

Reinhard Koch  
Fay Huang (Eds.)

LNCS 6469

# Computer Vision – ACCV 2010 Workshops

ACCV 2010 International Workshops  
Queenstown, New Zealand, November 2010  
Revised Selected Papers, Part II

2 Part II

 Springer

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Alfred Kobsa

*University of California, Irvine, CA, USA*

Friedemann Mattern

*ETH Zurich, Switzerland*

John C. Mitchell

*Stanford University, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

Oscar Nierstrasz

*University of Bern, Switzerland*

C. Pandu Rangan

*Indian Institute of Technology, Madras, India*

Bernhard Steffen

*TU Dortmund University, Germany*

Madhu Sudan

*Microsoft Research, Cambridge, MA, USA*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbruecken, Germany*

Reinhard Koch Fay Huang (Eds.)

# Computer Vision – ACCV 2010 Workshops

ACCV 2010 International Workshops  
Queenstown, New Zealand, November 8-9, 2010  
Revised Selected Papers, Part II

Volume Editors

Reinhard Koch  
Christian-Albrechts-University Kiel  
Computer Science Institute  
Olshausenstr. 40  
24098 Kiel, Germany  
E-mail: rk@informatik.uni-kiel.de

Fay Huang  
National Ilan University  
Institute of Computer Science and Information Engineering  
Shen-Lung Rd. 1  
26047 Yi-Lan, Taiwan R.O.C.  
E-mail: fay@niu.edu.tw

ISSN 0302-9743 e-ISSN 1611-3349  
ISBN 978-3-642-22818-6 e-ISBN 978-3-642-22819-3  
DOI 10.1007/978-3-642-22819-3  
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011936637

CR Subject Classification (1998): I.4, I.5, I.2.10, I.2.6, I.3.5, F.2.2, H.5, J.5

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

© Springer-Verlag Berlin Heidelberg 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

*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](http://www.springer.com))



# Preface

During ACCV 2010 in Queenstown, New Zealand, a series of eight high-quality workshops were held that reflect the full range of recent research topics in computer vision. The workshop themes ranged from established research areas like visual surveillance (the 10th edition) and subspace methods (third edition) to innovative vehicle technology (From Earth to Mars), from vision technology for world e-heritage preservation and mixed and augmented reality to aesthetic features in computational photography and human computer interaction.

From a total of 167 submissions, 89 presentations were selected by the individual workshop committees, yielding an overall acceptance rate of 53%. The reported attendance was quite attractive, between 40 and 60 participants in each of the workshops, sometimes over 70.

The two-volume proceedings contain a short introduction to each workshop, followed by all workshop contributions arranged according to the workshops.

We hope that you will enjoy reading the contributions which may inspire you to further research.

November 2010

Reinhard Koch  
Fay Huang

# Introduction to the 10th International Workshop on Visual Surveillance

Visual surveillance remains a challenging application area for computer vision. The large number of high-quality submissions is a testament to the continuing attention it attracts from research groups around the world. Within this area, the segmentation of the foreground (moving objects) from the background (residual scene) remains a core problem. Approximately half of the papers accepted for publication propose innovative segmentation processes. These include the modeling of photometric variations using local polynomials, the exploitation of geometric and temporal constraints, and the explicit modeling of foreground properties. The segmentation of foregrounds consisting of slowly moving objects is explored and there are two investigations into the improvements in segmentation that can be obtained using feedback from a subsequent tracking process.

Nonetheless, there is also an increasing interest in the detection of pedestrians, faces and vehicles using methods that do not rely on foreground-background segmentation. Several enhancements to the histogram of gradients method for pedestrian detection are proposed, leading to an improved efficiency and invariance of the results under rotations of the image. A method to improve the efficiency of the boosted cascade classifier is also proposed. A key problem for visual surveillance scene understanding is the tracking of pedestrians in arbitrarily crowded scenes across multiple cameras: there are several papers that offer contributions to the solution of this problem, including the modeling of pedestrian appearance as observed from multiple cameras in a network.

In the 12 years in which the Visual Surveillance workshops have been running, algorithms have become more sophisticated and more effective, more data sets have become available and experimental techniques and the reporting of results have improved. In spite of these advances, many of the classic problems in computer vision, such as optic flow estimation, object detection and object recognition, are still as relevant to the visual surveillance community as they have ever been.

The Workshop Chairs would like to thank the Program Committee for their valuable input into the reviewing process, and Reinhard Koch and Fay Huang for providing efficient liaison on behalf of the ACCV. The Chairs would also like to thank Graeme Jones, who dealt with many of the organizational aspects of this workshop.

November 2010

James Orwell  
Steve Maybank  
Tieniu Tan

## Program Committee

Francois Bremond	INRIA Sophia-Antipolis Research Unit, France
Andrea Cavallaro	Queen Mary, University of London, UK
Patrick Courtney	PerkinElmer Life and Analytical Sciences, UK
Roy Davies	Royal Holloway, University of London, UK
Rogério Feris	IBM Research, USA
Gustavo Fernandez Dominguez	AIT Austrian Institute of Technology Gmb, Austria
Gian Luca Foresti	University of Udine, Italy
Xiang Gao	Siemens Corporate Research, USA
Shaogang Gong	Queen Mary University London, UK
Riad Hammoud	Delphi Corporation, USA
R. Ismail Haritaoglu	Polar Rain Inc, USA
Janne Heikkila	Dept. of Electrical Engineering, Finland
Wei Ming Hu	NLPR, China
Kaiqi Huang	Institute of Automation CAS, China
Graeme Jones	DIRC, Kingston University, UK
Kyoung Mu Lee	Seoul National University, Korea
Peihua Li	Hei Long Jiang University, China
Stan Li	National Laboratory of Pattern Recognition, China
Xuelong Li	Birkbeck College, University of London, UK
Dimitrios Makris	Kingston University, UK
Steve Maybank	Birkbeck College, UK
James Orwell	Kingston University, UK
Vasudev Parameswaran	Siemens Corporate Research, USA
Federico Pernici	Università di Firenze, Italy
Justus Piater	Université de Liège, Belgium
Massimo Piccardi	University of Technology, Sydney, Australia
Ian Reid	University of Oxford, UK
Paolo Remagnino	Kingston University, UK
Gerhard Rigoll	Munich University of Technology, Germany
Neil Robertson	Heriot-Watt University, UK
Gerald Schaefer	Loughborough University, UK
Vinay Shet	Siemens Corporate Research, USA
Nils T Siebel	HTW University of Applied Sciences, Germany
Lauro Snidaro	Università degli Studi di Udine, Italy
Zoltan Szlavik	MTA SzTAKI, Hungary
Tieniu Tan	National Laboratory of Pattern Recognition, China
Dacheng Tao	Birkbeck College, University of London, UK

Geoffrey Taylor  
Stefano Tubaro  
Sergio Velastin  
Roberto Vezzani

Ramesh Visvanathan  
Guangyou Xu  
Wei Yun Yau

Fei Yin

ObjectVideo, Inc., USA  
Politecnico di Milano, Italy  
Kingston University, UK  
D.I.I. - University of Modena and R.E.,  
Modena, Italy  
Siemens Corporate Research, USA  
Tsinghua University, China  
Nanyang Technological University,  
Singapore  
Digital imaging Research Centre, UK

# Introduction to the Second International Workshop on Video Event Categorization, Tagging and Retrieval (VECTaR)

One of the remarkable capabilities of the human visual perception system is to interpret and recognize thousands of events in videos, despite a high level of video object clutter, different types of scene context, variability of motion scales, appearance changes, occlusions and object interactions. As an ultimate goal of computer vision systems, the interpretation and recognition of visual events is one of the most challenging problems and has increasingly become very popular in the last few decades. This task remains exceedingly difficult because of several reasons:

1. There still remain large ambiguities in the definition of different levels of events.
2. A computer model should be capable of capturing a meaningful structure for a specific event. At the same time, the representation (or recognition process) must be robust under challenging video conditions.
3. A computer model should be able to understand the context of video scenes to have meaningful interpretation of a video event. Despite these difficulties, in recent years steady progress has been made toward better models for video event categorization and recognition, e.g., from modeling events with a bag of spatial temporal features to discovering event context, from detecting events using a single camera to inferring events through a distributed camera network, and from low-level event feature extraction and description to high-level semantic event classification and recognition.

This workshop served to provide a forum for recent research advances in the area of video event categorization, tagging and retrieval. A total of 11 papers were selected for publication, dealing with theories, applications and databases of visual event recognition.

November 2010

Ling Shao  
Jianguo Zhang  
Tieniu Tan  
Thomas S. Huang

## Program Committee

Faisal Bashir  
Xu Chen  
Ling-Yu Duan  
GianLuca Foresti  
Kaiqi Huang  
Thomas S. Huang

Yu-Gang Jiang  
Graeme A. Jones  
Ivan Laptev  
Jianmin Li  
Xuelong Li  
Zhu Li  
Xiang Ma  
Paul Miller  
Shin'ichi Satoh  
Ling Shao  
Peter Sturm  
Tieniu Tan  
Xin-Jing Wang  
Tao Xiang  
Jian Zhang  
Jianguo Zhang

Heartland Robotics, USA  
University of Michigan, USA  
Peking University, China  
University of Udine, Italy  
Chinese Academy of Sciences, China  
University of Illinois at Urbana-Champaign,  
USA  
City University of Hong Kong, China  
Kingston University, UK  
INRIA, France  
Tsinghua University, China  
Chinese Academy of Sciences, China  
Hong Kong Polytechnic University, China  
IntuVision, USA  
Queen's University Belfast, UK  
National Institute of Informatics, Japan  
The University of Sheffield, UK  
INRIA, France  
Chinese Academy of Sciences, China  
Microsoft Research Asia, China  
Queen Mary University London, UK  
Chinese Academy of Sciences, China  
Queen's University Belfast, UK

# Introduction to the Workshop on Gaze Sensing and Interactions

The goal of this workshop is to bring researchers from academia and industry in the field of computer vision and other closely related fields such as robotics and human – computer interaction together to share recent advances and discuss future research directions and opportunities for gaze sensing technologies and their applications to human – computer interactions and human – robot interactions. The workshop included two keynote speeches by Ian Reid at the University of Oxford, UK, and Chen Yu at Indiana University, USA, who are world-leading experts on gaze – sensing technologies and their applications for interactions, and seven oral presentations selected from submitted papers by blind review. This workshop was supported by the Japan Science and Technology Agency (JST) and CREST. We would like to thank Yusuke Sugano, Yoshihiko Mochizuki and Sakie Suzuki for their support in organizing this event.

November 2010

Yoichi Sato  
Akihiro Sugimoto  
Yoshihiro Kuno  
Hideki Koike

## Program Committee

Andrew T. Duchowski  
Shaogang Gong  
Qiang Ji  
Kris Kitani

Yoshinori Kobayashi  
Yukie Nagai  
Takahiro Okabe  
Kazuhiro Otsuka

Ian Reid  
Yusuke Sugano  
Yasuyuki Sumi  
Roel Vertegaal

Clemson University, USA  
Queen Mary, University of London, UK  
Rensselaer Polytechnic Institute, USA  
The University of Electro-Communications,  
Japan  
Saitama University, Japan  
Osaka University, Japan  
University of Tokyo, Japan  
Nippon Telegraph and Telephone  
Corporation, Japan  
University of Oxford, UK  
University of Tokyo, Japan  
Kyoto University, Japan  
Queen's University, Canada

# Introduction to the Workshop on Application of Computer Vision for Mixed and Augmented Reality

The computer vision community has already provided numerous technical breakthroughs in the field of mixed reality and augmented reality (MR/AR), particularly in camera tracking, human behavior understanding, object recognition, etc. The way of designing an MR/AR system based on computer vision research is still a difficult research and development issue. This workshop focuses on the recent trends in applications of computer vision to MR/AR systems.

We were proud to organize the exciting and stimulating technical program consisting of ten oral presentations and five poster presentations. We were very happy to have a distinguished invited speaker, Hideyuki Tamura, who has led the MR/AR research field since the 1990s. Finally, we would like to thank all of the authors who kindly submitted their research achievements to ACVMAR 2010 and all members of the Program Committee for their voluntarily efforts.

ACVMAR 2010 organized in collaboration with SIG-MR(VRSJ) and the GCOE Program at Keio University.

November 2010

Hideo Saito  
Masayuki Kanbara  
Itaru Kitahara  
Yuko Uematsu



## Program Committee

Toshiyuki Amano	NAIST, Japan
Jean-Yves Guillemaut	University of Surrey, UK
Ryosuke Ichikari	Ritsumeikan University, Japan
Sei Ikeda	NAIST, Japan
Daiske Iwai	Osaka University, Japan
Yoshinari Kameda	University of Tsukuba, Japan
Hansung Kim	University of Surrey, UK
Kiyoshi Kiyokawa	Osaka University, Japan
Takeshi Kurata	AIST, Japan
Vincent Lepetit	EPFL , Switzerland
Walterio Mayol-Cuevas	University of Bristol, UK
Jong-Il Park	Hanyang University, Korea
Gerhard Reitmayr	TU Graz, Austria
Chiristian Sandor	South Australia University, Australia
Tomokazu Sato	NAIST, Japan
Fumihisa Shibata	Ritsumeikan University, Japan
Ryuhei Tenmoku	AIST, Japan
Yuki Uranishi	NAIST, Japan
Daniel Wagner	TU Graz, Austria
Woontack Woo	GIST, Korea

# Introduction to the Workshop on Computational Photography and Aesthetics

Computational photography is now well-established as a field of research that examines what lies beyond the conventional boundaries of digital photography. The newer field of computational aesthetics has seen much interest within the realm of computer graphics, art history and cultural studies. This workshop is intended to provide an opportunity for researchers working in both areas, photography as well as aesthetics, to meet and discuss their ideas in a collegial and interactive format.

The papers contained in these workshop proceedings make important contributions to our understanding of computational aspects of photography and aesthetics. The first paper, by Valente and Klette, describes a technique for blending artistic filters together. Their method allows users to define their own painting style, by choosing any point within the area of a triangle whose vertices represent pointillism, curved strokes, and glass patterns. The second paper, by Sachs, Kakarala, Castleman, and Rajan, describes a study of photographic skill whose purpose is to establish whether that skill can be identified in a double-blind manner. They show that human judges who are themselves expert photographers are able to identify up to four skill levels with statistical significance. The third paper, by Rigau, Feixas, and Sbert, applies the information theory of Shannon to model the channel between luminosity and composition. They show how changes in depth-of-field and exposure are reflected in the information channel, and formulate measures for saliency and “entanglement” in an image. The fourth paper, by Lo, Shih, Liu, and Hong, describes how computer vision may be applied to detect a classic error in photographic composition: objects which appear to protrude from a subject’s head. Their method is able to reliably detect protruding objects in a variety of lighting conditions and backgrounds, with a detection rate of 87% and false alarm rate of 12%. The fifth paper, by Constable, shows how traditional drawing methods such as incomplete perimeters, lines that suggest colors, and lines that suggest form, can inform and improve non-photorealistic rendering (NPR). This paper provides a valuable artistic perspective to illustrate how engineering and art work collaboratively in NPR.

The workshop was fortunate to have a keynote presentation by Alfred Bruckstein. He described the problem of emulating classic engraving using non-photorealistic image rendering, and proposed to use level-set-based shape from shading techniques. The problem contains interesting mathematical challenges in connecting essential contours in natural, flowing ways, which Professor Bruckstein described.

## Program Committee

Todd Sachs

Xuemei Zhang

Shannon Lee Castleman

Deepu Rajan

Soon-Hwa Oh

Philip Ogunbona

Aptina Imaging, USA

HP Labs, USA

Nanyang Technological University, Singapore

Nanyang Technological University, Singapore

Nanyang Technological University, Singapore

University of Wollongong, Australia

# Introduction to the Workshop on Computer Vision in Vehicle Technology: From Earth to Mars

Vision-based autonomous navigation of vehicles has a long history which goes back to the success story of Dickmanns in Munich and the Mechanical Engineering Laboratory of MITI in Japan in the 1980th. At the time, DARPA had asked us to compete with autonomous land vehicles in their GRAND Challenges. Today, computer vision techniques provide methodologies to assist in long-distance exploration projects using visual sensing systems such those with the Mars rover project. Modern cars are now driven with the assistance of various sensor data. These assisted driving systems are developed as intelligent transportation systems. Among the various types of data used for driving assistance and navigation, we find visual information as the interface between human drivers and vehicles.

Today, data captured by visual sensors mounted on vehicles provide essential information used in intelligent driving systems. For applications of computer vision methodologies in exploration, evaluation, and quality-control techniques in the absence of ground truth information, it is essential to design robust and reliable algorithms.

In this workshop, we focus on exchanging new ideas on applications of computer vision theory to vehicle technology. In computer vision for driving assistance, tracking, reconstruction, and prediction become important concepts. Furthermore, real-time and on-board processes for these problems are required.

We received 21 papers and selected 11 papers for publication based on the reviews by the Program Committee and by the additional reviewer Ali Al-Sarraf.

November 2010

Steven Beauchemin  
Atsushi Imiya  
Tomas Pajdla

## Program Committee

Hanno Ackermann	Leibniz Universität Hannover, Germany
Yousun Kang	National Institute of Informatics, Japan
Kazuhiko Kawamoto	Chiba University, Japan
Lars Krüger	Daimler AG, Germany
Norbert Krüger	University of Southern Denmark, Denmark
Ron Li	The Ohio State University, USA
Yoshihiko Mochizuki	Chiba University, Japan
Hiroshi Murase	Nagoya University, Japan
Gerhard Paar	Joanneum Research, Austria
Bodo Rosenhahn	University Hannover, Germany
Jun Sato	Nagoya Institute of Technology, Japan
Akihiko Torii	Tokyo Institute of Technology, Japan
Tobi Vaudrey	The University of Auckland, New Zealand

# Introduction to the Workshop on e-Heritage

Digitally archived world heritage sites are broadening their value for preservation and access. Many valuable objects have been decayed by time due to weathering, natural disasters, even man-made disasters such as the Taliban destruction of the great Buddhas in Afghanistan, or the recent destruction by fire of a 600-year-old South Gate in Seoul. Cultural heritage also includes music, language, dance, and customs that are fast becoming extinct as the world moves toward a global village. Furthermore, most of the sites still face a problem of accessibility. Digital access projects are necessary to overcome those problems.

Computer vision research and practices have, and will continue, to play a central role in such cultural heritage preservation efforts. The proposed Workshop on e-Heritage and Digital Art Preservation aims to bring together computer vision researchers as well as interdisciplinary researchers that are related to computer vision, in particular computer graphics, image and audio research, image and haptic (touch) research, as well as presentation of visual content over the Web and education.

In this workshop, seven contributions to the field of e-heritage were presented, covering the areas of on-site augmented-reality applications, three-dimensional modeling and reconstruction, shape and image analysis, and interactive haptic systems. All submissions were double-blind reviewed by at least two experts. We thank all the authors who submitted their work. It was a special honor to have In So Kweon (KASIT, Korea), Hongbin Zha (Peking University, China) and Yasuyuki Matsushita (Microsoft Research Asia) as the invited speakers at the workshop. We are especially grateful to the members of the Program Committee for their remarkable efforts and the quality of the reviews.

November 2010

Katsushi Lkevchi  
Takeshi Oishi  
Rei Kawakami  
Michael S. Brown  
Moshe Ben-Ezra  
Ryusuke Sagawa

## Program Committee

Yasutaka Furukawa	Google, USA
Luc Van Gool	ETH Zurich, Switzerland
Yi Ping Hung	National Taiwan University, Taiwan
Asanobu Kitamoto	NII, Japan
In So Kweon	KAIST, Korea
Kok-Lim Low	NUS, Singapore
Yasuyuki Matsushita	MSRA, China
Daisuke Miyazaki	Hiroshima University, Japan
Tomokazu Sato	NAIST, Japan
David Suter	University of Adelaide, Australia
Jun Takamatsu	NAIST, Japan
Ping Tan	NUS, Singapore
Robby T. Tan	University of Utrecht, The Netherlands
Lior Wolf	Tel Aviv University, Israel
Toshihiko Yamasaki	University of Tokyo, Japan
Naokazu Yokoya	NAIST, Japan
Hongbin Zha	Peking University, China

# Introduction to the Third International Workshop on Subspace Methods

We welcome you to the proceedings of the Third International Workshop of Subspace 2010 held in conjunction with ACCV 2010.

Subspace 2010 was held in Queenstown, New Zealand, on November 9, 2010. For the technical program of Subspace 2010, a total of 30 full-paper submissions underwent a rigorous review process. Each of these submissions was evaluated in a double-blind manner by a minimum of two reviewers. In the end, ten papers were accepted and included in this volume of proceedings.

The goal of the workshop is to share the potential of subspace-based methods, such as the subspace methods, with researchers working on various problems in computer vision; and to encourage interactions which could lead to further developments of the subspace-based methods. The fundamental theories of subspace-based methods and their applications in computer vision were discussed at the workshop.

Subspace-based methods are important for solving many theoretical problems in pattern recognition and computer vision. Also they have been widely used as a practical methodology in a large variety of real applications. During the last three decades, the area has become one of the most successful underpinnings of diverse applications such as classification, recognition, pose estimation, motion estimation. At the same time, there are many new and evolving research topics: nonlinear methods including kernel methods, manifold learning, subspace update and tracking. In addition to regular presentations, to overview these developments, we provided a historical survey talk of the subspace methods.

Prior to this workshop, we successfully organized two international workshops on subspace-based methods: Subspace 2007 in conjunction with ACCV 2007 and Subspace 2009 in conjunction with ICCV 2009. We believe that Subspace 2010 stimulated fruitful discussions among the participants and provided novel ideas for future research in computer vision.

November 2010

David Suter  
Kazuhiro Fukui  
Toru Tamaki



## Program Committee

Toshiyuki Amano	NAIST, Japan
Horst Bischof	TU Graz, Austria
Seiji Hotta	Tokyo University of Agriculture and Technology, Japan
Masakazu Iwamura	Osaka Prefecture University, Japan
Tae-Kyun Kim	University of Cambridge, UK
Xi Li	Xi'an Jiaotong University, China
Yi Ma	University of Illinois at Urbana Champaign, USA
Atsuto Maki	Toshiba Cambridge Research Lab, UK
Shinichiro Omachi	Tohoku University, Japan
Bisser Raytchev	Hiroshima University, Japan
Peter Roth	TU Graz, Austria
Hitoshi Sakano	NTT CS Laboratories, Japan
Atsushi Sato	NEC, Japan
Yoichi Sato	The University of Tokyo, Japan
Shin'ichi Satoh	National Institute of Informatics, Japan
Terence Sim	National University of Singapore, Singapore
Bjorn Stenger	Toshiba Cambridge Research Lab, UK
Qi Tian	University of Texas at San Antonio, USA
Fernando De la Torre	Carnegie Mellon University, USA
Seiichi Uchida	Kyushu University, Japan
Osamu Yamaguchi	Toshiba, Japan
Jakob Verbeek	INRIA Rhône-Alpes, Grenoble, France
Jing-Hao Xue	University College London, UK

## Table of Contents – Part II

<b>Workshop on Application of Computer Vision for Mixed and Augmented Reality</b>	
Computer Vision Technology Applied to MR-Based Pre-visualization in Filmmaking . . . . .	1
<i>Hideyuki Tamura, Takashi Matsuyama, Naokazu Yokoya, Ryosuke Ichikari, Shohei Nobuhara, and Tomokazu Sato</i>	
Model Based Pose Estimation Using SURF . . . . .	11
<i>Peter Decker and Dietrich Paulus</i>	
Real-Time Camera Tracking Using a Global Localization Scheme . . . . .	21
<i>Yue Yiming, Liang Xiaohui, Liu Chen, and Liu Jie</i>	
Visual Mapping and Multi-modal Localisation for <i>Anywhere</i> AR Authoring . . . . .	31
<i>Andrew P. Gee, Andrew Calway, and Walterio Mayol-Cuevas</i>	
Augmented Reality System for Visualizing 3-D Region of Interest in Unknown Environment . . . . .	42
<i>Sei Ikeda, Yoshitsugu Manabe, and Kunihiko Chihara</i>	
Interactive Video Layer Decomposition and Matting . . . . .	52
<i>Yanli Li, Zhong Zhou, and Wei Wu</i>	
Removal of Moving Objects and Inconsistencies in Color Tone for an Omnidirectional Image Database . . . . .	62
<i>Maiya Hori, Hideyuki Takahashi, Masayuki Kanbara, and Naokazu Yokoya</i>	
Shape Prior Embedded Geodesic Distance Transform for Image Segmentation . . . . .	72
<i>Junqiu Wang and Yasushi Yagi</i>	
Shortest Path Based Planar Graph Cuts for Bi-layer Segmentation of Binocular Stereo Video . . . . .	82
<i>Xiangsheng Huang and Lujin Gong</i>	
Color Information Presentation for Color Vision Defective by Using a Projector Camera System . . . . .	92
<i>Atsushi Yamashita, Rie Miyaki, and Toru Kaneko</i>	

## Workshop on Computational Photography and Aesthetics

Simulating Artworks through Filter Blending . . . . .	102
<i>Crystal Valente and Reinhard Klette</i>	
A Data-Driven Approach to Understanding Skill in Photographic Composition . . . . .	112
<i>Todd S. Sachs, Ramakrishna Kakarala, Shannon L. Castleman, and Deepu Rajan</i>	
Image Information in Digital Photography . . . . .	122
<i>Jaume Rigau, Miquel Feixas, and Mateu Sbert</i>	
Automatically Detecting Protruding Objects When Shooting Environmental Portraits . . . . .	132
<i>Pei-Yu Lo, Sheng-Wen Shih, Jen-Chang Liu, and Jen-Shin Hong</i>	
Artist-Led Suggestions towards an Approach in Content Aware 3D Non-photorealistic Rendering . . . . .	142
<i>Martin Constable</i>	

## Workshop on Computer Vision in Vehicle Technology: From Earth to Mars

Ground Truth Evaluation of Stereo Algorithms for Real World Applications . . . . .	152
<i>Sandino Morales and Reinhard Klette</i>	
Vehicle Ego-Localization by Matching In-Vehicle Camera Images to an Aerial Image . . . . .	163
<i>Masafumi Noda, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Yoshiko Kojima, and Takashi Naito</i>	
A Comparative Study of Two Vertical Road Modelling Techniques . . . . .	174
<i>Konstantin Schauwecker and Reinhard Klette</i>	
The Six Point Algorithm Revisited . . . . .	184
<i>Akihiko Torii, Zuzana Kukelova, Martin Bujnak, and Tomas Pajdla</i>	
Multi-body Segmentation and Motion Number Estimation via Over-Segmentation Detection . . . . .	194
<i>Guodong Pan and Kwan-Yee Kenneth Wong</i>	
Improvement of a Traffic Sign Detector by Retrospective Gathering of Training Samples From In-Vehicle Camera Image Sequences . . . . .	204
<i>Daisuke Deguchi, Keisuke Doman, Ichiro Ide, and Hiroshi Murase</i>	

