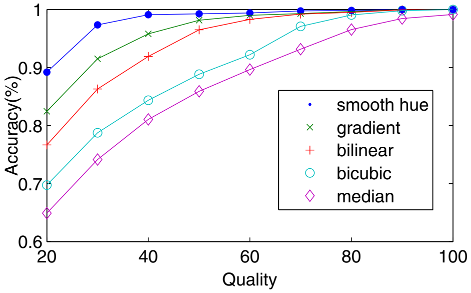


Fig. 2. Shown in the panel are the detection results of our method in different interpolations and different compression qualities

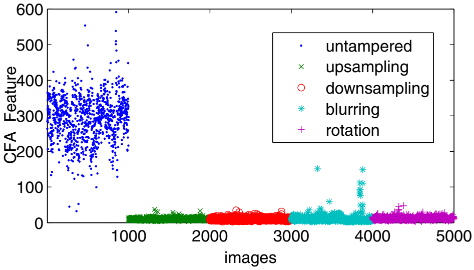
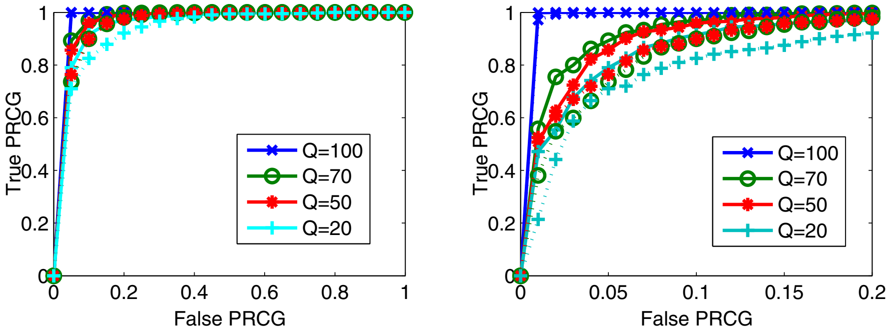


Fig. 3. Tamper Detection



(a) The original ROC curve (b)The magnified ROC curve

Fig. 4. The comparison of compression robustness between the proposed method and the existing one[12]. Each curve present the performance of distinguishing PRCG from untampered images in different compression levels. The proposed method is demonstrated with solid lines. The dashed lines present the method in [12]. The original curve is shown on the left and the magnified version is shown on the right.

To test the sensitivity, 1000 images are manipulated with different image editing tools(3% upsampling, 3% downsampling, blurring with 3 × 3 kernel and 2 degree rotation), making up of usual tampering operations. Applying proposed method to the tampered images, the detection result is shown in Fig.3. With a simple threshold, the tampered images can be separated from the original ones according to feature R. The performance of the proposed method when the different editing tools are used in tampering is presented in Table. 1. These

Table 1. Detection accuracy when different editing tools are used

Rotation	Up-sampling	Down-sampling	Blurring
100%	99.8%	99.9%	98.4%

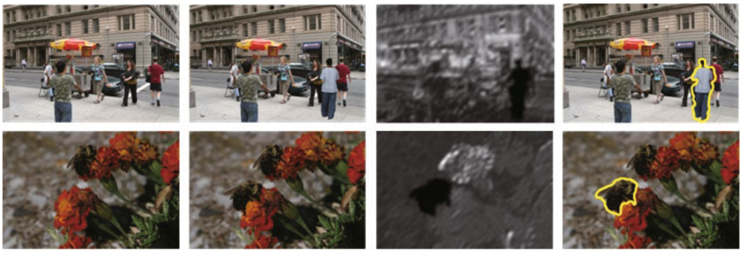


Fig. 5. Shown above are the results of local tamper detection. The first column shows the original images. The second column shows the tampered images. Column 3 shows the calculated R map. The final results are demonstrated in column 4. The boundaries of the tampered region are highlighted.

empirical results show that even slight manipulation will destroy the trace of CFA.

Finally, we compare the compression robustness between our method and the previous one with the best performance[12]. With the test set employed in Columbia's ADVENT dataset from [15], 1600 images (including 800 PIM and 800PRCG) are selected to be distinguished by the proposed method and the approach in [12] separately. As presented in Fig.4, the ROC curves of the proposed method have a better performance in different compression levels. The performance of distinguishing gradually decreases and is robust to JPEG compression.

3.2 Local Tamper Detection

To validate our approach in local tamper detection, the posterior probability map is calculated firstly as described in Sec.2.2. By applying sliding window, the feature R at each window center can be computed. In R map the tampered region can be detected with only a threshold. The tampered image is formed by inserting an image patch. Many editing tools, such as resizing, blurring, rotation and so on, are used to remove visual cues of tampering and make it credible in visual. Fig. 5 shows immediate results. In the calculated R map, tampered regions is darker than other areas. The tampered regions are completely detected.

One limitation of the the proposed method is that it is sensitive to smooth regions, especially in compressed images. The smooth regions of an image have less textures and is easily incorrect detected as area suffered blurring operation.

4 Conclusion

We propose a compression robust approach to expose tampering in images by detecting CFA correlation patterns in this paper. The proposed method can attenuate the damage of these CFA correlation patterns due to lossy JPEG compression. The resulted method not only provides a new way for these CFA pattern based image forensic tools to extend their usage, but also improves performance of previous methods. Experimental results on 1000 images show validity and efficiency of the proposed method.

Acknowlegement. This work was supported by the National Natural Science Foundation of China under Grant Nos.60875008 and 61075007, the National Basic Research Program of China (973 Program) under Grant No.2010CB327902.

References

1. Farid, H.: A survey of image forgery detection. IEEE Signal Processing Magazine 2, 16–25 (2009)
2. Popescu, A.C., Farid, H.: Exposing digital forgeries by detecting duplicated image regions. Dartmouth College Tech. Rep., TR2004-515 (2004)
3. Fridrich, J., Soukal, D., Lukáš, J.: Detection of Copy-Move Forgery in Digital Images. In: Proceedings of Digital Forensic Research Workshop (2003)

4. Hsu, Y.F., Chang, S.F.: Image splicing detection using camera response function consistency and automatic segmentation. In: IEEE International Conference on Multimedia and Expo, pp. 28–31 (2007)
5. Luka, J., Fridrich, J., Goljan, M.: Detecting digital image forgeries using sensor pattern noise. In: Proceedings of SPIE, vol. 6072, pp. 362–372 (2006)
6. Popescu, A.C., Farid, H.: Exposing digital forgeries by detecting traces of resampling. IEEE Transactions on Signal Processing 53, 758–767 (2005)
7. Popescu, A.C., Farid, H.: Exposing digital forgeries in color filter array interpolated images. IEEE Transactions on Signal Processing 53, 3948–3959 (2005)
8. He, J., Lin, Z., Wang, L., Tang, X.: Detecting doctored JPEG images via DCT coefficient analysis. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 423–435. Springer, Heidelberg (2006)
9. Farid, H.: Exposing digital forgeries from JPEG ghosts. IEEE Transactions on Information Forensics and Security 4, 154–160 (2009)
10. Johnson, M.K., Farid, H.: Exposing digital forgeries by detecting inconsistencies in lighting. In: The 7th Workshop on Multimedia and Security, pp. 1–10 (2005)
11. Johnson, M.K., Farid, H.: Exposing digital forgeries in complex lighting environments. IEEE Transactions on Information Forensics and Security 2, 450–461 (2007)
12. Gallagher, A.C., Chen, T.: Image authentication by detecting traces of demosaicing. In: CVPRW 2008 (2008)
13. Dirik, A.E., Memon, N.: Image tamper detection based on demosaicing artifacts. In: ICIP 2009, pp. 1497–1500 (2009)
14. Kirchner, M.: Efficient Estimation of CFA Pattern Configuration in Digital Camera Images. In: Proceedings of SPIE, the International Society for Optical Engineering (2010)
15. Kirchner, M.: Columbia photographic images and photorealistic computer graphics dataset. Columbia University, ADVENT Technical Report, 205–2004 (2004)

Robust Signal Generation and Analysis of Rat Embryonic Heart Rate in Vitro Using Laplacian Eigenmaps and Empirical Mode Decomposition

Muhammad Khalid Khan Niazi[1], Muhammad Talal Ibrahim[2],
Mats F. Nilsson[3], Anna-Carin Sköld[4], Ling Guan[2], and Ingela Nyström[1]

[1] Centre for Image Analysis, Uppsala University, Sweden
[2] Ryerson Multimedia Research Lab, Ryerson University, Toronto, Canada
[3] Department of Pharmaceuticals Biosciences, Drug Safety and Toxicology, Uppsala University, Sweden
[4] AstraZeneca R&D Södertälje, Safety Assessment, Sweden
{khalid,ingela}@cb.uu.se, muhammadtalal.ibrahi@ryerson.ca,
mats.nilsson@farmbio.uu.se, anna-carin.skold@astrazeneca.com,
lguan@ee.ryerson.ca

Abstract. To develop an accurate and suitable method for measuring the embryonic heart rate *in vitro*, a system combining Laplacian eigenmaps and empirical mode decomposition has been proposed. The proposed method assess the heart activity in two steps; **signal generation** and **heart signal analysis**. Signal generation is achieved by Laplacian eigenmaps (LEM) in conjunction with correlation co-efficient, while the signal analysis of the heart motion has been performed by the modified empirical mode decomposition (EMD). LEM helps to find the template for the atrium and the ventricle respectively, whereas EMD helps to find the non-linear trend term without defining any regression model. The proposed method also removes the motion artifacts produced due to the the non-rigid deformation in the shape of the embryo, the noise induced during the data acquisition, and the higher order harmonics. To check the authenticity of the proposed method, 151 videos have been investigated. Experimental results demonstrate the superiority of the proposed method in comparison to three recent methods.

1 Introduction

Rat whole embryo culture on gestation day (GD) 13 can be used to investigate if drugs may have an effect on the embryonic heart *in vitro*. If a drug cause reduced heart rate (bradycardia) and/or irregular heart rate (arrhythmia), it may have the potential to damage the embryo by causing periods of hypoxia (reduced oxygen tension). There is a clear relationship between hypoxia during mammalian development and an increased risk of birth defects and embryonic death [1]. For instance, in pregnant women a ∼20% reduction in fetal heart rate during the first trimester is associated with a markedly increased risk of spontaneous abortion [2].

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 523–530, 2011.

The *in vitro* method of investigating drug effects on the rat embryonic heart has improved over the years. However, it is just recently that feasible image analysis tools have been introduced to assess the effect on embryonic heart rate *in vitro* [4,6]. In comparison, a lot more work has been done to set up image analysis tools in the zebrafish *in vivo* culture system. For instance, in [7], the changes in average light intensity of each frame was combined with pixel differential computed across the frames to construct the heart signal. In [8], the heart signal was visualised as a waveform of dynamic pixels produced by the oscillatory movement of blood cells. Later on, short-time Fourier transform (STFT) was used to analyse the non-stationary heart signal. In STFT, it is assumed that some portion of a non-stationary signal is stationary. However, the determination of the stationary portion and its size is a dilemma in itself. In [9], fast differential interference contrast imaging carried out at 250 frames/sec was combined with autocorrelation to measure the heart activity. To have a true representation of the heart activity, the reference image should represent either of the extreme states in the heart. However, the selection of reference image, manual or automatic, is un-conclusive in [9].

In the current study, rat embryos are cultured with an open yolk sac. This, together with the vigorous heart beats at GD 13, results in a non-rigid deformation in the shape of the embryo. It can also result in translation of the embryo in the culture medium. In the case of an irregular heart rate, the motion of the heart can become non-stationary in nature.

To cope with the above mentioned issues, we carried out the assessment of the heart activity in two steps, **signal generation** and **heart signal analysis**. Here, signal generation is achieved by Laplacian eigenmaps (LEM) in conjunction with correlation co-efficient, while the signal analysis of the heart motion has been performed by empirical mode decomposition (EMD).

2 The Proposed Method

Existing methods to assess embryonic heart rate requires a skillful operator to perform the experiment by means of expensive imaging equipment [9,5,7]. The aim of this study is to present a simple and low cost solution to measure the rat embryonic heart rate *in vitro*. The proposed solution requires an ordinary light microscope with a video camera capable of capturing a video at 30 frames per second along with a desktop computer. The section to follow will present a robust method to generate a heart signal from cultured rat embryos.

2.1 Signal Generation

Laplacian eigenmaps (LEM) is a non-linear dimensionality reduction method which aims to find the lower dimensional manifold embedded in the higher dimensional space while preserving the spatial relationship [10]. To accomplish this task, LEM constructs a graph G in which every data point d_i is connected to its k-nearest neighbours. All edges between the connected data points in a graph G have a cost equal to one. It is followed by the construction of an adjacency matrix C_{ij} which have an entry **1** at location (i, j), if the data point d_i is among the k-nearest neighbours of d_j. The rest of the locations in C_{ij} are set to zero. LEM requires the construction of Laplacian matrix L, which can be computed as:

$$L_{ij} = \begin{cases} D_{ij} & \text{if } i = j \\ -C_{ij} & \text{if } d_i \text{ and } d_j \text{ are adjacent data points} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, D is a diagonal matrix computed as $D_{ij} = \sum_j C_{ij}$. The final step in the LEM method is to find the generalized eigenvector solution to:

$$L\mathbf{f} = \lambda D\mathbf{f}. \tag{2}$$

As L is a symmetric positive semidefinite matrix with $\lambda_0 = 0$ as a trivial eigenvalue, it implies that all eigenvalues can be ordered as: $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \cdots \leq \lambda_{N-1}$. Now, the mapping of d_i in a low-dimensional space m is computed by leaving the eigenvector $\mathbf{f}_0$ and using the next m eigenvectors as [10]:

$$d_i \longrightarrow (\mathbf{f}_1(i), \mathbf{f}_2(i), \cdots, \mathbf{f}_m(i)). \tag{3}$$

Here, the mapping $\mathbf{f}$ is the solution to the following minimization problem [10]:

$$\mathbf{f} = \arg \min_{\mathbf{f}} \sum_{(i,j)} C_{ij} \|\mathbf{f}(d_i) - \mathbf{f}(d_j)\|^2 \tag{4}$$

As, we are interested in measuring the heart rate in the atrium and the ventricle, respectively, the method requires manual selection of these regions. In our case, the area representing either chamber would be treated as d_i, as a vector in a higher dimensional space. It is intuitive that LEM on such data will provide quite logical representation by mapping the filled, semi-filled, emptied atrium frames close to filled, semi-filled, emptied atrium frames, respectively in the lower dimensional space. The same would be the case with the ventricle. K-medoids clustering of this low-dimensional mapping into three clusters yields intuitive classification, i.e., filled, semi-filled, emptied chamber classes. This mapping eases the selection of the template frame for the atrium and the ventricle. It is worth mentioning that each low-dimensional point has a one-to-one correspondence with the higher dimensional point. The class having the least average value (computed among the corresponding higher dimensional points) is considered as the state when the specific heart chamber is fully filled. It is mainly because the least light would pass through the blood in the case when the chamber is fully filled. The point in the higher dimensional space is considered as template (T) if the corresponding lower dimensional point represents the centre of this chamber class. The template selected using this method is relatively insensitive to outliers and noise [10]. To construct the adjacency matrix, we have used 5-nearest neighbours and for the low-dimensional embedding we have set $m = 2$. The choice of 5-nearest neighbours is motivated by the fact that it always resulted in one-connected component in the graph for our test videos.

Later on, cross-correlation of the template frame (representing the atrium filled class) with the area representing the atrium will yield the heart signal representing the atrium activity (A_s). The ventricle activity (V_s) is achieved similarly by using the template frame (representing the ventricle filled class) with the area representing the ventricle. For the sake of simplicity, we will use the term *heart signal* to refer to both, (A_s) and (V_s).

2.2 Heart Signal Analysis

Depending on the heart conditions, the associated signals are intrinsically non-linear and most of the time non-stationary in nature. It is also observed that the spatial movement of the embryo due to placement in a culture medium (liquid in nature) results in a trend term in the associated heart signal. In case of arrhythmia, the heart motion will change over time which demands for time-frequency analysis tool. Generally, wavelets and filter banks provide a pre-defined multi-scale representation of a non-stationary signal. The application of wavelets require the definition of mother wavelet as well as the total number of scales that are both user defined parameters. However, it will be more logical to select these parameters based on the frequency content of the underlying signal but they are a priori unknown. The recently developed EMD seems suitable, as it overcomes the above mentioned shortcomings.

EMD is an adaptive multi-scale representation that decomposes a non-linear and non-stationary signal into symmetrical oscillating functions known as intrinsic mode functions (IMF) and a less oscillating local mean [11,12]. Each IMF has the same number of extrema and zero-crossings or can differ at most by one. IMF should also be symmetric with respect to local zero mean. To extract each IMF, EMD uses an iterative procedure known as sifting process [11]. In [13], the sifting process was replaced with partial differential equations to improve the performance of EMD. Another approach to improve the performance of EMD is to better approximate the mean envelope as the conventional mean envelope can result in undershoot, overshoot, instabilities to noise, and erroneous detection of extrema [14].