Statistical Modeling of Long-Range Drift in Visual Odometry . . . . .	214
<i>Ruyi Jiang, Reinhard Klette, and Shigang Wang</i>	
Object Discrimination and Tracking in the Surroundings of a Vehicle by a Combined Laser Scanner Stereo System . . . . .	225
<i>Mathias Haberjahn and Ralf Reulke</i>	
Realistic Modeling of Water Droplets for Monocular Adherent Raindrop Recognition using Bézier Curves . . . . .	235
<i>Martin Roser, Julian Kurz, and Andreas Geiger</i>	
Illumination Invariant Cost Functions in Semi-Global Matching . . . . .	245
<i>Simon Hermann, Sandino Morales, Tobi Vaudrey, and Reinhard Klette</i>	
Relative Pose Estimation for Planetary Entry Descent Landing . . . . .	255
<i>Luca Zini, Francesca Odone, Alessandro Verri, Piergiorgio Lanza, and Alessandra Marcer</i>	
<b>Workshop on e-Heritage</b>	
AR Cultural Heritage Reconstruction Based on Feature Landmark Database Constructed by Using Omnidirectional Range Sensor . . . . .	265
<i>Takafumi Taketomi, Tomokazu Sato, and Naokazu Yokoya</i>	
Augmented Reality-Based On-Site Tour Guide: A Study in Gyeongbokgung . . . . .	276
<i>Byung-Kuk Seo, Kangsoo Kim, and Jong-Il Park</i>	
3D Reconstruction of a Collapsed Historical Site from Sparse Set of Photographs and Photogrammetric Map . . . . .	286
<i>Natchapon Futragoon, Asanobu Kitamoto, Elham Andaroodi, Mohammad Reza Matini, and Kinji Ono</i>	
Recognition and Analysis of Objects in Medieval Images . . . . .	296
<i>Pradeep Yarlagadda, Antonio Monroy, Bernd Carque, and Björn Ommer</i>	
3D Shape Restoration via Matrix Recovery . . . . .	306
<i>Min Lu, Bo Zheng, Jun Takamatsu, Ko Nishino, and Katsushi Ikeuchi</i>	
A Development of a 3D Haptic Rendering System with the String-Based Haptic Interface Device and Vibration Speakers . . . . .	316
<i>Kazuyoshi Nomura, Wataru Wakita, and Hiromi T. Tanaka</i>	

A Texture-Based Direct-Touch Interaction System for 3D Woven Cultural Property Exhibition . . . . .	324
<i>Wataru Wakita, Katsuhito Akahane, Masaharu Isshiki, and Hiromi T. Tanaka</i>	
<b>Workshop on Subspace Methods</b>	
High Dimensional Correspondences from Low Dimensional Manifolds – An Empirical Comparison of Graph-Based Dimensionality Reduction Algorithms . . . . .	334
<i>Ribana Roscher, Falko Schindler, and Wolfgang Förstner</i>	
Multi-label Classification for Image Annotation via Sparse Similarity Voting . . . . .	344
<i>Tomoya Sakai, Hayato Itoh, and Atsushi Imiya</i>	
Centered Subset Kernel PCA for Denoising . . . . .	354
<i>Yoshikazu Washizawa and Masayuki Tanaka</i>	
On the Behavior of Kernel Mutual Subspace Method . . . . .	364
<i>Hitoshi Sakano, Osamu Yamaguchi, Tomokazu Kawahara, and Seiji Hotta</i>	
Compound Mutual Subspace Method for 3D Object Recognition: A Theoretical Extension of Mutual Subspace Method . . . . .	374
<i>Naoki Akihiro and Kazuhiro Fukui</i>	
Dynamic Subspace Update with Incremental Nyström Approximation . . . . .	384
<i>Hongyu Li and Lin Zhang</i>	
Background Modeling via Incremental Maximum Margin Criterion . . . . .	394
<i>Cristina Marghes and Thierry Bouwmans</i>	
Trace Norm Regularization and Application to Tensor Based Feature Extraction . . . . .	404
<i>Yoshikazu Washizawa</i>	
Fast and Robust Face Recognition for Incremental Data . . . . .	414
<i>I. Gede Pasek Suta Wijaya, Keiichi Uchimura, and Gou Koutaki</i>	
Extracting Scene-Dependent Discriminant Features for Enhancing Face Recognition under Severe Conditions . . . . .	424
<i>Rui Ishiyama and Nobuyuki Yasukawa</i>	
A Brief History of the Subspace Methods . . . . .	434
<i>Hitoshi Sakano</i>	
<b>Author Index</b> . . . . .	437

# Table of Contents – Part I

## Workshop on Visual Surveillance

Second-Order Polynomial Models for Background Subtraction . . . . .	1
<i>Alessandro Lanza, Federico Tombari, and Luigi Di Stefano</i>	
Adaptive Background Modeling for Paused Object Regions . . . . .	12
<i>Atsushi Shimad, Satoshi Yoshinaga, and Rin-ichiro Taniguchi</i>	
Determining Spatial Motion Directly from Normal Flow Field: A Comprehensive Treatment . . . . .	23
<i>Tak-Wai Hui and Ronald Chung</i>	
Background Subtraction for PTZ Cameras Performing a Guard Tour and Application to Cameras with Very Low Frame Rate . . . . .	33
<i>C. Guillot, M. Taron, P. Sayd, Q.C. Pham, C. Tilmant, and J.M. Lavest</i>	
Bayesian Loop for Synergistic Change Detection and Tracking . . . . .	43
<i>Samuele Salti, Alessandro Lanza, and Luigi Di Stefano</i>	
Real Time Motion Changes for New Event Detection and Recognition . . . . .	54
<i>Konstantinos Avgerinakis, Alexia Briassouli, and Ioannis Kompatsiaris</i>	
Improving Detector of Viola and Jones through SVM . . . . .	64
<i>Zhenchao Xu, Li Song, Jia Wang, and Yi Xu</i>	
Multi-camera People Localization and Height Estimation Using Multiple Birth-and-Death Dynamics . . . . .	74
<i>Ákos Utasi and Csaba Benedek</i>	
Unsupervised Video Surveillance . . . . .	84
<i>Nicoletta Noceti and Francesca Odone</i>	
Multicamera Video Summarization from Optimal Reconstruction . . . . .	94
<i>Carter De Leo and B.S. Manjunath</i>	
Noisy Motion Vector Elimination by Bi-directional Vector-Based Zero Comparison . . . . .	104
<i>Takanori Yokoyama and Toshinori Watanabe</i>	
Spatio-Temporal Optimization for Foreground/Background Segmentation . . . . .	113
<i>Tobias Feldmann</i>	

Error Decreasing of Background Subtraction Process by Modeling the Foreground . . . . .	123
<i>Christophe Gabard, Laurent Lucat, Catherine Achard, C. Guillot, and Patrick Sayd</i>	
<i>Object Flow</i> : Learning Object Displacement . . . . .	133
<i>Constantinos Lalos, Helmut Grabner, Luc Van Gool, and Theodora Varvarigou</i>	
HOG-Based Descriptors on Rotation Invariant Human Detection . . . . .	143
<i>Panachit Kittipanya-ngam and Eng How Lung</i>	
Fast and Accurate Pedestrian Detection Using a Cascade of Multiple Features . . . . .	153
<i>Alaa Leithy, Mohamed N. Moustafa, and Ayman Wahba</i>	
Interactive Motion Analysis for Video Surveillance and Long Term Scene Monitoring . . . . .	164
<i>Andrew W. Senior, YingLi Tian, and Max Lu</i>	
Frontal Face Generation from Multiple Low-Resolution Non-frontal Faces for Face Recognition . . . . .	175
<i>Yuki Kono, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase</i>	
Probabilistic Index Histogram for Robust Object Tracking . . . . .	184
<i>Wei Li, Xiaoqin Zhang, Nianhua Xie, Weiming Hu, Wenhan Luo, and Haibin Ling</i>	
Mobile Surveillance by 3D-Outlier Analysis . . . . .	195
<i>Peter Holzer and Axel Pinz</i>	
Person Re-identification Based on Global Color Context . . . . .	205
<i>Yinghao Cai and Matti Pietikäinen</i>	
Visual Object Tracking via One-Class SVM . . . . .	216
<i>Li Li, Zhenjun Han, Qixiang Ye, and Jianbin Jiao</i>	
Attenuated Sequential Importance Resampling (A-SIR) Algorithm for Object Tracking . . . . .	226
<i>Md. Zahidul Islam, Chi-Min Oh, and Chil-Woo Lee</i>	
An Appearance-Based Approach to Assistive Identity Inference Using LBP and Colour Histograms . . . . .	236
<i>Sareh Abolahrari Shirazi, Farhad Dadgostar, and Brian C. Lovell</i>	
Vehicle Class Recognition Using Multiple Video Cameras . . . . .	246
<i>Dongjin Han, Jae Hwang, Hern-soo Hahn, and David B. Cooper</i>	

Efficient Head Tracking Using an Integral Histogram Constructing Based on Sparse Matrix Technology . . . . .	256
<i>Jia-Tao Qiu, Yu-Shan Li, and Xiu-Qin Chu</i>	
<b>Workshop on Video Event Categorization, Tagging and Retrieval (VECTaR)</b>	
Analyzing Diving: A Dataset for Judging Action Quality . . . . .	266
<i>Kamil Wnuk and Stefano Soatto</i>	
Appearance-Based Smile Intensity Estimation by Cascaded Support Vector Machines . . . . .	277
<i>Keiji Shimada, Tetsu Matsukawa, Yoshihiro Noguchi, and Takio Kurita</i>	
Detecting Frequent Patterns in Video Using Partly Locality Sensitive Hashing . . . . .	287
<i>Koichi Ogawara, Yasufumi Tanabe, Ryo Kurazume, and Tsutomu Hasegawa</i>	
Foot Contact Detection for Sprint Training . . . . .	297
<i>Robert Harle, Jonathan Cameron, and Joan Lasenby</i>	
Interpreting Dynamic Meanings by Integrating Gesture and Posture Recognition System . . . . .	307
<i>Omer Rashid Ahmed, Ayoub Al-Hamadi, and Bernd Michaelis</i>	
Learning from Mistakes: Object Movement Classification by the Boosted Features . . . . .	318
<i>Shigeyuki Odashima, Tomomasa Sato, and Taketoshi Mori</i>	
Modeling Multi-Object Activities in Phase Space . . . . .	328
<i>Ricky J. Sethi and Amit K. Roy-Chowdhury</i>	
Sparse Motion Segmentation Using Multiple Six-Point Consistencies . . . . .	338
<i>Vasileios Zografos, Klas Nordberg, and Liam Ellis</i>	
Systematic Evaluation of Spatio-Temporal Features on Comparative Video Challenges . . . . .	349
<i>Julian Stöttinger, Bogdan Tudor Goras, Thomas Pöntiz, Allan Hanbury, Nicu Sebe, and Theo Gevers</i>	
Two-Probabilistic Latent Semantic Model for Image Annotation and Retrieval . . . . .	359
<i>Nattachai Watcharapinchai, Supavadee Aramvith, and Supakorn Siddhichai</i>	
Using Conditional Random Field for Crowd Behavior Analysis . . . . .	370
<i>Saira Saleem Pathan, Ayoub Al-Hamadi, and Bernd Michaelis</i>	



## Workshop on Gaze Sensing and Interactions

Understanding Interactions and Guiding Visual Surveillance by Tracking Attention . . . . .	380
<i>Ian Reid, Ben Benfold, Alonso Patron, and Eric Sommerlade</i>	
Algorithm for Discriminating Aggregate Gaze Points: Comparison with Salient Regions-Of-Interest . . . . .	390
<i>Thomas J. Grindinger, Vidya N. Murali, Stephen Tetreault, Andrew T. Duchowski, Stan T. Birchfield, and Pilar Orero</i>	
Gaze Estimation Using Regression Analysis and AAMs Parameters Selected Based on Information Criterion . . . . .	400
<i>Manabu Takatani, Yasuo Arika, and Tetsuya Takiguchi</i>	
Estimating Human Body and Head Orientation Change to Detect Visual Attention Direction . . . . .	410
<i>Ovgu Ozturk, Toshihiko Yamasaki, and Kiyoharu Aizawa</i>	
Can Saliency Map Models Predict Human Egocentric Visual Attention? . . . . .	420
<i>Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki</i>	
An Empirical Framework to Control Human Attention by Robot . . . . .	430
<i>Mohammed Moshiul Hoque, Tomami Onuki, Emi Tsuburaya, Yoshinori Kobayashi, Yoshinori Kuno, Takayuki Sato, and Sachiko Kodama</i>	
Improvement and Evaluation of Real-Time Tone Mapping for High Dynamic Range Images Using Gaze Information . . . . .	440
<i>Takuya Yamauchi, Toshiaki Mikami, Osama Ouda, Toshiya Nakaguchi, and Norimichi Tsumura</i>	
Evaluation of the Impetuses of Scan Path in Real Scene Searching . . . . .	450
<i>Chen Chi, Laiyun Qing, Jun Miao, and Xilin Chen</i>	
<b>Author Index</b> . . . . .	461

# Computer Vision Technology Applied to MR-Based Pre-visualization in Filmmaking

Hideyuki Tamura<sup>1</sup>, Takashi Matsuyama<sup>2</sup>, Naokazu Yokoya<sup>3</sup>,  
Ryosuke Ichikari<sup>1</sup>, Shohei Nobuhara<sup>2</sup>, and Tomokazu Sato<sup>3</sup>

<sup>1</sup> College of Information Science and Engineering, Ritsumeikan University

<sup>2</sup> Graduate School of Informatics, Kyoto University

<sup>3</sup> Graduate School of Information Science, Nara Institute of Science and Technology

**Abstract.** In this talk, we introduce the outline of the MR-PreViz Project performed in Japan. In the pre-production process of filmmaking, PreViz, pre-visualizing the desired scene by CGI, is used as a new technique. In its advanced approach, we propose MR-PreViz to utilize mixed reality technology as in current PreViz. MR-PreViz makes it possible to merge the real background and the computer-generated humans and creatures in an open set or at an outdoor location. Computer vision technologies are required for many aspects of MR-PreViz. For capturing an actor's action, we applied 3D Video, which is a technology that allows one to reconstruct an image seen from any viewpoint in real time from video images taken by multiple cameras. As the other application of CV, we developed a vision based camera tracking method. The method collects environmental information required for tracking efficiently using a structure-from-motion technique before the shooting. Additionally, we developed a relighting technique for lighting design of MR-PreViz movie.

## 1 Introduction

Mixed reality (MR) which merges real and virtual worlds in real-time, is an advanced form of virtual reality (VR) [1]. The word "augmented reality" (AR) has the same meaning as MR. In AR space, the real world is dominant, and it is electronically augmented and enhanced. On the other hand, MR is based on the concept of fusing the real world and virtual world by almost treating them equally. In terms of visual expression, VR deals with completely Computer-Generated Images (CGI). By contrast, AR/MR superimposes the CGI onto real scenes. Therefore, capturing the elements and analyzing and understanding attributes of the real world are necessary for AR/MR. Consequently, VR requires computer graphics technology; computer vision (CV) plays an important role for AR/MR.

AR/MR has a variety of applications. It already has been applied to medicine and welfare, architecture and urban planning, industrial design and manufacturing, art and entertainment, etc. This paper describes an application of MR technology for filmmaking, particularly the pre-visualization process. In the post-production stage of feature films, visual effects, or composing the CGIs with



**Fig. 1.** Conceptual image of MR-PreViz

live action images, is used routinely. Since this is operated as an off-line procedure, redoing and time-consuming processes are allowed. On the other hand, pre-visualization using MR technology in the pre-production stage requires real-time interactive merging of live actions and CGIs. This is a very difficult and challenging topic. In hopes of obtaining many fruits from this challenging theme, we are promoting the "CREST/MR-PreViz Research Project" [2][3]. Fig.1 shows the conceptual image of the MR-PreViz project.

AR/MR is a newly emerged attractive application field for CV technology. At the same time, filmmaking is a worthwhile application field to tackle. The theme "MR-based Pre-visualization in Filmmaking" makes stages for CV technology, we will introduce examples of 3 of these stages in this paper.

## 2 Overview of the CREST/MR-PreViz Project

### 2.1 Significance of MR Technology for PreViz

Recently, pre-visualization (PreViz, also described as "PreVis" or "pre-vis") has been used to further develop a storyboard. PreViz is a technique based on computer-generated images for visualizing action scenes, camera angles, camera blockings, lighting conditions, and other situations and conditions before the actual shoot. Compared with the conventional PreViz, which previsualizes the desired movie scene with only CGI, our MR-PreViz has significant differences:

1. MR-PreViz utilizes real backgrounds, such as sound stages, open sets, and location sites. CG objects imitate actors or creatures such as dinosaurs and aliens, which are superimposed onto the background. In terms of VR, this is an MR composite from a camera view point. In terms of filmmaking, this is on-site real-time 3D match moving.
2. Compared with virtual studios used in TV productions, MR-PreViz is a generic form which can be used in outdoor environment.
3. MR-PreViz can be used in multiple stages of PreViz from Pitch-Vis to Post-Vis, and it is especially suitable for camera rehearsals and set simulations.

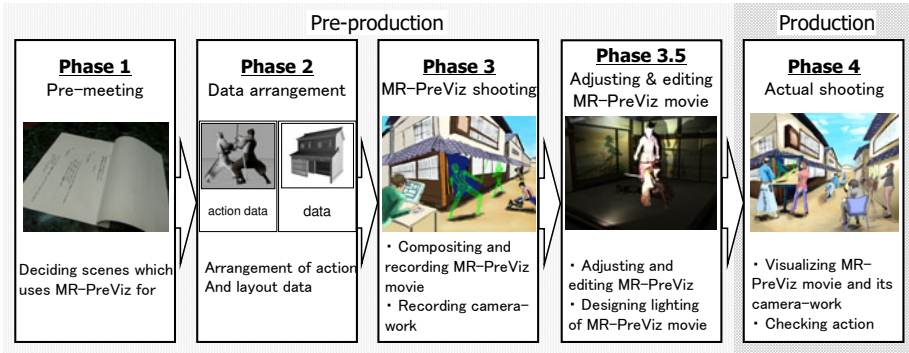


Fig. 2. Workflow of MR-PreViz

## 2.2 Workflow of Filmmaking with MR-PreViz

The workflow of filmmaking with MR-PreViz is as follows (Fig.2).

**Phase 1:** Selecting scenes suitable for MR-PreViz; Scenes that should be checked using MR-PreViz, are selected.

**Phase 2:** Arrangement of action and layout data; We collect CG character data, animation setting data, and action data before making MR-PreViz movies.

**Phase 3:** MR-PreViz shooting; MR-PreViz movies are shot at the shooting site using Camera-Work Authoring Tools with a professional digital cinema camera.

**Phase 3.5:** Editing and Adjusting MR-PreViz movie; A high definition version of the MR-PreViz movie can be obtained by off-line rendering. Additionally, illumination on the MR-PreViz movie can be changed by relighting.

**Phase 4:** Application to actual shooting; The results of MR-PreViz shooting are applied to the actual shooting. Actors and staff can share ideas and images by using a MRP browser.

Recently, the processes in Phase 3.5 become more important, because the function of editing and adjusting after the MR-PreViz shooting were appreciated in experimental shootings.

## 3 CV Technologies in MR-PreViz (1)—3D Video

This section describes 3D video technology [4], which is used in Phase 2. While wearing a special suit has been necessary for capturing action in the past, 3D video has an advantage that the actor's actions can be captured while wearing a normal costume for the real shooting. We describe some technical highlights of our 3D video technology as computer vision research and then discuss its advantage against other possible approaches as a data source of MR-PreViz.

### 3.1 Introduction of 3D Video

The 3D video is media which records visual dynamic events as is. It records the actor's 3D shape, texture and motion without attaching any extra devices or markers to the object. Unlike 3D-TV, which only gives 2D stereo appearances of the scene to the human brain, 3D video explicitly estimates full 3D geometry and texture. It first captures 2D multi-view videos of the object and then estimates its 3D information purely from acquired 2D videos. Once 3D shapes are obtained from 2D videos, the original 2D images are mapped onto the 3D surface as its texture. Since this produces a conventional g3D surface geometry + textureh style output, we can render it from an arbitrary viewpoint even with other virtual objects.

### 3.2 Technical Highlights of 3D Video as a Computer Vision Research

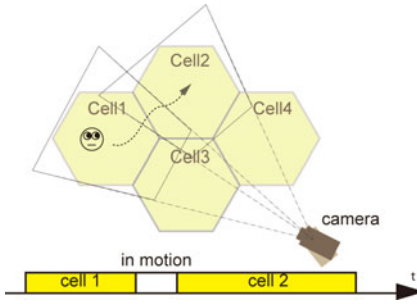
3D video technology consists of lots of challenging computer vision research topics including (1) object tracking and calibration and (2) 3D kinematic motion estimation for further editing.

**(1) Object tracking and calibration:** The fundamental criteria for the 3D video capture are twofold. All regions of the object surface must be observed from at least two cameras, and the intrinsic and extrinsic parameters of the cameras must be calibrated accurately. As long as we can satisfy these requirements, we can choose any combinations of cameras and their arrangement, which controls the resolution and captures area size of the system. To achieve the best combination of the resolution and capture area with a fixed image resolution, we have developed an active (pan-tilt-zoom) camera system named gcell-based 3D video captureh (Fig.3) which tracks and captures the object on a cell-to-cell basis [5]. In this approach, we can reformulate the original online tracking and calibration problem as a cell arrangement and tracking problem.

**(2) 3D kinematic motion estimation:** 3D video consists of a time-series of 3D surface geometry and texture information and does not have any information on the kinematic structure of the object. The goal here is to estimate the kinematic structure and posture of the object by observing its surfaces. Fig.4 shows a result for a complex posture. The key point here is explicit management of the 3D surface areas invisible from any cameras which have less accuracy on the surface geometry [6].

### 3.3 3D Video for MR-PreViz

The key point of 3D video technology for MR-PreViz is its geometry-based representation of the scene. This property brings (1) seamless integration with other 3D virtual objects, and (2) free-viewpoint rendering for pre-visualization. In addition, once we obtain the posture information of the captured object, we can edit them with conventional CG techniques.



**Fig. 3.** Cell-based 3D video capture. T cameras cooperatively track the object a cell-to-cell



**Fig. 4.** Complex posture estimation

## 4 Camera Tracking Using Landmark Database

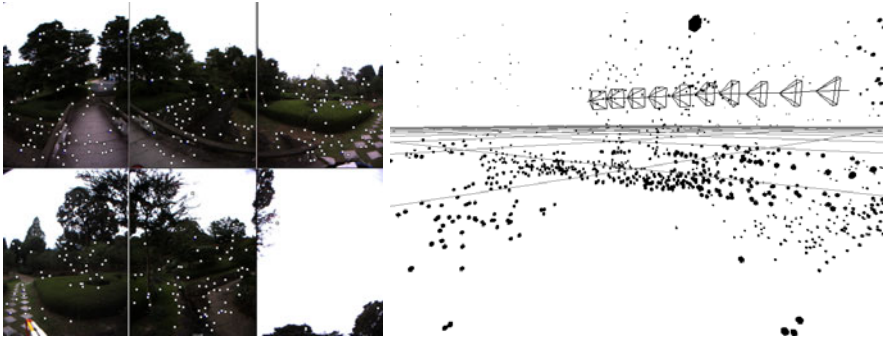
This section describes a camera tracking method used in Phase 3. The proposed method has a significant role to play in composing CGI onto real background in out-door environments as an on-site real-time 3D matchmove.

### 4.1 Registration Method Using Landmark Database

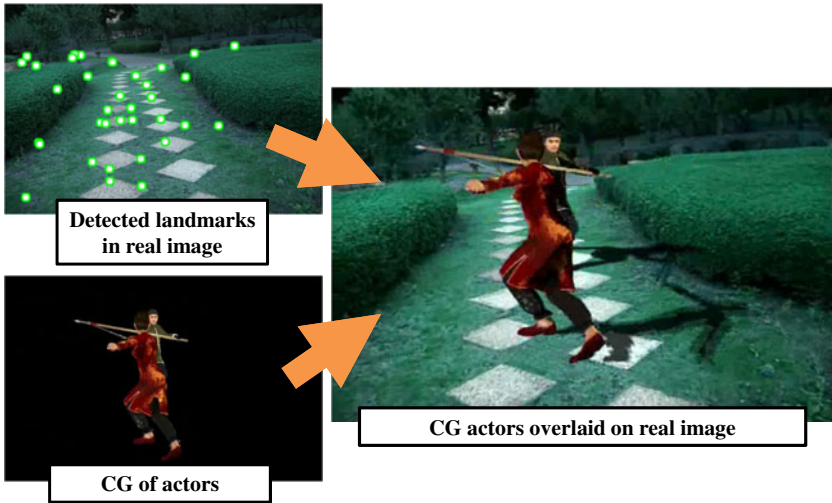
In order to overlay CG actors at geometrically correct positions in captured images, position and posture parameters of a video camera are necessary. High-accurate measurements of these parameters has been achieved by using combinations of sensors attached to the video camera and emitters arranged in the target environment, such as ultrasonic, magnetic or optical sensors/emitters. Although these methods work well in a small environment like a TV studio, it is not a realistic scenario to use them in a large outdoor environment due to the difficulty in arranging and calibrating those emitters. A combination of GPS and other sensors is one of possibilities for an outdoor environment but its accuracy has not reached the practical level of geometric registration in MR.

On the other hand, vision-based methods can estimate camera parameters without external sensors. The PTAM [7] is one of the famous methods that estimates camera parameters in real-time by tracking feature points on input images. This method obtains relative camera motion and 3-D positions of feature points simultaneously without prior knowledge. One problem in the PTAM for MR-PreViz is that absolute position and posture information have never been recovered. This limitation makes pre-arrangement of CG objects difficult.

Landmark database (LMDB) [8] is one of the promising approaches to this problem. In this method, as an offline stage, the target environment is captured by using an omnidirectional camera, and feature points tracked in this video sequence are registered to the database as landmarks. 3-D coordinate of feature points are estimated by using a structure-from-motion technique for omnidirectional video that simultaneously estimates 3-D positions of feature points and camera motion [9].



**Fig. 5.** Detected feature points in omnidirectional image (left) and their 3-D positions estimated by structure from motion (right)



**Fig. 6.** Geometric registration between CG object and real scene using LMDB

Fig.5 shows detected positions of feature points on the omnidirectional video sequence and estimated 3-D positions of feature points. Absolute 3-D positions of landmarks are recovered by using several reference points whose absolute 3-D positions are manually measured. Visual information of each landmark is also stored to the database from the omnidirectional video.

In the online stage, pre-registered landmarks are searched for from each image of the input video using its visual information. After finding corresponding pairs of landmarks and feature points, absolute position and posture of the video camera are estimated by minimizing re-projection errors of these pairs. By using estimated camera parameters, CG characters are rendered from the appropriate viewpoint and they are finally merged into the input image as shown in Fig.6.



## 4.2 Rehearsal Path Method: Refinement of the Registration Method Using LMDB

The registration method using LMDB was originally developed for a general scenario of MR. We can utilize constraints of the usage in filmmaking for refinement [10]. In particular, it is assumed that a rough camera path is known in filmmaking. We have developed a method called "Rehearsal Path Method; RPM" which refines efficiency and accuracy of the registration method by restricting the moving range of camera to the camera path during construction of LMDB. The RPM automatically gathers information of the shooting site by pre-shooting and constructs a landmark database (LMDB). The RPM consists of two phases as shown in Fig.7.

**Rehearsal Phase:** In this phase, RPM utilizes a video sequence with a fiducial marker captured during the rehearsal as a learning sequence. The geometry of the site is estimated using a structure-from-motion technique. In particular, the positions of feature points in 3D space are first estimated by using epipolar geometry on several frames in the video sequence. Secondly, 6DOF parameters of the camera and positions of new feature points are simultaneously calculated by tracking the feature points. Finally, the coordinates of the 3D points are transformed into real world coordinates by recognizing the marker. SIFT, a local invariant, of each landmark is calculated and entered into our LMDB with 3D positions.

**Shooting Phase:** In this phase, the fiducial marker is removed for MR-PreViz shooting. The registration of virtual world and real world is realized by correlating the 2D feature points in the images with the 3D points in the LMDB. The RPM is possible to automatically obtain an initial position and recover from tracking failure by using SIFT matching between a present frame and keyframes on the camera path of the LMDB prepared in advance. The RPM successfully refined efficiency and accuracy of the registration method by using the knowledge of camera path. Besides PreViz, The RPM is also applicable to other purposes which have the same constraint.

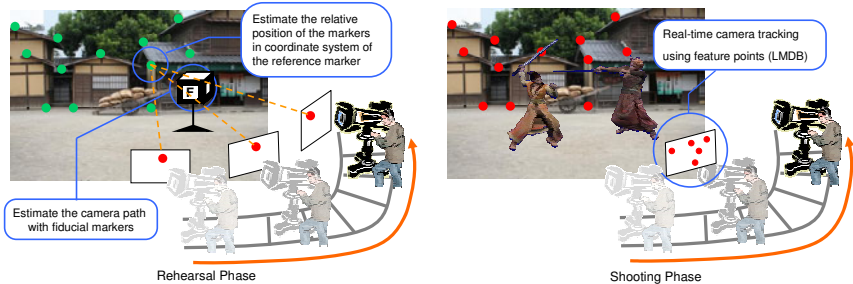


Fig. 7. Rehearsal Path Method (RPM)



## 5 CV Technologies in MR-PreViz (3) - IBL and Relighting

This section describes the method assumed to be used in Phase 3.5, which is the additional process after the MR-PreViz shooting. In addition to improving the speed for processing, this method also can be used for the MR-PreViz shooting in Phase 3. Photometric consistency between virtual objects and a real scene is one of the most important issues in the CG research area. "Virtual Cinematography" [11], developed by Debevec et. al., is a famous example which applied illumination technology researched in an academic field to film making. This work is based on image based lighting (IBL). IBL enables the lighting condition in the desired scene to be reconstructed by utilizing a series of images or omnidirectional images stored in advance.

Light Stage is the dedicated equipment for translating the IBL concept into reality as shown in Figure 6. Light Stage enables illumination onto actors to be changed freely by systematically illuminating structured lights. Previously, several versions of light stage have been developed. One group received the Academy Award in 2010 for their technical contributions to "Spiderman 2" and "Avatar". Light Stage was originally used in the postproduction process for keeping photometric consistency. By contrast, we are developing a visualization method for "Look", which refers to the feeling of an image provided by illumination and color tone. For visualizing the look that directors and cinematographers imagine, we developed relighting technology for MR-PreViz movies. Light Stage realizes relighting for the actor who appears on the stage. On the other hand, the target of our relighting for MR-PreViz is the background of the indoor-outdoor location. Therefore the approach using structured lights can not be used in our case.

In the MR research area, there are many works for photometric consistency between virtual objects and real objects. Lighting conditions are estimated for illuminating the virtual objects on the same condition of the real scene. Our research takes a lateral approach. As a next step, we will focus on challenging trials to Look-Change of MR space. In this section, we introduce a relighting method for the Look-Change [12]. The proposed method allows an MR space to have additional virtual illumination for the Look-Change. The effects of virtual illumination are applied to both real objects and virtual objects while keeping photometric consistency. There is a trade-off between the quality of the lighting effect and the processing time. Therefore, the challenge is to create an efficient model of the lighting condition of real scenes. Our method adopts a simple and approximate approach to realize indoor-outdoor relighting as shown in Fig.8.

### Step 1: Preparing images without distinct shadows

If a distinct shadow, exists in the background images, it may cause a paradox between real and virtual shadows in the relighting process. We should prepare background images without shadows. We can use physical lighting equipment or image processing methods for removing shadows in the images.

### Step 2: Adjusting color tone This process approximates environmental light.

The color tone is adjusted by multiplying arbitrary values with respect to



**Fig. 8.** Flow of our relighting method

each color channel. As a result of this process, we can change an image of daylight into an image of a night scene.

**Step 3:** Adding lighting effects to MR-PreViz images

After color correction, real and virtual objects in the MR-PreViz images are illuminated by virtual lighting. To optically correct illuminate objects, we estimate reflectance properties and geometry of the real objects. A relationship between pixel value and illuminance is obtained as a reflecting property. Illuminance is automatically calculated under several lighting conditions by using a reference marker where the reflecting property of the marker is known. Geometry of the site is estimated using a structure-from-motion technique.

The final MR-PreViz images of relighting are shown in Step 3 of Figure 8. Nevertheless the proposed method is developed for changing Look in a MR-PreViz movie. It is also applicable for lighting effects in MR attractions.

## 6 Conclusions

In this paper, we picked mixed reality as a useful target application of CV technology, and introduced mixed reality based pre-visualization in filmmaking with 3 elemental CV technologies.

These elemental technologies, as outlined below, have been steadily improved by being used many times in the actual workflow of filmmaking during the projects 5 years (since Oct 2005).

- (i) The 3D video has verified its utility for the purpose of PreViz, though it does not have enough quality as a final sequence.
- (ii) The camera tracking method has steadily improved its practicality by utilizing constraints of PreViz in filmmaking. Specifically, we refined the method under the assumption that a rough camera-path is decided before the shooting. We called this method "Rehearsal Path Method." The method is not omnipotent because it depends on the target scene and weather. We can improve efficiency and reliability of the method by gathering experience and setting modes based on the situations.
- (iii) At first the relighting method had been used only after the MR-PreViz shooting in Phase 3.5 as an off-line process. It also can be used in Phase 3 as a real-time process.

Even conventional PreViz composed of only CG was not popular 5 years ago. However, the number of the use of PreViz in feature films was rapidly increased since then. Today, there are PreViz studios which specialize in PreViz, and "The Pre-visualization Society" has been established [13]. Although PreViz is not able to contribute to the quality of the final movie, it enables filmmakers to inspire their creativity by facilitating the process of trial-and-error in the pre-production or production stages of filmmaking. Additionally, PreViz is able to contribute to reducing the total production costs. The processes of the PreViz are subdivided into pitch-vis, tech-vis, post-vis, etc. Accordingly, MR-PreViz continues to receive much attention and makes progress for CV technology.

## References

1. Tamura, H., Yamamoto, H., Katayama, A.: Mixed reality: Future dreams seen at the border between real and virtual worlds. *IEEE Computer Graphics & Applications* 21, 64–70 (2001)
2. Tenmoku, R., Ichikari, R., Shibata, F., Kimura, A., Tamura, H.: Design and prototype implementation of MR pre-visualization workflow. In: *DVD-ROM Proc. Int. Workshop on Mixed Reality Technology for Filmmaking*, pp. 26–30 (2006)
3. Ichikari, R., Tenmoku, R., Shibata, F., Ohshima, T., Tamura, H.: Mixed reality pre-visualization for filmmaking: On-set camera-work authoring and action rehearsal. *Int. J. Virtual Reality* 7, 25–32 (2008)
4. Starck, J., Maki, A., Nobuhara, S., Hilton, A., Matsuyama, T.: The multiple-camera 3-D production studio. *IEEE Trans. on Circuits and Systems for Video Technology* 19, 856–869 (2009)
5. Yamaguchi, T., Nobuhara, S., Matsuyama, T.: Cell-based object tracking method for 3D shape reconstruction using multi-viewpoint active cameras. In: *IEEE Int. Workshop on Visual Surveillance, VS 2009* (2009)
6. Mukasa, T., Miyamoto, A., Nobuhara, S., Maki, A., Matsuyama, T.: Complex human motion estimation using visibility. In: *IEEE Int. Conf. Series on Automatic Face and Gesture Recognition* (2008)
7. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: *Proc. ISMAR*, pp. 225–234 (2007)
8. Taketomi, T., Sato, T., Yokoya, N.: Real-time geometric registration using feature landmark database for augmented reality applications. In: *Proc. SPIE* (2009) 723804–723804–9
9. Sato, T., Ikeda, S., Yokoya, N.: Extrinsic camera parameter recovery from multiple image sequences captured by an omni-directional multi-camera system. In: *Proc. European Conf. on Computer Vision*, vol. 2, pp. 326–340 (2004)
10. Toishita, W., Momoda, Y., Tenmoku, R., Shibata, F., Tamura, H., Taketomi, T., Sato, T., Yokoya, N.: A novel approach to on-site camera calibration and tracking for MR pre-visualization procedure. In: *Proc. Human-Computer Interaction International (HCI 2009)*, pp. 492–502 (2009)
11. Debevec, P.: Virtual cinematography: Relighting through computation. *IEEE Computer* 39, 57–65 (2006)
12. Ichikari, R., Hatano, R., Ohshima, T., Shibata, F., Tamura, H.: Designing cinematic lighting by relighting in MR-based pre-visualization. In: *ACM SIGGRAPH ASIA 2009 Posters* (2009)
13. <http://www.previssociety.com/>

# Model Based Pose Estimation Using SURF

Peter Decker and Dietrich Paulus

Active Vision Group  
University of Koblenz-Landau  
Universitätsstr. 1  
56070 Koblenz, Germany  
{decker,paulus}@uni-koblenz.de

**Abstract.** Estimation of a camera pose (position and orientation) from an image, given a 3d model of the world, is a topic of great interest in many current fields of research. When aiming for a model based pose estimation approach, several questions arise: What is the model? How do we acquire a model? How is the image linked to the model? How is a pose computed and verified using the latter information? In this paper we present a new approach towards model based pose estimation based solely on SURF features. We give a formal definition of our model, show how to build such a model from image data automatically, how to integrate two partial models, and how pose estimation for new images works.

## 1 Introduction

Computing the pose of a camera given an image and a model of the world is an important task in computer vision. There are many different approaches using all kind of different models and matching techniques. Most are feature based, some use features which provide a descriptor for easier matching. SURF [1] is popular because of its invariance properties and high distinctiveness of the descriptor, as well as its speed. We present an approach towards model based pose estimation based solely on SURF features.

The rest of the paper is organized as follows: In the next section we give an overview of related work. We define our model in section 3 and show how to generate such a model automatically from images in section 4. In section 5, we demonstrate how a camera pose is computed from the model and a query image. Section 6 describes an algorithm to integrate two models which partially overlap into a single model. Evaluation takes place in section 7. Section 8 concludes the paper.

## 2 Related Work

Zhang and Kosecka discuss pose estimation in urban environments [2]. They store a number of GPS localized images and extracted SIFT features. Then the images most similar to a query image are identified and possible motion models

are computed. For a final pose triangulation the two best fitting views are taken into account. Schindler et. al. focus on large scale databases and present an approach which is able to handle over 100 million SIFT features using vocabulary trees [3]. More recently, Wu et. al. introduced a new method of matching so called VIP features, which greatly increased the number of correct matches from query images [4]. The system described by Snavely et al. [5] covers much more aspects than pose estimation of images alone. Not only do they show how to compute structure and camera position from a large number of unstructured, uncalibrated images, but they also cover means of how to visualize and navigate the result. Irschara et al. introduce the idea of synthetic views to handle images taken from new viewpoints [6].

In our approach, we reconstruct 3d data directly from the images using SURF, thus skipping the search for the best fitting image as for example in [2]. We establish a connection between the features from a query image and the model directly, thus enabling direct pose estimation.

### 3 Model Definition

A model  $\mathcal{M}$  is defined as a tuple

$$\mathcal{M} = (\mathcal{P}, \mathcal{F}, \mathcal{S}, g, h) . \quad (1)$$

It consists of a set of *world points*  $\mathcal{P}$ , *frames*  $\mathcal{F}$ , *SURF features*  $\mathcal{S}$  and the relations  $g \subseteq \mathcal{P} \times \mathcal{S}$  as well as  $h \subseteq \mathcal{S} \times \mathcal{F}$ . A world point  $\mathbf{p}^w$  is a simple point in three dimensional Euclidian space:  $\mathbf{p}^w \in \mathbb{R}^3$ . A frame  $\mathbf{f}$  represents a three dimensional Euclidian transformation. It describes the position of a camera by the rotation and translation applied to a world point before projection to the image plane, thus  $\mathbf{f} \in SE(3)$ . A surf feature  $\mathbf{s} = (x, y, \sigma, \theta, \mathbf{d})$  consists of its location  $(x, y)$  in the image, detection scale  $\sigma$ , orientation  $\theta$  and a 64 dimensional descriptor  $\mathbf{d}$ , as described in [1]. The relation  $g$  holds information, which SURF feature  $\mathbf{s}$  is connected to which world point  $\mathbf{p}^w$ .  $h$  connects the surf features to the frames from which they originate. For easier notation, we define the set of features  $\mathcal{S}_{\mathbf{f}_i}$  extracted from frame  $\mathbf{f}_i \in \mathcal{F}$  as

$$\mathcal{S}_{\mathbf{f}_i} := \{ \mathbf{s}_j \in \mathcal{S} \mid (\mathbf{s}_j, \mathbf{f}_i) \in h \} . \quad (2)$$

The set of all world points  $\mathcal{P}_{\mathbf{f}_i}$  visible in frame  $\mathbf{f}_i \in \mathcal{F}$  is defined using both relations  $g$  and  $h$ :

$$\mathcal{P}_{\mathbf{f}_i} := \{ \mathbf{p}_j^w \in \mathcal{P} \mid \exists \mathbf{s} \in \mathcal{S} : (\mathbf{p}_j^w, \mathbf{s}) \in g \wedge \mathbf{s} \in \mathcal{S}_{\mathbf{f}_i} \} . \quad (3)$$

With these relations a number of queries to the model are possible, e.g.

- In which frames has world point  $\mathbf{p}^w$  been recognized?
- Which surf descriptors are connected to a given world point  $\mathbf{p}^w$ ?
- Is there already a world point associated with feature  $\mathbf{s}$ ?

These are important during the model building process.

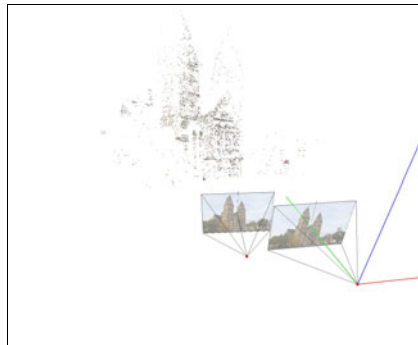
## 4 Automatic Model Generation from Images

Automatic model generation from images is split into two phases. The first phase initializes the model using stereo geometry, while the second phase iteratively expands and improves the model. We assume the images to be in an order in which the first two images have a sufficient overlap for stereo processing. All further images have to be taken roughly from a direction of any preceding image, so they show a detail of the world which has already been covered to some extent.

We assume a geometrically calibrated camera with known intrinsic parameters. All images are undistorted beforehand. This allows us to use image coordinates directly, which simplifies the structure from motion process.

### 4.1 Initialization

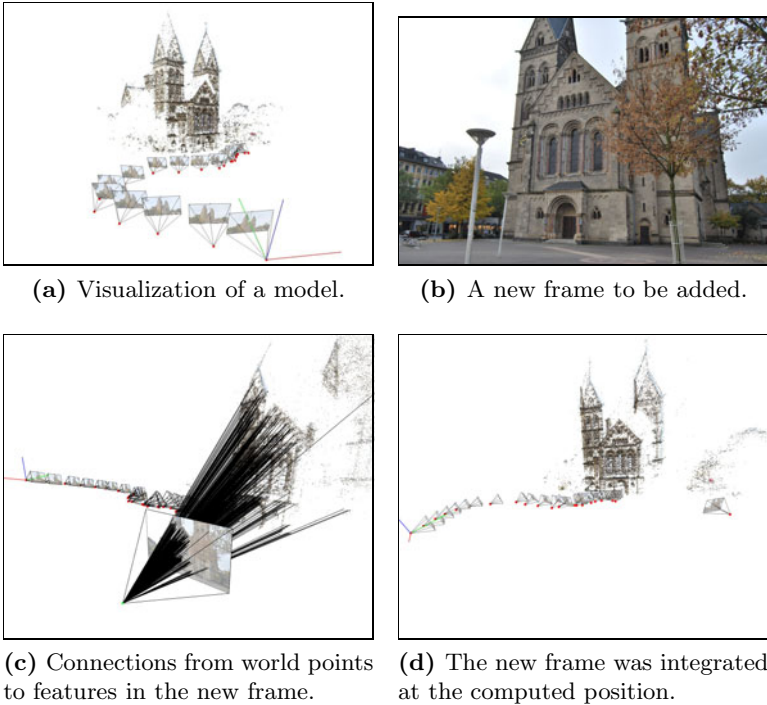
For the two initial frames, we extract SURF features and compute possible correspondences using nearest neighbour matching of the descriptors along with a distance ratio threshold as proposed by Lowe [7]. RANSAC [8] with an adaptive termination criterion [9] is used to estimate the best fitting epipolar geometry using Nistér’s five-point algorithm [10]. We then extract the camera movement from the essential matrix to get the pose of the second camera [9]. Next, all inlier to the epipolar constraint are triangulated and tested for their reprojection error in both images (which can in fact differ). For triangulation, we used the sum of the squared reprojection errors as minimization criterion. We also found it necessary to test if the triangulated point is in front of both cameras, since with very many extracted SURF features it eventually happens, that outlier correspondences are inlier to the epipolar geometry by chance but reconstruct a point behind the cameras. This is also known as the cheirality constraint [10]. Fig. 1 shows an initial stereo pair and the reconstructed world points.



**Fig. 1.** Initialization of the model with a stereo image pair

## 4.2 Incremental Expansion of the Model

The model is expanded frame by frame. First, the pose of the new frame  $\mathbf{f}_{n+1}$  with respect to the model's coordinate system is computed as will be shown in section 5. The new frame is added to  $\mathcal{F}$ , features extracted in the frame are added to  $\mathcal{S}$ . The inlier correspondences from existing world points to features are then added to  $g$ .



**Fig. 2.** Integration of a new frame. Features are extracted and correspondences to descriptors connected to world points of the model are computed. From these 2d/3d correspondences, the pose of the new frame is computed.

After adding a frame, we apply global bundle adjustment to all estimated world points and frames. We do so by optimizing these with respect to the reprojection error using the sparse bundle adjustment software [11] based on a Levenberg Marquard implementation [12], which both are publicly available. We parametrize our world points as  $\mathbf{p}^w \in \mathbb{R}^3$  and the camera location as the vector of three parameters representing translation and another three parameters representing the rotation axis. The length of the rotation axis defines the amount of rotation. Jacobians are computed using finite differences.

After global structure and motion optimization, new world points are created by computing the epipolar geometry between  $\mathbf{f}_{n+1}$  and any other frame we want

to consider. It makes sense to restrict the choice of these frames using constraints concerning their relative pose to each other, thus omitting frames which are too far away from each other or have too different viewing directions.

Assuming  $\mathbf{R}_i$  and  $\mathbf{t}_i$  to be the rotation and translation of frame  $i$  from its projection matrix, an essential matrix between two frames  $\mathbf{f}_i$  and  $\mathbf{f}_j$  can be computed by

$$\mathbf{E}_{ij} = [\mathbf{R}_j(\mathbf{R}_i^T(-\mathbf{t}_i) - \mathbf{R}_j^T(-\mathbf{t}_j))]_{\times} \mathbf{R}_j \mathbf{R}_i^T. \quad (4)$$

Inlier matches to the epipolar geometry between these frames can create new world points, if they also satisfy geometric and reprojection constraints as in section 4.1. These are added to  $\mathcal{P}$ , their connections are added to  $g$ . If an inlier contains a feature from the model which is already connected to a world point, no new world point is created. Instead, the new feature is connected to the existing world point by adding the relation to  $g$ , thus increasing the number of SURF features describing the particular world point.

## 5 Model Based Pose Estimation

We can compute a camera pose given the model and SURF features extracted from a new frame directly.

### 5.1 Matching

First, the descriptors from the new image are matched against all descriptors from the model which are connected to world points. For each feature we consider the two best matches. If they pass the distance ratio threshold as in 7 or they are connected to the same world point, we create a 2d/3d correspondence. Passing the distance ratio threshold means, that the first match is distinct enough from the second best. If both matches are connected to the same world point it means, that the feature from the new image describes this particular world point very well. There is still the possibility that several descriptors connected to the same world point are matched to different features in the query image. In that case we would create contradicting 2d/3d correspondences, where a world point is projected to different points in the image - we therefore consider these matches unstable and drop them all.

For faster nearest neighbour queries on the descriptors, we use the Fast Library for Approximate Nearest Neighbors FLANN [13], which is publicly available.

### 5.2 Pose Estimation

The resulting 2d/3d correspondences are passed to a RANSAC procedure encapsulating Fiore's linear pose estimation algorithm [14]. Since we work in image coordinates at this point, the result is an Euclidian transformation in  $\mathbb{R}^3$ , describing the pose of the new frame with respect to the model's world coordinate system.



## 6 Model Integration

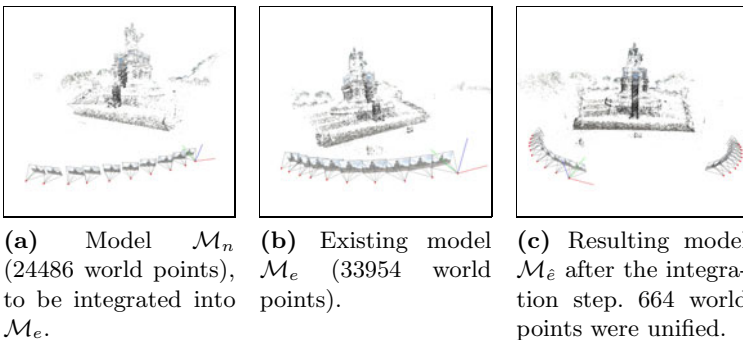
We developed a method to integrate a model  $\mathcal{M}_n$  into an existing model  $\mathcal{M}_e$ . The models need to have some overlapping areas to do so. We do not need an initial guess of the position or correspondences. The result is a model  $\mathcal{M}_{\hat{e}}$ , which includes world points, frames, features, and connection information from both models.

First, we need to identify the set of features in each model which are connected to a world point:  $\mathcal{S}_{p_n^w}$  and  $\mathcal{S}_{p_e^w}$ . These sets are matched against each other, again taking into account the distance ratio and discarding contradicting correspondences as in section 5.1. The result is a set of 3d/3d correspondences. These correspondences are passed to a RANSAC procedure encapsulating an absolute orientation estimation, which estimates the Euclidian transformation  $\mathbf{T} \in SE(3)$  between the points of the models, as well as an overall scale  $\sigma_{\mathbf{T}}$  between the models. We use the algorithm proposed by Umeyama [15] to do so. After the transformation  $\mathbf{T}$  and scale  $\sigma_{\mathbf{T}}$  between the models have been determined, all frames from  $\mathcal{M}_n$  are transformed into the coordinate system of  $\mathcal{M}_e$ :

$$\begin{aligned}\mathcal{F}_{\hat{e}'} &= \mathcal{F}_n \mathbf{T}^{-1} \\ \mathcal{F}_{\hat{e}}^t &= \mathcal{F}_{\hat{e}'}^t \sigma_{\mathbf{T}} \\ \mathcal{F}_{\hat{e}}^R &= \mathcal{F}_{\hat{e}'}^R,\end{aligned}\tag{5}$$

where  $\mathcal{F}^t$  denotes the translation part of the frame's transformation,  $\mathcal{F}^R$  the rotation. Feature positions do not need to be transformed, they stay in the coordinate system of the according frame.  $h_e$  is joined with  $h_n$ :

$$h_{\hat{e}} = h_e \cup h_n.\tag{6}$$



**Fig. 3.** Model integration. We built two models from frames 1 – 10 ( $\mathcal{M}_n$ ) and 21 – 30 ( $\mathcal{M}_e$ ) of the Deutsches Eck sequence and automatically integrated them to a single model ( $\mathcal{M}_{\hat{e}}$ ) afterwards.

World points  $\mathbf{p}^w$  and the relation  $g$  need to be treated differently, depending on whether a world point was an inlier to the result of RANSAC or not. If it was an outlier, it is transformed and added to the model:

$$\mathbf{p}_{\hat{e}}^w = \sigma_{\mathbf{T}}(\mathbf{T}\mathbf{p}_n^w). \quad (7)$$

If it was an inlier, it needs to be *unified* with its corresponding world point. The unification of two world points  $\mathbf{p}_e^w$  and  $\mathbf{p}_n^w$  creates a new world point  $\mathbf{p}_{\hat{e}}^w$  at the position of  $\mathbf{p}_e^w$ . All connections in  $g_n$  and  $g_e$  from the specific world point are then inserted into  $g_{\hat{e}}$ , with  $\mathbf{p}_{\hat{e}}^w$  substituting  $\mathbf{p}_e^w$  or  $\mathbf{p}_n^w$ .

Fig. 3 shows two models and the result of the integration. A run of the global bundle adjustment should always follow the model integration step to minimize possible errors resulting from unified world points.

## 7 Evaluation

In this section we evaluate our model generation and pose estimation approach. We first show some exemplary models, before we take a look at model accuracy and the robustness of the pose estimation algorithm. All images presented here were taken with a 10 megapixel consumer camera.

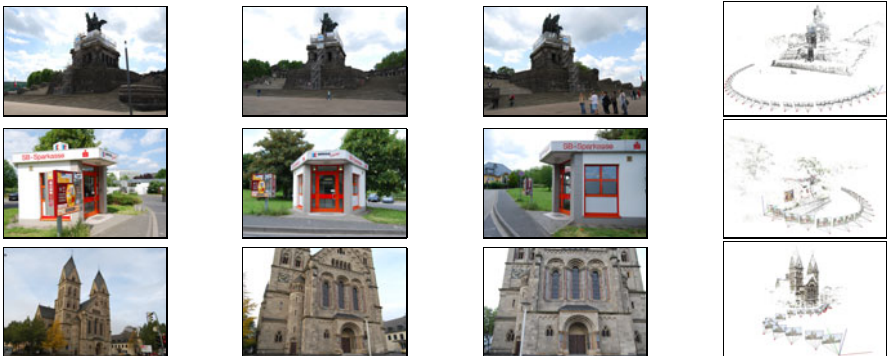
### 7.1 Exemplary Models

We considered three different image sequence:

**Deutsches Eck:** An image series of the Deutsches Eck monument. Images are taken in regular intervalls while circling the monument, focusing the same point approximately.

**ATM:** An image series of the ATM building. Images are taken in regular intervalls while circling the building, focusing the same point approximately.

**Herz Jesu:** An image series of the Herz Jesu church. Images are taken in irregular intervalls while approaching the church, focusing on different parts of the building.