To overcome the above mentioned problems, we have followed the same lines as discussed in [15] and [11] along with a new relaxation in the stopping criterion which helps in faster convergence. The proposed method (Algorithm 1) provides the steps to remove the trend term and noise from the heart signal.

Algorithm 1. Empirical Mode Decomposition for Heart Signal Analysis

```
$f_r \longleftarrow$ frame rate of the camera
$n_f \longleftarrow$ total number of frames
$f_h \longleftarrow 6\,Hz$ (Maximum heart beat rate)
$i \longleftarrow -1$, $\varepsilon_1 \longleftarrow 0.05$, $\varepsilon_2 \longleftarrow 0.10$, $\alpha \longleftarrow 2$
repeat
    Find all extrema locations ($t_k$) in $z(t)$.
    Compute the centroid ($\bar{z}_k(\bar{t}_k)$) between two consecutive extremum as:
        $\bar{z}_k \longleftarrow \frac{1}{t_{k+1} - t_k} \int_{t_k}^{t_{k+1}} z(t)dt$
        $\bar{t}_k \longleftarrow \frac{\int_{t_k}^{t_{k+1}} t \mid z(t) - \bar{z}_k \mid^2 dt}{\int_{t_k}^{t_{k+1}} \mid z(t) - \bar{z}_k \mid^2 dt}$
    Find a mean envelope $m(t)$ by fitting a cubic spline through all centroids $(\bar{z}_k(\bar{t}_k))$.
    if $(z(t) - m(t))$ is an $imf$ then
        $i \longleftarrow i + 1$
        Find the IMF, $h_i(t)$:
            $h_i(t) \longleftarrow z(t) - m(t)$
        Find the residual, $r(t)$:
            $r(t) \longleftarrow z(t) - h_i(t)$
            $z(t) \longleftarrow r(t)$
        Find all extrema locations $t_k$ in $h_i(t)$.
        Find the symmetry ratio, $s_r$:
            $\lambda_i(k) \longleftarrow \frac{h_i(t_{k+1}) - h_i(t_{k-1})}{t_{k+1} - t_{k-1}}(t_k - t_{k-1}) + h_i(t_{k-1})$
            $s_r \longleftarrow \frac{\mid h_i(t_k) + \lambda_i(k) \mid}{\mid h_i(t_k) - \lambda_i(k) \mid} \leq \varepsilon_1$
    else
            $z(t) \leftarrow z(t) - m(t)$
        Find all extrema locations $t_k$ in $z(t)$.
        Find the symmetry ratio $s_r$ as given below:
            $\lambda(k) \longleftarrow \frac{z(t_{k+1}) - z(t_{k-1})}{t_{k+1} - t_{k-1}}(t_k - t_{k-1}) + z(t_{k-1})$
            $s_r \longleftarrow \frac{\mid z(t_k) + \lambda(k) \mid}{\mid z(t_k) - \lambda(k) \mid} \leq \varepsilon_1$
    end if
    Find the number of extrema $n_e$ in $z(t)$.
until $((s_r \geq \varepsilon_1) \wedge (n_e > \alpha))$
Find the cut-off frequency $cut_{off}$ to remove noise and higher harmonics:
    $cut_{off} \longleftarrow \frac{2f_h}{f_r} + \varepsilon_2$
Convolve $h_0$ with a FIR low-pass filter $fil$, with a cut-off frequency of $cut_{off}$:
    $h_0(t) \longleftarrow \sum_k h_0(k) fil(t - k)$
Drop the trend term to compensate for the spatial heart motion in the culture medium:
    $z_{comp}(t) \longleftarrow \sum_{i=0}^{n} h_i(t)$
```

Here, the sifting process enables EMD in separating intrinsic modes of oscillatory components with their frequency ratio up to 0.8, thus greatly mitigating the effect of mode mixing and enhancing the frequency resolving power [14]. The proposed method helps to find the non-linear trend term without defining the regression model. The non-rigid shape deformation produced due to the vigorous heart motion, the noise induced during the data acquisition, and the higher order harmonics (produced due to unequal stay of blood in each of the heart chambers) are catered during step 22 to 24 of Algorithm 1. Taking benefit from the zero mean property of the IMF, we have defined a beat in a different way. We count it as a beat if there is a negative minimum between two positive maxima. Here, maxima is defined as the highest positive value between every two negative minima. The positive minimum separating the two positive maxima is consider as false minima and the negative maxima separating the two minima is considered as false maxima. This methodology avoids false maxima/minima detection and also avoids the computation of differential for maxima/minima detection. Such a definition assures that every beat has substantial amplitude and also justifies the removal of the trend term.

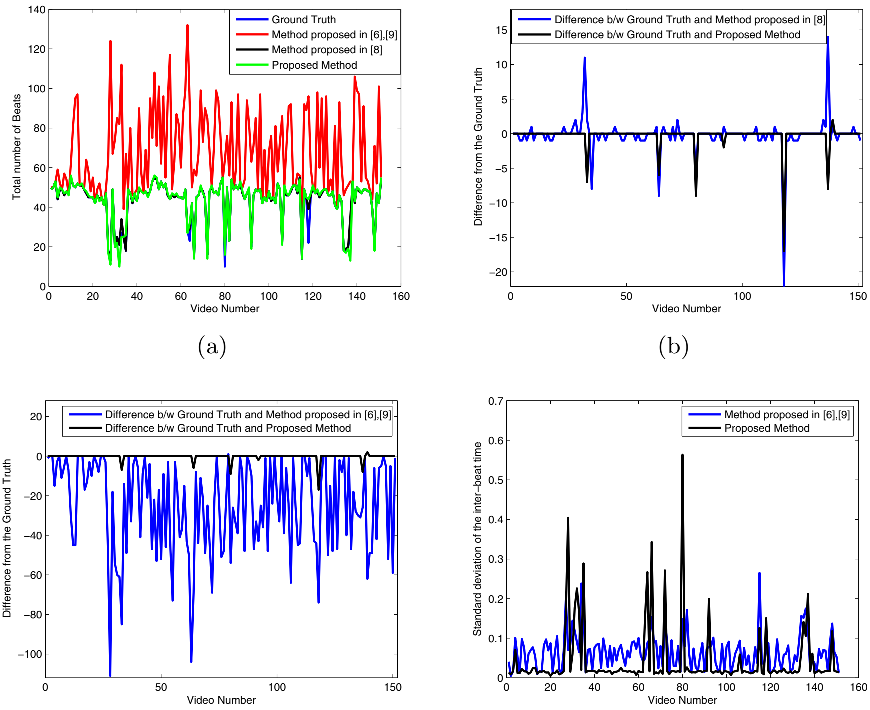
3 Experimental Results and Discussion

For evaluation of our proposed method, we have used 151 videos from an ongoing project where the effects of pharmaceutical drugs with ion channel-blocking activity on the heart were investigated. To generate the videos, we used a culture system (with some modification) previously published in [3]. In the current study each embryo was cultured in 25 ml bottles containing 4 ml of Dulbeccos Modified Eagles Medium (DMEM, Ref. No. D1145, Sigma Chemical Co., St. Louis, MO, USA). Intermittent gassing (95% O_2, 5% CO_2) for 2 minutes instead of continuous gassing was used, and the bottles were gassed after addition of the drugs to the culture media, and after every recording of the heart rate.

After 1 hour of incubation, each embryo still in its bottle, was examined under a light microscope (Olympus SZ-40) equipped with a camera (uEye UI-2210-M/C). A 30-second video of the embryos was recorded to be used for later analysis. Damaged and dead embryos or embryos with a heart rate less than 160 beats per minute were discarded. After recording the embryonic heart rate, the test compound (or vehicle serving as control) was added to the culture medium and the bottles were gassed for 2 minutes with the same gas mixture as above. The embryos were then incubated for 1 hour before being re-examined again as described above. In total, we have used 151 videos for the evaluation of our proposed method.

We opted to compare our heart signal analysis method with the analysis methods mentioned in [6,8,9]. We skipped the comparison between the proposed signal generation method and the signal generation methods mentioned in [8,9], because they have used different imaging technique to acquire images. However, it seems fair to compare the signal analysis methods as there is no difference in the generated signals. For validation of our method, we have generated a ground truth (G_t) where videos were stepped through frame by frame and the exact heart rate was determined visually. Fig. 1(a) shows the number of heart beats counted by each of these methods in 151 different videos. It is clear from Fig. 1(a) that the proposed method is quite accurate in counting the heart beats in comparison to the others.

Fig. 1(b) shows the difference between the number of G_t beats and the number of beats counted by the proposed method for all 151 videos. It also shows the difference between G_t beats and the beats counted by [8]. Results in Fig. 1(b) demonstrate that the difference between the total number of heart beats and the G_t is quite small as compared to [8]. Similar results are evident in Fig. 1(c) where the difference between the G_t and the number of heart beats counted by [6,9] is quite higher. It is interesting to mention that the method in [8] provides quite accurate count of the heart beats by finding the fundamental frequency in the frequency domain but fails to localize the location of the peaks in the time domain. Fig. 1(d) shows the standard deviation of the inter-beat time which is often used by biologists to judge the heart condition. Here, the higher values are often associated with the irregularity of the heart motion. It is evident from the results that the proposed method outperforms the other methods but still there are 5 instances when the proposed method fails to correctly count the heart beats.



(a)

(b)

(c)

(d)

Fig. 1. a) Number of heart beats counted by different methods on 151 videos. b) Difference in the number of the heart beats between the ground truth and the other methods. c) Similar to (b) but using the method proposed in [6,9]. d) Standard deviation of the inter-beat time computed individually for every video.

4 Conclusion

The proposed method selects the template for each of the heart chambers using Laplacian eigenmaps which is robust against noise and outliers. The results clearly show the accuracy of the proposed method in counting and localizing different heart states. The efficiency and the robustness of the proposed method makes it more attractive for the biologists to analyse the heart motion without the need for parameter tweaking and expensive imaging equipments.

References

1. Webster, W.S., et al.: The effect of hypoxia in development. Birth Defects Res. C. Embryo Today 81, 215–228 (2007)
2. Doubilet, P.M., et al.: Outcome of first-trimester pregnancies with slow embryonic heart rate at 6-7 weeks gestation and normal heart rate by 8 weeks at US. Radiology 236, 636–643 (2005)
3. Abela, D., et al.: The effects of drugs with ion channel-blocking activity on the early embryonic rat heart. Birth Defects Res., Part B 89, 429–440 (2010)
4. Sköld, A.C., et al.: Teratogenicity of the IKr-Blocker Cisapride: Relation to Embryonic Cardiac Arrhythmia. Reprod. Toxicol. 16(4), 333–342 (2002)
5. Robkin, M.A., et al.: A new in vitro culture technique for rat embryos. Teratology 5, 367–376 (1972)
6. Khan, M., et al.: Fully automatic heart beat rate determination in digital video recordings of rat embryos. In: Perner, P., Salvetti, O. (eds.) MDA 2008. LNCS (LNAI), vol. 5108, pp. 27–37. Springer, Heidelberg (2008)
7. Fink, M., et al.: A new method for detection and quantification of heartbeat parameters in drosophila, zebrafish, and embryonic mouse hearts. BioTechniques 46, 101–113 (2009)
8. Chan, P.K., et al.: Noninvasive technique for measurement of heartbeat regularity in zebrafish (Danio rerio) embryos. BMC Biotechnol 9, 1–10 (2009)
9. Zhu, J.T., et al.: Fast differential interference contrast imaging combined with autocorrelation treatments to measure the heart rate of embryonic fish. J. Biomed. Opt. 13(2) (2008)
10. Belkin, M., et al.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15, 1373–1396 (2003)
11. Oberlin, T., et al.: An Alternative Formulation for the Empirical Mode Decomposition. Tech. Rep. hal-00553107, HAL (January 2011)
12. Huang, N.E., et al.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In: Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, vol. 454, pp. 903–995 (March 1998)
13. Hadji, S.E., et al.: Analysis of Intrinsic Mode Functions: A PDE Approach. IEEE Signal Processing Letters 17, 398–401 (2010)
14. Hong, H., et al.: Centroid-based sifting for empirical mode decomposition. Journal of Zhejiang University - Science C 12, 88–95 (2011)
15. Hong, H., et al.: Local integral mean-based sifting for empirical mode decomposition. IEEE Signal Processing Letters 16, 841–844 (2009)

Radial Symmetry Guided Particle Filter for Robust Iris Tracking

Francis Martinez, Andrea Carbone, and Edwige Pissaloux

Université Pierre et Marie Curie (UPMC)
CNRS, UMR 7222, Institut des Systèmes Intelligents et de Robotique (ISIR)
4 place Jussieu, 75005 Paris, France
martinez@isir.upmc.fr

Abstract. While pupil tracking under active infrared illumination is now relatively well-established, current iris tracking algorithms often fail due to several non-ideal conditions. In this paper, we present a novel approach for tracking the iris. We introduce a radial symmetry detector into the proposal distribution to guide the particles towards the high probability region. Experimental results demonstrate the ability of the proposed particle filter to robustly track the iris in challenging conditions, such as complex dynamics. Compared to some previous methods, our iris tracker is also able to automatically recover from failure.

Keywords: iris tracking, particle filter, radial symmetry.

1 Introduction

Visual gaze tracking is of particular interest in many applications [1], ranging from human-computer interactions to cognitive studies. A first step to estimate the direction of the gaze often relies on tracking eye features such as the pupil or the iris. Then, the point-of-regard can be computed thanks to a mapping between the eye and the environment, usually a planar surface. We refer to [2] for an extensive survey on eye tracking and gaze estimation.
In this paper, we address the problem of tracking the iris in close-up images with a low-cost camera and under uncontrolled conditions. In comparison with pupil tracking [3], iris tracking remains still a challenging task due to *eye-* (motion, shape variation, eyelid/eyelash occlusion, pupillary dilation, ethnicity), *camera-* (focus, resolution) and *environment-* (light variation, passive illumination, external occlusion) dependent factors. It can be categorized into two separate classes: *non-probabilistic* [4][5][6][7] and *probabilistic* [8][9] approaches.

Non-probabilistic approaches. They are mostly improvements of the *Starburst* algorithm proposed in [3] originally designed for pupil tracking. These methods rely on feature detection followed by a RANSAC [10] ellipse fitting. Extensions of the algorithm include : a distance filter and constraints about directions of rays [4], constraints about size of iris and number of inliers [5], a limbus-pupil switching mechanism using an adaptive threshold to account for

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 531–539, 2011.

light variations [6] and a constrained search of strong gradients along normal lines under the implicit assumption of smooth movements [5][7].

Probabilistic approaches. Hansen *et al.* [8] was the first to propose an iris tracking algorithm based on particle filter. The contour log-likelihood ratio is modelled thanks to a Generalized Laplacian to approximate the distribution of gray-level differences and a Gaussian distribution to consider the deformations along the measurement lines [11]. In [9], Wu *et al.* present an iris and eyelids tracking method based on particle filter and a 3D eye model. An intensity cue along with edge computation is employed to update the weights of the particles.

Relying solely on feature detection and ellipse fitting is often prone to failure, even if strong constraints are applied in both stages. Moreover, the iris motion is constrained by the eyeball rotation and hence, a bounded state space could be accordingly defined. However, in practice, such a strategy requires a large number of particles to sample from. Iris distortions also become more important in close-up images and should be considered in the algorithm. In this work, a more general framework is provided to, at the same time, guide and constrain the feature detection and the ellipse fitting. Our contribution is to combine an iris-based detector, namely *radial symmetry*, and a *Sequential Monte Carlo* approach to robustify iris tracking and improve state-of-the-art approaches.

2 Radial Symmetry Transform

The radial symmetry transform used in our work is the one proposed by Loy *et al.* [12]. We briefly review the algorithm and for more details, the reader is referred to [12]. The idea behind the radial symmetry transform is to accumulate orientation and magnitude contributions at different radii from a position $\mathbf{p} = \{x, y\}$ in the direction of the gradient:

$$\mathbf{p}_{\pm} = \mathbf{p} \pm round\left(\frac{\mathbf{g}(\mathbf{p})}{||\mathbf{g}(\mathbf{p})||} r\right) \tag{1}$$

where $r \in N_r$ is the radius with N_r being the discrete set of radii and $\mathbf{g}(\cdot)$ is the gradient computed thanks to the 3x3 Sobel operator. At each step, orientation and magnitude projection images are updated depending on positively- or negatively-affected pixels:

$$O_r(\mathbf{p}_{\pm}) = O_r(\mathbf{p}_{\pm}) \pm 1 \qquad M_r(\mathbf{p}_{\pm}) = M_r(\mathbf{p}_{\pm}) \pm ||\mathbf{g}(\mathbf{p})|| \tag{2}$$

For each radius r, the radial symmetry transform then becomes:

$$S_r = F_r * A_r \quad with \quad F_r = \frac{M_r}{k_r}\left(\frac{|\tilde{O}_r|}{k_r}\right)^{\alpha} \tag{3}$$

where A_r is Gaussian filter that allows spreading the symmetry contribution, α is the radial-strictness parameter and k_r is a normalizing factor so that M_r and O_r can be represented on a similar scale. $\tilde{O}_r(\mathbf{p})$ and k_r are given by:

$$\tilde{O}_r(\mathbf{p}) = \begin{cases} O_r(\mathbf{p}) \text{ if } O_r(\mathbf{p}) < k_r \\ k_r \quad \text{else} \end{cases} \quad with \quad k_r = \begin{cases} 9.9 \text{ if } r > 1 \\ 8 \quad \text{else} \end{cases} \tag{4}$$

Finally, the symmetry contributions can be averaged over the set of radii:

$$\tilde{S} = \frac{1}{|N_r|} \sum_{r \in N_r} S_r \tag{5}$$

3 Proposed Iris Tracker

Particle filter. Let $\mathbf{x}_t \in \mathcal{X}$ denote the hidden state and $\mathbf{z}_t \in \mathcal{Z}$ the observation at time step t. The principle of the particle filter is to infer the marginal posterior distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ of the state $\mathbf{x}_t$ given the observation sequence $\mathbf{z}_{1:t}$. Given Bayes' theorem and the Chapman-Kolmogorov equation, it can be expressed by the following Bayesian recursive formula under a first-order Markovian assumption:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{z}_{1:t}) &= \tfrac{p(\mathbf{z}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{z}_{1:t-1})}{p(\mathbf{z}_t|\mathbf{z}_{1:t-1})} \\ &\propto p(\mathbf{z}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1} \end{aligned} \tag{6}$$

However, in practice, because the integral is intractable, this distribution is approximated by a set of weighted samples $\{\mathbf{x}_t^{(n)}, w_t^{(n)}\}_{n=1}^N$:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \approx \sum_{n=1}^{N} w_t^{(n)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}) \tag{7}$$

where $\delta(\cdot)$ denotes the Dirac function and the unormalized weights are sequentially updated according to:

$$\tilde{w}_t^{(n)} = w_{t-1}^{(n)} \frac{p(\mathbf{z}_t|\mathbf{x}_t^{(n)}) p(\mathbf{x}_t^{(n)}|\mathbf{x}_{t-1}^{(n)})}{q(\mathbf{x}_t^{(n)}|\mathbf{x}_{0:t-1}^{(n)}, \mathbf{z}_{1:t})} \tag{8}$$

Typically, the proposal distribution is set equal to the prior transition leading to a simplified update scheme known as the Bootstrap filter or Condensation algorithm [13]: $\tilde{w}_t^{(n)} = w_{t-1}^{(n)} \, p(\mathbf{z}_t|\mathbf{x}_t^{(n)})$.

State space. The iris shape is represented by an ellipse : $\mathbf{x} = [\, c_x \; c_y \; a \; b \; \theta]^T \in \mathbb{R}_{5\text{x}1}$ where (c_x, c_y) are the center coordinates, (a, b) the semi-minor/major axis and θ the angle of the ellipse with respect to the horizontal axis.

Dynamic model. The idea of including a detector within the particle filter to boost the tracking, is not new and was proven to be very effective [14]. Furthermore, radial symmetry transform is well-suited to detect pupil and aid to locate iris in non-ideal conditions [15]. It relies only on the boundaries which makes it suitable to ignore specular reflections (often located inside or outside the iris). In our work, we apply the radial symmetry transform at a low resolution on a restricted set of radii based upon the iris size at $t-1$. Moreover, because the iris is darker than its surrounding, only negatively-affected pixels $\mathbf{p}_{-}$ resulting from strong gradients are considered. To integrate the radial symmetry detection, the proposal distribution is modelled by a mixture of Gaussian distributions:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) = \alpha \, q_{obs}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t) + (1-\alpha) \, p(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{9}$$

where α is the mixture coefficient. If $\alpha = 0$, the proposal distribution equals the prior transition *i.e.* the particle filter is only driven by a random walk. The

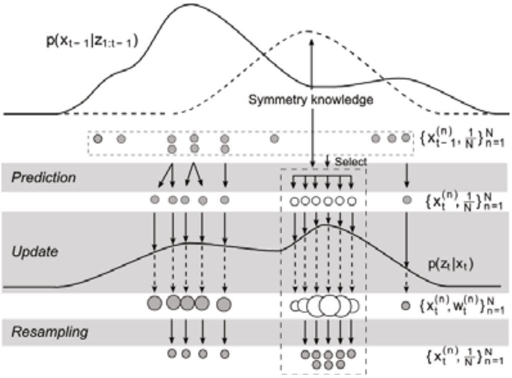


Fig. 1. Radial symmetry guided particle filter. The gray particles are generated according to $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ while the white particles are propagated by $q_{obs}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t)$.

advantage of the proposed state evolution is that it makes use of the current observation thanks to the strong detector in order to generate particles in the region of high probability and allows recovering from failure. Particles are generated thanks to:

$$\begin{aligned}\mathbf{x}_t^{(n)} &\sim q_{obs}(\mathbf{x}_t|\mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_t^d(\mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t), \mathbf{\Sigma}_1)\\ \mathbf{x}_t^{(n)} &\sim p(\mathbf{x}_t|\mathbf{x}_{t-1}^{(n)}) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_{t-1}^{(n)}, \mathbf{\Sigma}_2)\end{aligned} \tag{10}$$

where $\mathbf{x}_t^d(\mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t)$ is the new proposed state obtained by radial symmetry transform and $\mathbf{\Sigma}_k = diag\{\sigma^2_{c_x,k}, \sigma^2_{c_y,k}, \sigma^2_{a,k}, \sigma^2_{b,k}, \sigma^2_{\theta,k}\}, k \in \{1,2\}$. $\mathbf{\Sigma}_1$ favors iris deformations and $\mathbf{\Sigma}_2$ puts more emphasize on translational movements. The principle of the radial symmetry guided particle filter is depicted in Fig.1. Defining $\mathbf{x}_t^d(\mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t)$ is not straightforward as compared to [14]. The radial symmetry detector only provides an estimate of the center. For the remaining parameters of the state model, two main strategies are possible : either the shape and orientation of (i) each particle or (ii) the estimated state at $t-1$ are kept. These strategies have a non-negligible impact on tracking : the latter one allows much more freedom in deformation than considering each particle independently. Based on this observation, we intuitively chose the first approach because the iris keeps the same size, even if it is occluded. This intuition was then confirmed experimentally.

Likelihood estimation. As a trade-off between tracking robustness and computational cost, we use a simple yet effective likelihood model based on contour information. Under conditional independence assumption, we define the joint measurement density by:

$$p(\mathbf{z}_t|\mathbf{x}_t^{(n)}) = \prod_{u \in \Omega} p(\mathbf{z}_t(u)|\mathbf{x}_t^{(n)}) \tag{11}$$

with Ω a discrete set of N_{ML} measurement locations of length L and:

$$p(\mathbf{z}_t(u)|\mathbf{x}_t^{(n)}) \propto |\nabla I_p(\nu_m(u))|^2 \exp(-\frac{||\nu_m(u) - \nu_o(u)||^2}{2\sigma_c^2}) \tag{12}$$

Algorithm 1. `Proposed iris tracker`

Input - Set of weighted samples $\{\mathbf{x}_{t-1}^{(n)}, w_{t-1}^{(n)}\}_{n=1}^{N}$

Resampling : Duplicate high-weighted particles and eliminate low-weighted particles to form a new set $\{\mathbf{x}_{t-1}^{(n)}, \frac{1}{N}\}_{n=1}^{N}$.

Prediction : $\{\mathbf{x}_t\} = \{\hat{\mathbf{x}}_t\} \cup \{\hat{\mathbf{x}}_t^d\}$

1. Randomly select $N_1 = \alpha N$ particles from $\{\mathbf{x}_t^d\}$ defined by :
   $$\mathbf{x}_t^d = [\, c_x^d(\mathbf{z}_t)\ c_y^d(\mathbf{z}_t)\ a_{t-1}^{(n)}\ b_{t-1}^{(n)}\ \theta_{t-1}^{(n)}\,]^T$$
   where $(c_x^d(\mathbf{z}_t), c_y^d(\mathbf{z}_t))$ are the center coordinates obtained by the radial symmetry detector. Then, generate $\{\hat{\mathbf{x}}_t^d\}$ according to (10).
2. Randomly select $N_2 = (1-\alpha)N$ particles from $\{\mathbf{x}_{t-1}\}$ and generate $\{\hat{\mathbf{x}}_t\}$ based on (10).

Update : Evaluate the weights according to the likelihood (11) and normalize them: $w_t^{(n)} = \frac{\tilde{w}_t^{(n)}}{\sum_{i=1}^{N} \tilde{w}_t^{(i)}}$.

Output - Estimated state given by (15).

where ν_o and ν_m denote respectively the reference and the detected feature point on the normal line to a hypothesised contour. The gradient magnitude projected on the normal line at ν_m is also incorporated in the likelihood to provide higher weights to boundaries having larger magnitudes. [8] instead assumes a homogeneous intensity distribution inside and outside the object through the Generalized Laplacian.

Confidence measure. In some cases, the detector is not reliable and is attracted by spurious distractors. However, the likelihood evaluation sometimes alleviates the problem of false positives. To further mitigate tracking failure, we introduce a confidence measure acting on the guidance strategy:

$$conf(\tilde{S}_t^m, d_t) = \begin{cases} 1 & \text{if} \quad \tilde{S}_t^m > S^{ref} \vee p_{\mathcal{N}}(\frac{\tilde{S}_t^m - S^{ref}}{S^{ref}}) p_{\mathcal{N}}(\frac{d_t}{M_{t-1}^{a,b}}) > \tau \\ 0 & \text{else} \end{cases} \tag{13}$$

where $p_{\mathcal{N}}(x) \triangleq \mathcal{N}(x; 0, \sigma^2)$, S^{ref} and $\tilde{S}_t^m$ are respectively the reference and the current maximum magnitude of the symmetry contribution given by (5), d_t is the distance between the current center detected by the radial symmetry transform and the estimated center at time step $t-1$, $M_{t-1}^{a,b} = max\{a_{t-1}^{MAP}, b_{t-1}^{MAP}\}$ and τ is a constant threshold set to 0,25. The underlying assumptions are that : (i) if the symmetry contribution is low, the output of the detector is not reliable and (ii) if, furthermore, the distance to the detector is high, the output of the detector is ignored. The product of the Gaussian distributions allows assuming an uncertainty about the high/low transition of the symmetry contribution. The hard decision given by the confidence measure serves then to modify the mixture coefficient α of the proposal distribution:

$$\alpha = \alpha_d \cdot conf(\tilde{S}_t^m, d_t) \tag{14}$$

with α_d a user-defined threshold reflecting the mixture strategy.

Table 1. Experimental results obtained for moderate and abrupt motion

	Moderate motion			Abrupt motion	
Subject	*A*	*B*	*C*	*D*	*E*
Nb. frames	763	524	854	244	351
Max amp. (pixels)	63,39	47,01	42,72	95,71	81,56
P_m	3,90 %	2,87 %	2,70 %	16,87 %	14,72 %
P_s	79,46 %	76,86 %	85,92 %	65,61 %	70,11 %
P_d	47,94 %	47,80 %	40,21 %	62,55 %	55,28 %

Local refinement. For visualization, the point estimate is given by the *maximum a posteriori* (MAP) state estimate:

$$\hat{\mathbf{x}}_t^{MAP} = \arg\underset{\mathbf{x}_t}{max}\ p(\mathbf{x}_t|\mathbf{z}_{1:t}) \approx \arg\underset{\mathbf{x}_t}{max} \sum_{n=1}^{N} w_t^{(n)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(n)}) \qquad (15)$$

which is refined to locally capture boundaries. Feature points are detected along normals to the ellipse and the best fitting ellipse is computed using a RANSAC method such as in [4]. Within the ellipse fitting step, the evaluation of the ellipse follows 2 criteria:

1. Similarity in size : $|a_t - a_{t-1}| < \delta_a$, $|b_t - b_{t-1}| < \delta_b$
2. Bounded size ratio : $R_{min} < \frac{a_t}{b_t} < R_{max}$

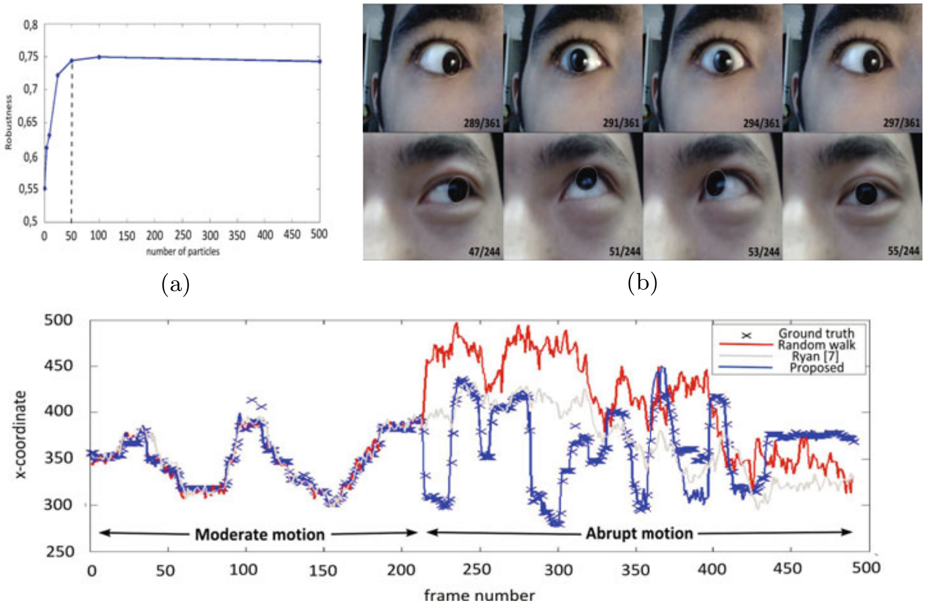
Tracking initialization. In order to initialize the tracking algorithm, the prior density $p(\mathbf{x}_0|\mathbf{z}_0) \triangleq p(\mathbf{x}_0)$ has to be defined. We represent it as a Dirac function centered at the position given by the non-probabilistic approach described in [4] with the seed point obtained by the radial symmetry transform.

4 Experimental Results

The algorithm was implemented in C (non-optimized) and tested on an Intel Core i5 CPU 2,27 GHz and runs in real-time at less than 40 fps with a resolution of 640x480. For visual assessment, the reader is invited to see video demos at: `http://people.isir.upmc.fr/martinez`.

Datasets. Unfortunately, there exists no publicly available database of iris videos. For this reason, experiments were carried on a self-made database consisting of 5 subjects (3 Europeans, 1 African and 1 Asian). Videos were captured with a webcam running at 30 fps and providing unfocused images. Hand labeled ground truths, c_x and c_y, were estimated for each video. Please note that annotation of noisy iris images introduces an error highly dependent on the iris size.

Performance measure. The robustness was experimentally evaluated instead of the accuracy (being similar to the other iris trackers). We define a robustness measure by the percentage of successful tracked frames : $P_s = \frac{N_s}{N_T}$ with N_S the number of successful tracked frames and N_T the total number of frames. A successfull track is determined by setting a threshold τ_s relatively to the average



(a)

(b)

(c)

Fig. 2. (a) Robustness as a function of the number of particles N, (b) Snapshots and corresponding point estimate for the dataset with abrupt motion (Notice the high appearance change between some adjacent frames), and (c) Comparison of x-coordinate trajectory of the iris center between state-of-the-art and proposed iris trackers

semi-minor/major axis. The percentage of abrupt motion, P_m, and the percentage of detector-generated $\hat{\mathbf{x}}_t^{MAP}$, P_d, are also indicated in the table.

Parameters setting. All parameters have been experimentally set, but most stayed unchanged for the whole dataset. N, N_{ML}, L and α_d were respectively set to : 50, 30, 20 and 0.5. The optimal number of particles was computed by averaging P_s over the 5 subjects and is given in Fig.2 (a).

Results. Table 1 shows the results obtained for the 5 subjects. Fig.2 (b) shows snapshots of the abrupt motion dataset. Although the videos exhibit large iris movements and significant blur, the proposed method is able to track the iris. According to P_d, one can notice that the more abrupt motion occurs, the more the MAP estimate is generated by the detector. The presented particle filter was also compared against other methods to evaluate its robustness : (i) the non-probabilistic tracker proposed in [7], (ii) our tracker *without* radial symmetry knowledge for $N = 100$ and $\mathbf{\Sigma}_2$ slightly increased (it actually has a behaviour similar to the one described in [8]) and (iii) our *full* tracker with radial symmetry knowledge. Results showed that all approaches perform well under smooth motion. However, state-of-the-art trackers are not able to track the iris when large-amplitude motion occurs. An example of such behaviour is illustrated in Fig.2 (c). The reason of track loss are that state-of-the-art methods do not incorporate specific knowledge in order to model the iris and especially rely on modelling and detecting the contour. On the contrary, the radial symmetry transfom provides two iris-specific information guiding the tracking: (i) a *close-to-radial* shape due to the contribution of convergent gradient directions and (ii) a *sharp* limbus by only keeping negatively-affected pixels. Furthermore, even if our tracker fails in some situations, such as eyelid occlusion, it still can quickly recover from failure which was never the case with the other methods. This last crucial point opens the door towards long-term and fully-automated tracking systems, not yet handled by state-of-the-art iris trackers.