**Fig. 4.** Examples from the Deutsches Eck, ATM and Herz Jesu image sequences and the resulting models.

**Table 1.** Number of frames, world points and the mean reprojection error for all models

model	$\ \mathcal{F}\ $	$\ \mathcal{P}\ $	$\mu(\Delta p^{\mathcal{P}})$
Deutsches Eck	31	41403	1.51057
ATM	18	53367	1.19294
Herz Jesu	20	62886	1.76852

## 7.2 Model Accuracy

To analyze the accuracy of our models, we consider the reprojection error  $\Delta p^{\mathcal{P}}$  of world points, that is the error between the feature location and the corresponding world point reprojected to the image plane. Table 1 lists the mean reprojection errors  $\mu(\Delta p^{\mathcal{P}})$  of all world points and their corresponding features, measured in pixels. Note that a 2 pixel error is less than 0.05% of the image’s diagonal.

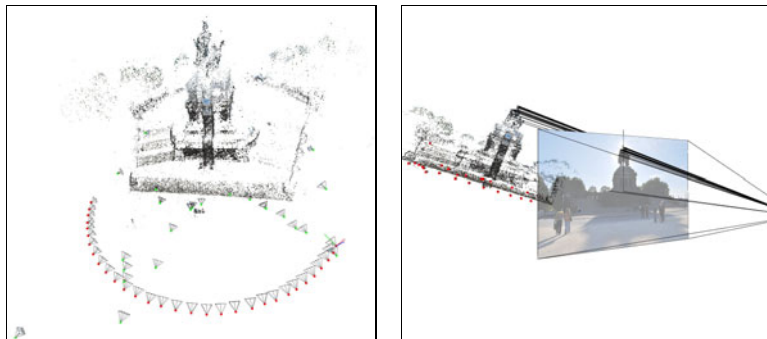
## 7.3 Robustness of the Pose Estimation

To test the robustness of our approach towards changes in the environment, we took a second sequence of images from the Deutsches Eck. The second sequence includes another 31 images, taken from very different positions and viewing angles than the first sequence. They were taken several weeks later, when the scaffolding was removed from the monument and the weather condition was very different. The main challenges are a difficult lighting situation with backlighting, the missing scaffolding which contributed many features to the model, as well as the very wide baseline of several images towards the first sequence. Fig. 5 shows two exemplary images from the second sequence, for the first sequence see Fig. 4.

The results of our pose estimation applied to the images of the second sequence is visible in Fig. 6a. Note that the images of the second sequence were *not* used to enhance the model iteratively.

Since there was no ground truth of the image sequences, we had to determine manually if the computed pose was correct or not. From the 31 images, 23 times the pose was computed correctly, in 3 cases there was only a small error and for 5 images the pose estimation produced erroneous results. In most cases, this

**Fig. 5.** Examples from the second image sequence of the Deutsches Eck



(a) Green cameras mark the positions of frames from the second sequence, red cameras mark the positions of frames from the first sequence which was used to build the model.

(b) An example of a wrong estimated pose due to degenerate data and a single outlier.

**Fig. 6.** Pose estimation applied to the second Deutsches Eck sequence

was due to some degenerate point configuration which we do not detect and handle correctly yet, as in Fig. 6b. There, a degenerate set of points allows the pose estimation to include a single outlier (at the bottom of the image) and still appear valid.

## 8 Conclusion

In this paper we presented a formalism for a model suitable for image based pose estimation. The model uses SURF features solely. We showed how to create a model from images automatically, and how pose estimation on a model works. We also formulated an algorithm to integrate two models which partially overlap into a single model.

Our evaluation revealed a high accuracy of the automatically generated models, with a mean reprojection error of world points less than 0.05% of the image's diagonal. We showed that the proposed model can be used for pose estimation even for images taken under different, more difficult lighting situations with large changes in viewpoint and a partially changed world. It is therefore suitable for initialization of pose tracking or similar applications in changing outdoor environments.

In future work we would like to address the scalability of our approach and refine the detection of degenerate configurations to make pose estimation even more robust. We would also like to test our model building and pose estimation on images with ground truth information.

This work was supported by the DFG under grant PA 599/7 and PR 161/12.

## References

1. Bay, H., Ess, A., Tuytelaars, T., van Gool, L.: Speeded-up robust features (surf). *Journal of Computer Vision* 110, 346–359 (2008)
2. Zhang, W., Kosecka, J.: Image based localization in urban environments. In: 3DPVT 2006: Proceedings of the Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT 2006), pp. 33–40. IEEE Computer Society, Washington, DC (2006)
3. Schindler, G., Brown, M., Szeliski, R.: City-scale location recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–7 (2007)
4. Wu, C., Fraundorfer, F., Frahm, J.M., Pollefeys, M.: 3d model search and pose estimation from single images using vip features. In: *Computer Vision and Pattern Recognition Workshop*, pp. 1–8 (2008)
5. Snavely, N., Seitz, S.M., Szeliski, R.: Modeling the world from internet photo collections. *Int. J. Comput. Vision* 80, 189–210 (2008)
6. Irschara, A., Zach, C., Frahm, J., Bischof, H.: From structure-from-motion point clouds to fast location recognition, pp. 2599–2606 (2009)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110 (2004)
8. 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)
9. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press, Cambridge (2004)
10. Nistér, D.: An efficient solution to the five-point relative pose problem. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 195–202 (2003)
11. Lourakis, M., Argyros, A.: Sba: A software package for generic sparse bundle adjustment. *ACM Trans. Math. Software* 36, 1–30 (2009)
12. Lourakis, M.: levmar: Levenberg-marquardt nonlinear least squares algorithms in c/c++ (2004) (accessed on January 31, 2005)
13. Muja, M.: Flann, fast library for approximate nearest neighbors (2009), <http://mloss.org/software/view/143/>
14. Fiore, P.D.: Efficient linear solution of exterior orientation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 140–148 (2001)
15. Umeyama, S.: Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 376–380 (1991)

# Real-Time Camera Tracking Using a Global Localization Scheme

Yue Yiming, Liang Xiaohui, Liu Chen, and Liu Jie

State Key Lab. of Virtual Reality Technology and Systems,  
Beihang University, Beijing, 100191, P.R. China

**Abstract.** Real-time camera tracking in previously unknown scene is attractive to a wide spectrum of computer vision applications. In Recent years, Simultaneous Localization and Mapping (SLAM) system and its varieties have shown extraordinary camera tracking performance. However, the robustness of these systems to rapid and erratic camera motion is still limited because of the typically used Local Localization scheme. To overcome this limitation, we present an efficient online camera tracking algorithm using a Global Localization scheme which matches features in a global way through two steps: First, coarse matches are obtained through nearest feature descriptor search. Afterwards, a Game Theoretic approach is exploited to eliminate the incorrect matches and the left correct matches can be used to estimate the camera pose. Result shows our camera tracking algorithm has significantly improved the robustness of camera tracking system to rapid and erratic camera motion.

## 1 Introduction

Vision-based camera tracking aims to estimate the pose (6-DOF parameters) of a camera relative to its surroundings based on the input image sequence or live video. This is attractive for many computer vision applications, e.g., 3D reconstruction, video registration and enhancement. Traditionally, this problem is solved by the offline Structure from Motion (SfM) methods [1,2,3]. However, in some practical applications, such as Augmented Reality (AR) and Autonomous Navigation, the urgent camera pose is of great necessity. In such cases, the offline method could not satisfy the efficiency requirements, and therefore online real-time camera tracking has drawn more attentions in recent years.

A camera tracking system called Simultaneously Localization and Mapping (SLAM) and its varieties [4,5,6,7,8,9] have shown extraordinary camera tracking performance. With little or even no prior knowledge of the scene, the SLAM systems can estimate the immediate camera pose accurately and efficiently. This extends the applicable field of the camera tracking technique. However, with less prior knowledge, it also leads to the weakness of the agility and robustness of the camera tracking systems. Here, we use [9]'s definition of the agility: the robustness of the camera tracking system to rapid camera motion.

We observed that a key factor which constrains the robustness and agility of the prevalent SLAM varieties is the local localization (LL) scheme which matches

the map and the keypoints extracted from input image in a local way. In most SLAM varieties, a motion model is used to predict the current camera pose, and afterwards the optimization in LL scheme will converge near the prediction. Consequently, the final result highly depends on the initial state (the predicted camera pose). Once the prediction is not reliable, the local optimization tends to converge to an incorrect state, and subsequently leads to the failure of tracking. In fact, this unreliable case is common for many reasons, e.g. sudden camera move, error accumulation. As a result, using a LL scheme, the robustness and agility of the camera tracking system is limited.

In contrast to the local localization scheme, the global localization (GL) scheme [10,11] which matches the map points and features in a global way can overcome this limitation. However, as [10] point out, there are two common problems in prior GL scheme works. First, it is difficult to achieve real-time performance due to expensive feature extraction and matching, even in a relatively small working space. Second, these methods rely excessively on the feature distinctiveness, which cannot be guaranteed when the scale of the scene is large or the scene contains repeated structures. [10] proposed a camera tracking algorithm using a GL scheme, which involves an offline process for space abstraction using features and an online step for feature matching. On the one hand, this strategy transfers the expensive computation of the map building to the offline process, and therefore makes the global feature matching possible to be achieved in the online step. On the other hand, however, the practical workspace of this camera tracking algorithm is consequently constrained to the area where map has been built in the offline step and could not be extended online.

In this paper, we propose a completely online real-time camera tracking algorithm using a GL scheme to achieve robust and efficient camera tracking performance without any prior knowledge of the scene. To overcome the efficiency problem in GL scheme, we exploit a signature feature descriptor which is designed to be fast enough to train and match online based on statistical learning techniques. To reduce the dependence of the feature distinctiveness, we adopt a novel map maintenance and selection strategy and a game-theoretic based global matching approach in which the global geometry information of the scene is brought into the matching process to eliminate mismatches. Our GL scheme retains the correct matches that are compatible to a rigid transformation and uses these matches to calculate the camera pose.

## 2 SLAM and Its Varieties

Monocular SLAM [4], the first successful application of the SLAM methodology in real-time camera tracking field, was demonstrated by Davison in 2003. Since then, there have been attempts to improve the scalability, robustness and agility of monocular SLAM. [5,6] attach descriptors to map points either to reduce data association error or to relocalize the camera in case of failure. Eade and Drummond [7] employ a different statistical framework which allows denser maps to improve tracking quality. However, because of the low time efficiency of these

descriptors and the LL scheme, these improvements did not ameliorate the camera tracking agility and robustness either directly or significantly. Georg Klein [8] et.al proposed a new PTAM system which split the tracking and mapping into two separate tasks. This allows the use of computationally expensive batch optimization technologies in mapping and thus leads to extraordinary accurate tracking result. However, this system still suffers from the problem caused by the LL scheme that when camera moves to a new scene, the agility is limited until the new map has been established. This leads to the failure in many practical applications in which cameras exploring new scene is a common task.

Different from the SLAM systems, Dong et.al [10] proposed a keyframe-based camera tracking algorithm using a two step strategy. The first offline step is to extract sparse invariant features (using SIFT [12] in the implementation) from the captured reference images and the successive online step matches them with the features extracted from the captured live video frame. For the reason that the offline step has built the whole map of the scene, the online step has less computation burden and thus could employ the high accuracy but low efficiency SIFT descriptor to obtain high tracking agility. Though this algorithm leads to a robust and accurate camera tracking performance, it would be preferable if the camera tracking can be completely online, especially when the offline step is complex and time-consuming.

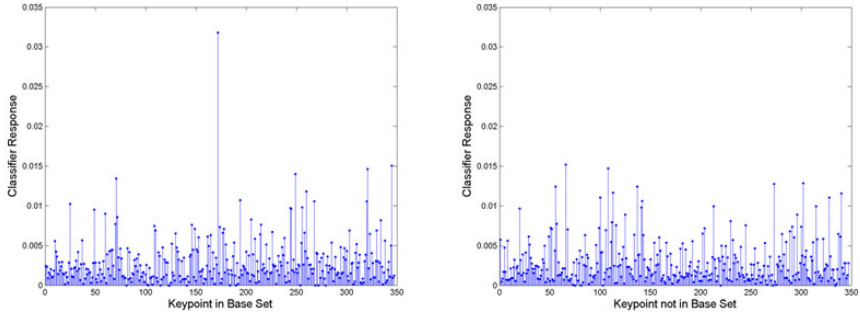
### 3 Global Localization Scheme

In this section, we describe our global localization scheme. Its purpose is to improve the robustness of the feature matching procedure which is a key step of most camera tracking systems. It includes two steps: First, a coarse matching procedure is adopted to form a preliminary match set; Afterwards, a game theoretic approach is used to eliminate the incorrect matches in this set. The core of the first step is a signature feature descriptor which could be trained and matched efficiently. The formulation and evolution of a game theoretic framework is the focus of the second step.

#### 3.1 Signature Descriptor

Though many feature descriptors, e.g. SIFT in [5], Random list in [6], have been employed in SLAM varieties, few of them is used directly in the matching procedure because of their low efficiency. To ensure real-time performance, the efficiency of the matching between the map points in the built map and the keypoints extracted from the input image is extremely important. Thus, many efforts have been made in recent years to speed up the matching procedure of feature descriptors. Among these, a group of statistical learning technique based descriptors [13,14,15,16,17] have attracted our attention because they are designed to achieve fast matching whereas preserve high recognition rate. Though these feature descriptors are much more efficient than the traditional feature descriptor [12,18], they still hardly achieve real-time performance. Take advantage of the multi-cores in modern computers, we exploit the parallel computing





**Fig. 1.** LEFT: the response of Randomized Tree classifiers to a keypoint in the base set. There is only one spike that represent the keypoint is recognized as a keypoint in the base set. RIGHT: the response of Randomized Tree classifiers to a keypoint not in the base set. There is no especially high response of classifiers and this is the signature of this keypoint.

technique to speed up the feature training and matching to satisfy the real-time requirement.

In our algorithm, we use the signature descriptor proposed by Michael Calonder et.al [13] to match the keypoints because its training phase is also efficient. The signature descriptor relies on the fact that if we train a Randomized Tree classifier [14] to recognize a number of keypoints extracted from an image database, all other keypoints can be characterized in terms of their response to these classification trees. Given a few training images, a set of keypoints could be extracted and organized as a base set. Then a Randomized Tree classifier [14] can be trained to recognize the keypoint in the base set under arbitrary perspective, scale and light condition. Given a new keypoint that is not in the base set, we show below that the classifier responds to it in a way that is also stable to changes in scale, perspective, and lighting. We therefore take this response to be the compact and fast-to-compute signature we are looking for.

Each keypoint  $u_i \in \mathbb{R}^2$  in the base set is related to exactly one point  $k_i$  in 3D. Given a set of  $N$  points  $K = k_1, \dots, k_N, k_i \in \mathbb{R}^3$ ,  $N$  is the base size, we then build a classifier based on Randomized Trees that is able to recognize the  $k_i$  under varying conditions. Let  $p_i$  be the patch centered on  $u_i$ . Then the classifier provides a function  $C(p_i)$  mapping a patch  $p_i$  to a vector in  $\mathbb{R}^N$ . Using the notation  $C^{(j)}(p_i)$  to refer the  $j$ -th element of the vector  $C(p_i)$ ,  $1 \leq j \leq N$ , we can state a special property of  $C$ :

$$C^{(j)}(p_i) \text{ is } \begin{cases} \text{large} & \text{if } j = i \\ \text{small} & \text{otherwise} \end{cases} \quad (1)$$

This is shown in Fig.1 for  $i = 177$  and  $N = 350$ .

Furthermore, let  $T(p, \theta)$  be a transformation of an image patch  $p$  under viewing condition change  $\theta$ .  $\theta$  typically encodes changes in illumination, viewpoint, or scale. If the classifier has been trained well, we can assume that

$$\forall \Theta : C(p) \approx C(T(p, \Theta)) \quad (2)$$

When we consider a new 3D-point  $k$  that does not belong to  $K$  and center a patch  $q$  on the keypoint corresponding to  $k$ , we can define the signature of the patch  $q$  simply as

$$\text{Signature}(q) = C(q) \quad (3)$$

A patch  $q'$  centered on the keypoint of  $k$  in another image can be written as  $T(q, \Theta)$ , for some  $\Theta$ . Under the assumption of Equ. (2), the signature of  $q'$  is equal to the signature of  $q$  because

$$\text{Signature}(q') = C(q') = C(T(q, \Theta)) = C(q) = \text{Signature}(q) \quad (4)$$

In other words, the signature is stable under changes in viewing conditions. Thus, we can exploit the signature descriptor to formulate the preliminary match set. In detail, the map is projected into the image plane using the pose of the last input frame and a map point selection strategy is adopted to select those map points which are most probably appears in the input frame. These selected map points are organized as a base set and a feature tree is trained to facilitate the afterwards nearest feature search of the keypoint extracted from the input frame. Then for each keypoint, there is a corresponding map point. We organize these correspondence as the preliminary match set.

### 3.2 Game Theoretic Approach

It is illustrated in Fig.2 that there exist mismatches that will lead to the accuracy of the estimate of the camera pose degenerate significantly. In many computer vision applications, RANSAC [19] has been used to eliminate outliers. However, it is not suitable to our case. RANSAC can only estimate one model for a particular dataset, and therefore can hardly address the problem with more than 50% outliers. Due to the real-time requirement of a camera tracking algorithm, outliers in the preliminary match set are frequently more than 50% of total matches. This leads to the useless of exploiting RANSAC in our GL scheme, and on the other hand validates the effectiveness of our game theoretic approach which is demonstrated below.

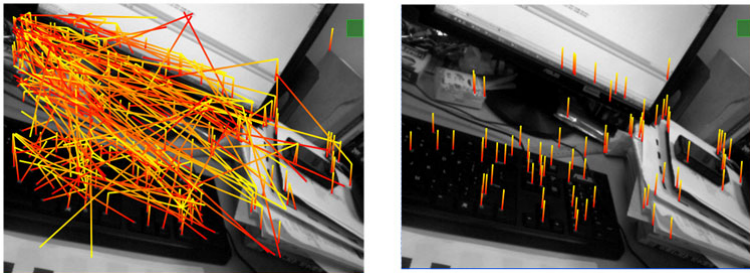
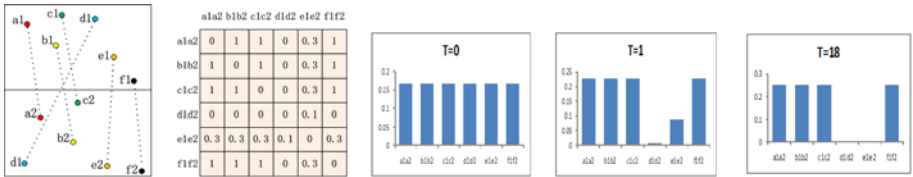


Fig. 2. Matches before and after Game Theoretic Approach

We exploit a Game Theoretic approach which brings the global geometry information into the matching procedure to eliminate the mismatches. In the camera tracking applications, a general accepted assumption is that the scene must remain static. Under this assumption, the transformation between the points observed in two different frames is a rigid transformation. Since all rigid transformation preserves Euclidean distances, we take advantage of this property to eliminate the mismatches and ensure the accuracy of the estimated camera pose. This Game Theoretic approach is proposed by [20] to solve the surface registration problem. We first introduce the underlying idea of this approach and then show how it works in our algorithm.

The key idea of the approach is selecting the sets of point-correspondences that are mutually compatible with a single rigid transformation. Fundamental to this approach is the fact that requiring the compatibility to a single transformation is equivalent to requiring that there exists a compatible transformation for each pair of mates. Following [20], we model the mismatch elimination procedure in a Game Theoretic framework, where two players extracted from a large population select a pair of corresponding points from the base set and the extracted keypoints. If there exists a rigid transformation that moves both his point and the other player’s point close to the corresponding mates, then both players receive a high payoff, otherwise the payoff will be low. In general, as the game is repeated, players will adapt their behavior to prefer matings that yield larger payoffs, driving all inconsistent hypotheses to extinction, and settling for an equilibrium where the pool of mates from which the players are still actively selecting their associations forms a cohesive set with high mutual support. Within this formulation, the solutions of the matching problem correspond to evolutionary stable states (ESS’s), a robust population-based generalization of the notion of a Nash equilibrium. An illustration of the elimination procedure is shown in Fig.3.



**Fig. 3.** An example of the evolutionary process. 6 point correspondences have been matched by the feature finding step, and 6 mating strategy are selected for initial hypothesis. The matrix shows the compatibilities between pairs of mating strategies according to a one-to-one rigidity enforcing payoff function. Initially (at  $T=0$ ) the population is set to the barycenter of the simplex. After just one iteration,  $(d1, d2)$  and  $(e1, e2)$  have lost a significant amount of support. After eight iterations ( $T=18$ ), the evolution has converged, the matches that are more coherent to rigidity  $((a1, a2), (b1, b2), (c1, c2), (f1, f2))$  have high weights while the weights of the matches that do not coherent to rigidity  $((d1, d2), (e1, e2))$  evolve to 0.

In practice, for each extracted keypoint in the input frame, we use all the correspondences in the preliminary match set as mating strategies. On the other hand, to enforce the rigid constraint of the correctly matched pairs, we need to assign a rigidity-enforcing payoff function. Typically there are two candidates: the negative exponentiation of the difference between the distances of the model and data points and the ratio between the min and the max distance. We observed that the first one is too steep that some correct matches will also been eliminated while the second one is too shallow. Thus, we choose a compromise which is the  $N$  times power of the ratio between the min and the max distance.

Then we could start the non-cooperative mating game in which the search for a stable state is performed by simulating the evolution of a natural selection process. The evolution procedure is described in [20]. Once the population has reached a local maximum, all the non-extincted mating strategies can be used to calculate the estimate camera pose of the input frame. However, to improve the accuracy of estimated camera pose, we set a threshold to select the matches with a higher weight. With the selected 3D to 2D matches, we could estimate the camera pose of the input frame.

## 4 Experiments and Results

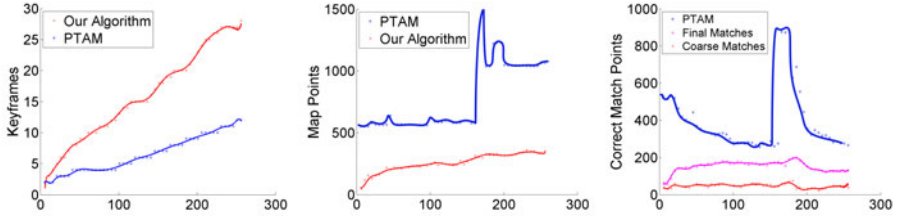
All the results in this section are produced on a computer with a Intel Core2 Quad 2.66GHZ CPU. The reference images are captured by a Microsoft VX-6000 web camera with  $71^\circ$  wide-angle.

We compare our camera tracking performance with a landmark of SLAM varieties, the publicly accessible code PTAM [8] with a real indoor data sequence. Our comparison includes the three following aspects: matching performance, time efficiency and tracking result.

### 4.1 Matching Performance

First we compare the matching procedure of our algorithm with the PTAM implementation. The PTAM implementation produces a relatively dense map which contains thousands of features at most, and the accuracy of estimated camera pose relies on large amount of features. On the contrary, for the reason that we exploit the Game-Theoretic approach, we do not try to find a good estimate of camera pose by means of a vote of large number of matches; instead we take advantage of the internal coherence between the feature points and therefore fewer matches are needed to estimate a more accurate camera pose. Thus our map contains fewer map points than the PTAM implementation and this benefits the efficiency of the map expansion procedure.

Fig.4 shows the evolution of the map in these two systems. As the correct matches in the PTAM implementation varies sharply, the correct matches in our algorithm is much more stable. This leads to a more robust estimate of the camera pose. Furthermore, less map points speed up the Bundle Adjustment (BA) used in the mapping thread and more keyframes could be handled by the



**Fig. 4.** The map comparison of our algorithm with the PTAM implementation. As the camera explores in the scene, the map points and keyframes both increase. The increase of keyframes in our algorithm is much faster than that in PTAM while the increase of map points is on a opposite way. In the right figure, we can observe that the matches in our algorithm are more stable than PTAM.

system. Since the keyframe is the representation of known scene, more keyframes in the map means more area of the scene is covered by the map and further leads to a low possibility of tracking failure.

## 4.2 Timing

Time efficiency is a preliminary requirement of real-time camera tracking systems. Table 1 shows the average time spent in each step of our algorithm and the PTAM implementation. We note that the average overall time spent in each input frame is 56.1ms in our algorithm. That means the frame rate is near 18 frames per second. Comparing to the PTAM implementation, the efficiency of our algorithm is slightly lower. However, this does not significantly affect the applicability of our algorithm because they both satisfy the real-time requirement.

Table 1 Time efficiency of our algorithm and PTAM implementation. The left table shows the time spent with a single thread and multi-threads. The right one shows that of the PTAM implementation.

**Table 1.** Time Efficiency

Our Algorithm	Single thread	Multi-threads	PTAM	Time Spent
Image Retrieve	12.2ms	12.1ms	Image Retrieve	12.1ms
Base Set Selection	2.6ms	2.3ms	Keyframe preparation	2.2ms
Nearest Feature Searching	35.1ms	10.2ms	Feature projection	3.5ms
Game Theoretic Approach	84.2ms	30.5ms	Patch search	9.8ms
Pose Estimate	1.0ms	1.0ms	Iterative pose update	3.7ms
Overall	135.1ms	56.1ms	Overall	31.3ms

### 4.3 Tracking Result

Tracking agility is one of the most important factors that evaluate the performance of a camera tracking system. Since our algorithm exploits the global localization scheme, it could be used in applications in which the camera moves more erratically or suddenly. It is hard to convey the behavior of a real-time tracking system on paper, so we encourage the reader to refer to the attached results video which demonstrates the operation of the system. Subjectively, we note that an obvious change in our algorithm is the improvement of the tolerance to elastic camera motion, especially those motions that do not obey the prediction of motion model. In addition, we observed that when the camera explore to an unknown scene, our algorithm extends the map much more quickly than the PTAM. This property makes our algorithm more suitable in applications in which the camera needs to explore new scene frequently and therefore extends the applicable field of real-time camera tracking technique.

## 5 Conclusion

In this paper, we proposed a real-time camera tracking algorithm using an efficient global localization scheme. This GL scheme significantly ameliorates the dependence of the prediction of camera pose in prior SLAM varieties and leads to a camera tracking algorithm that is more robust to erratic and fast motion in the previous unknown scene of a camera. Of course, there still exist limitations in our algorithm. Though the efficiency of the signature descriptor is high, its distinctiveness is relatively low, at least in our implementation. This leads to the decrease of correct matches when the base set is large. As many new efficient feature descriptors are proposed recently, the signature descriptor can be replaced to improve the performance of feature matching.

**Acknowledgement.** This paper is supported by the National High-Tech Research & Development 863 Program of China under Grant No. 2009AA012103, the National Natural Science Foundation of China under Grant No. 60873159, and the Beijing Municipal Natural Science Foundation under Grant No. 4112032.

## References

1. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press, Cambridge (2004), ISBN: 0521540518
2. Wong, K.Y.K., Cipolla, R.: Structure and motion from silhouettes. In: *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV 2001)*, vol. 2, pp. 217–222 (2001)
3. Pollefeys, M., Gool, L.V., Vergauwen, M.F.V., Cornelis, K., Tops, J., Koch, R.: Visual modeling with a handheld camera. *International Journal of Computer Vision* 59, 207–232 (2004)
4. Davison, A., Reid, I., Molton, N., Stasse, O.: Monoslam: Real-time single camera slam. *IEEE Trans. Pattern Analysis and Machine Intelligence* 29, 1052–1067 (2007)

5. Chekhlov, D., Pupilli, M., Mayol-Cuevas, W., Calway, A.: Real-time and robust monocular slam using predictive multi-resolution descriptors. In: Proceedings of the 2nd International Symposium on Visual Computing, pp. 276–285 (2006)
6. Williams, B., Klein, G., Reid, I.: Real-time slam relocalisation. In: Proceedings of 11th IEEE International Conference on Computer Vision (ICCV 2007), pp. 1–8 (2007)
7. Eade, E., Drummon, T.: Scalable monocular slam. In: Proceedings of IEEE Intl. Conference on Computer Vision and Pattern Recognition (CVPR 2006), pp. 469–476 (2006)
8. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR 2007), pp. 1–10 (2007)
9. Klein, G., Murray, D.: Improving the agility of keyframe-based slam. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 802–815. Springer, Heidelberg (2008)
10. Dong, Z.L., Zhang, G.F., Jia, J.Y., Bao, H.J.: Keyframe-based real-time camera tracking. In: Proceedings of IEEE International Conference on Computer Vision (ICCV 2009), pp. 1538–1545 (2009)
11. Se, S., Lowe, D., Little, J.: Vision-based global localization and mapping for mobile robots. *IEEE Transactions on Robotics* 21, 364–375 (2005)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–100 (2004)
13. Calonder, M., Lepetit, V., Fua, P.: Keypoint signatures for fast learning and recognition. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 58–71. Springer, Heidelberg (2008)
14. Lepetit, V., Fua, P.: Keypoint recognition using randomized trees. *Transactions on Pattern Analysis and Machine Intelligence* 28, 1465–1479 (2006)
15. Ozuysal, M., Fua, P., Lepetit, V.: Fast keypoint recognition in ten lines of code. In: Proceedings of 20th Conference on Computer Vision and Pattern Recognition (CVPR 2007), pp. 1–8 (2007)
16. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast keypoint recognition using random ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 448–461 (2010)
17. Calonder, M., Lepetit, V., Fua, P.: Pareto-optimal dictionaries for signatures. In: Proceedings of 23rd Conference on Computer Vision and Pattern Recognition, CVPR 2010 (2010)
18. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)* 110, 346–359 (2008)
19. Fischler, A.M., Bolles, C.R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24, 381–395 (1981)
20. Albarelli, A., Rodol, E., Torsello, A.: A game-theoretic approach to fine surface registration without initial motion estimation. In: Proceedings of 23rd Conference on Computer Vision and Pattern Recognition, CVPR 2010 (2010)

# Visual Mapping and Multi-modal Localisation for *Anywhere* AR Authoring

Andrew P. Gee, Andrew Calway, and Walterio Mayol-Cuevas

Dept. of Computer Science, University of Bristol, UK

**Abstract.** This paper presents an Augmented Reality system that combines a range of localisation technologies that include GPS, UWB, user input and Visual SLAM to enable both retrieval and creation of annotations in most places. The system works for multiple users and enables sharing and visualizations of annotations with a control centre. The process is divided into two main steps i) global localisation and ii) 6D local mapping. For the case of visual relocalisation we develop and evaluate a method to rank local maps which improves performance over previous art. We demonstrate the system working over a wide area and for a range of environments.

## 1 Anywhere Authoring

Most Augmented Reality (AR) systems to date can be categorized by either having high levels of accuracy in small scale spaces, as provided by 3D visual simultaneous localisation and mapping (SLAM), or systems covering larger areas but resorting to approximate location, as offered by GPS. The former systems are capable of delivering accurate 3D object registration in unprepared environments and the latter well suited to deliver, for example, audio AR outdoors.

The vast majority of systems have also concentrated on the *retrieval* rather than the *input* of content, and therefore an AR application is often described solely as a system where annotations are visualized when the user is at the right location. To differentiate an AR system's ability to both retrieve and input content in any area, we use the term *anywhere authoring*. This is an ability needed in applications that aim to take AR to the next level of impact e.g. a fine-grained city maintenance system, worldwide AR encyclopedias or wide area forensics.

To combine GPS and local visual mapping may appear to be sufficient for anywhere authoring. Unfortunately this is not the case, in part because users spend most of the time indoors where GPS positioning is unreliable at best. This seriously hampers AR for most of the places that can be annotated and places high requirements on the visual mapping that can work indoors. In order to offer truly wide and robust anywhere authoring it appears likely that a range of localisation technologies from GPS to indoor positioning systems jointly with visual mapping have to operate seamlessly as the user moves in and out of areas. This, combined with an adequate framework for the propagation of both



existing and newly created content, are crucial for enabling fluid AR interactions anywhere.

To our knowledge, a system that seamlessly combines these many levels and modalities of localisation accuracies and the ability to enable users to retrieve and input AR content anywhere has not been presented before.

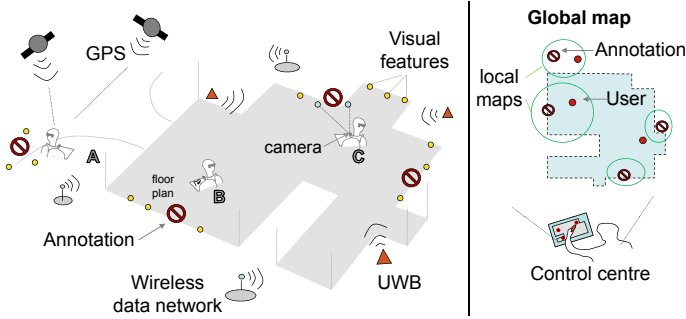
## 2 Related Work

The combination of global and local sensing has been explored in the related field of ubiquitous computing for some time. As an example using computer vision, the works in [12] use visual feature descriptors to provide accurate object detection while GPS helps in the gating of the objects' database based on location. In both examples, the objects of interest are buildings whose facades are usually distinctive, relatively large, and less prone to perspective and occlusion problems. To extend the area of operation for AR outdoors, GPS was also the natural choice and this was the case for early systems e.g. [3]. The further addition of inertial sensors, markerless visual tracking and aerial photographs to GPS in [4] has achieved higher accuracy annotation of large, outdoor scenes. More recently, in [5] GPS combined with inertial sensors is shown to be able to deliver relatively good visualization of underground pipes outdoors despite not using visual methods.

For wide area indoor AR, systems have used localisation methods that include ultrasonic positioning [6] or odometry recovered from the user's steps [7], as well as visual tags from the ARToolkit or similar to provide well localized annotations [8,9,10]. Another recent alternative indoors is Ultra Wide Band (UWB) which in [11] has been combined with fiducial markers to provide extended indoor operation. The combination of inertial sensors and visual markers has been used in [12]. In the case of [13] ultrasound and GPS are combined with visual SLAM and demonstrated in a small scale environment.

The use of a global reference provided by any of the above methods helps to improve the localisation results and prepares the scene for integration of technologies with different accuracy granularities. When the global frame of reference is not built-in, the extreme alternative is to use the visual appearance of each area of interest as the way to position the user. This is the case in [14] where a visual SLAM system creates small submaps that are kept disjointed and that are compared against an input image to detect that the user is in the same area once again. Assuming that no area looks exactly the same, this is a viable possibility, however the scalability of a system based on purely visual (even when combined with geometric) appearance, and disregarding any global reference, appears unrealistic. Furthermore, a system that can deliver true anywhere authoring is likely to encounter areas where no global reference either from indoor positioning or GPS is available and this demands an alternative referencing method.

While some of the above systems combine a few localisation techniques, none seems to have the seamless interaction over the different areas that we are after. Importantly, none of them appear to be built with a multi-user and robust



**Fig. 1.** System overview showing multiple users authoring a scene with AR annotations and using different localisation methods. See text for detailed explanation.

communications infrastructure for the input of annotations, as needed for anywhere authoring.

### 3 Operational Overview

Figure 1 shows an overview of the overall system in operation featuring three different modes of localisation: A) GPS, B) floor plan maps and C) UWB. The insertion and retrieval of annotations is made locally accurate by using visual SLAM. Figure 1 shows the SLAM features represented by yellow circles which serve as anchor points for the annotations. A communications infrastructure (in our case using WiFi and TETRA [15]), links users and allows visualization in a global map at a control centre. Local SLAM maps (green circles in global map) are positioned in the global reference with different accuracies depending on the positioning method at the time of authoring. However, visible annotations will always be displayed with local accuracy relative to the camera thanks to the automatic SLAM relocalisation, even if the location of the annotation in the global map is metrically inaccurate. The global map is used primarily as a topological representation for gating and rough navigational guidance.

### 4 Multi-modal Positioning

In this paper we divide the overall operation of the system into two main steps: i) locate the user in 3D space and ii) use a 6D referencing method to position accurately local AR annotations. The positioning of the user helps to establish a frame of reference that can later be used to provide only the relevant information for the immediate environment. This is the idea of location-based gating mentioned before. User positioning needs to be achieved on indoor and outdoor areas before we can combine it with an accurate local frame of reference.

**GPS and UWB.** As with other systems, we employ GPS, when available, to provide an accepted alignment with an absolute frame of reference. Our GPS uses a Teseo GPS chipset to provide 3D positioning accurate to 2m with a 50% confidence limit. For the indoors case we employ a UWB positioning system

composed of multiple transponders [16]. These can be located indoors or outdoors and self-calibrate once they are active. In a typical indoor environment, the UWB system provides 3D positioning to at least metre-level accuracy, enabling the visualization of paths and places that users have visited. Accuracy varies according to the coverage of the UWB base units, which is affected by obstructions in the lines of sight between units and reflective surfaces in the environment that add multipath effects.

A rigid transformation can be found to align the UWB transponders with a reference from GPS, however when a global map is not required this alignment is not necessary. This is because even if these two references (GPS and UWB) are kept separate, the system can still determine at any instance if there is coverage by one or the other system and a decision can be made as to which reference will be used with priority (in our case it is UWB). Recall that an external reference is sought only for the task of gating which annotations should be near the user. This does not require an absolute or aligned set of frames of reference. In our system, switching between the UWB and GPS is transparent to users.

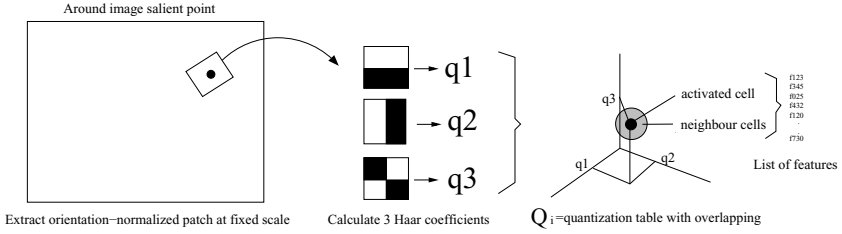
**Interactive input.** In contrast to previous systems for wide-area AR, we employ user interactivity as a bridge to operate between the areas covered by GPS and UWB. For the case of a system designed for people, user input is a sensible alternative for positioning almost anywhere. When the user wishes to create an annotation, and when neither UWB nor GPS are available, the system prompts the user to refine location on a 2D map shown centered on the last trusted position fix. The user can then simply select an approximate location in this map. Our system uses street maps showing only the outlines of buildings (Fig. 7), but nothing prevents the use of more detailed map representations. The maps can also potentially be extended to include architectural floor plans if available.

By combining automatic referencng with the interactive user input we are able in principle to perform authoring anywhere.

## 5 Visual Mapping and Relocalisation

The requirement for working in unprepared, untagged environments has favoured the use of visual SLAM methods. Indeed, it was the construction of a local map for an AR scenario that was the first application of real-time visual SLAM. That system was based around an EKF process [17]. The PTAM system [18], a more recent take on the problem, uses bundle adjustment and splits the tasks of mapping and tracking to make gains from parallelization while delivering impressive results.

While the framework for mapping is important to the achievable accuracy, it is the way in which the system will re-localise in a previously visited area which is more critical for the application we are considering in this paper. Anywhere authoring demands a method that is able to work with efficiency over many local submaps while providing unambiguous camera pose recovery. This is important because although location-based gating helps to reduce ambiguity, a truly robust system should be able to work when there is large uncertainty in the location



**Fig. 2.** The process of relocalisation from an input image in a single local map is based on the computation of three appearance coefficients per saliency point to approximate a nearest neighbor search using a quantization table

of the user, perhaps when entering an area annotated using interactive input as described above, or if one of the other positioning systems fail.

In [19] a method is presented for visual SLAM relocalisation that uses randomized trees for re-detecting features, combined with a RANSAC verification step for pose estimation. Randomized trees are generated offline and use relatively large storage space — about 1.3MB per map point [14]. The PTAM system [18] uses a relocalisation method based on low resolution keyframes which has been used in the work of [14] for localisation over multiple maps. This approach is better from the point of view of data storage, however, in our experience, keyframe based localisation is prone to false positives, in particular when operating in roughly similar areas.

Another popular alternative is to use visual codebooks as used in [20] to match image frames. Visual codebooks are usually found after an optimization process of clustering and are therefore not easily updated on the fly, something which is corrected in [21].

In this paper we use the method for relocalisation and mapping described in [22]. This method uses robust visual descriptors and geometry consistency checks. The relocalisation is based on a quantization table which is small in comparison to other description approaches (e.g. using randomized trees) and can be updated on the fly. The method described in [22] was designed to work on a single map but in this work we extend that approach to work more efficiently with multiple maps as needed here and as described in Sec. 5.2. Furthermore, our method differs from the previous multiple map relocalisation work in [21] both in the smaller size of the descriptors used and in our use of a relatively small quantization table created only from the 3D features in our SLAM maps.

## 5.1 Single Map Relocalisation

Relocalisation assumes that a map  $M_i$  of features has been built previously and the 3D geometry of features together with their visual descriptors is available. To attempt to relocalize, a saliency detector is run on the input image. Around all image areas above a saliency threshold, a fixed-size window is used to obtain a rough estimate of local orientation. This local orientation allows extraction of a fixed-size patch from which three Haar coefficients are computed. These

coefficients encode the rough appearance of that patch in  $x$ ,  $y$ , and  $xy$ . These numbers are used to index a quantization table  $Q_i$  where descriptors of other similar patches have been stored jointly with their 3D position, i.e. a cell  $c_{ij}$  in  $Q_i$  contains a list of features  $F = \{f_k, \dots, f_m\}$  generated by visual SLAM at the time  $M_i$  was created. In relocalisation, only the descriptors in  $c_{ij}$  and neighboring cells are compared with the input patch’s descriptor. The process is illustrated in Fig. 2. The use of a fixed size patch here does not prevent working at different scales since the system builds a multi-scale stack of descriptors for every feature in a background process [22], and these are indexed too via  $Q_i$ .

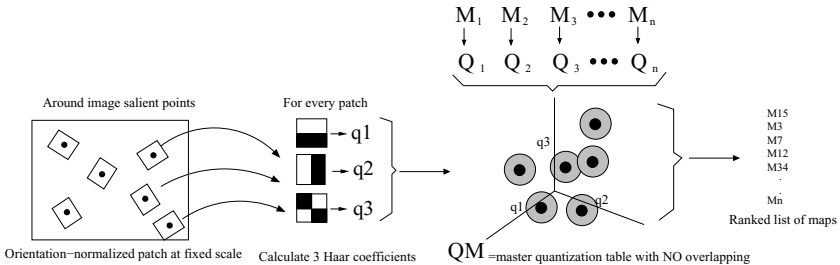
After candidate matches are found with this procedure, a RANSAC method attempts to compute a consistent camera pose. If successful, and if an annotation linked to  $M_i$  is visible in the current frame, it will be displayed as an AR object.

In our tests, this approach uses only about 3% of the comparisons needed by an exhaustive search. The whole process is also fast, usually relocalizing within 50 – 300ms.

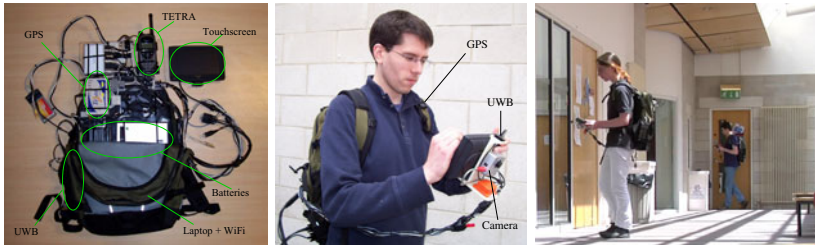
## 5.2 Multiple Maps Relocalisation

When considering many local maps, the naive approach would be to run the above process in every  $M_i$  individually, perhaps gated by location. When the number of maps in a vicinity is small, that process may be sufficient but in general we would need to be prepared to run relocalisation on many maps to ensure robustness. To this end, we developed a system of map ranking based on the single map method described in Sec. 5.1 by combining the information of the individual  $Q_i$ s as follows.

We create a master quantization table  $QM$  based on all the quantization tables  $Q_i$  from the local maps. This  $QM$  uses the same input as needed in the single map relocalisation. The process therefore starts with three Haar coefficients extracted around every salient point in the input image but in this case the coefficients are first used to index cells in  $QM$ . Every cell in  $QM$  keeps a list of the index  $i$  of all the maps  $M$  that have features in that cell. Therefore if a cell in  $QM$  is activated by an input patch, a list of all possible  $M_i$ s that have to be searched is obtained. In addition, each cell is weighted by the **tf-idf** measure in a similar way as introduced in [20] to reflect the uniqueness of a cell. In this way cells



**Fig. 3.** When multiple maps have to be searched to attempt relocalisation, a master quantization table  $QM$  assists in the ranking of the maps to speed up the process



**Fig. 4.** Hardware components and multiple users exploring and annotating an area

that activate for every map will have a lower weight than those that activate for fewer maps. By combining the weighted lists generated for every patch on the image it is possible to rank all maps according to the cosine similarity score between the **tf-idf** vectors for each map and the current image.

The process is illustrated in Fig. 3 and is very fast as we only need to look at the weighted frequency of  $i$  indices and rank them. The rank establishes the order in which relocalisation in the individual maps is to be attempted as per Sec. 5.1. When the first relocalisation is successful the process stops and switches to AR visualization, since in our experience the method does not produce false positives in real applications.

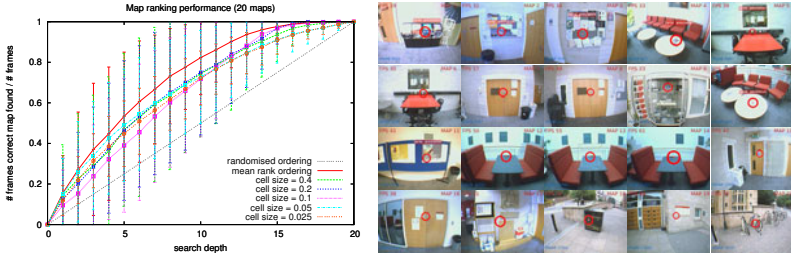
## 6 Experiments with Multiple Maps Relocalisation

Each hardware unit integrates components around a dual core Centrino laptop worn on a backpack as shown in Fig. 4. The interface with the user is displayed on a handheld touchscreen which has a firewire camera with a horizontal FOV of  $80^\circ$  rigidly attached to a 3D orientation sensor (which is not used in this work). The touchscreen also has the UWB antenna attached to it so that the most accurate sensors are close together. The GPS antenna is worn on the backpack’s shoulder strap to enhance reception strength.

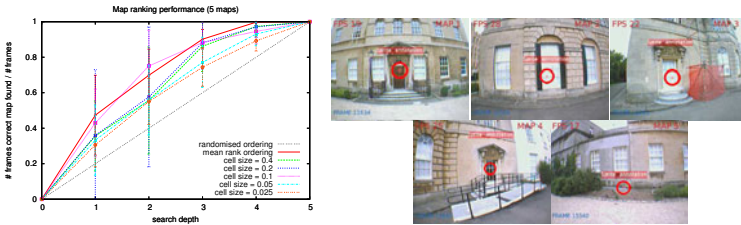
We performed experiments on the performance of the relocalisation in multiple maps. For this we assume the worst case where no location based gating is available. Experiments were conducted for an indoor scenario with 20 maps and an outdoor scenario with 5 maps, as shown in Figs. 5 and 6 respectively. We do not need to consider more maps than this, since the location based gating in the real system will always place a relatively low bound on the number of maps that need to be checked.

The performance of the map ranking was evaluated using camera tracking and exhaustive single-map relocalisation to provide a ground-truth estimate of the correct map for each frame. This was matched against the multiple map relocalisation ranking computed at each frame and used to plot the cumulative distribution function of the ranking. The results of the ranking method were then compared against the baseline case of a randomized sort of the maps.

Five different cell sizes for  $QM$  were tested. Although average performance was better than the baseline in all cases, the results showed that no single cell size



**Fig. 5.** Twenty maps were generated over a large indoor space incorporating many similar areas (several tables with red chairs). The cumulative distribution function of the ranking shows the improvement in performance achieved by the proposed method.



**Fig. 6.** Five maps were generated over a local outdoor area within a 10m radius representative of GPS accuracy. The cumulative distribution function of the ranking shows the improvement in performance achieved by the proposed method.

gave good results for all maps. Sorting the maps by their mean rank over the five different cell sizes improved the average performance and reduced the number of individual maps that performed worse than than the baseline. Alternative methods of combining the ranks from the different cell sizes, such as the median, minimum or maximum rank, were also considered but provided less performance improvement than the mean rank method.

In all cases, exhaustive relocalisation over all maps provided just a single positive match to the correct submap. This is despite the fact that the test sequences contain several instances of maps with very similar appearance. This supports the claim that the single map relocalisation method produces very low false positive rates in real scenes.

## 7 Demonstration

The performance of the system was demonstrated by building multiple maps over a  $0.1\text{km}^2$  area containing a mixture of indoor and outdoor locations. The scenario mimics a maintenance task where users label multiple objects to be revisited by other users at a later time. In some indoor locations a UWB positioning system was available to provide absolute position. The full set of 16 maps is shown in Fig. 7.





**Fig. 7.** Sixteen maps were created over an area containing a mixture of indoor and outdoor locations and with a mixture of GPS, UWB and User Input positioning. The 20m search radius reflects that the user is currently in an area using the interactive input positioning.

In areas with UWB coverage, a 2m distance threshold was used and the separation of the constructed maps was such that a maximum of one candidate map was returned for relocalisation. In one of the maps (map 3), the UWB accuracy was degraded by the surrounding furniture, producing position measurements outside the expected distance threshold and preventing automatic relocalisation. However, single map relocalisation was successful when the map was selected manually from the user interface.

Areas with GPS coverage used a 10m distance threshold and returned a maximum of two candidate maps for relocalisation. In areas requiring interactive input to define absolute position, the 20m distance threshold returned between two and six candidate maps. The multiple map relocalisation method found the correct map within the first two maps tested on each of the six occasions it was used.

## 8 Conclusions

This paper has presented a novel system that combines a range of positioning technologies with local visual SLAM to enable the retrieval and creation of AR annotations. We have developed and evaluated a method for the efficient ranking of visual maps to improve performance and demonstrated the system operating over various areas in a maintenance-like scenario where multiple users cover an area finding and labelling objects practically anywhere in the environment.

**Acknowledgement.** This work was funded by the UK Technology Strategy Board and the UK Engineering and Physical Sciences Research Council. The authors wish to thank all partners in the ViewNet project for their discussions and participation in this work. Ordnance Survey mapping © Crown copyright.



## References

1. Fritz, G., Seifert, C., Paletta, L.: A mobile vision system for urban object detection with informative local descriptors. In: *Int. Conf. on Computer Vision Systems* (2006)
2. Hutchings, R., Mayol-Cuevas, W.: Building recognition for mobile devices: incorporating positional information with visual features. Technical Report CSTR-06-017, Dept. of Computer Science, University of Bristol (2005)
3. Höllerer, T.: *User Interfaces for Mobile Augmented Reality Systems*. PhD thesis. Columbia University (2004)
4. Höllerer, T., Wither, J., Diverdi, S.: Anywhere augmentation: Towards mobile augmented reality in unprepared environments. In: *Loc. Based Services and TeleCartography* (2007)
5. Schall, G., Mendez, E., Kruijff, E., Veas, E., Junghanns, S., Reitingner, B., Schmalstieg, D.: Handheld augmented reality for underground infrastructure visualization. *Personal and Ubiquitous Computing* 13 (2009)
6. Newman, J., Ingram, D., Hopper, A.: Augmented reality in a wide area sentient environment. In: *Int. Symp. on Augmented Reality* (2001)
7. Kourog, M., Sakata, N., Okuma, T., Kurata, T.: Indoor/outdoor pedestrian navigation with an embedded GPS/Rfid/self-contained sensor system. In: *Int. Conf. on Artificial Reality and Telexistence* (2006)
8. Wagner, M.: Building wide-area applications with the AR toolkit. In: *Int. Augmented Reality Toolkit Workshop* (2002)
9. Reitmayr, G., Schmalstieg, D.: Location based applications for mobile augmented reality. In: *Australasian User Interface Conference* (2003)
10. Nakazato, Y., Kanbara, M., Yokoya, N.: Localization system for large indoor environments using invisible markers. In: *ACM Symp. on Virtual Reality Software and Tech.* (2008)
11. Newman, J., Schall, G., Barakonyi, I., Andreas, S., Schmalstieg, D.: Wide-area tracking tools for augmented reality. In: *Int. Conf. on Pervasive Computing* (2006)
12. Wormell, D., Foxlin, E., Katzman, P.: Advanced inertial-optical tracking system for wide area mixed and augmented reality systems. In: *Int. Immersive Projection Tech. Workshop/Eurographics Workshop on Virtual Environments* (2007)
13. Banwell, T., Calway, A.: Combining absolute positioning and vision for wide area augmented reality. In: *Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications* (2010)
14. Castle, R., Klein, G., Murray, D.: Video-rate localization in multiple maps for wearable augmented reality. In: *Int. Symp. on Wearable Computers* (2008)
15. Efthymiou, C., Gormus, S., Fan, Z., Calway, A., Mayol-Cuevas, W., Doufexi, A.: Application of multiple-wireless to a visual localisation system for emergency services. In: *IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications* (2010)
16. Harmer, D., Russell, M., Frazer, E., Bauge, T., Ingram, S., Schmidt, N., Kull, B., Yarovoy, A., Nezirović, A., Xia, L., Dizdarević, V., Witrisal, K.: EUROPCOM: emergency ultrawideband radio for positioning and communications. In: *IEEE Conf. on Ultra-Wideband* (2008)
17. Davison, A., Mayol, W., Murray, D.: Real-time localisation and mapping with wearable active vision. In: *Int. Symp. on Mixed and Augmented Reality* (2003)
18. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: *Int. Symp. on Mixed and Augmented Reality* (2007)

19. Williams, B., Klein, G., Reid, I.: Real-time SLAM relocalisation. In: Int. Conf. on Computer Vision (2007)
20. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Int. Conf. on Computer Vision (2003)
21. Eade, E., Drummond, T.: Unified loop closing and recovery for real time monocular SLAM. In: British Machine Vision Conf. (2008)
22. Chekhlov, D., Mayol-Cuevas, W., Calway, A.: Appearance based indexing for relocalisation in real-time visual SLAM. In: British Machine Vision Conf. (2008)

# Augmented Reality System for Visualizing 3-D Region of Interest in Unknown Environment

Sei Ikeda, Yoshitsugu Manabe, and Kunihiro Chihara

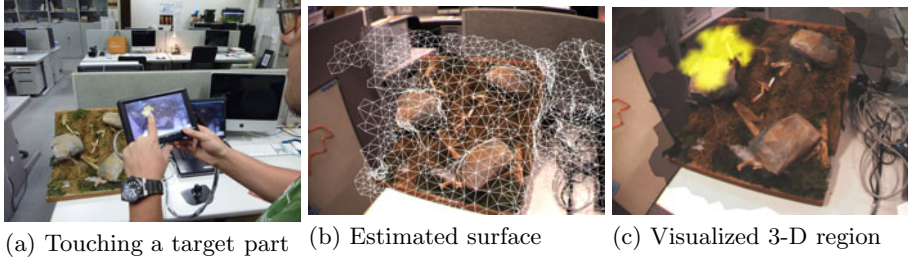
Nara Institute of Science and Technology,  
8916-5 Takayama, Ikoma, Nara, Japan  
{sei-i,manabe,chiara}@is.naist.jp

**Abstract.** This paper presents a novel augmented reality system which allows a user to visualize 3-D region of interest to share with other users in a real environment. To allocate the region, user specifies a point on the target object through a mobile display. The most remarkable difference from the existing works is that semantic information of the environment is not given. This kind of augmented reality application is still few though vision tracking techniques without prior knowledge about environment are coming into practical use. By realizing minimum set of our concept, we could found several concrete future works, most of which are computer vision problems.