5 Conclusion

The key idea presented in this paper is that iris tracking can be enhanced by embedding radial symmetry knowledge into the particle filter. Experimental results have shown the effectiveness of the proposed approach compared with state-of-the-art approaches to cope with complex dynamics and track loss. However, there is still room for improvement. Iris side-appearance can significantly affect the radial symmetry detector and should be better modelled to handle stronger elliptical shape and partial occlusions. Future works could be also to model the eye state (open/closed) to handle eyelid occlusion while tracking and to integrate a gaze estimation method to infer the point-of-regard.

Acknowledgements. This work was supported by the European Comission under the 7th FWP project AsTeRICS (Assistive Technology Rapid Integration & Construction Set, G.A.No. 247730).

References

1. Duchowski, A.T.: A breadth-first survey of eye-tracking applications. Behavior Research Methods, Instruments & Computers 34, 455–470 (2002)
2. Hansen, D.W., Ji, Q.: In the eye of the beholder: A survey of models for eyes and gaze. IEEE Trans. on PAMI 32, 478–500 (2010)
3. Li, D., Winfield, D., Parkhurst, D.J.: Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In: Proc. of CVPR, pp. 79–86 (2005)
4. Li, D., Parkhurst, D.: Open-source software for real-time visible-spectrum eye tracking. In: COGAIN, pp. 18–20 (2006)
5. Colombo, C., Comanducci, D., Del, B.: Robust iris localization and tracking based on constrained visual fitting. In: Proc. of ICIAP, pp. 454–460 (2007)
6. Ryan, W.J., Duchowski, A.T., Birchfield, S.T.: Limbus/pupil switching for wearable eye tracking under variable lighting conditions. In: Proc. of ETRA, pp. 61–64 (2008)
7. Ryan, W.J., Duchowski, A.T., Vincent, E.A., Battisto, D.: Match-moving for area-based analysis of eye movements in natural tasks. In: Proc. of ETRA, pp. 235–242 (2010)
8. Hansen, D.W., Pece, A.E.C.: Iris tracking with feature free contours. In: Proc. of AMFG, pp. 208–214 (2003)

9. Wu, H., Kitagawa, Y., Wada, T., Kato, T., Chen, Q.: Tracking iris contour with a 3D eye-model for gaze estimation. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part I. LNCS, vol. 4843, pp. 688–697. Springer, Heidelberg (2007)
10. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24, 381–395 (1981)
11. Pece, A.E.C., Worrall, A.D.: Tracking with the em contour algorithm. In: Proc. of ECCV, pp. 3–17 (2002)
12. Loy, G., Zelinsky, A.: Fast radial symmetry for detecting points of interest. IEEE Trans. on PAMI 25, 959–973 (2003)
13. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. IJCV 29, 5–28 (1998)
14. 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)
15. Zhang, W., Li, B., Ye, X., Zhuang, Z.: A robust algorithm for iris localization based on radial symmetry. In: Proc. of CISW, pp. 324–327 (2007)

Spatio-Temporal Stereo Disparity Integration

Sandino Morales and Reinhard Klette

The *.enpeda..* Project, The University of Auckland
Tamaki Innovation Campus, Auckland, New Zealand
pmor085@aucklanduni.ac.nz

Abstract. Using image sequences as input for vision-based algorithms allows the possibility of merging information from previous images into the analysis of the current image. In the context of video-based driver assistance systems, such temporal analysis can lead to the improvement of depth estimation of visible objects. This paper presents a Kalman filter-based approach that focuses on the reduction of uncertainty in disparity maps of image sequences. For each pixel in the current disparity map, we incorporate disparity data from neighbourhoods of corresponding pixels in the immediate previous and the current image frame. Similar approaches have been considered before that also use disparity information from previous images, but without complementing the analysis with data from neighbouring pixels.

Keywords: disparity map, Kalman filter, temporal propagation, stereo analysis, driver assistance.

1 Introduction

Disparity (depth) estimation obtained from stereo algorithms is commonly used to provide basic spatial data for complex vision-based applications [Zhihui 2008], such as in vision-based driver assistance [Franke et al. 2005]. Many applications require a very high accuracy of disparity values, often depending on the application context (e.g. accurate disparity discontinuities in driver assistance, and accurate disparities within object regions in 3D modelling).

The following three approaches have been followed to improve disparity estimation. First, the design of new or the improvement of existing strategies for stereo matchers (e.g. local, global, semi-global, or hierarchical). Second, improvements of cost functions (e.g., sum of absolute differences, census transform, or gradient-based cost functions) or smoothness constraints (e.g. Potts model, or truncated linear penalty) used within such a matching strategy. Third, some manipulation of input images (e.g. residuals with respect to some smoothing operator, or edge detection; see [Vaudrey and Klette 2009]) or some post-processing of the obtained results (e.g. consistency checks or mean filters; see [Atzpadin et al. 2004]).

In this work we discuss post-processing of the stereo analysis results (i.e. disparity maps) using available spatial and temporal information in the context

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 540–547, 2011.

of vision-based *driver assistance systems* (DAS). The input of such a system is a continuous stream of images, thus it is possible to use the information contained in the *temporal domain* by incorporating disparity values from previous frames into measured disparity values in the current frame. The intention is to reduce the influence of miscalculated disparities in the overall disparity map.

Information contained in the temporal domain has been used before. For example, in [Morales et al. 2009], an 'alpha-blending' of disparity values calculated at current and previous image frames lead to an improved performance for the currently measured values. This model did not yet consider the *ego-motion* induced by the *ego-vehicle* (i.e. the vehicle where the system is operating in; see Fig. 1, left), nor the independent motion of other objects in the scene. [Badino et al. 2007] merged previous and current information using an iconic (pixel-wise) Kalman-based approach [Matthies et al. 1989] and ego-motion information (yaw and speed rate). This approach was designed to improve disparity measurements in regions of input images where visible motion was only induced by ego-motion, and not by motion of other objects. The latter method was extended in [Vaudrey et al. 2008a] by adding a *disparity rate* term to the Kalman filter. The goal was to improve disparity measurements for objects that move relatively to the ego-vehicle in longitudinal direction.

Post processing of disparity maps can also aim at optimizing disparity values by considering disparity values within a neighbourhood of the current pixel (i.e. using data from the *spatial domain*). [Morales et al. 2009] used the same alpha-blending approach for modifying the disparity value at a pixel by using disparity values of the 'north' and 'south' neighbours. Again, no ego- or independent motion was taken into account. In this paper we improve disparity maps by reducing the uncertainty in calculated values using information from both, the spatial and the temporal domain. The Kalman filter used in [Badino et al. 2007, Vaudrey et al. 2008a] is now modified such that data from the spatial domain can also be used. Spatial information is used from the previous and the current disparity maps, and can be taken from an arbitrarily defined neighbourhood. (Actually, those should be not too large.) We perform experiments with a computer-generated sequence (i.e. with available ground truth) to

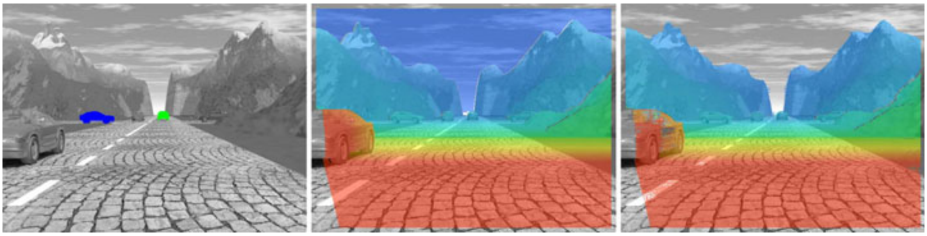


Fig. 1. Sample frame from the used data set. *Left*: Segmented objects used for different approaches. The blue car is static w.r.t the ground while the green car moves away from the ego-vehicle. – Colour encoded (red close, blue far) ground truth (*Middle*) and a disparity map obtained with a semi-global matching algorithm (*Right*).

compare results using both data domains separately for filtering, or both in combination. The rest of this paper is structured as follows. We start in Section 2 with recalling briefly the structure of Kalman filters. In Section 3 we describe the proposed approach. Section 4 reports and discusses results obtained in our experiments. Section 5 concludes.

2 Kalman Filter

We briefly recall the linear Kalman filter [Kalman 1960] as commonly used for a discrete dynamic system [Kuo and Golnaraghi 2002], with states $\mathbf{x}_t$, measurements $\mathbf{y}_t$, state transition matrix $\mathbf{A}$, process noise $\mathbf{v}$, a control matrix $\mathbf{B}$, and measurement noise $\mathbf{w}$. Given the system

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A} \cdot \mathbf{x}_{t-1} + \mathbf{B} \cdot \mathbf{u}_t + \mathbf{v} \\ \mathbf{y}_t &= \mathbf{B} \cdot \mathbf{x}_{t-1} + \mathbf{w} \end{aligned} \tag{1}$$

The Kalman filter is defined in two steps. First, the information from the previous step is incorporated into the filter by generating a *predicted* state

$$\begin{aligned} \mathbf{x}_{t|t-1} &= \mathbf{A} \cdot \mathbf{x}_{t-1|t-1} \\ \mathbf{P}_{t|t-1} &= \mathbf{A} \cdot \mathbf{P}_{t-1|t-1} \cdot \mathbf{A}^T + \mathbf{Q} \end{aligned} \tag{2}$$

We use the notation $t|t-1$ to denote that we are in an intermediate step between t and $t-1$, while $t-1|t-1$ denotes the state obtained with the Kalman filter at time $t-1$. Matrix $\mathbf{Q}$ represents the process noise variance (obtained from vector $\mathbf{v}$) and $\mathbf{P}_{t|t-1}$, denotes the covariance matrix of the error of $\mathbf{x}_{t|t-1}$ compared to the true value $\mathbf{x}_{t|t}$.

In the second step, the predicted state is corrected using data from the current state via the measurement vector $\mathbf{y}_t$ and the predicted matrix $\mathbf{P}_{t|t-1}$:

$$\mathbf{x}_{t|t} = \mathbf{x}_{t|t-1} + \mathbf{K}_t \left(\mathbf{y}_t - \mathbf{H} \cdot \mathbf{x}_{t|t-1} \right) \tag{3}$$

where

$$\mathbf{K}_t = \mathbf{P}_{t|t-1} \cdot \mathbf{H}^T \left(\mathbf{H}_t \cdot \mathbf{P}_{t|t-1} \cdot \mathbf{H}^T + \mathbf{R} \right) \tag{4}$$

is the *Kalman gain* as derived in [Kalman 1960]. It follows that

$$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t \cdot \mathbf{H}^T)\mathbf{P}_{t|t-1} \tag{5}$$

This is an iterative process. The initial state $x_{0|0}$ and the initial covariance matrix $P_{0|0}$ need to be selected.

3 Approach

The basic idea of our approach is to incorporate disparity information contained in the spatial domain, as reported in [Morales et al. 2009], into the Kalman filter approaches presented in [Badino et al. 2007, Vaudrey et al. 2008a]. These previous methods were designed to handle different kinds of moving objects.

In [Badino et al. 2007] the interest was on the improvement of disparity values for static objects in the scene, while in [Vaudrey et al. 2008a] the model was about longitudinal movements with respect to ego-motion. We aim at enhancing both methods by adding data from the disparity spatial domain.

By means of a Kalman filter, we merge temporal and spatial information. For simplicity we use an *iconic* Kalman filter [Matthies et al. 1989] (i.e. we use an individual Kalman filter for each pixel under consideration). In each iteration we obtain a new disparity value for each pixel using both spatial and temporal information. For doing so it is necessary to consider the motion of 3D world points that define pixels in the disparity map. Each pixel in an initial disparity map will be followed as the image sequence advances in time (as long as still visible in the current frame). This is done by considering ego-motion and disparity rate.

Let p be a pixel of a disparity map calculated at time t, and $p_1, \ldots, p_n$ adjacent pixels of it with $n \geq 1$. Our presentation of formulas is for $n = 2$ only; it generalizes for arbitrary neighbourhoods. The pixel p represents a 3D world point P projected into the image plane at time t. Let p also denote the projection of P at time $t + 1$ besides knowing that its position on the image plane is actually different due to ego-motion or possible independent motion of P. The disparity value assigned to p is also expected to change through the sequence.

Consider the dynamic system defined at time t by the state vector $\mathbf{x}_t$ and the transition matrix $\mathbf{A}$, given by

$$\mathbf{x}_t = \begin{pmatrix} d_p \\ d_1 \\ d_2 \\ \dot{d} \end{pmatrix} \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} \alpha & \beta & \gamma & \Delta t \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6}$$

where d_* denotes disparity values corresponding to p and two adjacent pixels p_1 and p_2; value $\dot{d}$ is the *disparity rate* as introduced in [Vaudrey et al. 2008a]. Values α and β control the interaction of disparity values of pixels p, p_1, and p_2. (In our experiments, we use $\alpha = 0{\cdot}8$ and $\beta = \gamma = 0{\cdot}1$). The parameter Δt represents the time elapsed between two consecutive frames. We assume that the noise vector $\mathbf{v}$ associated to the system (see Section 2) is Gaussian with zero mean and standard deviation σ_d for all disparity values and $\sigma_{\dot{d}}$ for disparity rate.

Measurement data (i.e. data from the current disparity map) is not available for the disparity rate. But, it contains all the involved disparity values, thus we can also consider the spatial information in a frame at time t. Therefore, the dimension of the measurement vector $\mathbf{y}_t$ is $n + 1$, and matrix $\mathbf{H}$ equals

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \tag{7}$$

The noise vector $\mathbf{w}$ associated to the measurements taken from the system is assumed to be the same for all of its coordinates: Gaussian with zero mean and with a standard deviation σ_v.

To start the filtering process we need to define the initial state and the initial covariance matrix. The initial state is defined using the disparity values of p

and its neighbours calculated with a given stereo algorithm at time $t = 0$. The disparity rate is set to be zero. The covariance matrix is defined by

$$\mathbf{P}_{0|0} = \begin{pmatrix} \sigma_d^2 & \sigma_{dd_1} & \sigma_{dd_2} & \sigma_{d\dot{d}} \\ \sigma_{dd_1} & \sigma_{d_1^2} & \sigma_{d_1 d_2} & \sigma_{d_1 \dot{d}} \\ \sigma_{dd_2} & \sigma_{d_1 d_2} & \sigma_{d_2^2} & \sigma_{d_2 \dot{d}} \\ \sigma_{d\dot{d}} & \sigma_{d_1 \dot{d}} & \sigma_{d_2 \dot{d}} & \sigma_{\dot{d}^2} \end{pmatrix} \tag{8}$$

where, for example, $\sigma_{d\dot{d}} = \sigma_d \cdot \sigma_{\dot{d}}$. Recall that we assumed that all the calculated disparity values have the same variance.

Once the filter has been initialized, we can start the iteration process. Assume that we have already calculated $t-1$ steps and that $\mathbf{x}_{t-1|t-1}$ is available. After the prediction step at time t, the first coordinate of $\mathbf{x}_{t|t-1}$ contains information about a neighbourhood (a 2-neighbourhood in this case) of the disparity map calculated at time $t-1$.

Before applying the Kalman update process it is necessary to update the position of pixel p (i.e. the visible movement of the 3D world point P between $t-1$ and t). This is done by calculating the relative motion of P with respect to the ego-vehicle, and this is done in tow steps.

First, the motion induced by the disparity rate is incorporated into P. This motion is away from the ego-vehicle and in the same direction. Thus, only the Z coordinate will be modified. Second, the positional change induced by ego-motion is considered. We use only speed and yaw rate (i.e. the bicycle model [Franke et al. 2005]) and assume that they are noise free. Roll and pitch are not included into our model, but both do have a minor influence for vision-augmented vehicles.

Once the position of p in frame t is known, we calculate the measurement vector $\mathbf{y}_t$ from the disparity map at time t. This vector is then used to generate the corresponding Kalman gain. Note that from the design of our system [see Equation (2)], disparity information from neighbouring pixels of p at time t is included in the measurement vector.

This measurement vector can now be used to calculate the Kalman gain in order to obtain the updated state, so the next iteration at time t can be performed. To avoid noisy states in both steps of the Kalman process, we follow the validation rules suggested in [Vaudrey et al. 2008a].

For adding temporal information to the approach of [Badino et al. 2007] it is necessary to remove the disparity rate term from the state and modify accordingly the rest of the dynamic system presented in Equation (6).

4 Experiments and Results

We perform experiments using the Sequence 1 [Vaudrey et al. 2008b] from the Set 2 from [EISATS 2011]. It is a computer generated sequence representing a driving scenario with available stereo ground truth; see Figure 1, middle. The ego-vehicle drives straight through the whole sequence. Thus we assume a constant yaw rate of 0 degrees and a speed of 6·99 m/s, calculated from the ground truth and an assumed frame rate of 25 frames per second.

As stated in Section 3, the method proposed in [Vaudrey et al. 2008a] (*dynamic* method from now on) was designed to improve the calculated disparity values from objects moving away but in the same driving direction of the ego-vehicle. We segmented from the test sequence (100 frames) a vehicle whose movement fulfils such requirements. For the experiments using the approach introduced in [Badino et al. 2007] (now called the *static* method) we segmented a static vehicle (53 frames) with respect to the ground. In Figure 1 left, the green vehicle is moving away from the ego vehicle, while the blue one is totally static.

For generating the disparity data we use a semi-global matching (SGM) stereo algorithm [Hirschmüller 2005], with a four-path configuration and the census transform as cost function (see [Herman et al. 2011]). See Figure 1, right, for a sample disparity map. As expected with noise-free stereo pairs, the SGM results were 'very close' to the ground truth, letting almost no room for improvement. All the filtered results were slightly worse than the raw stereo results, as already reported in [Morales et al. 2009]. However, this previous study reported that using either spatial or temporal post-processing leads to improvements when using noisy sequences which resemble real-world data.

Experiments were made for neighbourhoods with $n = 1, 2, 4, 8$ adjacent pixels, both for the static and the dynamic method. For each frame, we calculate the mean of all the disparity values within the corresponding region of interest (i.e. the moving vehicle for the dynamic method and the static vehicle for the static method). We then compare with available ground truth.

The parameters to initialize the Kalman filter for the dynamic method are as follows (for $n = 2$ and analogously for the other cases): The initial state was filled up with data from the disparity map calculated for the first available stereo pair; except for the disparity rate term, which was set to zero.

For the matrix $\mathbf{A}$, let $\alpha = 0{\cdot}8$, $\beta = \gamma = 0{\cdot}1$ and $\varDelta t = 0{\cdot}04$ (i.e. an assumed frame rate of 25 frames per second).

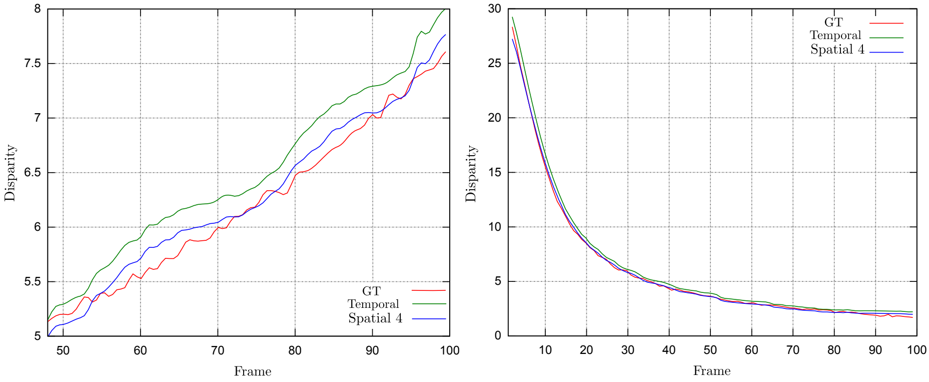


Fig. 2. Average disparity in the region of interest for the static (left) and the dynamic (right) method. Both plots show ground truth (GT), the results using only data from the temporal domain (temporal), and when using the spatio-temporal approach (Spatial 4) with $n = 4$.

The entries of the covariance matrix $\mathbf{P}_{0|0}$ were initialized with the following values: $\sigma^2_{d*} = 0{\cdot}3$, assuming an imperfect disparity map. $\sigma_{d*d*} = 0{\cdot}5$, a relative large value to represent large correlation between the pixel under analysis and its neighbours. $\sigma_{d*\dot{d}} = 0{\cdot}0001$, to show a low correlation between the disparity values and the disparity rate. $d*$ represent either d, d_1, or d_2. Finally, $\sigma_{\dot{d}^2} = 1$, a relative large value to express high uncertainty in the initial disparity rate. - The parameter initialization for the static method is analogous. It is only necessary to remove the terms where the disparity rate is involved and modify the matrices accordingly.

The results for the dynamic method show an improvement when using data from the spatio-temporal domain (for $n = 1$, 2, or 4) compared to when just using the temporal domain. See Figure 2. The results improve according to the size of the neighbourhood, being the best for $n = 4$. For $n = 8$, the results were not satisfactory. As expected, a large neighbourhood degrades the final disparity value. The results are summarized in Table 1, presenting the average deviation from the ground truth for the whole sequence and for all the considered settings.

For the static method we obtained similar results. The spatio-temporal approach (for $n = 1$, 2, or 4) shows a better performance than the temporal method. See Figure 2. The best performance was archived when using $n = 4$, and the worst was measured for $n = 8$. However, in this case, for $n = 8$ the results were still better than when just using the temporal approach. Interestingly, the average deviation for the whole sequence for SGM and $n = 4$ is almost the same. See Table 1. For some frames, the average disparity value was closer to the ground truth when using the spatio-temporal approach than with the original SGM algorithm.

Table 1. Average deviation from the ground truth for each one of the methods. SGM stands for the raw stereo results. Temporal is for the exclusively temporal approach. The rest are for the spatio-temporal methods, where n indicates the size of the neighbourhood.

	SGM	Temporal	$n = 1$	$n = 2$	$n = 4$	$n = 8$
Mobile	0·11	0·41	0·18	0·16	0·15	3·14
Static	0·10	0·28	0·12	0·12	0·10	0·22

5 Conclusions

In this paper we present a method for post-processing disparity maps using a spatio-temporal approach. We use an iconic Kalman filter approach for merging data from both domains.

Obtained experimental results (this short paper only discusses one sequence given with ground truth) showed improvements compared to original or either only temporal or only spatial post-processing. We suggest to use the combined approach for filtering out noisy values in disparity maps generated for stereo sequences recorded in the real-world.

Future work will quantify improvements on real-world sequences using the prediction-error approach as discussed, for example, in [Morales et al. 2009].

References

[Atzpadin et al. 2004] Atzpadin, N., Kauff, P., Schreer, O.: Stereo analysis by hybrid recursive matching for real-time immersive video stereo analysis by hybrid recursive matching for real-time immersive video conferencing. IEEE Trans. Circuits Systems Video Techn. 14, 321–334 (2004)

[Badino et al. 2007] Badino, H., Franke, U., Mester, R.: Free space computation using stochastic occupancy grids and dynamic programming. In: Proc. Dynamic Vision Workshop for ICCV (2007)

[EISATS 2011] .enpeda.. Image Sequence Analysis Test Site, EISATS (2011), http://www.mi.auckland.ac.nz/EISATS/

[Franke et al. 2005] Franke, U., Rabe, C., Badino, H., Gehrig, S.: 6D-vision: Fusion of stereo and motion for robust environment perception. In: Kropatsch, W.G., Sablatnig, R., Hanbury, A. (eds.) DAGM 2005. LNCS, vol. 3663, pp. 216–223. Springer, Heidelberg (2005)

[Herman et al. 2011] Hermann, S., Morales, S., Klette, R.: Half-resolution semi-global stereo matching. In: Proc. IEEE Intelligent Vehicles Symp. (2011)

[Hirschmüller 2005] Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: Proc. Computer Vision Pattern Recognition, vol. 2, pp. 807–814 (2005)

[Kalman 1960] Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Engineering 82, 35–45 (1960)

[Kuo and Golnaraghi 2002] Kuo, B., Golnaraghi, F.: Automatic Control Systems. John Wiley and Sons Inc., New York (2002)

[Matthies et al. 1989] Matthies, L., Kanade, T., Szeliski, R.: Kalman filter-based algorithms for estimating depth from image sequences. Int. J. Computer Vision 3, 209–238 (1989)

[Morales et al. 2009] Morales, S., Vaudrey, T., Klette, R.: Robustness evaluation of stereo algorithms on long stereo sequences. In: Proc. IEEE Intelligent Vehicles Symposium, pp. 347–352 (2009)

[Vaudrey et al. 2008a] Vaudrey, T., Badino, H., Gehrig, S.: Integrating disparity images by incorporating disparity rate. In: Sommer, G., Klette, R. (eds.) RobVis 2008. LNCS, vol. 4931, pp. 29–42. Springer, Heidelberg (2008)

[Vaudrey et al. 2008b] Vaudrey, T., Rabe, C., Klette, R., Milburn, J.: Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. In: Proc. IVCNZ, pp. 1–6 (2008)

[Vaudrey and Klette 2009] Vaudrey, T., Klette, R.: Residual images remove illumination artefacts for correspondence Algorithms! In: Denzler, J., Notni, G., Süße, H. (eds.) DAGM. LNCS, vol. 5748, pp. 472–481. Springer, Heidelberg (2009)

[Zhihui 2008] Zhihui, X.: Computer Vision. InTech, online books (2008)

Refractive Index Estimation Using Polarisation and Photometric Stereo

Gule Saman and Edwin R. Hancock*

Department of Computer Science, University of York, UK
{saman,erh}@cs.york.ac.uk

Abstract. This paper describes a novel approach to the estimation of refractive indices of surfaces using polarisation information. We use a moments estimation method for computing the polarisation image components from intensity images obtained using multiple polariser angles. This yields estimates of the mean-intensity, polarisation and phase at each pixel location.The surface normals are estimated using the photometric stereo. Using the Fresnel theory at each pixel we estimate the refractive index of the surface from the zenith angle of the surface normal and the measured polarisation. The method has been applied to determine the variations in paintings, human skin refractive indices and also for inspecting fruit surfaces. To test the effectiveness of the method, we coat a variety of objects with a layer of transparent liquid of known refractive index. Experiments on naturally occurring surfaces (e.g. human skin and fruits) and manufactured objects such as a plastic balls and paintings illustrate the effectiveness of this method in estimating refractive indices.

Keywords: Naturally occurring surfaces, Manufactured surfaces, Polarisation Information, Photometric stereo, Fresnel Theory, Refractive index estimation

1 Introduction

The physics of light has been widely used for surface analysis. The optical properties of surfaces prove to be useful for assessing the quality of surfaces (natural and manufactured), finding widespread application in biomedical optics. One important surface property is the refractive index. However, determining the refractive index of naturally occurring surfaces is challenging due to the fact that planar samples are rarely available and analysis is complicated by the intrinsic shape of the object under study. The Fresnel equations describes the reflection and refraction of incident light as it passes an interface between two media ([1], [11]). In this paper, we aim to use the scattering properties of light in the visible spectrum for estimating refractive index.

Refractive index is the ratio of the speed of light in two media, which determines the transmission properties of light in a material. Polarisation measurements and the Fresnel theory can be used to measure the refractive index. Computer vision deals with a wide variety of problems for surface inspection and reconstruction, where polarisation information has been used for developing algorithms to solve them. The Fresnel

* Edwin Hancock is supported by the EU FET project SIMBAD and by a Royal Society Wolfson Research Merit Award.

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 548–554, 2011.

theory of light has been used to model the transmission and reflection of the parallel and perpendicular field components of incident light. In particular, Wolff [2] has used the Fresnel theory for developing methods for identifying dielectric and metal surfaces based on polarisation. Moreover, it is easier to model dielectrics in comparison with metals since when an electromagnetic field is incident on a metal surface it induces surface currents which can not be explained by the Fresnel theory alone.

When a polaroid filter is used as an analyser the degree of polarisation and phase can be measured for both specular and diffuse polarisation. According to the Fresnel theory, the refractive index is determined by the the degree of polarisation and the zenith angle between the remitted light and the surface. In this paper we aim to exploit this relationship to estimate refractive index. We commence by using the method of moments to estimate the components of polarisation image [3]. To acquire the zenith angle of the remitted light, we use photometric stereo. With polarisation and zenith angles at hand, we use the the Fresnel theory for diffuse reflectance to estimate the refractive index at each pixel. This allows for the refractive index to be estimated across curved surfaces. To test the accuracy of the method we coat curved objects with layers of transparent fluid (olive oil and vaseline) of known refractive index.

2 Components of Polarisation Image

For computing the components of the polarisation image, the method proposed by [3] has been used which uses robust moment estimators for diffuse reflectance, in which incident light undergoes subsurface reflections before being re-emitted. Here the Fresnel theory provides the relationship between the degree of polarisation and angle of reflection of the scattered light.

2.1 Data Set

We have used a geodesic light dome constructed along the lines described in [10] to collect a set of images with different orientations of the analyser polaroid for the measurement of the polarisation state. The object under study is placed at the centre of the geodesic light dome. A Nikon D200 camera is used to collect images with fixed exposure and aperture settings. The light sources are unpolarised, while a polariser has been placed in front of the camera to analyse the polarisation state of the reflected light. The analyser is rotated from 0 ° to 180 ° in increments of 10 ° to give 19 images per object.

The experiments were conducted in a room with matte black walls and surfaces so as to minimise inter-reflections. The geodesic light dome is calibrated and allows for accurate measurement of the angle of the incident light source. Examples of the acquired image data set are shown in Fig.(4) and(1) for a coated and uncoated terracotta object.

2.2 Polarisation Image

The conventional method for estimating the three components of the polarisation image is based on least-squares fitting which yields, the mean-intensity I_0, degree of polarisation ρ and phase ϕ [4]. Instead, we have used a robust moments estimation method [3], which gives improved estimates of the components of the polarisation image. Suppose that the analyser angle is given by β_i, where i is the analyser angle index. The predicted brightness at the pixel indexed p with the analyser angle indexed i, is give by:

$$I_p^i = \hat{I}_p \left\{ 1 + \rho_p \cos(2\beta_i - 2\phi_p) \right\}. \tag{1}$$

where $\hat{I}_p$, ρ_p and ϕ_p are the mean intensity, polarisation and phase at the pixel indexed p.

3 Surface Normal Estimation

We use photometric stereo to estimate the surface normal directions across the surface of the object under study. We acquire images with three different light source directions. The surface normal directions are estimated using the method of Woodham [5] and later used by [6]. The mean-intensity estimated from the polarisation image has been used as input to the photometric stereo for computation of the surface normals.

Let $S_m = (S_1|S_2|S_3)$ be the matrix with the three light source vectors as columns, N_p the surface normal at the pixel indexed p and $\hat{J}_p = (\hat{I}1_p, \hat{I}2_p, \hat{I}3_p)^T$ be the vector of the three mean brightness values recorded at the pixel indexed p with the three different light source directions. Under the assumption of Lambertian reflectance, at pixel p we have:

$$\hat{J}_p = S_m N$$

. The surface normal can be computed from the vector of brightness values J_p and the inverse of the source matrix S_m. The reflectance factor, R, is computed by taking the magnitude of the right side of equation (2) as the surface normal, N, is assumed to have unit length

$$R_p = |[S_m]^{-1} \hat{J}_p|. \tag{2}$$

The unit normal vector is computed as follows:

$$N_p = (1/R) * [S_m]^{-1} \hat{J}_p. \tag{3}$$

The images taken across the polarizer angles are reconstructed using the following equation:

$$J_p^i = S_m.N(1 + \rho_p \cos(2\beta_i - 2\phi_p)). \tag{4}$$

The angle of incidence between the light source indexed l and the surface normal at the pixel index p is $\theta_p^l = N_p.S_l$.

4 Estimating the Refractive Index

The Fresnel theory of light predicts that light incident on a surface is partially polarised and refracted while penetrating the surface. The structure of the reflecting surface changes the polarisation state of the incident light. The remitted light is refracted into the air and is partially polarised in the process. This model as has been reported by

[4], can be used to compute the relationship between the degree of diffuse polarisation, the angle between the surface normal and the remitted light θ and the refractive index n:

$$\rho = \frac{(n-\frac{1}{n})^2 \sin^2\theta}{2+2n^2-(n+\frac{1}{n})^2\sin^2\theta+4\cos\theta\sqrt{n^2-\sin^2\theta}}. \tag{5}$$

We have used the above equation to compute the refractive index n, from measured values of ρ and θ at each pixel. The refractive index is given by the solution of the quartic equation:

$$A^2n^4 + (2AC-1)n^3 + (2AB + C^2 + \sin^2\theta)n^2 + 2BCn + B^2 = 0. \tag{6}$$

where $A = ((1+\rho)\sin^2\theta - \frac{2\rho}{4\rho\cos\theta})$, $B = (\frac{(1+\rho)\sin^2\theta+\rho}{4\rho\cos\theta})$ and $C = -\frac{2((1-\rho)\sin^2\theta+\rho)}{4\rho\cos\theta}$. Newton-Raphson method has been used to compute the roots of the equation (6). 10 iterations have been used as the method converges before that.

Alternative methods have been reported in the literature for estimating the refractive index. These include the use of multi-spectral polarisation imagery from a single viewpoint [7]. The spectral dependence of the refractive index has been studied by [8], [9]. The novelty of our method is that we have used the Fresnel theory along with photometric stereo and estimates from the polarisation image for estimating the refractive indices.

5 Experiments

The data set was acquired in a room with matte black walls and working surfaces to avoid errors form environmental reflections. The object and camera are placed on the same axis while the light sources are at three different angles from the object in the geodesic light dome. We have conducted experiments with both objects of known refractive index and objects of unknown refractive index. The objective is to understand whether the method can robustly deal with variations in surface composition and shape for computing the refractive index. The estimated values for the refractive indices fall in the range $1 < n < 2.5$. Some of the non-physical outlier values have been filtered out in the computation process (i.e. those less than unity). The refractive index images for the objects under study have been shown in Fig.(2).