## 1 Introduction

Building an information society can be rephrased as digitizing everything possible in our real world. In such society, everything must have ID number related to the semantic information: what it is, and additionally where it is or how it is. In general, it is thought that most of useful augmented reality (AR) applications except for games and art also require this semantic information. However, even though our environment is exhaustively digitized, there certainly will remain essentially difficult things to be digitized such as collapsed structures and new objects in the making. Since proceeding digitization increases a kind of gaps between digitized and non-digitized environments, research on AR techniques dealing with interactions between human and non-digitized environment becomes more important rather than digitized one.

In order to simplify the explanation, an example scenario about interactions between human and non-digitized environment is introduced. Imagine a scene that a leader of a rescue corps is briefing to other members and pointing out a target part of a mudslide area. In many cases, since the viewpoint of the leader is close to the members' ones, they can recognize the indicated target without moving to the leader's back if they are well-trained members. However, this communication task becomes extremely difficult if the members are standing far from the leader's position. Transmitting each other's views taken from head-mounted cameras, the member can understand the target from leader's view or



**Fig. 1.** Visualizing 3-D regions of interest

the leader can directly indicate the target in the members' views [1]. In this method, however, users must find correspondences between their own view and transmitted view. Only computer instead of the members can both observe and visually inform the target at the respective viewpoints by using wireless network.

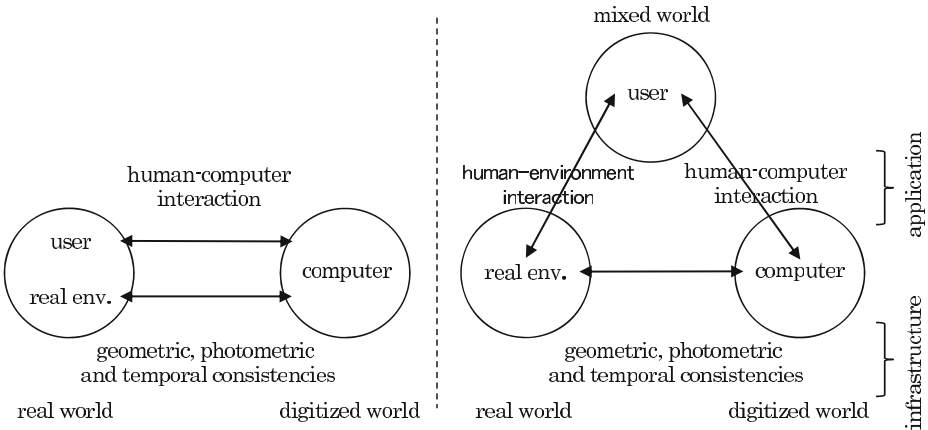
To clarify problems in this scenario, we should consider from what kinds of information the members recognize the region indicated by the leader. The leader provides direction and trajectory of the end of a finger and context of the leader's explanation. The members can have prior knowledge about the target. The target itself provides them with information about shape and texture of the target [2]. These kinds of information depend on position of the viewpoint to the target. However, it is difficult to consider all of such information in design of a communication tool supporting the rescue corps,

This paper presents a novel AR application system which supports users' activities in unknown environments not digitized with semantic information. Information used in this system is three kinds: target's surface shape, a indicated direction and a user's view point as at least necessary information. This system allows a user to visualize 3-D region of interest (ROI) to share with other users in a real environment, as shown in Fig. 1. The system estimates target's surface and segments it into distinct parts. To allocate the region, user specifies a point on the target object through a mobile display. Once the 3-D ROI is determined, this can presented to other users standing at different positions.

Our current prototype system consists of quite simple algorithms of computer vision and has many ad-hoc parameters. Furthermore, there are many problems described in Section 5 for practical use. Despite of these faults, the reason why we publish this work is because it is worth enough to discuss one of the applications treating few remarked but important research field described in the next section.

## 2 Relation to Other Works

One of common problems in AR is geometric consistency between real and digitized worlds [3]. For solving the geometric consistency problem in various situations, many researchers might have thought that vision tracking without any sensors, markers or prior knowledge is the most fundamental technique, and this kind of tracker is ideal in the sense that it is similar to human vision. PTAM [4],



**Fig. 2.** Problems on augmented reality. In the left figure, left and right circles represent information sources in real and digitized worlds, respectively. In the right figure, user and environment are separated as different sources.

which is a neat implementation of SLAM or vision tracker and does not require other information, clearly gave us a prospect of practical use of vision tracking for a static surrounding though many similar techniques [5,6,7,8] already had been published. However, there are still few applications effectively using this kind of vision tracker [9] except for art and games.

What we have to consider is what we can do with this kind of vision tracking. For that purpose, we consider again position of geometric consistency problem in AR. AR is a technique enabling us to seamlessly treat information in both real and digitized worlds. In this sense, many researchers have thought two things<sup>1</sup>: real and digitized worlds, as shown in the left of Fig. 2. The real world includes user and other objects, namely the environment. The digitized world is whole information in the computer memory such as 3-D models, annotation data and other semantic information. Computer supports us by detecting interactions among these things. Detecting interaction between CG model and real environment is consistency problem. Detecting interaction between information in user’s head and information in computer is human-computer interaction.

In contrast to the left figure, we consider user and environment are essentially different sources, as shown in the right of Fig. 2. We can inevitably find out there is another problem: human-environment interaction. This is interaction between user and non-digitized world (real environment). SLAM is compatible with this third problem because it does not require any pre-digitized semantic information. The application described in this paper is one of applications

<sup>1</sup> The reason why we consider many researchers imagine the left figure is because the recent survey paper [9] on AR focuses on three technical problems: tracking, interaction and display as important topics. Each of them is one of consistency problems or human-computer interaction problems.

[10] mainly treating this third problem. The main contribution of this paper is to show feasibility of the proposed application system and to confirm concrete future works.

### 3 Prototype System Visualizing 3-D Region of Interest

#### 3.1 System Overview

The prototype system consists of a computer (Toshiba, Qosmio G30 97A) and a touch panel display (Hanwa-Japan, HM-TL7T) attached with a camera (Point Grey Research, DragonFly), as shown in Fig. 3. The system performs in parallel two kinds of processes: recognition of unknown environment and recognition of user's action, both of which are implemented based on a free SLAM software PTAM [4].

Our first approach to environment recognition is to perform a structure-from-motion method. To acquire images taken from different viewpoints, a user must move with the see-through display. This action corresponds to a preliminary survey in our scenario.

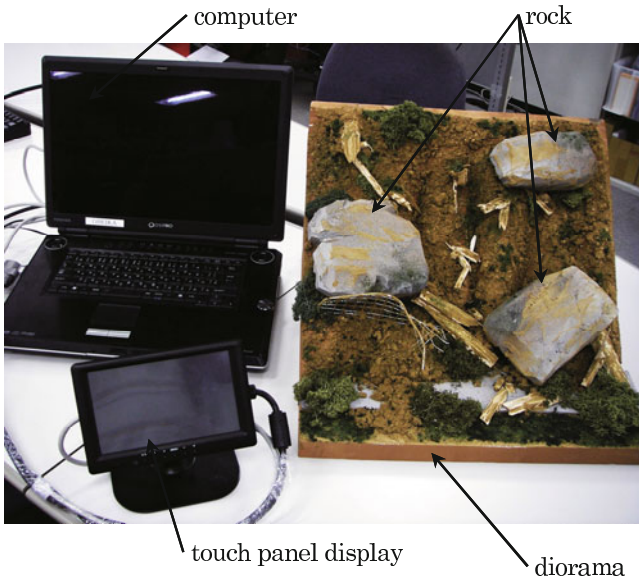
As recognition of user's action, in the current system, direction of the end of user's finger is calculated from user's viewpoint and the clicked position on the display. Given the user's clicked position and 3-D surface mesh, we can simply determine a surface point as a intersection between the surface and the line of sight corresponding to the clicked pixel. Each surface point is labeled by the surface labeling method described in the next section. Then, the system emphasizes the selected region [11] as a set of the same-labeled points.

#### 3.2 Detail of Surface Labeling

- (1) **Eliminate outliers:** The surface reconstruction method described in Step (3) is sensitive to outliers of surface points. Before reconstructing a surface, outliers are eliminated by the following two criteria. The first one is related to confidence of each point. If a point has not been observed in many frames, the point is eliminated as a mis-tracked one. More concretely, if ratio of the number of observations to the total number of frames is less than a threshold (= 0.25), the point is eliminated. The second one is whether the point is isolated or not. If distance between the target point and the nearest one is less than a threshold, the point is also eliminated.
- (2) **Estimate initial surface normals:** In the surface reconstruction of the next step, a set of points on the target surface and their surface normals are required. Surface normal  $\mathbf{n}_j$  of each point  $j$  is estimated from positions of cameras which observed its point.

$$\mathbf{n}_j = \frac{\sum_i (\mathbf{c}_i - \mathbf{p}_j)}{|\sum_i (\mathbf{c}_i - \mathbf{p}_j)|}, \quad (1)$$

where  $\mathbf{p}_i$  and  $\mathbf{c}_i$  represent positions of surface point  $j$  and camera at the frame  $i$ , respectively. The above method does not work well when distribution of camera positions is lopsided.



**Fig. 3.** Prototype system and target diorama

- (3) **Reconstruct surface:** From the oriented points, surface mesh is estimated simply by a Poisson surface reconstruction method [12]. Since this method assumes the target can be represented as a closed surface, a closed surface mesh is generated even if a set of observed points of the target are not distributed like a closed surface. In our case, the target is environment, not a small object or a closed surface. In our method, such unwanted parts of the surface are removed. More concretely, each vertex of the mesh is eliminated if there are no feature points in a constant distance from the vertex.
- (4) **Estimate surface normal:** Given a surface mesh, we can estimate curvature of each vertex directly by fitting a general quadratic surface. In our prototype, however, in order to decrease parameters of a fitting function, a surface normal of each vertex is calculated as a normalized vector of the third principal component of a set of its neighbors in a constant radius centered at the vertex. The sign of the normal can be given from the initial surface normal.
- (5) **Fit a quadratic surface:** A quadratic surface represented as the following function  $z$  of variables  $(x, y)$  is fitted to the set of neighbors of each vertex.

$$z(x, y) = ax^2 + by^2 + cxy + dx + ey. \quad (2)$$

$z$  axis is first determined as the surface normal of the reference point.  $x$  and  $y$  axes are determined as the first and second principal components, respectively. The signs of  $x$  and  $y$  are determined so that  $x - y - z$  is the right-hand system.

- (6) **Calculate curvature:** Surface curvature is calculated as mean curvature of the function  $z(x, y)$  at the point  $(0, 0)$  by the following equation.

$$H = \frac{b(1 + d^2) + a(1 + e^2) - cde}{(1 + d^2 + e^2)^{\frac{3}{2}}}. \quad (3)$$

- (7) **Segmentation:** For general mesh segmentation, as summarized in [13], we can chose a segmentation cue from various features such as curvature [14,15] and difference in normals of vertices [16] as local feature, geodesic distances [17] as global feature. Global features cannot be applied to environments because environment cannot be represented as a closed surface. We selected mean curvature [18] based on the idea that arbitrary objects can be represented by logical disjunction of multiple convex hulls [19]. A watershed segmentation [18] using this feature is performed to the obtained mesh.

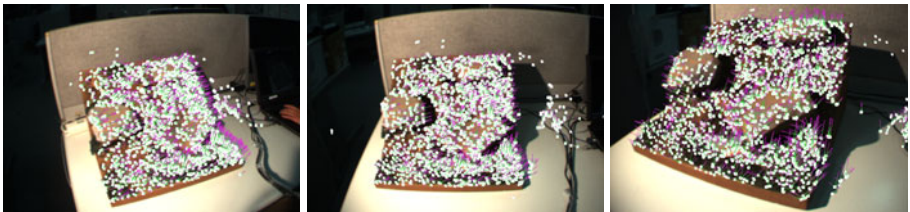
## 4 Experiment

We have tested the prototype system using a diorama, shown in Fig. 3, where there are three rocks on a mudslide slope. Fig. 4,7 show parts of results of the steps described in the previous section.

Fig. 4 shows the surface points and their initial normals reconstructed by the structure-from-motion. We can confirm the reconstructed surface points distributing on the object surface. However, directions of their normals are irregular even though the surface is smooth.

Fig. 5 and Fig. 6 show the generated surface mesh and curvature map, respectively. From these figures, we can confirm the generated mesh roughly fitted on the target surface. However, the generated mesh does not represent several edges such as contact borders between the ground and a rock.

Fig. 7 shows the visualized 3-D ROI. The system could emphasize the part including the surface point corresponding to the clicked position. However, border of the part is unnaturally jagged because of roughness of the mesh. Problems about density of reconstructed points will be described in the next section.



**Fig. 4.** Reconstructed surface points and initial normals. The surface points are represented by white dots. The normals are represented by line segments colored from green to purple.





Fig. 5. Generated surface mesh

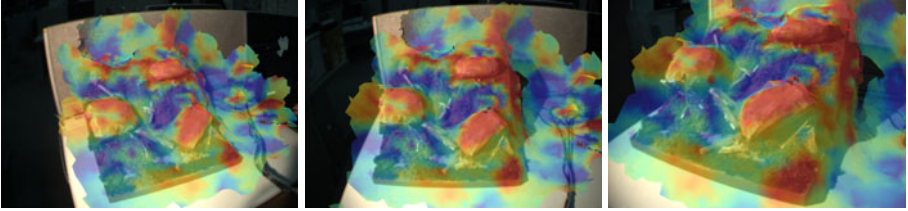


Fig. 6. Mean curvature map. Relative curvature value to curvature distribution is represented by color. Color graduation from green to red represents a range of  $m \pm \sigma$ , where  $m$  and  $\sigma$  are average and standard deviation of the distribution, respectively.



Fig. 7. Selected 3-D ROI. The selected region is visualized by a set of yellow polygons, superimposed on the interested object.

## 5 Steps toward the Practical Use

### 5.1 System Initialization

To acquire multiple images for structure-from-motion, user must intentionally move the viewpoint. Though this action can be considered as a preliminary survey, there exist situations in which users can not widely move so as to cover the target. Furthermore, since PTAM is used as SLAM in the current system, user must specify two frames for a simple stereo method to calculate initial value of position of feature points.

Both initializations should be made unnecessary since they require intentional operations by user who knows about structure-from-motion well. For this purpose, we can apply structure-from-motion among multiple users standing at different positions (leader and other members in our scenario) via wireless network.

## 5.2 Density of Reconstructed Points

Density of reconstructed feature points must be high enough in order to generate exact surface. However, it is lower than density of feature points detected in a single image because a bare minimum number of highly confident points are required to estimate exact camera pose and unconfident points are eliminated in general SLAM. Moreover, density of detected feature points is not uniform because it depends on lighting and texture of the target surfaceD

One of measures against to both problems is to reconstruct other surface points than tracked ones in order to cover low density parts. For this purpose, the mesh can be refined by a multi-baseline stereo method [20] using key-frame images.

## 5.3 Scale-Dependency of Coefficients

In the prototype system, there are coefficients dependent of environment scale. This is a problem because in structure-from-motion, scale cannot be determined only from images without prior knowledge. In this system, the scale is tentatively determined with an unmeaningful value in initialization of PTAM.

One of methods solving this problem is to determine the scale a meaningful value. We can use other information independent of environment such as accelerometer outputs and prior knowledge about user's motion.

The other method is to make those coefficients scale-independent. This is important future work to exclude dependency of environment from the system. However, it is not always effective because calculating statistics value of the target environment needs often high cos.

## 5.4 Mesh Segmentation Cue

The current system uses surface curvature as a geometric segmentation cue. However, human may recognize objects from their shadow and texture as photometric cues. Photometric cue can be also introduced and combined with the geometric one in the same framework. The simplest one of this kind of photometric cues is edge intensity.

## 5.5 ROI Recognition Cue

The current system simply calculate line of sight from clicked position and view-point in order to decide 3-D ROI. As mentioned above, trajectory of the end of a finger, context of the leader's explanation and member's common knowledge can be also ROI recognition cues. Trajectory of finger is easy to be integrated in the same framework. If a user touches the display carefully, finger's motion may be static or slow. Then, visually small part must be selected by this system. In contrast, if the same user touches the display roughly, the finger's motion may become fast. Then, visually large part must be selected.

## 6 Conclusion

This paper presents a novel augmented reality application system which allows a user to visualize 3-D region of interest to share with other users in an unknown environment. This kind of application is classified into a research field on user-environment interaction, which has hardly been addressed in augmented reality. SLAM without prior knowledge is a suitable technique for understanding user and environment. By prototyping a minimum set of our concept, we have found several concrete future works, most of which are computer vision problems.

## References

1. Kurata, T., Sakata, N., Kouroggi, M., Kuzuoka, H., Billinghurst, M.: Remote collaboration using a shoulder-worn active camera/laser. In: Proc. ISWC 2004, pp. 62–69. IEEE Computer Society, Washington, DC (2004)
2. Yu, Y., Ferencz, A., Malik, J.: Extracting objects from range and radiance images. *IEEE Transactions on Visualization and Computer Graphics* 7, 351–364 (2001)
3. Azuma, R.T.: A survey of augmented reality. *Presence: Teleoperators and Virtual Environments* 6, 355–385 (1997)
4. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proc. 6th IEEE and ACM Int. Symp. on Mixed and Augmented Reality (ISMAR 2007), pp. 1–10 (2007)
5. Sim, R., Dudek, G.: Learning and evaluating visual features for pose estimation. In: Proc. 1999 IEEE Int. Conf. on Computer Vision, vol. 2, pp. 1217–1222 (1999)
6. Davison, A.J.: Real-time simultaneous localisation and mapping with a single camera. In: Proc. 9th IEEE Int. Conf. on Computer Vision, vol. 2, pp. 1403–1410 (2003)
7. Burschka, D., Hager, G.D.: V-GPS (SLAM): Vision-based inertial system for mobile robots. In: Proc. 2004 IEEE Int. Conf. on Robotics and Automation, pp. 409–415 (2004)
8. Gordon, I., Lowe, D.G.: Scene modelling, recognition and tracking with invariant image features. In: Proc. 3rd IEEE and ACM Int. Symp. on Mixed and Augmented Reality, pp. 110–119 (2004)
9. Zhou, F., Duh, H.B.L., Billinghurst, M.: Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. In: ISMAR 2008: Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 193–202. IEEE Computer Society, Washington, DC (2008)
10. Reitmayr, G., Eade, E., Drummond, T.W.: Semi-automatic annotations in unknown environments. In: Proc. ISMAR 2007, Nara, Japan, pp. 67–70 (2007)
11. Tenmoku, R., Kanbara, M., Yokoya, N.: Annotating user-viewed objects for wearable AR systems. In: Proc. IEEE and ACM Int. Sympo. on Mixed Augmented Reality (ISMAR 2005), pp. 192–193 (2005)
12. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proc. Symposium on Geometry Processing, pp. 61–70 (2006)
13. Shamir, A.: A survey on mesh segmentation techniques. In: *Computer Graphics Forum*, vol. 27, pp. 1539–1556 (2008)
14. Lavoue, G., Dupont, F., Baskurt, A.: New CAD mesh segmentation method a, based on curvature tensor analysis 37, 975–987 (2005)

15. Mortara, M., Patané, G., Spagnuolo, M., Falcidieno, B., Rossignac, J.: Plumber: A method for a multi-scale decomposition of 3D shapes into tubular primitives and bodies. In: Proc. ACM Symposium on Solid Modeling and Applications, pp. 139–158 (2004)
16. Kalvin, A., Taylor, R.: Superfaces: Polygonal mesh simplification with bounded error. *IEEE Computer Graphics and Applications* 16 (1996)
17. Zhang, E., Mischaikow, K., Turk, G.: Feature based surface parameterization and texture mapping, vol. 24, pp. 1–27 (2005)
18. Mangan, A.P., Whitaker, R.T.: Partitioning 3D surface meshes using watershed segmentation. *IEEE Transactions on Visualization and Computer Graphics* 5, 308–321 (1999)
19. Zheng, B., Takamatsu, J., Ikeuchi, K.: 3D model segmentation and representation with implicit polynomials. *IEICE Trans. Information and Systems* E91-D, 1149–1158 (2008)
20. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 1362–1376 (2010)

# Interactive Video Layer Decomposition and Matting

Yanli Li<sup>1,2</sup>, Zhong Zhou<sup>1,2</sup>, and Wei Wu<sup>1,2</sup>

<sup>1</sup> State Key Lab. of Virtual Reality Technology and Systems, Beihang University

<sup>2</sup> School of Computer Science and Engineering, Beihang University

liy1@vrlab.buaa.edu.cn

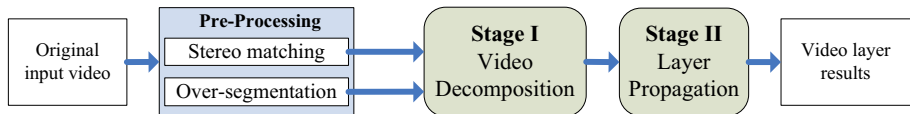
**Abstract.** The problem of accurate video layer decomposition is of vital importance in computer vision. Previous methods mainly focus on the foreground extraction. In this paper, we present a user-assisted framework to decompose videos and extract all layers, which is built on the depth information and over-segmented patches. The task is split into two stages: i) the clustering of over-segmented patches; ii) the propagation of layers along the video. Correspondingly, this paper has two contributions: i) a video decomposition method based on greedy over-segmented patches merging; ii) a layer propagation method via iteratively updating color Gaussian Mixture Models(GMM). We test this algorithm on real videos and verify that it outperforms state-of-the-art methods.

## 1 Introduction

Video decomposition is one of the most fundamental vision tasks. It extracts multiple layers from videos, which can be further used for kinds of Augmented Reality applications. Generally, it solves two problems: layer clustering and layer segmentation. The first problem, which has been extensively studied in the space clustering field, is to estimate the number of layers in every frame. The second problem is to assign each pixel to the corresponding layer.

For the last decades, researchers have presented various approaches for this task, which lie in the fields of motion segmentation and figure-ground separation. However, most motion segmentation methods fail to accurately separate layers, mainly due to an improper energy formulation and the unreliable optical flow fields, while most figure-ground separation approaches only focus on the foreground object extraction, they seldom consider the multi-layer separation.

In this paper, we provide an interactive framework to decompose videos. It combines the merits of motion segmentation and figure-ground separation. With only several clicks, the user can accurately decompose the video into multiple layers. Comparing with motion segmentation methods, our algorithm can segment more “meaningful” layers, and the fine information is preserved well. Compared to figure-ground separation methods, our algorithm is less labor-intensive and can soft-segment every layer. Fig. 1 provides a high level overview of the algorithm pipeline. Our algorithm is based on the depth information and over-segmented patches. In Stage I, we utilize a greedy bottom-to-up scheme to merge



**Fig. 1.** The framework of decomposing static scene videos

patches into layers, and provide a User Interface(UI) for users to refine layers. In Stage II, a layer propagation method is employed to extract the layer sequence along the entire video. We will explain the details in following sections.

The remainder of the paper is organized as follows: Section 2 reviews related work; Section 3 explains the video decomposition scheme; Section 4 describes the User Interface; Section 5 gives a description of the layer propagation method; Section 6 demonstrates the experimental results; Finally, Section 7 discusses our algorithm’s limitations and further works.

## 2 Related Work

The idea of video decomposition was introduced by Darrell et al. [1]. Wang et al. [2] present the first precise mathematical formulation for this problem. Since then, researchers have developed various models for effective motion segmentation, such as Linear Subspace [3]. Although the clustering number can be successively decided, the pixel assignments are unsatisfactory in most cases.

In parallel, the accurate figure-ground separation is being extensively studied. Y. Boykov et al. [4] formulate the problem as a global energy function in Markov Random Fields(MRF) and solve it with Graph Cuts [5]. C. Rother et al. [6] extend the graph-cut approach [4] by developing a more powerful, iterative version-“GrabCut”. These segmentation approaches are both based on uniform color information. There are some other methods using texture cue [7], or symmetry cue [8]. More recently, S. Bagon et al. [9] present an approach which unifies those cues into a framework. It is based on the concept of “Segmentation by Composition”. By developing a description for a segment and maximizing the difference in description lengths, they extract good figures.

Although the above figure-ground separation methods can be extended to videos and obtain more accurate results, it is hard to optimize boundaries of some objects, such as hairs, as they mix the background and foreground colors. Therefore, image matting [10], as a soft segmentation technique, evolves to accurately extract foreground objects. It is extensively used to recover the foreground per-pixel opacity from the background.

However, neither figure-ground separation methods nor image matting focuses on multi-layer extraction, thus their extension to video object extraction mainly lies in foreground separation, including the moving object extraction and the static foreground layer separation. The former such as [11] [12] extracts objects by tracking and optimizing the boundaries, when applied to the occluded background layer, it fails as some boundaries disappear in subsequence frames; The

latter such as [13] utilizes scene depth as a cue and can automatically extract the foreground layer, but they are constrained to bi-view separation.

The most similar works to us are J. Xiao et al. [14] and G. Zhang et al. [15]. J. Xiao et al. [14] present an algorithm of motion layer extraction and matting for short video clips. They first establish a novel MRF framework to solve the motion segmentation problem, and then the Poisson matting [16] is employed to refine the foreground segmentation. However, it is impossible to separate several objects in the same background layer. G. Zhang et al. [15] present a general re-filming system, in which foreground layer matting is extracted by an interactive tool and the cut out information is propagated from key frames to the other frames automatically. They improve the optical-flow-based Bayesian video matting [17] by geometry projection of depth information. Just as the authors stated, the limitation of their system lies on depth ambiguity in extremely textureless regions (such as the clear blue sky). We overcome these problems by combining the color and depth information.

What's more, video matting approaches such as [12] are cumbersome, even in the process of initial key frame matting. While the easy-to-use GrabCut [6] often fails to construct a satisfactory result when the foreground colors are similar to the background colors. Here we developed a more robust and easy-manipulated framework for layer extraction.