For computing the refractive index the potential sources of error are: a) dents in the surface of the naturally occurring objects leading to inter-reflections and sub-surface reflections, b) in the case of coated objects sub-surface reflections from the base object, and c) noise and camera jitter. For the human subject the main source of error has been the image alignment due to difficulties in keeping the subject stationary. The misalignment of the intensity images leads to inaccuracies in the polarisation image computations.

5.1 Estimation for Known Refractive Indices

Here we have coated a plastic ball, an orange, a human face and a terracotta object with Vaseline and olive oil. For the human face, we have analysed patches from the face of the forehead and the cheeks and nose area. The estimated refractive index values along with the percentage error are given in Table(1). It can be seen in Fig.(2) that in case of the coated terracotta object the different materials can be identified by the presence of a distinct boundary and in the case of the painting the coatings and different paints have different effect on the estimation. The reference values for refractive index of olive oil and vaseline are:1.4677 and 1.480, respectively.

5.2 Estimation for Unknown Refractive Indices

The objects studied include a human face, a painting of a house and a painting of mountains and a pool of water. For the the second painting we have analysed patches from it. The estimated refractive index values are given in Table(2). It can be seen from the table that the values of refractive indices fall in a similar range, these paintings have been made using acrylic paints.

Table 1. The estimated refractive indices and estimation error of known substances

Object	Estimated Refractive index	Percentage Error
Coated Ball (Olive Oil)	1.2738	13.21
Coated Terracotta Object (Olive oil)	1.2709	13.41
Coated Orange (Vaseline)	1.3053	11.80
Cheek and Nose (Vaseline)	1.4756	0.30

Table 2. The estimated refractive indices of unknown substances

Object	Estimated Refractive index
Island Painting	1.2016
Tree and Water Painting	1.2536
House Painting	1.2093
Terracotta	1.2743
Forehead	1.2467
Cheek and Nose	1.4150

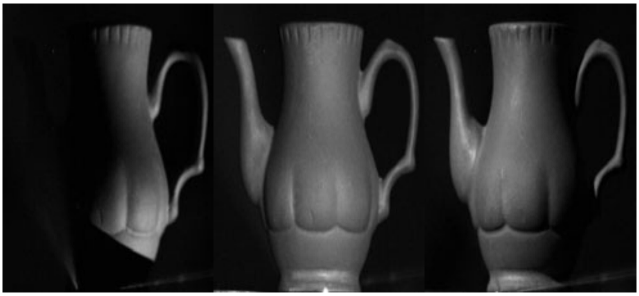


Fig. 1. The scene for a terracotta object using three different light source directions

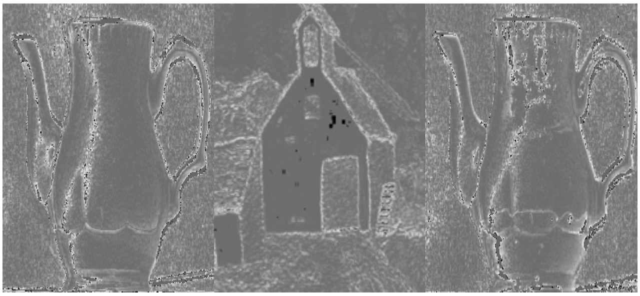


Fig. 2. The refractive index images for a terracotta object, a painting and a coated terracotta object

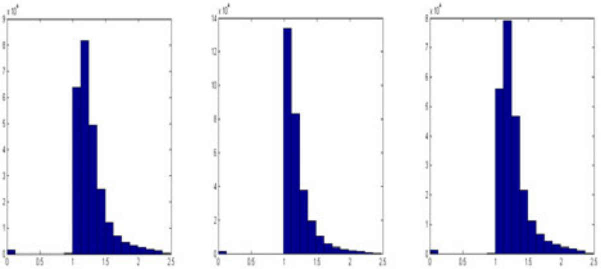


Fig. 3. The histograms for a terracotta object, a painting and a coated terracotta object

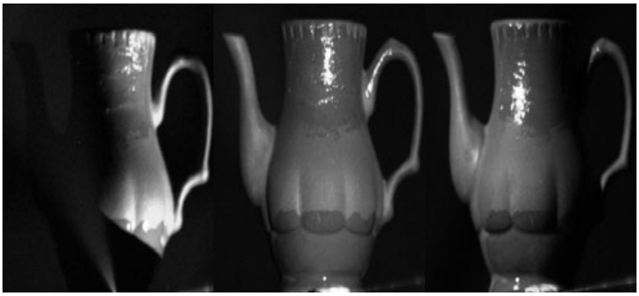


Fig. 4. The scene for a terracotta object coated with vaseline(top) and olive oil(bottom) using three different light source directions

6 Conclusions

In this paper, we have explored the use of polarisation information and Fresnel theory for estimating refractive index. Our approach has been to use the polarisation image information estimated using the robust moments method and photometric stereo for estimating the surface normals. This information is used as an input for the Fresnel equation for diffuse reflectance for computing the refractive index. We have given a set of results with error estimates for substances of known refractive indices and also a set of values for unknown refractive index values. It can be seen that the values do not exhibit any shape bias.

References

1. Hecht, E.: Optics, 4th edn. Addison-Wesley, Reading (2002)
2. Wolff, L.B., Boult, T.E.: Constraining Object Features using a Polarisation Reflectance Model. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 635–657 (1991)
3. Saman, G., Hancock, E.R.: Robust Computation of the Polarisation Image. In: International Conference on Pattern Recognition (2010)
4. Atkinson, G., Hancock, E.R.: Robust estimation of reflectance functions from polarization. Springer, Heidelberg (2007)
5. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical Engineering 19(1) (1980)
6. Coleman, E.N., Jain, R.: Obtaining 3-Dimensional Shape of textured and Specular surfaces using four-source photometry. Computer Graphics and Image Processing 18(4), 309–328 (1982)
7. Huynh, C.P., Robles-Kelly, A., Hancock, E.R.: Shape and Refractive Index Recovery from single-view polarisation images. In: IEEE Conference on Computer Vision and Pattern Recognition (2010)
8. Chang, H., Charalampopoulos, T.T.: Determination of the wavelength dependence of refractive indices of flame soot. Royal Society (1990)
9. Ding, H., Lu, J.Q., Wooden, W.A., Kragel, P.J., Hu, X.: Refractive indices of human skin tissues at eight wavelengths and estimated dispersion relations between 300 and 1600nm. Journal Physics in Medicine and Biology 51(6) (2006)
10. Cooper, P., Thomas, M.: Geodesic Light Dome. Department of Computer Science, University of York, UK (March 2010), http://www-users.cs.york.ac.uk/-pcc/Circuits/dome (Accessed On: September 10, 2010)
11. Born, M., Wolf, E.: Principles of Optics, 7th edn. Cambridge university Press, Cambridge (1999)
12. Yee, K.: Numerical Solutions of initial boundary value problems involving Maxwell's equations in instotropic media. IEEE Transactions on Antennas Propagation AP-14, 302–307 (1966)
13. Dunn, A., Richards-Kortum, R.: Three-Dimensional Computation of Light Scattering From Cells. IEEE Journal of Selected topics in Quantum Electronics 2(4) (1996)
14. Nieminen, T.A., Rubinsztein-Dunlop, H., Heckenberg, N.R.: Calculation of the T-matrix: general considerations and application of the point-matching method. Journal of Quantitative Spectroscopy and Radiative Transfer 79-80, 1019–1029 (2003)

3D Gestural Interaction for Stereoscopic Visualization on Mobile Devices

Shahrouz Yousefi, Farid Abedan Kondori, and Haibo Li

Digital Media Lab.,
Department of Applied Physics and Electronics, Teknikhuset,
Umeå University, 901 87, Umeå, Sweden
{shahrouz.yousefi,farid.kondori,haibo.li}@tfe.umu.se

Abstract. Number of mobile devices such as smart phones or Tablet PCs has been dramatically increased over the recent years. New mobile devices are equipped with integrated cameras and large displays which make the interaction with device more efficient. Although most of the previous works on interaction between humans and mobile devices are based on 2D touch-screen displays, camera-based interaction opens a new way to manipulate in 3D space behind the device in the camera's field of view. In this paper, our gestural interaction heavily relies on particular patterns from local orientation of the image called *Rotational Symmetries*. This approach is based on finding the most suitable pattern from a large set of rotational symmetries of different orders which ensures a reliable detector for hand gesture. Consequently, gesture detection and tracking can be hired as an efficient tool for 3D manipulation in various applications in computer vision and augmented reality. The final output will be rendered into color anaglyphs for 3D visualization. Depending on the coding technology different low cost 3D glasses will be used for viewers.

Keywords: 3D mobile interaction, rotational symmetries, gesture detection, SIFT, gesture tracking, stereoscopic visualization.

1 Introduction

Nowadays gesture detection, recognition or tracking are terms which have frequently been encountered in discussions of human computer interaction. Gesture recognition enables humans to interact with computers and makes input devices such as keyboard, joystick or touch screen panels redundant. Having more effective interaction with mobile devices could be the most significant reason behind the manufacturing of devices with larger screens during the recent years. Although the idea of working with larger touch screen displays helps users to have a better interaction with device, their limitations in 2D space manipulation remain an unsolved issue. A novel solution for limitations of 2D touch-screen displays is taking advantage of 3D space behind the camera. Manipulation in the camera's field of view provides a chance for users to work with any mobile device regardless of the screen size or touch sensitivity. As it is shown in Fig. 1, the user's hand in farther distances from the camera occupies smaller place in the

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 555–562, 2011.

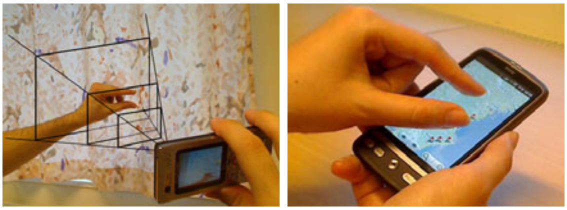


Fig. 1. 3D interaction vs. 2D interaction. Left: User capability to move in depth and change the finger size according to the distance to the camera. Right: Limited area for user's fingers in 2D interaction.

screen which is a positive point to compensate the limited area for fingers on 2D displays. Moreover, behind the camera users are capable of moving in depth which could handle a lot of difficulties in various applications. Our experiments on mobile applications reveal that in 2D interaction on the screen, users have limitations in moving in depth, zooming in to observe the details of an image, map, text or zooming out to skim through a data. Even in more complicated applications rotations around different axes are unavoidable.

Here the question is whether it is possible to solve these limitations by introducing a new interaction environment? Our experiment shows that in 3D interaction under the mobile phone's camera, fingertips approximately occupy $10-20\%$ area of the touch-screen display. This observation reveals that interaction resolution in 3D space is $5-10$ times higher than 2D displays. Moreover, regarding the higher degrees of freedom in 3D space, limitations in rotation will be handled. Therefore, this new interaction environment can be introduced to significantly improve the efficiency and effectiveness of the human mobile device interaction. Our gesture recognition method is based on the *rotational symmetries* detection in video input from the camera. This method finds patterns from local orientation of the image. The implemented operator searches for particular features in local images and detects expected patterns associated with the human gesture. Tracking the detected gesture enables humans to interact with mobile phone in 3D space and manipulate in various applications. Reliable human gesture detection algorithm raises a strong possibility of extracting the human gesture motion. By finding the corresponding keypoints in consecutive frames, the relative rotation and translation between two image frames can be computed. In order to convey the depth illution to viewrs, stereoscopic techniques are perfomed for visualization. Our system adjusts the output stereo channels and renders them in different types of color anaglyphs. The 3D output can be displayed on the mobile device and users simply need to use low cost anaglyph eyeglasses to view in 3D.

2 Related Work

A common method for gesture detection is marker-based approach. Most of the augmented reality applications are based on marked gloves for accurate and

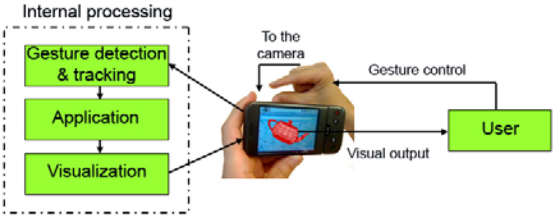


Fig. 2. System overview

reliable fingertip tracking [3, 10]. However, in marker-based methods users have to wear special inconvenient markers. Moreover, some strategies rely on object segmentation by means of shape or temperature [5, 6, 8]. Robust finger detection and tracking could be gained by using a simple threshold on the infrared images. Despite the robustness, thermal-based approaches require expensive infrared cameras which are not provided to mobile devices. In addition, feature-based algorithms for gesture tracking have been employed in various applications [2, 4, 8, 12]. Model-based approaches are also being used in this area [20, 21].

Almost all these techniques are generally computationally expensive which is not suitable for our purposes. Another set of methods for hand tracking are based on color segmentation in appropriate color space [1, 5]. Color-based techniques are always sensitive to lighting conditions which degrades the quality of recognition and tracking. Other approaches such as template matching and contour-based methods often work for specific hand gestures [11, 15]. For 3D visualization, stereoscopic techniques using 3D glasses, 3D displays without glasses and other technologies have been introduced and used for many years.We propose a way to take advantage of circular patterns and in general rotational symmetries associated with the model of the hand gesture [7]. We suggest an accurate algorithm to estimate the relative gesture motion in image sequences with the aid of stable features [16]. Our 3D coding is based on stereopsis approach which is known for many years [14, 18]. For 3D visualization we performed two different methods called red-cyan and colorcode(amber-blue) anaglyphs [13, 19, 22].

3 System Description

In this part the proposed 3D camera-based interaction approach is presented. As user moves his/her hand gesture in the camera's field of view behind the mobile device, the device captures a sequence of images. Then this input will be processed in gesture detection block. As a result, the user gesture will be detected and localized. Afterwards, stable features in each frame are extracted to compute the relative rotation and translation of the hand gesture between two frames. Finally, this information can be used to facilitate the user-mobile device interaction and manipulation of virtual objects on the screen (see Fig.3). In the visualization part we explain how stereoscopic techniques are used to convey the illution of depth to the viewer's eyes.

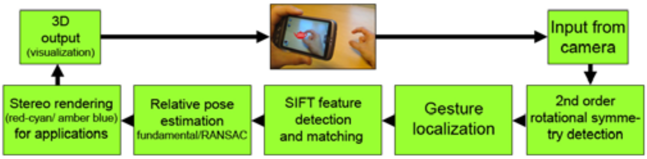

3D output (visualization)
Input from camera
2nd order rotational symmetry detection
Gesture localization
SIFT feature detection and matching
Relative pose estimation fundamental/RANSAC
Stereo rendering (red-cyan/ amber blue) for applications

Fig. 3. System description

3.1 Gesture Detection and Tracking

The main idea behind the rotational symmetries theory is to use local orientation to detect complex curvatures in double-angle representation [7]. Using a set of complex filters on the orientation image will result in detection of number of features in different orders, such as curvatures, circular and star patterns [7].

Our gesture detection system takes advantage of the rotational symmetries to localize the user's gesture in image sequences, which leads to differentiate between fingers and other features even in complicated backgrounds. Since the natural and frequently used gesture to manipulate objects in 3D space is similar to Fig.4, this model can satisfy our expectations for different applications. Our experiments based on various test images of different scales and backgrounds revealed that the user's gesture substantially responds to the second order rotational symmetry patterns (circular patterns). Thus our gesture detection system is designed to detect circular pattern in the video input. The double-angle representation of a given image can be computed as: $z(x) = (f_x(\mathbf{x}) + if_y(\mathbf{x}))^2$ where local orientation is defined as, $f(\mathbf{x}) = (f_x(\mathbf{x})\ f_y(\mathbf{x}))^T$. Eventually, to detect the 2nd order rotational symmetries in an image, the double-angle image should be correlated with the complex filter $a(\mathbf{x})b_2(\mathbf{x})$, where $b_2(\mathbf{x}) = e^{i2\varphi}$ is the 2nd order symmetry basis function, and $a(\mathbf{x})$ is the weight window for the basis function. In each local region in an image we compute the scalar product: $S_2 = \langle ab_2, z\rangle$ High magnitudes in the result S_2 indicate the higher probability of 2nd order rotational symmetries patterns in the image. Our observation shows that searching for the second order rotational symmetries in image frames by a suitable filter size will result in high probability of responses of user's gesture in different scales. Consequently, this will result in a proper localization of the user's gesture.

3.2 3D Structure from Motion

By localizing the hand gesture, a region of interest is defined around the user's hand. In order to analyze the gesture motion and estimate its rotation and translation, we need to extract keypoints in the ROI(region of interest). Among different feature detectors, SIFT [16] feature detector is used for its robustness and invariance to image transformation. Next we find feature point correspondences by matching feature points between consecutive frames, as it is depicted in Fig. 4. Then a fundamental matrix for each image pair is computed using robust iterative RANSAC [9] algorithm. Due to the fact that the matching part might be degraded by noise, the RANSAC algorithm is used to detect and remove

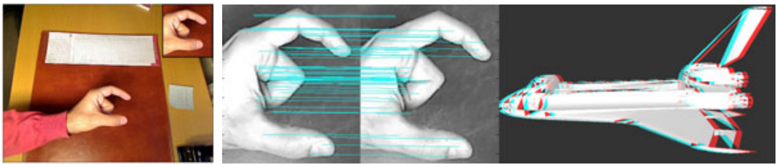


Fig. 4. Left: The localized gesture from input image. Middle: Feature matching in two consecutive frames where 54 point correspondences are detected. Right: Rendered graphical model according to the gesture motion (red-cyan anaglyph).

the wrong matches(outliers) and improve the performance. Running RANSAC algorithm, the candidate fundamental matrix is computed based on the 8-point algorithm [17]. Each point correspondence provides one linear equation in the entries of the fundamental matrix F. If the intrinsic parameters of the cameras are known, the essential matrix E can be introduced. Once the essential matrix is known, the relative translation and rotation matrices, t and R can be recovered [17]. The recovered information can be used in different mobile applications. For instance, in Fig. 4 the relative rotation between two consecutive images are $X = -0.4, Y = 1.7$, and $Z = -1.5$ degree.

3.3 3D Coding and Visualization

Stereoscopy or 3D imaging is the enhancement of conveying the illusion of depth in photos or videos. This effect can be presented by transmission of slightly different image to each eye. One common and low cost group of stereoscopic methods are color anaglyphs. In this method users wear special glasses with two different left and right colors, each for filtering the corresponding layer from the stereoscopic image or video. The difference in perceived images from each eye is the source of depth perception and 3D illusion. In order for left eye signal to be different from right eye the absorption curve has to be different. Furthermore, due to the parallax in stereoscopic image pair it requires at some points the luminance of one channel be greater than the other and vice-versa [19]. Hence the absorption curves should satisfy the non-overlapping bands and luminance condition. Based on the above discussion, we implemented two different color anaglyphs known as *red-cyan* and *color code (amber-blue)*. In the implementation the so-called optimized red-cyan anaglyph [22] is used. The optimized anaglyph discards the red component of the original image and replaces that with the red channel derived from the weighted green and blue components. The cyan channel is directly made of green and blue components. The improved method with gamma correction is suggested in [22].

The idea behind color code algorithm is that if one eye perceives a view which is in color and the other eye sees the view in monochrome, most likely the fusion between these two channels contains the full color range perception. Therefore, the amber color allows most of the colors to go through the channel and dark blue provides the monochrome image for the other eye channels [19].

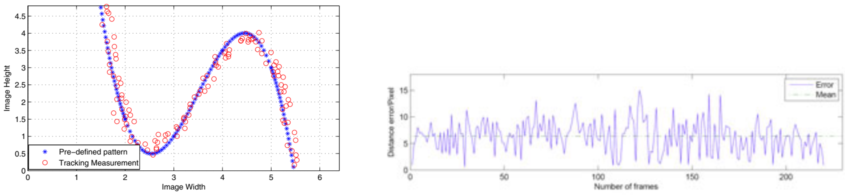


Fig. 5. Left: System performance in gesture tracking. Right: Error of the tracking in a sequence of images.

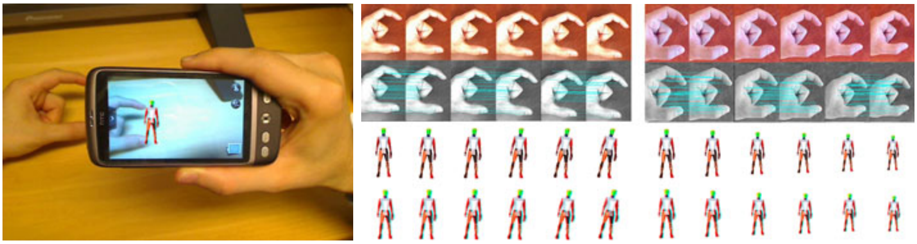


Fig. 6. Examples of gestural interaction in manipulation of the graphical model (rotation and zoom out). Rotation: Gesture localization (first row), feature detection and matching (second row), rendered model in normal mode (third row), and red-cyan anaglyph (fourth row). Next columns show the same steps for zoom out.

Based on the above discussion, we can provide our output model in color anaglyph channels. Before the visualization part, where the graphical model will be rendered on the display, our anaglyph channels shoud be relatively translated on their horizontal baseline. This geometrical shift is the source of depth perception and 3D illution [14, 18]. Our experiment revealed that 15 units of horizontal shift between left and right images is required to perform the stereoscopic views. Afterwards, both channels will be merged and the final output will be cropped for visualization.

4 Experimental Results

As a matter of fact, for a particular gesture behind the mobile device's camera, users have freedom to move in a reasonable distance. Moreover, depending on the application, they are free to rotate in different angles. Our observation indicates that the effective interaction happens in the area between $15 - 25cm$ from the camera. Interaction in the area beyond $25cm$ does not seem to be convenient for users. Clearly, for distances below $15cm$, gesture occupies a large area in the screen and degrades the interaction. Due to the fact that we aimed to benefit from this approach in real-time applications, we implemented the software in $C++$. Gesture detection, tracking and structure from motion parts are conducted in OpenCV environment and 3D coding and graphical visualization are implemented in OpenGL. Fig.5 illustrates the system performance in the tracking of the particular curve on a complex background. In this example the user is asked to follow the defined curve drawn on the screen. Circles mark the position of the detected gesture corresponding to each image frame. The error in the tracking of the original curve for more than 200 frames is plotted. The mean value of the error (6.59 pixels) shows the slight difference between the original curve and the one plotted by the tracked gesture which is quite satisfying. Fig.6 shows the 3D manipulation of a virtual object on the screen by user's gesture. The result is rendered by both red-cyan and amber-blue methods.

5 Conclusion

In this paper we presented a novel approach for 3D camera-based gesture interaction with mobile devices. Rotation, translation, and manipulation of virtual objects on the screen are not limited as 2D interaction any more. Our detection algorithm can estimate the position of the user's gesture in consecutive frames. This algorithm is computationally efficient and work well in practice. The relative pose estimation method can accurately extract the relative rotation and translation of the gesture between two consecutive frames, which could be used to facilitate the user-mobile device interaction. Robustness and simplicity are the main advantages of this approach that rely on the low level operations and equipment-free gestural interaction. Depending on the user's interest, the final output might be rendered by 3D techniques. For depth perception users simply need to wear low cost anaglyph glasses. System performance in gesture detection and tracking is quite satisfying. Our system can process more than twenty frames per second. Although wrong detected features caused by quite complex backgrounds are rare, but future work can concentrate on improvement and optimization of the algorithm. Moreover, we can more focus on graphical design.

References

[1] Bencheikh, M., Bouzenada, M., Batouche, M.C.: A New Method of Finger Tracking Applied to the Magic Board. In: Conf. on Industrial Technology (2004)

[2] Bretzner, L., Laptev, I., Lindeberg, T.: Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In: Proc. of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, (2002)

[3] Dorfmueller-Ulhaas, K., Schmalstieg, D.: Finger Tracking for Interaction in Augmented Environments. In: 2nd ACM/IEEE Symposium on Augmented Reality (2001)

[4] Erol, A., Bebis, G., Nicolescu, M., Boyle, R., Twombly, X.: Vision-based hand pose estimation: A review. Computer Vision and Image Understanding, (2007)

[5] Hardenberg, C.V., Berard, B.: Bare-hand human-computer interaction. In: Proceedings of the 2001 Workshop on Perceptive User Interfaces, Orlando, Florida. ACM International Conference Proceeding Series, vol. 15 archive (2001)

[6] Iwai, D., Sato, K.: Heat Sensation in Image Creation with Thermal Vision. In: ACM SIGCHI Int. Conf. on Advances in Computer Entertainment Technology (2005)

[7] Johansson, B.: Low Level Operations and Learning in Computer Vision. Linkoping Studies in Science and Technology Dissertation, Linkopings universitet (2004)

[8] Kolsch, M., Turk, M.: Fast 2D Hand Tracking with Flocks of Features and Multi-Cue Integration. In: Proc. CVPR Workshop (2004)

[9] Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Readings in Computer Vision: Issues, Problems, Principles, and Paradigms, 726–740 (1987)

[10] Maggioni, C.: A novel gestural input device for virtual reality. In: IEEE Annual International Symposium on Virtual Reality, pp. 118–124 (1993)

[11] Rehg, J., Kanade, T.: Digiteyes Vision-based Human Hand Tracking. Technical Report CMU-CS-TR-93-220, Carnegie Mellon University (1993)

[12] Laptev, I., Lindeberg, T.: Tracking of Multi-state Hand Models Using Particle Filtering and a Hierarchy of Multi-scale Image Features. In: Proc. Scale-Space and Morphology in Computer Vision (1999)

[13] Dubois, E.: A Projection Method To Generate Anaglyph Stereo Images. In: Proc. IEEE Int. Conf. Acoustics Speech Signal Processing (2001)

[14] Jones, G., Lee, D., Holliman, N., Ezra, D.: Controlling Perceived Depth in Stereoscopic Images. Stereoscopic Displays and Virtual Reality Systems VIII (2001)

[15] Zhou, H., Ruan, Q.: Finger Countour Tracking Based on Model. In: Conf. on Computers, Comunications, Control and Power Engineering (2002)

[16] Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)

[17] Hartley, R., Zisserman, A.: Multiple View Geometry. Cambridge University Press, Cambridge (2004)

[18] Holliman, N.: Mapping perceived depth to regions of interest in stereoscopic images. In: Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems XI (2004)

[19] Tran, V. M.: New methods for rendering of anaglyph stereoscopic images on CRT displays and photo-quality ink-jet printers. Ottawa-Carleton Institute for Electrical and Computer Engineering, SITE (2005)

[20] Stenger, B., Thayananthan, A., Torr, P., Cipolla, R.: Model-based hand tracking using a hierarchical Bayesian filter. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1372–1384 (2006)

[21] Yang, R., Sarkar, S.: Gesture Recognition using Hidden Markov Models from Fragmented Observations. In: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2006)

[22] Mcallister, D.F., Zhou, Y., Sullivan, S.: Methods for computing color anaglyphs (2010)

Statistical Tuning of Adaptive-Weight Depth Map Algorithm

Alejandro Hoyos[1], John Congote[1,2], Iñigo Barandiaran[2],
Diego Acosta[3], and Oscar Ruiz[1]

[1] CAD CAM CAE Laboratory, EAFIT University, Medellin, Colombia
{ahoyossi,oruiz}@eafit.edu.co
[2] Vicomtech Research Center, Donostia-San Sebastián, Spain
{jcongote,ibarandiaran}@vicomtech.org
[3] DDP Research Group, EAFIT University, Medellin, Colombia
dacostam@eafit.edu.co

Abstract. In depth map generation, the settings of the algorithm parameters to yield an accurate disparity estimation are usually chosen empirically or based on unplanned experiments. A systematic statistical approach including classical and exploratory data analyses on over 14000 images to measure the relative influence of the parameters allows their tuning based on the number of `bad_pixels`. Our approach is systematic in the sense that the heuristics used for parameter tuning are supported by formal statistical methods. The implemented methodology improves the performance of dense depth map algorithms. As a result of the statistical based tuning, the algorithm improves from 16.78% to 14.48% `bad_pixels` rising 7 spots as per the Middlebury Stereo Evaluation Ranking Table. The performance is measured based on the distance of the algorithm results vs. the Ground Truth by Middlebury. Future work aims to achieve the tuning by using significantly smaller data sets on fractional factorial and surface-response designs of experiments.

Keywords: Stereo Image Processing, Parameter Estimation, Depth Map.

1 Introduction

Depth map calculation deals with the estimation of multiple object depths on a scene. It is useful for applications like vehicle navigation, automatic surveillance, aerial cartography, passive 3D scanning, automatic industrial inspection, or 3D videoconferencing [1]. These maps are constructed by generating, at each pixel, an estimation of the distance between the screen and the object surface (depth).

Disparity is commonly used to describe inverse depth in computer vision, and also to measure the perceived spatial shift of a feature observed from close camera viewpoints. Stereo correspondence techniques often calculate a disparity function $d(x, y)$ relating target and reference images, so that the (x, y) coordinates of the disparity space match the pixel coordinates of the reference image. Stereo methods commonly use a pair of images taken with known camera geometry to

A. Berciano et al. (Eds.): CAIP 2011, Part II, LNCS 6855, pp. 563–572, 2011.

generate a dense disparity map with estimates at each pixel. This dense output is useful for applications requiring depth values even in difficult regions like occlusions and textureless areas. The ambiguity of matching pixels in heavy textured or textureless zones tends to require complex and expensive overall image processing or statistical correlations using color and proximity measures in local support windows.

Most implementations of vision algorithms make assumptions about the visual appearance of objects in the scene to ease the matching problem. The steps generally taken to compute the depth maps may include: (i) matching cost computation, (ii) cost or support aggregation, (iii) disparity computation or optimization, and (iv) disparity refinement.

This article is based on work done in [1] where the principles of the stereo correspondence techniques and the quantitative evaluator are discussed. The literature review is presented in section 2, followed by section 3 describing the algorithm, filters, statistical analysis and experimental set up. Results and discussions are covered in section 4, and the article is concluded in section 5.

2 Literature Review

The algorithm and filters use several user-specified parameters to generate the depth map of an image pair, and their settings are heavily influenced by the evaluated data sets [2]. Published works usually report the settings used for their specific case studies without describing the procedure followed to fine-tune them [3,4,5], and some explicitly state the empirical nature of these values [6]. The variation of the output as a function of several settings on selected parameters is briefly discussed while not taking into account the effect of modifying them all simultaneously [3,2,7]. Multiple stereo methods are compared choosing values based on experiments, but only some algorithm parameters are changed not detailing the complete rationale behind the value setting [1].

2.1 Conclusions of the Literature Review

Commonly used approaches in determining the settings of depth map algorithm parameters show all or some of the following shortcomings: (i) undocumented procedures for parameter setting, (ii) lack of planning when testing for the best settings, and (iii) failure to consider interactions of changing all the parameters simultaneously.

As a response to these shortcomings, this article presents a methodology to fine-tune user-specified parameters on a depth map algorithm using a set of images from the adaptive weight implementation in [4]. Multiple settings are used and evaluated on all parameters to measure the contribution of each parameter to the output variance. A quantitative accuracy evaluation allows using main effects plots and analyses of variance on multi-variate linear regression models to select the best combination of settings for each data set. The initial results are improved by setting new values of the user-specified parameters, allowing the algorithm to give much more accurate results on any rectified image pair.

3 Methodology

3.1 Image Processing

In the adaptive weight algorithm ([3,4]), a window is moved over each pixel on every image row, calculating a measurement based on the geometric proximity and color similarity of each pixel in the moving window to the pixel on its center. Pixels are matched on each row based on their support measurement with larger weights coming from similar pixel colors and closer pixels. The horizontal shift, or disparity, is recorded as the depth value, with higher values reflecting greater shifts and closer proximity to the camera.

The strength of grouping by color ($f_s\left(c_p, c_q\right)$) for pixels p and q is defined as the Euclidean distance between colors (Δc_{pq}) by Equation (1). Similarly, grouping strength by distance ($f_p\left(g_p, g_q\right)$) is defined as the Euclidean distance between pixel image coordinates (Δg_{pq}) by Equation (2). Where γ_c and γ_p are adjustable settings used to scale the measured color delta and window size respectively.

$$f_s\left(c_p, c_q\right) = exp\left(-\frac{\Delta c_{pq}}{\gamma_c}\right) \tag{1}$$

$$f_p\left(g_p, g_q\right) = exp\left(-\frac{\Delta g_{pq}}{\gamma_p}\right) \tag{2}$$

The matching cost between pixels shown in Equation (3) is measured by aggregating raw matching costs, using the support weights defined by Equations (1) and (2), in support windows based on both the reference and target images.

$$E\left(p, \bar{p}_d\right) = \frac{\sum_{q\in N_p, \bar{q}_d \in N_{\bar{p}_d}} w\left(p,q\right) w\left(\bar{p}_d, \bar{q}_d\right) \sum_{c\in\{r,g,b\}} \left|I_c\left(q\right) - I_c\left(\bar{q}_d\right)\right|}{\sum_{q\in N_p, \bar{q}_d \in N_{\bar{p}_d}} w\left(p,q\right) w\left(\bar{p}_d, \bar{q}_d\right)} \tag{3}$$

where $w\left(p,q\right) = f_s\left(c_p, c_q\right) \cdot f_p\left(g_p, g_q\right)$, $\bar{p}_d$ and $\bar{q}_d$ are the target image pixels at disparity d corresponding to pixels p and q in the reference image, I_c is the intensity on channels red (r), green (g), and blue (b), and N_p is the window centered at p and containing all q pixels. The size of this movable window N is another user-specified parameter. Increasing the window size reduces the chance of bad matches at the expense of missing relevant scene features.