### 3 Video Decomposition

The dense correspondence we established is initialized by a quasi-dense correspondence method [18], also called point propagation. Although lots of pixels are matched after point propagation, there are still some unmatched pixels. To obtain a total dense correspondence, we take the problem as a global cost function in Markov Random Fields, formulate a MAR-MRF model which is same to [19] and solve it by the max-flow algorithm [5].

The Pedro's method [20] is adopted for over-segmentation. Based on the depth map and over-segmentation patches, we employ a graph-based scheme to merge patches into layers. Taking each patch  $\nu_i$  as a vertex, we construct an undirected weighed graph  $G = \langle V, E \rangle$ . The edge  $(\nu_i, \nu_j) \in E$  connects two adjacent patches, its weight is defined as:

$$\omega(i, j) = \gamma_1 \omega_c(i, j) + \gamma_2 \omega_d(i, j) + \gamma_3 \omega_s(i, j) \quad (1)$$

$\omega_c(i, j)$  measures the similarity of color information, which is defined as:

$$\omega_c(i, j) = \exp\left(-\frac{\min(\|\mu_c(i) - \mu_c(j)\|_2, T_c)}{\sigma_c}\right) \quad (2)$$

$\omega_d(i, j)$  measures the similarity of depth information, which is defined as:

$$\omega_d(i, j) = \exp\left(-\frac{\min(|\mu_d(i) - \mu_d(j)|, T_d)}{\sigma_d}\right) \quad (3)$$

$\omega_s(i, j)$  measures the minimum size of the two regions, it is defined as:

$$\omega_s(i, j) = 1 - \min\left(\sqrt{\frac{\mu_s(i)}{S}}, \sqrt{\frac{\mu_s(j)}{S}}\right) \quad (4)$$

where  $\gamma_1, \gamma_2, \gamma_3$  are weighting values, they are all in the range of  $[0, 1]$  and  $\gamma_1 + \gamma_2 + \gamma_3 = 1.0$ ;  $\mu_c(\cdot), \mu_d(\cdot)$  are the mean color and depth values of the region;  $T_c, T_d$  are truncation values (empirically set to 15 and 1.7);  $\sigma_c = 255, \sigma_d$  is taken as the maximum disparity value;  $S$  is the image size,  $S = width * height$ , and  $\mu_s(\cdot)$  is the region size.

It is obvious that the edge weight formulation defined above encourages to priorly join two adjacent regions with similar color information, or/and with similar depth information, or/and with smaller size.

We use a greedy scheme to merge patches one by one. Each time, we select the edge with the maximum weight value and unite its two patches. And then the weight of the edges which connect either of the two newly united patches are recomputed. This step repeats until all patches are merged into one. We record the clustering process, so that the user can rebroadcast the process and select a satisfactory clustering result by a granular value. The maximum granular value is just the number of over-segmented patches.

## 4 User-Assisted Layer Refinement

Although plenty of over-segmented patches are clustered into compact components under a granularity, components of the same object may still be isolated, e.g. the two sides of an occluded wall, while further adjusting the granularity may lead to under-segmentation. Therefore, an interactive User Interface (UI) is necessary to increase the diversity of “meaningful” segmentation.

The user interactions in our system involve two stages. The first stage is to merge components into a complete layer. And the second one is to refine the layer. In the first stage, the user needs to choose a granularity with a slider first, and then click several components to acquire a complete layer. In the second stage, the user should refine the layer by clicking a button first, and then draw scribbles to refine boundaries if needed.

Here, we employ the approach of Lazy Snapping [21] to refine the layer. To apply to our task, we make some adjustments. Lazy Snapping solves the problem by formulating a global “Gibbs” energy function in patch level, while we built a similar model in pixel level. The input of Lazy Snapping is a rectangle, while our input is a layer mask with arbitrary shape. The layer mask is created by enlarging the contour of the original layer mask outwards with 3 pixels size. What’s more, we use the depth information in our model, which is unavailable in Lazy Snapping.

The “Gibbs” energy function is formulated as follows:

$$E = \sum_{i \in V} E_1(l(x_i)) + \lambda \sum_{(i,j) \in Neigh} E_2(l(x_i), l(x_j)) \quad (5)$$



It consists of a data term  $E_1$  and a smooth term  $E_2$ ,  $\lambda$  is the weighting value. The data term  $E_1$  is defined as:

$$E_1(l(x)) = \begin{cases} \frac{d_f(x)}{d_f(x) + d_b(x)}, & l(x) = 0 \\ \frac{d_b(x)}{d_f(x) + d_b(x)}, & l(x) = 1 \end{cases} \quad (6)$$

where  $l(x) = 1$  indicates  $x$  locates in the foreground layer, while  $l(x) = 0$  indicates  $x$  locates in the background layer;  $d_f(x) = \max_k \|I(x) - C_k^F\|_2$ ,  $d_b(x) = \max_k \|I(x) - C_k^B\|_2$ ,  $k = 1 \dots 5$ ;  $\{C_k^B\}$  and  $\{C_k^F\}$  are the centroids of GMM (Gaussian Mixed Models) of the background and foreground colors, which are obtained by the K-Means method. If the user draws some scribbles, the data term of the marked pixels is taken as following:

$$\begin{cases} E_1(0) = 0 & E_1(1) = \infty, & \text{if } (x \in \text{"foreground scribbles"}) \\ E_1(0) = \infty & E_1(1) = 0, & \text{if } (x \in \text{"background scribbles"}) \end{cases} \quad (7)$$

The smooth term  $E_2$  is defined as:

$$E_2(l(x), l(y)) = \begin{cases} \frac{\|I(x) - I(y)\|_2}{\varepsilon + 1} |l(x) - l(y)| & \text{if } (D(x) = D(y)) \\ \frac{\|I(x) - I(y)\|_2}{\varepsilon + 1} (1 - |l(x) - l(y)|) & \text{if } (D(x) \neq D(y)) \end{cases} \quad (8)$$

where  $D(\cdot)$  stands for the depth value. The above function  $E$  is a two-labels ‘‘Gibbs’’ function. The max-flow algorithm [5] is invoked to minimize it. Now we obtain a refined layer mask. A trimap is generated by automatically dilating the layer boundaries with 5 pixels. Then Bayesian matting [22] is applied to soft segment the layer using the trimap.

The above stages are all repeatable. The user manually clicks the slider to control the clustering granularity, clicks components to merge or separate them, adds some strokes on the layer to refine boundaries, and clicks a button to examine the matte until satisfied.

## 5 Spatial-Temporal Layer Propagation

To extract layers along the entire video, we require a trimap in every frame. Constructing the trimaps manually is a tedious and time-consuming work. Moreover, layer matting applied frame-by-frame produces temporally incoherency as the small errors are stochastic in each individual frame. In Section 3, we have built dense correspondences for pairwise frames. Based on the spatial-temporal depth maps, we propagate the trimaps from the key frames to the rest of the frames.

For a video clip  $\hat{I} = \{I_i, i = 1 \dots n\}$  with a depth map sequence  $\hat{D} = \{D_i, i = 1 \dots n\}$ , we assume the key frames are sampled and soft segmented. We denote the trimap sequence by  $\hat{T} = \{T_i, i = 1 \dots n\}$  and successively propagate the

trimaps based on the depth maps. Suppose the trimap  $T_i$  of the frame  $I_i$  is available, we first create a tri-labeling mask  $L_{i+1}$  for the frame  $I_{i+1}$ :

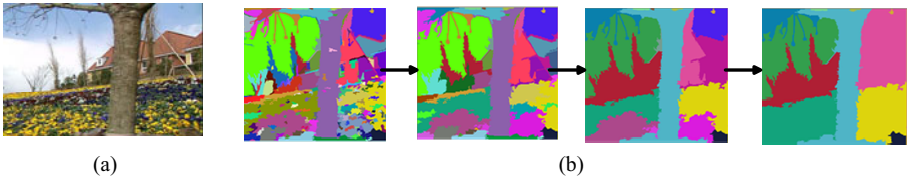
$$L_{i+1}(x) = \begin{cases} \text{'F'}, & \text{if } x' + D_i(x') = x \text{ and } T_i(x') = \text{'F'} \text{ for at most one } x' \in I_i \\ \text{'B'}, & \text{if } x' + D_i(x') = x \text{ and } T_i(x') = \text{'B'} \text{ for at most one } x' \in I_i \\ \text{'U'}, & \text{otherwise} \end{cases}$$

Then we build a foreground GMM(Gaussian Mixture Model) and a background GMM. The foreground GMM is built with pixels whose  $L_{i+1} = \text{'F'}$  and the background GMM is built with pixels whose  $L_{i+1} = \text{'B'}$ . The GMM components  $\{C_k^B\}$  and  $\{C_k^F\}$  are individually computed by the K-Means clustering, where  $k = 1 \dots 5$ . By optimizing a global ‘‘Gibbs’’ energy function which is the same to formula (8) for pixels whose  $L_{i+1} = \text{'U'}$ , we obtain a foreground/background mask  $M_{i+1}$  for the layer. Note that the layer refinement in formula (5) is only applied for the layer mask, while it is applied for the whole image here. The data terms of definite labeled pixels, i.e,  $L_{i+1}(x) = \text{'F'}$  or  $\text{'B'}$ , are taken as formula (7). Finally, the boundaries of the mask  $M_{i+1}$  are dilated with 5 pixels size to generate the trimap. Bayesian matting is further applied to soft-segment the layer.

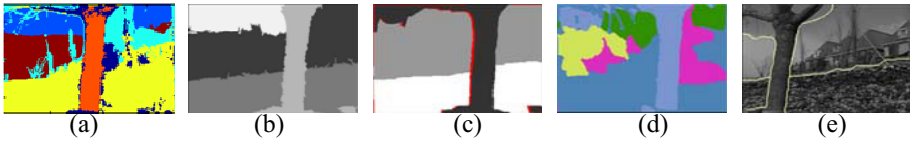
Generally speaking, this scheme takes advantage of spatial-temporal depth information and involves one stage of optimization. The depth map is used to preserve the intra-frame trimap coherence and the optimization process is used for inner-frame refinement.

## 6 Experimental Results

We apply our interactive layer decomposition algorithm to a number of video clips, involving of indoor and outdoor scenes. For the outdoor scenes, we demonstrate our result on the standard flower garden sequence. For the indoor scenes, we test several video clips from the multi-view stereo dataset [23]. Fig. 2 shows four clustering results for a still frame of the flower garden sequence. The results in video form are available in supplementary material. The granularity is defined as the number of components here. Just as Fig. 2 shows, the components always keep semantically consistent when they are merged, mainly due to our method taking account of both the color and depth information.



**Fig. 2.** The clustering results under different granularity. (a) Original frame. (b) Four clustering results, consisting of 203, 89, 16 and 10 components individually.



**Fig. 3.** Previous results for the flower garden sequence. (a) Result of S. Khan et al. [24]. (b) Result of Q. Ke et al. [3]. (c) Result of J. Xiao et al. [25]. (d) Result of R. Dupont et al. [26]. (e) Result of T. Schoenemann et al. [27].

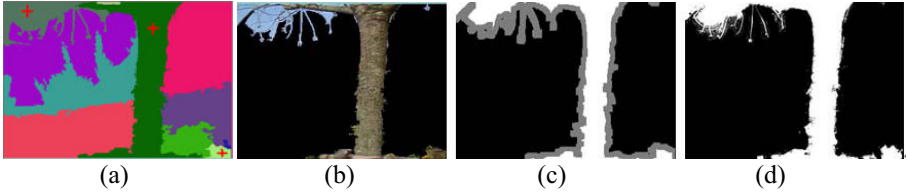


**Fig. 4.** Results using GrabCut. (a) The initial input is a red rectangle. (b) Result after drawing the rectangle. (c) The additional inputs are some scribbles, in which the yellow scribbles indicate the foreground and the blue scribbles indicate the background. (d) Result after applying those scribbles.

We compare the result with five motion segmentation methods [24] [3] [25] [26] [27]. As showed in Fig. 3, methods of [3] [25] [27] (shown in Fig. 3(b)(c)(e)) all fail to extract the red house, and they do not separate some tree trunks from the sky. [24]’s method presents too many noises in the whole image, while [26]’s method under-segments several components, e.g., the tree is extracted without the bottom root, and portions of the red house are merged into the flower bed.

Compared to their results, ours preserves the layer integrity well. As demonstrated in Fig. 2 (or the videos in supplementary material), the red house is always isolated from the sky until the number of components is lower than 10, and the thin tree trunks are always preserved until the granularity is lower than 4. Taking an edge value defined in formula(1) as a threshold, our method will automatically generate a clustering result too. The drawback of our method is that it fails to merge two sides of the same occluded layer, such as the flower bed of the flower garden sequence. This is because we only merge two adjacent components each time.

We verify the UI efficiency of our algorithm by comparing with GrabCut [6]. To extract the tree layer in the flower garden scene through GrabCut, we first draw a bounding rectangle covering the tree, and then draw scribbles to refine the foreground layer. Fig. 4(b) is the layer extraction result after we draw a rectangle (Fig. 4(a)). Fig. 4(d) is the refined layer result after we draw several scribbles (Fig. 4(c)). It is obviously cumbersome to fulfill this task through GrabCut. In contrast, our method extracts a more satisfactory layer using only several clicks (Fig. 5).



**Fig. 5.** Layer decomposition and matting for a flower garden frame. (a)The inputs are several clicks. (b)Refined results of the tree layer, in which some boundaries artifacts are removed after applying the layer refinement. (c)The generated trimap. (d)Matted of the layer.



**Fig. 6.** Results of layer propagation on the teddy sequence. The teddy sequence consists of 9 frames, we only show the 1st, 2nd, 5th, 8th and 9th frame. The first row shows the original frames. The second row displays the composition results of two extracted toys on the flower garden clips.

**Table 1.** Timings for each stage of the algorithm

Sequence (352*240)	Over- Segmentation	Stereo Matching	Layer Clustering	Layer Refining	Layer Matting	Total Time
Flower Garden (20 frames)	2.67sec	73.60sec	3.32sec	12.80sec	45.40sec	137.79sec
Cone (9 frames)	1.19sec	44.58sec	0.48sec	3.52sec	14.84sec	64.61sec
Teddy (9 frames)	1.22sec	42.74sec	0.50sec	6.80sec	10.32sec	61.58sec

Fig. 6 demonstrates the layer propagation results on the teddy sequence. The two toys are extracted manually in the first frame, and the layer results propagate to the rest frames automatically. Even if there are some newly appeared regions, including the image borders and the previous occluded regions, our method can still extract the whole layer in the rest frames.

The running time is shown in Table 1, which is tested on an Intel 3.0GHz CPU with 2.0G RAM. All frames are reduced to 352\*240. The key frames are sampled at every 8 frames. Other interactions all give real-time feedbacks. Obviously, the

bottlenecks are the stereo matching and layer matting. It is well known that the Bayesian matting[22] in layer matting and the max-flow solution[5] in the stereo matching are both time-consuming. In the pre-processing of our framework(as shown in Fig. 1), we compare our stereo matching method with the method[19], reporting that the method[19] costs 88.39sec, 53.56sec and 54.53sec for three sequences respectively, and our method can find stronger local minima.

## 7 Conclusion

In this paper, we proposed an interactive algorithm for decomposing and soft-segmenting various complicated videos. The major contribution of our algorithm is the easy-manipulated framework to fulfill the task, which is built on the depth information and over-segmented patches. We also speed up the global bi-view stereo solution via point propagation. By dynamically updating the foreground and background color models in a global energy formulation, our algorithm can handle the occluded layer matting problem well. The limitation of our algorithm is that it is constrained to stabilized videos. In the future, we will incorporate the multi-view stereo into our framework for applying to various hand-hold videos. We will also exploit multi-layer matting solutions in order to simultaneously soft-segment multiple layers of videos.

## Acknowledgement

This work is supported by the Industry-Academy-Research Program of Guangdong Province and the Ministry of Education(2008A090400020), National Science and Technology Supporting Program(2008BAH37B08), and the Fundamental Research Funds for the Central Universities of China.

## References

1. Darrell, T., Pentland, A.: Cooperative robust estimation using layers of support. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 17, 474–487 (1991)
2. Wang, J., Adelson, E.: Representing moving images with layers. *IEEE Trans. on Image Processing Special Issue: Image Sequence Compression* 3, 625–638 (1994)
3. Ke, Q., Kanade, T.: Robust subspace clustering by combined use of knnd metric and svd algorithm. In: *CVPR* (2004)
4. Boykov, Y., Jolly, M.: Interactive graph cuts for optimal boundary region segmentation of objects in n-d images. In: *ICCV* (2001)
5. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26, 1222–1239 (2001)
6. Rother, C., Kolmogorov, V., Blake, A.: “grabcut”: Interactive foreground extraction using iterated graph cuts. *ACM Trans. on Graphics* 23, 309–314 (2004)
7. Galun, M., Sharon, E., Basri, R., Br, A.: Texture segmentation by multiscale aggregation of filter responses and shape elements. In: *ICCV* (2003)

8. Riklin-raviv, T., Kiryati, N., Sochen, N.: Segmentation by level sets and symmetry. In: CVPR (2006)
9. Bagon, S., Boiman, O., Irani, M.: What is a good image segment? A unified approach to segment extraction. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 30–44. Springer, Heidelberg (2008)
10. Rhemann, C., Rother, C., Wang, J., Gelautz, M., Kohli, P., Rott, P.: A perceptually motivated online benchmark for image matting. In: CVPR (2009)
11. Bai, X., Wang, J., Simons, D., Sapiro, G.: Video snapcut: Robust video object cutout using localized classifiers. In: SIGGRAPH (2009)
12. Li, Y., Sun, J., Shum, H.: Video object cut and paste. *ACM Trans. on Graphics* 24, 595–600 (2005)
13. Zhu, J., Liao, M., Yang, R., Pan, Z.: Joint depth and alpha matte optimization via fusion of stereo and time-of-flight sensor. In: CVPR (2009)
14. Xiao, J., Shah, M.: Accurate motion layer segmentation and matting. In: CVPR (2005)
15. Zhang, G., Dong, Z., Jia, J., Wan, L., Wong, T., Bao, H.: Refilming with depth-inferred videos. *IEEE Trans. on Visualization and Computer Graphics* 15, 828–840 (2009)
16. Sun, J., Jia, J., Tang, C., Shum, H.: Poisson matting. *ACM Trans. on Graphics* 23, 315–321 (2004)
17. Chuang, Y., Agarwala, A., Curless, B., Salesin, D., Szeliski, R.: Video matting of complex scenes. *ACM Trans. on Graphics* 21, 243–248 (2002)
18. Lhuillier, M., Quan, L.: Match propagation for image-based modeling and rendering. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24, 1140–1146 (2002)
19. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions via graph cuts. In: ICCV, pp. 508–515 (2001)
20. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. *Int. J. of Computer Vision* 70, 109–131 (2004)
21. Li, Y., Sun, J., Tang, C., Shum, H.: Lazy snapping. *ACM Trans. on Graphics* 23, 303–308 (2004)
22. Chuang, Y.Y., Curless, B., Salesin, D.H., Szeliski, R.: A bayesian approach to digital matting. In: CVPR (2001)
23. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. of Computer Vision* 47, 7–42 (2002)
24. Khan, S., Shah, M.: Object based segmentation of video using color, motion and spatial information. In: CVPR (2001)
25. Xiao, J., Shah, M.: Motion layer extraction in the presence of occlusion using graph cut. In: CVPR (2004)
26. Dupont, R., Paragios, N., Keriven, R., Fuchs, P.: Extraction of layers of similar motion through combinatorial techniques. In: Rangarajan, A., Vemuri, B.C., Yuille, A.L. (eds.) EMMCVPR 2005. LNCS, vol. 3757, pp. 220–234. Springer, Heidelberg (2005)
27. Schoenemann, T., Cremers, D.: High resolution motion layer decomposition using dual-space graph cuts. In: CVPR (2008)

# Removal of Moving Objects and Inconsistencies in Color Tone for an Omnidirectional Image Database

Maiya Hori, Hideyuki Takahashi, Masayuki Kanbara, and Naokazu Yokoya

Nara Institute of Science and Technology (NAIST),  
8916-5 Takayama, Ikoma, Nara, Japan  
{maiya-h,kanbara,yokoya}@is.naist.jp

**Abstract.** This paper proposes a method for removing image inconsistencies which occur by an existence of moving objects or a change of illumination condition when an omnidirectional image database is generated. The database is used for archiving an outdoor scene in wide areas or generating novel view images with an image-based rendering approach. In related work, it is difficult to remove moving objects in an outdoor environment where illumination condition drastically changes, and to remove inconsistencies of color tone of images which included moving objects. The proposed method iterates the two processes which are the estimation of candidate region of moving objects and the achievement of color consistency to split regions. The color consistency is achieved by estimating linear color transformation parameters which change a histogram of an input image to that of the standard image.

## 1 Introduction

In a panoramic image view system such as Google Street View, a user can see images from a street using omnidirectional images. Some studies [1,2] which use omnidirectional images can also generate a novel view with an image-based rendering (IBR) approach in an outdoor environment. These studies use an image database which consists of many images captured with an omnidirectional camera. When the image database is generated from many images which are captured at different position and time, these images have inconsistencies which occur by an existence of moving objects or a change of illumination condition.

As a method for removing moving objects in images, a technique of compensation using images which is captured at different time is used usually. A color tone differs only in the complemented regions when a simple compensation approach is applied using an image whose color tone is different from the original image. Shadow in an image is also treated as a region of moving object. A technique of removing shadows corresponding to change of illumination [3] is proposed by estimating a light source condition. However, it is difficult to detect an object whose color is similar to the color of background.

On the other hand, as one of the methods for removing an inconsistency in color tone of images, a technique which handles an image whose color tone is

different locally [4] is proposed. In this technique, color consistency in images is achieved by splitting the input image to small regions. However, the method is difficult to be applied when moving objects exist in an image, because a static environment is assumed in this method. The work [5] which removes moving objects after correcting a color tone detects moving objects with a slight change of illumination condition. This study cannot be applied to the case that an illumination condition changes drastically such as an outdoor scene.

If the conventional approaches are applied to the images captured in an outdoor environment, there are many problems such as the existence of moving objects and the change of illumination conditions. Furthermore, in order to capture in an outdoor environment efficiently, when omnidirectional camera is used, moving objects are easy to be captured and change of illumination condition is large. This paper proposes a method for removing inconsistencies among omnidirectional images captured at different positions and times. This research assumes an outdoor environment is a target of an omnidirectional image database. We use omnidirectional images which are captured several times along similar paths with a car-mounted omnidirectional camera. We assume that these images add position and posture information, and are captured densely. To remove inconsistencies of omnidirectional images, the proposed method iterates the two processes which are removal of moving objects and achievement of color consistency of images. The iteration of two processes can narrow down a region of moving objects and omnidirectional images with consistency of color tone are generated.

## 2 Removal of Moving Object and Inconsistencies in Color Tone for Omnidirectional Image Database

### 2.1 Outline of the Proposed Method

This section describes a method for generating an omnidirectional image database with consistency of images. Fig. 1 shows the flow diagram of the proposed method. The proposed method consists of three principal processes.

First, omnidirectional image sequences with camera positions and postures are acquired in an outdoor environment. As the pre-processing, we remove the regions which cannot be corrected by linear color transformation in phase (A). In the iteration processing (phase (B)-(D)), two processes which are estimation of linear color transformation parameters in phase (B) and estimation of candidate region of moving object in phase (C) are iterated by splitting regions in phase (D). Finally, as the post-processing, candidate regions of moving objects are compensated in phase (E). Details of each phase are given below.

### 2.2 Pre-processing: Removal of Exceptive Regions for Iteration Process

When inconsistencies in color tone are removed, we assume that color tone of images can be changed by linear color transformation except in a region where



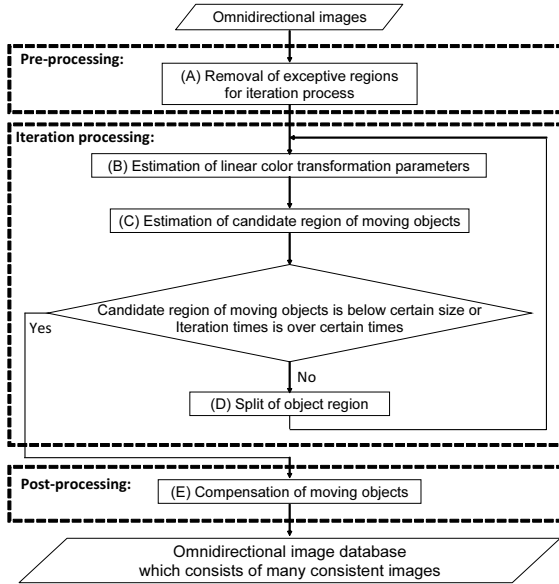


Fig. 1. Flow diagram of proposed method

moving object is observed. It cannot assume that color of the sky region is transformed to the color of standard image by linear transformation since intensity of the pixel in the region is often saturated or background image cannot be defined due to cloud. In this study, the sky region in the omnidirectional images is detected and removed in advance. These can be realized by using the previous methods [6, 7].

### 2.3 Iteration Processing: Estimation of Candidate Region of Moving Object and Linear Color Transformation Parameters

This section explains the method to realize a consistency of omnidirectional images with the following iteration processes.

**Estimation of linear color transformation parameters for color correction.** A color tone of input omnidirectional images is corrected with a standard image. The standard image which is suitable for an IBR approach is selected manually. Since input images have a few disparities when they are captured with a moving vehicle, the color transformation parameters cannot be estimated for every pixel. In this research, to reduce the influence of the disparities, histograms of the standard image and the input image are used for estimating color transformation parameters.

We assume that it is possible to correct the color tone of an image with a linear color transformation if a different appearance depends on changes in illuminate conditions. An equation which changes intensity in image is shown in Eq. (1).

$$I'(x, y) = p_a I(x, y) + p_b, \quad (1)$$

where linear color transformation parameters are  $p_a, p_b$ ,  $I(x, y)$  is intensity at  $(x, y)$  of input image and  $I'(x, y)$  is intensity at  $(x, y)$  after correcting color. The color transformation parameters are estimated in such way that the similarity value of histogram between the input image and the standard image becomes the highest. In this method, Bhattacharyya coefficient (2) is used as a similarity of histograms  $\gamma$ . Bhattacharyya coefficient has the advantage of robustness for outlier by using inner product.

$$\gamma = \sum_i \sqrt{h_{A(i)} h_{B(i)}}, \quad (2)$$

where  $h_{A(i)}$  shows a frequency of intensity  $i$  in image A, and  $h_{B(i)}$  shows that in image B. Histograms are generated in each spectrum.

**Color correction based on robust estimation.** Color of moving objects can not be corrected with the linear color transformation. Then, after removing moving objects, to correct a color tone is desired. The region of the moving objects is difficult to extract from one image or some images which have a different color tone. In this research, we try to estimate color transformation parameters by eliminating the moving objects with a LMedS [8] approach as a robust estimation. In order to estimate color transformation parameters based on the LMedS method, the candidate region of moving objects needs to be less than half of an object region. If regions of moving objects are less than half of an object region, the color transformation parameters can be estimated by iterating a random sampling.