Post-Processing Filters. Algorithms based on correlations depend heavily on finding similar textures at corresponding points in both reference and target images. Bad matches happen more frequently in textureless regions, occluded zones, and areas with high variation in disparity. The winner takes all approach enforces uniqueness of matches only for the reference image in such a way that points on the target image may be matched more than once, creating the need to check the disparity estimates and fill any gaps with information from neighboring pixels using post-processing filters like the ones shown in Table 1.

Table 1. User-specified parameters of the adaptive weight algorithm and filters

Filter	**Function**	**User-specified parameter**
Adaptive Weight [3]	Disparity estimation and pixel matching	γ_{aws}: similarity factor, γ_{awg}: proximity factor related to the W_{AW} pixel size of the support window
Median	Smoothing and incorrect match removal	W_M: pixel size of the median window
Cross-check[8]	Validation of disparity measurement per pixel	Δ_d: allowed disparity difference
Bilateral[9]	Intensity and proximity weighted smoothing with edge preservation	γ_{bs}: similarity factor, γ_{bg}: proximity factor related to the W_B pixel size of the bilateral window

Median Filter. They are widely used in digital image processing to smooth signals and to remove incorrect matches and holes by assigning neighboring disparities at the expense of edge preservation. The median filter provides a mechanism for reducing image noise, while preserving edges more effectively than a linear smoothing filter. It sorts the intensities of all the q pixels on a window of size M and selects the median value as the new intensity of the p central pixel. The size M of the window is another of the user-specified parameters.

Cross-check Filter. The correlation is performed twice by reversing the roles of the two images and considering valid only those matches having similar depth measures at corresponding points in both steps. The validity test is prone to fail in occluded areas where disparity estimates will be rejected. The allowed difference in disparities is one more adjustable parameter.

Bilateral Filter. Is a non-iterative method of smoothing images while retaining edge detail. The intensity value at each pixel in an image is replaced by a weighted average of intensity values from nearby pixels. The weighting for each pixel q is determined by the spatial distance from the center pixel p, as well as its relative difference in intensity, defined by Equation (4).

$$O_p = \frac{\sum_{q \in W} f_s\left(q-p\right) g_i\left(I_q - I_p\right) I_q}{\sum_{q \in W} f_s\left(q-p\right) g_i\left(I_q - I_p\right)} \tag{4}$$

where O is the output image, I the input image, W the weighting window, f_s the spatial weighing function, and g_i the intensity weighting function. The size of the window W is yet another parameter specified by the user.

3.2 Statistical Analysis

The user-specified input parameters and output accuracy measurements data is statistically analyzed measuring the relations amongst inputs and outputs with correlation analyses, while box plots give insight on the influence of groups of settings on a given factor. A multi-variate linear regression model shown in Equation (5) relates the output variable as a function of all the parameters to find the equation coefficients, correlation of determination, and allows the analysis of variance to measure the influence of each parameter on the output variance. Residual analyses are checked to validate the assumptions of the regression model like constant error variance, and mean of errors equal to zero, and if necessary, the model is transformed. The parameters are normalized to fit the range $(-1, 1)$ as shown in Table 2.

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon \tag{5}$$

where $\hat{y}$ is the predicted variable, x_i are the factors, and β_i are the coefficients.

3.3 Experimental Set Up

The depth maps are calculated with an implementation developed for real time videoconferencing in [4]. Using well-known rectified image sets: Cones from [1], Teddy and Venus from [10], and Tsukuba head and lamp from the University of Tsukuba. Other commonly used sets are also freely available [11,12]. The sample used consists of 14688 depth maps, 3672 for each data set, like the ones shown in Figure 1.

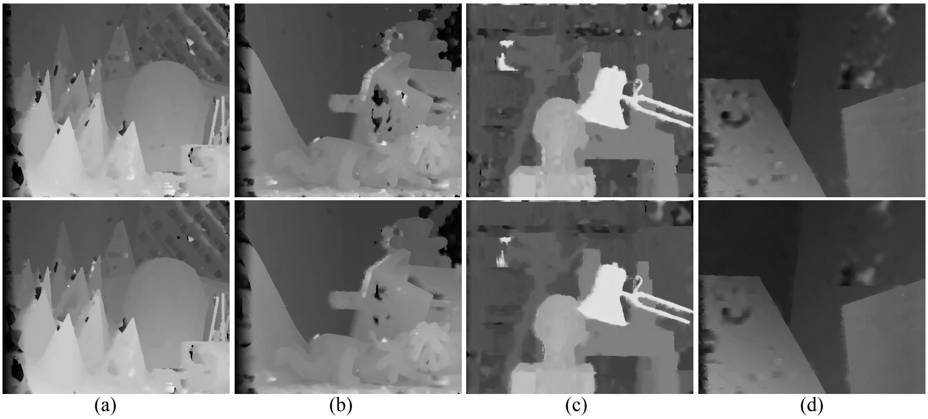


Fig. 1. Depth Map Comparison. Top: best initial, bottom: new settings. (a) Cones, (b) Teddy, (c) Tsukuba, and (d) Venus data set.

Many recent stereo correspondence performance studies use the Middlebury Stereomatcher for their quantitative comparisons [2,7,13]. The evaluator code, sample scripts, and image data sets are available from the Middlebury stereo vision site[1], providing a flexible and standard platform for easy evaluation.

[1] http://vision.middlebury.edu/stereo/

Table 2. User-specified parameters of the adaptive weight algorithm

Parameter	**Name**	**Levels**	**Values**	**Coding**
Adaptive Weights Window Size	*aw_win*	4	[1 3 5 7]	[-1 -0.3 0.3 1]
Adaptive Weights Color Factor	*aw_col*	6	[4 7 10 13 16 19]	[-1 -0.6 -0.2 0.2 0.6 1]
Median Window Size	*m_win*	3	[*N/A* 3 5]	[*N/A* -1 0.2 1]
Cross-Check Disparity Delta	*cc_disp*	4	[*N/A* 0 1 2]	[*N/A* -1 0 1]
Cross-Bilateral Window Size	*cb_win*	5	[*N/A* 1 3 5 7]	[*N/A* -1 -0.3 0.3 1]
Cross-Bilateral Color Factor	*cb_col*	7	[*N/A* 4 7 10 13 16 19]	[*N/A* -1 -0.6 -0.2 0.2 0.6 1]

The online Middlebury Stereo Evaluation Table gives a visual indication of how well the methods perform with the proportion of bad pixels (`bad_pixels`) metric defined as the average of the proportion of bad pixels in the whole image (`bad_pixels_all`), the proportion of bad pixels in non-occluded regions (`bad_pixels_nonocc`), and the proportion of bad pixels in areas near depth discontinuities (`bad_pixels_discont`) in all data sets.

4 Results and Discussion

4.1 Variable Selection

Pearson correlation of the factors show that they are independent and that each one must be included in the evaluation. On the other hand, a strong correlation amongst `bad_pixels` and the other outputs is detected and shown in Figure 2. This allows the selection of `bad_pixels` as the sole output because the other responses are expected to follow a similar trend. Other output are explain in the Table 3.

Table 3. Result metrics computed by the Middlebury Stereomatcher evaluator

Parameter	**Description**
rms_error_all	Root Mean Square (RMS) disparity error (all pixels)
rms_error_nonocc	RMS disparity error (non-occluded pixels only)
rms_error_occ	RMS disparity error (occluded pixels only)
rms_error_textured	RMS disparity error (textured pixels only)
rms_error_textureless	RMS disparity error (textureless pixels only)
rms_error_discont	RMS disparity error (near depth discontinuities)
bad_pixels_all	Fraction of bad points (all pixels)
bad_pixels_nonocc	Fraction of bad points (non-occluded pixels only)
bad_pixels_occ	Fraction of bad points (occluded pixels only)
bad_pixels_textured	Fraction of bad points (textured pixels only)
bad_pixels_textureless	Fraction of bad points (textureless pixels only)
bad_pixels_discont	Fraction of bad points (near depth discontinuities)
evaluate_only	Read specified depth map and evaluate only
output_params	Text file logging all used parameters
depth_map	Evaluated image

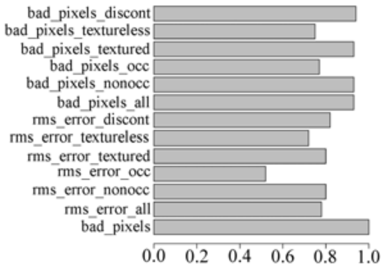
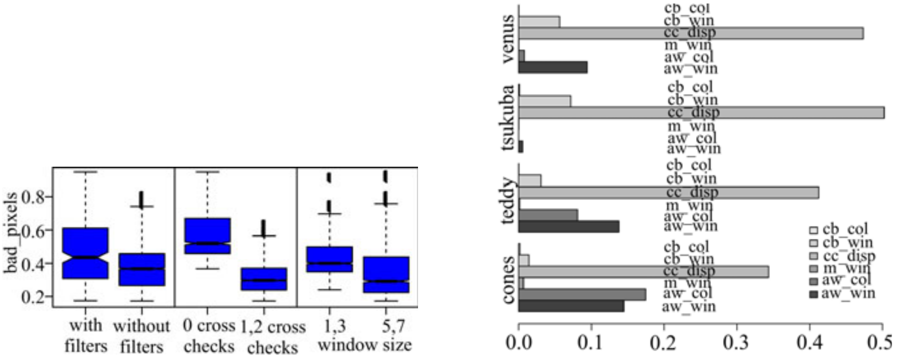


Fig. 2. `bad_pixels` and other output correlation

4.2 Exploratory Data Analysis

Box plots analysis of `bad_pixels` presented in Figure 3(a) show lower output values from using filters, relaxed cross-check disparity delta values, large adaptive weight window sizes, and large adaptive weight color factor values. The median window size, bilateral window size, and bilateral window color values do not show a significant influence on the output at the studied levels.

The influence of the parameters is also shown on the slopes of the main effects plots of Figure 4 and confirms the behavior found with the ANOVA of the multi-variate linear regression model. The settings to lower `bad_pixels` from this analysis yields a result of 14.48%.



(a) Box Plots (b) ANOVA proportion of `bad_pixels`

Fig. 3. (a) Box Plots of `bad_pixels`. (b) Contribution to the `bad_pixels` variance by parameter.

4.3 Multi-variate Linear Regression Model

The analysis of variance on a multi-variate linear regression (MVLR) over all data sets using the most parsimonious model quantifies the parameters with the most influence as shown in Figure 3(b). *cc_disp* is the most significant factor accounting for a third to a half of the variance on every case.

Interactions and higher order terms are included on the multi-variate linear regression models to improve the goodness of fit. Reducing the number of input images per dataset from 3456 to 1526 by excluding the worst performing cases corresponding to $cc\_disp = 0$ and $aw\_col = [4, 7]$, allows using a cubic model with interactions and an R^2 of 99.05%.

The residuals of the selected model fail to follow a normal distribution. Transforming the output variable or removing large residuals does not improve the residuals distribution, and there are no reasons to exclude any outliers from the image data set. Nonetheless, improved algorithm performance settings are found using the model to obtain lower bad_pixels values comparable to the ones obtained through the exploratory data analysis (14.66% vs. 14.48%).

In summary, the most noticeable influence on the output variable comes from having a relaxed cross-check filter, accounting for nearly half the response variance in all the study data sets. Window size is the next most influential factor, followed by color factor, and finally window size on the bilateral filter. Increasing the window sizes on the main algorithm yield better overall results at the expense of longer running times and some foreground loss of sharpness, while the support weights on each pixel have the chance of becoming more distinct and potentially reduce disparity mismatches. Increasing the color factor on the main algorithm allows better results by reducing the color differences, and slightly compensating minor variations in intensity from different viewpoints.

A small median smoothing filter window size is faster than a larger one, while still having a similar accuracy. Low settings on both the window size and the color factor on the bilateral filter seem to work best for a good balance between performance and accuracy.

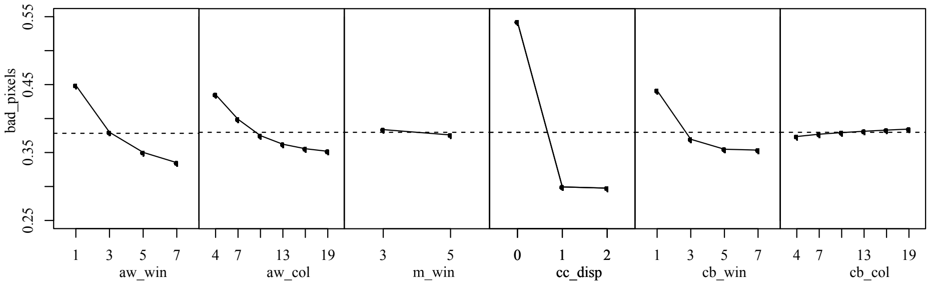


Fig. 4. Main Effects Plots of each factor level for all data sets. Steeper slopes relate to bigger influence on the variance of the bad_pixels output measurement.

The optimal settings in the original data set are presented in Table 4 along with the proposed combinations. **Low settings** comprise the depth maps with all their parameter settings at each of their minimum tested values yielding 67.62% bad_pixels. **High settings** relates to depth maps with all their parameter settings at each of their maximum tested values yielding 19.84% bad_pixels. **Best initial** are the most accurate depth maps from the study data set yielding

Table 4. Model comparison. Average bad_pixels values over all data sets and their parameter settings.

Run Type	bad_pixels	aw_win	aw_col	m_win	cc_disp	cb_win	cb_col
Low Settings	67.62%	1	4	3	0	1	4
High Settings	19.84%	7	19	5	2	7	19
Best Initial	16.78%	7	19	5	1	3	4
Exploratory analysis	14.48%	9	22	5	1	3	4
MVLR optimization	14.66%	11	22	5	3	3	18

16.78% bad_pixels. **Exploratory analysis** corresponds to the settings determined using the exploratory data analysis based on box plots and main effects plots yielding 14.48% bad_pixels. **MVLR optimization** is the extrapolation optimization of the classical data analysis based on multi-variate linear regression model, nested models, and ANOVA yielding 14.66% bad_pixels.

The exploratory analysis estimation and the MVLR optimization tend to converge at similar lower bad_pixels values using the same image data set. The best initial and improved depth map outputs are shown in Figure 1.

5 Conclusions and Future Work

This work presents a systematic methodology to measure the relative influence of the inputs of a depth map algorithm on the output variance and the identification of new settings to improve the results from 16.78% to 14.48% bad_pixels. The methodology is applicable on any group of depth map image sets generated with an algorithm where the relative influence of the user-specified parameters merits to be assessed.

Using design of experiments reduces the number of depth maps needed to carry out the study when a large image database is not available. Further analysis on the input factors should be started with exploratory experimental fractional factorial designs comprising the full range on each factor, followed by a response surface experimental design and analysis. In selecting the factor levels, analyzing the influence of each filter independently would be an interesting criterion.

Acknowledgments. This work has been partially supported by the Spanish Administration Agency CDTI under project CENIT-VISION 2007-1007, the Colombian Administrative Department of Science, Technology, and Innovation; and the Colombian National Learning Service (COLCIENCIAS-SENA) grant No. 1216-479-22001.

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vision 47(1-3), 7–42 (2002)
2. Gong, M., Yang, R., Wang, L., Gong, M.: A performance study on different cost aggregation approaches used in real-time stereo matching. Int. J. Comput. Vision 75, 283–296 (2007)

3. Yoon, K., Kweon, I.: Adaptive support-weight approach for correspondence search. IEEE Trans. Pattern Anal. Mach. Intell. 28(4), 650 (2006)
4. Congote, J., Barandiaran, I., Barandiaran, J., Montserrat, T., Quelen, J., Ferrán, C., Mindan, P., Mur, O., Tarrés, F., Ruiz, O.: Real-time depth map generation architecture for 3d videoconferencing. In: 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010, pp. 1–4 (2010)
5. Gu, Z., Su, X., Liu, Y., Zhang, Q.: Local stereo matching with adaptive support-weight, rank transform and disparity calibration. Pattern Recogn. Lett. 29, 1230–1235 (2008)
6. Hosni, A., Bleyer, M., Gelautz, M., Rhemann, C.: Local stereo matching using geodesic support weights. In: Proceedings of the 16th IEEE Int. Conf. on Image Processing (ICIP), pp. 2093–2096 (2009)
7. Wang, L., Gong, M., Gong, M., Yang, R.: How far can we go with local optimization in real-time stereo matching. In: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT 2006), pp. 129–136 (2006)
8. Fua, P.: A parallel stereo algorithm that produces dense depth maps and preserves image features. Machine Vision and Applications 6(1), 35–49 (1993)
9. Weiss, B.: Fast median and bilateral filtering. ACM Trans. Graph. 25, 519–526 (2006)
10. Scharstein, D., Szeliski, R.: High-accuracy stereo depth maps using structured light. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 195–202 (2003)
11. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
12. Hirschmuller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
13. Tombari, F., Mattoccia, S., Di Stefano, L., Addimanda, E.: Classification and evaluation of cost aggregation methods for stereo correspondence. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)

Author Index