The color transformation parameters in every region which is extracted by a random sampling are estimated with an evaluation function based on a histogram's similarity. If there are no moving objects in the region, the same color transformation parameters should be estimated in each region. Then, color transformation parameters which are estimated in each region are applied to other regions. If there are no moving objects in the applied regions, the similarity value of histograms between the color corrected image and the standard image becomes higher. On the contrary, if there is the region which includes moving objects, the similarity value of histograms becomes lower. If an area where moving objects do not exist is more than half of the region, similarity value of histograms is calculated without the effect of moving objects by extracting a median of similarity values. Finally, color transformation parameters  $p_a, p_b$  when a similarity value is the highest in all regions are applied for the input image which is pre-processed in section 2.2. As a result, consistency of color tone is possible by removing the moving object. However, if an occupied rate of moving objects is more than half of the region, iteration process which is explained below is needed.

**Estimation of candidate region of moving object.** Regions of moving objects are estimated by calculating a difference of intensity between the corrected image and the standard image. Since the omnidirectional images are captured

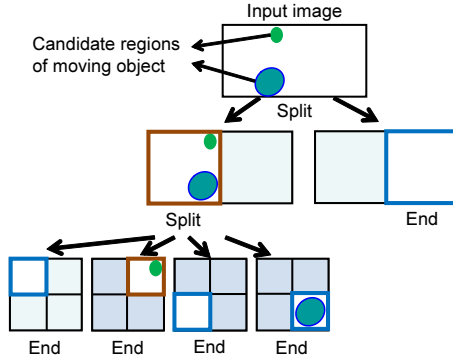


Fig. 2. Split of object region

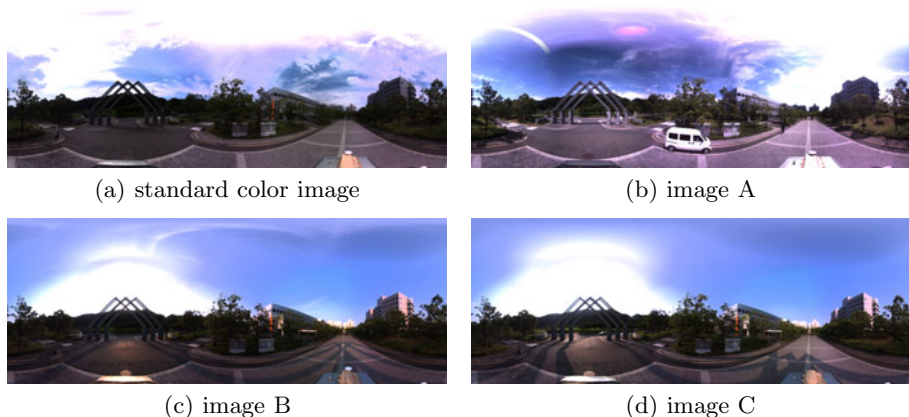
with a motion omnidirectional camera, a few disparities are existed between input images. Therefore, when the difference of an intensity value is calculated, a template matching approach is performed for every region and a candidate region of moving objects is estimated in consideration of the disparities between omnidirectional images. Here, an image which is masked to the region of moving objects is generated and is used for the region split in the following paragraph.

**Split of object region.** In each region for processing, when an occupied rate of moving objects is more than a fixed rate, the object region is split and color transformation parameters are estimated. This is based on an idea that areas with different transformation parameters are existing in one region for processing. Appearance of re-splitting the input image is shown in Fig. 2. The color transformation parameters of a major object in the region can be estimated with this approach.

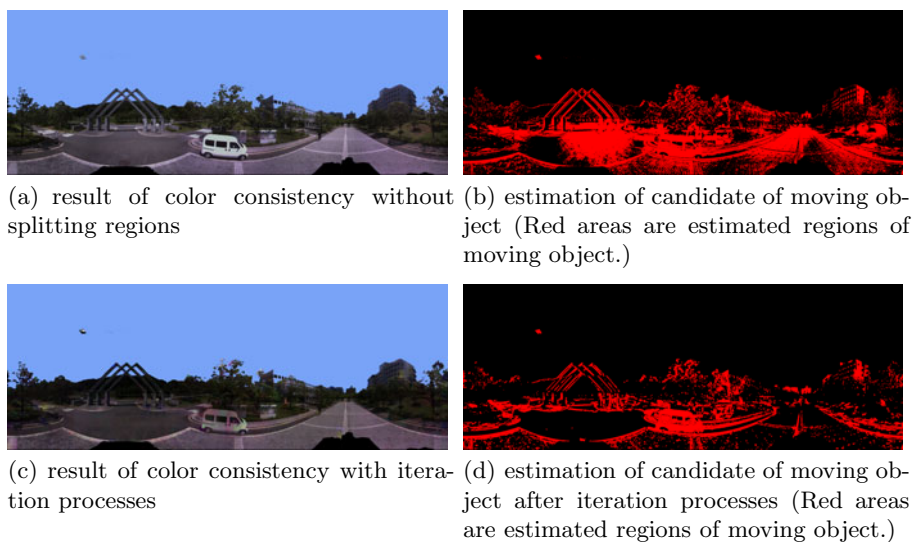
By iterating these processes, color consistency of images except in a region where the moving object is observed is achieved. Estimation of color transformation parameters which are robust to the influence of disparity is possible by maintaining a certain size of the split region.

## 2.4 Post-processing: Compensation of Moving Objects

Even if the calculated transformation parameters are applied for the input image, the moving objects which exist in the image can not be removed only by iterating the color consistency processing. In our work, the candidate object regions are compensated using other corrected omnidirectional images which are captured at near positions. Here, there is an assumption that a background of the moving object exists in the corrected input images. When the regions of moving objects are compensated, to consider the disparities between input images, the corresponded region with the area of moving objects are searched from the input images.



**Fig. 3.** Examples of input images which are captured at different times in nearby positions



**Fig. 4.** Result of color consistency with iteration processes to image A in Fig. 3

### 3 Experiments

#### 3.1 Experimental Environment

In the experiment, we used omnidirectional images which are captured with a car-mounted omnidirectional camera as input images and removed inconsistencies among them. We used an omnidirectional multi-camera system (Ladybug2,

Point Grey Research) for capturing in an outdoor environment. 5 omnidirectional images which were captured at near positions were used for input images. Since each image is captured at different time, color tones differ respectively. In this research, we assume that camera positions and postures can be acquired with some sensors [1] or by a vision-based approach [9]. Fig. 3 shows examples of input images. There are some moving objects in each image and it turns out that color tones are different due to a change of illumination condition. Each image has disparities due to a difference of captured positions.

### 3.2 Experimental Results

The standard image which has no moving object and are suitable for an input of an IBR approach are selected as shown in Fig. 3(a). The regions which are not necessary for color consistency, like sky region and equipment of capturing system were removed in advance from the input images.

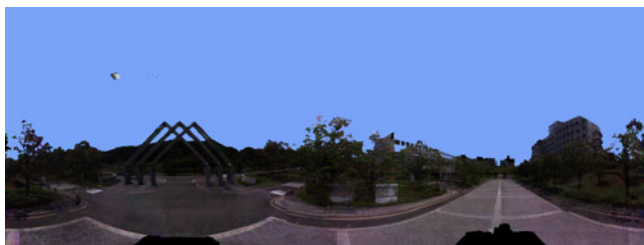
Fig. 4 shows the intermediate results of color consistency using LMedS method to image A. The result of color consistency without splitting regions is shown in Fig. 4(a). Here, since a set of transformation parameters were estimated to whole image and were applied to the input image, it turns out that some regions has a different color tone between the input image and the standard image. The difference of intensity value between the color corrected image and the standard image was computed as shown in Fig. 4(b). In this figure, regions which have large difference of intensity are masked red. This result shows that it is difficult to correct color tone only with a set of transformation parameters. Input image should be split into small regions and color transformation parameters should be estimated in each region. Next, the region which is estimated to be a moving object is split and the color transformation parameters based on robust estimation are estimated recursively. The result of color consistency with iteration processes is shown in Fig. 4(c). The comparison of the results as shown in Fig. 4(d) shows that iteration processes are effective for removing inconsistencies among omnidirectional images. Fig. 5 shows the results of compensation of moving objects using the corrected images. In each image, it turn out that moving objects are removed by the compensation.

**Table 1.** Similarity value of histogram with color standard image (1 is the highest similarity value.)

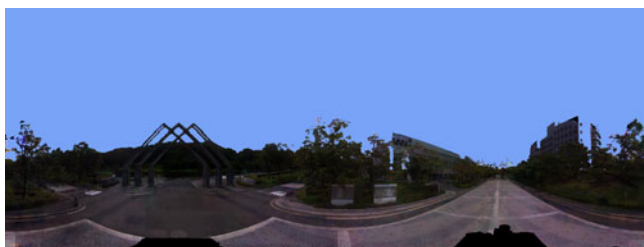
		R	G	B
Image A	input image	0.8429	0.8996	0.8877
	proposed method	0.9945	0.9925	0.9916
Image B	input image	0.9585	0.9703	0.9669
	proposed method	0.9943	0.9921	0.9918
Image C	input image	0.9717	0.9740	0.9697
	proposed method	0.9955	0.9935	0.9934
Image D	input image	0.9036	0.9265	0.9251
	proposed method	0.9944	0.9909	0.9920
Image E	input image	0.9203	0.9346	0.9306
	proposed method	0.9913	0.9817	0.9856



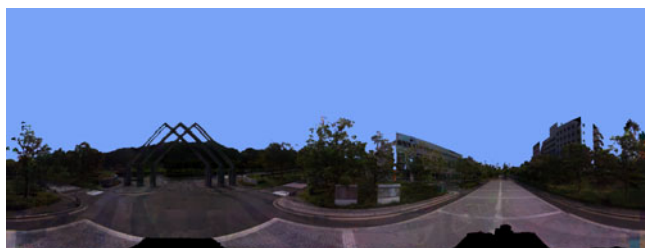
(a) standard color image



(b) image A



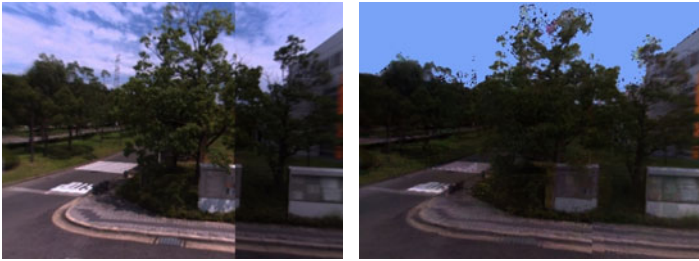
(c) image B



(d) image C

**Fig. 5.** Result for removing moving object and inconsistencies of color tone to input images in Fig. 3

To verify the validity of the proposed method, we conducted a quantitative evaluation. Similarity of histogram between the standard image and the color corrected image was used for evaluation. Bhattacharyya coefficient was used as a similarity of histograms. Table. 1 shows the similarity value of histogram with the standard image. It turned out that the similarity of the histogram with the standard image was improved for every result.



(a) Generation result from original omnidirectional images. (b) Generation result from proposed omnidirectional image database.

**Fig. 6.** Novel view images which are generated from omnidirectional images with the IBR approach.

Finally, as an application, novel view images were generated with the IBR approach [1] using the generated omnidirectional image database. Fig. 6 shows the novel view images which are generated by using omnidirectional images. The result as shown in Fig. 6(a) has inconsistencies which occur by a change of illumination condition. Fig. 6(b) shows good result by using the proposed omnidirectional image database.

## 4 Conclusion

In this paper, we have proposed the method for removing inconsistencies which occur by an existence of moving objects or a change of illumination condition when an omnidirectional image database is generated. Our approach has realized consistency of images with the iteration processes which are the estimation of candidate region of moving objects and the achievement of color consistency to split regions. Consistency of color tone is realized by estimating linear color transformation parameters which change histogram of the input image to that of the standard image. In experiments, we have confirmed that the proposed method can remove inconsistencies among omnidirectional images for an image database. As a future work, we have to make a large-scale omnidirectional image database.

## Acknowledgement

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (A), 19200016.

## References

1. Hori, M., Kanbara, M., Yokoya, N.: Novel stereoscopic view generation by image-based rendering coordinated with depth information. In: Ersbøll, B.K., Pedersen, K.S. (eds.) SCIA 2007. LNCS, vol. 4522, pp. 193–202. Springer, Heidelberg (2007)
2. Sato, R., Ono, S., Kawasaki, H., Ikeuchi, K.: Real-time image-based rendering system for virtual city based on image compression technique and eigen texture method. In: International Conference on Pattern Recognition (2008)
3. Finlayson, G., Hordley, S., Drew, M.: Removing shadows from images. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 823–836. Springer, Heidelberg (2002)
4. Moroney, N.: Local color correction using non-linear masking. *Color Science and Engineering Systems, Technologies, Applications*, 108–111 (2000)
5. Tsuchida, M., Kawanishi, T., Murase, H., Takagi, S.: Sequential monte-carlo estimation of background image for background subtraction under changing illumination. In: Int. Conf. on Visualization, Imaging, and Image Processing, pp. 421–425 (2003)
6. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut" - interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics* 23, 309–314 (2004)
7. Zafarifar, B., With, P.: Blue sky detection for picture quality enhancement. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2006. LNCS, vol. 4179, pp. 522–532. Springer, Heidelberg (2006)
8. Massart, D.L., Kaufman, L., Rousseeuw, P.J., Leroy, A.: Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta* 187, 171–179 (1986)
9. Sato, T., Ikeda, S., Yokoya, N.: Extrinsic camera parameter recovery from multiple image sequences captured by an omni-directional multi-camera system. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004, Part II. LNCS, vol. 3022, pp. 326–340. Springer, Heidelberg (2004)



# Shape Prior Embedded Geodesic Distance Transform for Image Segmentation

Junqiu Wang and Yasushi Yagi

The Institute of Scientific and Industrial Research  
Osaka University  
8-1 Mihogaoka, Ibaraki, Osaka, Japan  
jerywangjq@gmail.com

**Abstract.** Image segmentation is able to provides elements for enhancing a physical real-world environment. Although many existing segmentation methods have achieved impressive performances, they face problems where multiple similar objects are in close proximity to one another. We improve geodesic distance transform and define a symmetric morphology filter for segmentation. We embed shape prior knowledge into this geodesic distance transform filter. The proposed geodesic distance transform filter considers three factors simultaneously: the geometric distance, weighted gradients, and the distance to the boundary of the shape priors. As a result, it provides segmentation in line with the real shape of a particular kind of object. Positive results are demonstrated for several images and video sequences.

## 1 Introduction

Augmented reality enhances a physical real-world environment using virtual computer generated imagery. Image segmentation is helpful in creating an augmented environment since it can separate images into meaningful elements. Although image segmentation finds applications in augmented reality, it is a rather challenging problem in real images. We wish to provide a partial solution to this problem.

Image segmentation algorithms can be classified into two categories, namely, fully automatic segmentation and interactive segmentation. The algorithms in the former category are prone to failure in many cases as there are often ambiguities in the low-level intensity or color information of a given image. It is therefore advantageous to exploit the guidance obtained from user interaction or high-level knowledge about the expected objects in an attempt to disambiguate the low-level information. In this work, we incorporate shape priors into geodesic distance transform segmentation. The guidance of shape priors can be helpful to obtain a target labeling for a particular task that is too difficult to achieve using other methods. The algorithm proposed here can be applied in interactive segmentation, or automatic segmentation where shape priors are provided by a preprocessor. Whereas semi-automatic segmentation can provide elements for

augmented reality, the automatic segmentation can be applied into augmented reality more appropriately.

We employ a geodesic distance transform in our algorithm. The distance transform is a general fundamental operator that is widely applicable in computer vision and graphics. It maps each image pixel to the smallest distance to a region of interest. Two efficient distance transform methods have been proposed to speed up the computation: *ordered propagation* and *raster scanning*. The first method computes the smallest-distance information starting from the seeds and progressively propagating the information to other pixels in order of increasing distance. The second method, raster scanning, uses kernels to guide the processing of pixels from left to right, top to bottom and then from right to left, bottom to top.

An ordered propagation-based distance transform has been applied in colorization [1]. The idea has been extended to interactive image segmentation [2] based on roughly placed user scribbles. Image segmentation and matting is improved by computing weighted geodesic distances to the user-provided scribbles using spatial and/or temporal gradients [3].

Criminisi *et al.* [4] proposed geodesic segmentation in which image gradients are included in the distance transform to encourage spatial regularization and contrast-sensitivity. Their idea is similar to the algorithm in [3], except that they use raster scanning for the distance transform. Furthermore, the geodesic filter acts only on the energy unaries, and not on the user scribbles.

Although these segmentation approaches [3,4] have achieved impressive performance in many examples, the filter may fail in images in which the quality of the likelihood images is not satisfactory, or where multiple similar objects are located close to one another. Optimization relying solely on low level image data is subject to many local optima representing irregular segmentations. One possible solution to this problem is to incorporate prior knowledge into the segmentation. In this work, we embed shape priors in an image segmentation algorithm based on a geodesic distance transform. In contrast to other works, we consider geometric distance, image gradients, and shape priors simultaneously in the computation of the geodesic transform, which is especially important when likelihood images are not satisfactory. Shape prior knowledge is incorporated in a distance transform-based morphology, and hence, the segmentation achieves good performance by adding appropriate regularization terms to the functions. Our approach is applicable to both interactive segmentation and automatic segmentation, where a tracking algorithm [5,6] provides likelihood images and shape priors.

This paper is structured as follows. The remainder of this section gives a brief review of prior works. In Section 2, we introduce geodesic distance transforms in an integrated framework. We also describe image segmentation that makes use of a geodesic distance transform in this section. We develop a geodesic distance transform by embedding shape priors in the transform in Section 3. The performance of the proposed method is evaluated in Section 4. We conclude the paper in Section 5.

## 1.1 Related Work

There is a great deal of literature on image segmentation. Level sets methods evolve user placed contours to the boundary at local energy minima. The implementation is difficult due to the specification of the many free parameters and the difficulty in providing progressive user guidance. Shape statistics have been integrated into a Mumford-Shah based segmentation process [7]. The segmentation can incorporate shape prior knowledge, making the segmentation process robust, however, it inherits the disadvantages of level set algorithms.

The graph cuts technique [8] has achieved impressive success thanks to the efficient computation of max-flow/min-cut. In this technique, an image is viewed as a graph, weighted to reflect intensity changes. The segmentation problem is transformed into an energy minimization in a conditional random field. The technique returns the smallest cut separating the seeds provided by the user. Unfortunately, perceptual grouping in certain areas may not correspond to the global minimum because Markov random fields (MRFs) provide a poor prior for specific shapes [9].

Image segmentation has been augmented by using shape priors in the min-cut [9,10,11,12,13], level-set methods [14,15], watershed segmentation [16], random walk segmentation [17], and the Mumford-Shah [18] based process for segmentation [19,7,20]. Freedman and Zhang [9] added shape priors into energy minimization using min-cut. Shape priors can be incorporated into level sets methods, which evolve user placed contours to the boundary at local energy minima. However, the implementation of such methods is difficult due to the specification of the many free parameters and the difficulty in providing progressive user guidance. Cremers et al. integrated shape statistics into a Mumford-Shah based segmentation process [7]. The segmentation can incorporate shape prior knowledge, making the segmentation process robust. They estimate the translation, scaling and rotation of the shape before applying their density estimate using a method similar to [21]. In their method, the correspondences between two point sets are assumed known before the segmentation. Recently, the problem of registration of two point sets is solved by using a polynomial transform under an affine transformation [22]. In [22], the correspondences are not known on the point-level. Nevertheless, it is still necessary to know the correspondences on the moment-level.

## 2 Distance Transform

In image processing applications, a distance transform is usually performed on a regular grid,  $\mathcal{G}$ . Based on a set of points  $P$  ( $P \subset \mathcal{G}$ ) (The point set usually contains certain structuring information.) on the grid, the distance transform is defined as

$$D_P(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{G}} (d(\mathbf{x}, \mathbf{y}) + 1(\mathbf{y})), \quad (1)$$

where  $d(\mathbf{x}, \mathbf{y})$  is a particular measure of the distance between  $\mathbf{x}$  and  $\mathbf{y}$ ,  $1(\mathbf{y}) = 0$  if  $\mathbf{y} \in P$  and  $\infty$  otherwise. The distance transform finds a point  $\mathbf{y}$  that is nearest

to  $\mathbf{x}$ . In other words, the distance transform computes the shortest path in all possible paths in  $\mathcal{G}$  between  $\mathbf{x}$  and  $\mathbf{y}$ .

## 2.1 Generalized Distance Transform

Substituting a function  $f(\mathbf{y})$  into Eq. 1, the distance transform is generalized:

$$D_f(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{G}} (d(\mathbf{x}, \mathbf{y}) + f(\mathbf{y})). \quad (2)$$

$D_f(\mathbf{x})$  depends the specific definition of the function  $f(\mathbf{y})$ . For each point  $\mathbf{x}$ , it finds a point  $\mathbf{y}$  that is close to  $\mathbf{x}$ , and for which  $f(\mathbf{y})$  is small.

The generalized distance transform has many variations according to the definition of the distance measure  $d(\mathbf{x}, \mathbf{y})$  and the definition of  $f(\mathbf{y})$ .

## 2.2 Geodesic Distance Transform Using Image Gradients

The essence of the geodesic distance transform lies in encoding the knowledge of image gradients or other information available.

Each path between two points  $\mathbf{x}$  and  $\mathbf{y}$  on image lattices is composed of many discrete steps. In Geodesic distance transform, we compute a weight factor for each step according to image gradients  $\nabla I$ .

The distance in Eq. 2 is now defined as an accumulation of weighted step distances

$$d_G(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{x} \rightarrow \mathbf{y}} \sqrt{(1 + \rho^2 (\nabla I \cdot \tilde{\mathbf{u}})^2)} \Delta \mathbf{u}, \quad (3)$$

where  $\Delta \mathbf{u}$  is the step distance,  $\tilde{\mathbf{u}}$  is the unit vector that is tangential to the direction of a step in one of the possible paths from  $\mathbf{x}$  to  $\mathbf{y}$ , and the factor  $\rho$  weights the contribution of the image gradient versus the spatial distance.

Based on the user input in the image, we compute probabilities that a pixel belongs to the background ( $p(I(\mathbf{y})|BK)$ ) and foreground ( $p(I(\mathbf{y})|FG)$ ). After that, we compute likelihood ratio  $l(\mathbf{y}) = \log \frac{p(I(\mathbf{y})|BK)}{p(I(\mathbf{y})|FG)}$ . Then, we compute a sigmoid function

$$f_L(\mathbf{y}) = \frac{1}{1 + \exp(-l(\mathbf{y})/\mu_l)}, \quad (4)$$

where  $\mu_l$  is coefficient and set to 5 experimentally.  $f_l(\mathbf{y})$  gives structuring information in a probabilistic form.

Considering the weighted distance  $d_G(\mathbf{x}, \mathbf{y})$  and the probabilistic structuring information  $f_l(\mathbf{y})$ , we can apply geodesic distance transform

$$D_{GL}(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{G}} (d_G(\mathbf{x}, \mathbf{y}) + p_L f_L(\mathbf{y})), \quad (5)$$

where  $p_L$  is the confidence on the likelihood ratios.  $p_L$  is set interactively [4].

### 3 Shape Prior Embedded Geodesic Segmentation

#### 3.1 Shape Prior Embedded Geodesic Distance Transform

Each shape prior is defined by a region  $z^R$  and a contour  $z^S$  ( $z^S = \partial z^R$ ). Given a shape prior, we apply a Euclidean distance transform to the contour  $z^S$  based on Eq. 1. We get distance transform result  $T_S(\mathbf{y}) = D_S(\mathbf{y})$ . Then, we assign  $T_{S,R}(\mathbf{y}) = -T_S(\mathbf{y})$  if  $\mathbf{y}$  is in the foreground region defined by  $z^R$ ; and  $T_{S,R}(\mathbf{y}) = T_S(\mathbf{y})$  if  $\mathbf{y}$  is in the background region. To encode the structuring information in the shape prior, we compute another sigmoid function

$$f_S(\mathbf{y}) = \frac{1}{1 + \exp(-T_{S,R}(\mathbf{y})/\mu_T)}, \quad (6)$$

where  $\mu_T$  determines the confidence of the shape priors.

Integrating the geodesic distance in Eq. 3 and the structuring information defined in Eq. 4 and Eq. 6, we compute shape prior embedded geodesic distance transform

$$D_{GLS}(\mathbf{x}) = \min_{\mathbf{y} \in \mathcal{G}} (d_G(\mathbf{x}, \mathbf{y}) + p_L f_L(\mathbf{y}) + p_S f_S(\mathbf{y})), \quad (7)$$

where  $p_L$  and  $p_S$  indicate our confidence on the likelihood ratios and shape priors, respectively.  $p_S$  is set to the probability of the selected shape prior.

In the shape prior embedded geodesic distance transform, we consider the gradients, the likelihood ratios and the shape prior knowledge simultaneously. This strategy improves the performance of our geodesic segmentation, as confirmed by the experimental results presented in our experimental results.

#### 3.2 Geodesic Morphology Operators

It is well known that distance transform results can be used for image morphology [23]. Based on the distance transform in Eq. 7, we define a signed distance

$$D_s(\mathbf{x}; \nabla I, f_L, f_S) = D_{GLS}(\mathbf{x}; \nabla I, f_L, f_S) - D_{GLS}(\mathbf{x}; \nabla I, \overline{f_L}, \overline{f_S}), \quad (8)$$

where  $\overline{f_L} = 1 - f_L$  and  $\overline{f_S} = 1 - f_S$ . Different from [4], we consider the geometric distance, weighted gradients, and distance to the boundary of shape priors in the signed distance.

We extract structuring information using erode and dilate morphology techniques. Thresholding the signed distance in Eq. (8), we get structuring information

$$P_e = \delta_{\mathbf{x}}(D_s(\mathbf{x}; \nabla I, f_L, f_S) > -\theta_e), \quad (9)$$

and

$$P_d = \delta_{\mathbf{x}}(D_s(\mathbf{x}; \nabla I, f_L, f_S) > \theta_d), \quad (10)$$

where  $\delta(\cdot)$  is a delta function,  $\delta(\cdot) = 1$  when the function inside is true; otherwise  $\delta(\cdot) = 0$ ;  $P_e$  is the structuring information by erode and  $P_d$  by dilate;  $\theta_e$  and  $\theta_d$  control the smoothness in erode and dilate operations, respectively ( $\theta_d > 0$  and  $\theta_e > 0$ ). The setting of  $\theta_e$  and  $\theta_d$  will be discussed in Section 3.3.

Based on Eq. (9) and Eq. (10), we compute structuring information  $P_o$  and  $P_c$  by applying open and close operations,

$$P_o = \delta_{\mathbf{x}}(D_s(\mathbf{x}; P_e) > \theta_d), \quad (11)$$

and

$$P_c = \delta_{\mathbf{x}}(D_s(\mathbf{x}; \overline{P_d}) < \theta_e), \quad (12)$$

where  $\overline{P_d} = 1 - P_d$ .

We get another signed distance  $D_s^s$  for segmentation using structuring information  $P_e$  and  $\overline{P_d}$ ,

$$D_s^s(\mathbf{x}) = D_s(\mathbf{x}; P_e) - D_s(\mathbf{x}; \overline{P_d}) + \Delta\theta, \quad (13)$$

where  $\Delta\theta = \theta_d - \theta_e$ .

For image segmentation, we define a symmetric morphology operator

$$P_s(\mathbf{x}) = \delta_{\mathbf{x}}(D_s^s(\mathbf{x}) > 0), \quad (14)$$

where  $P_s$  is the structuring information.

### 3.3 Segmentation

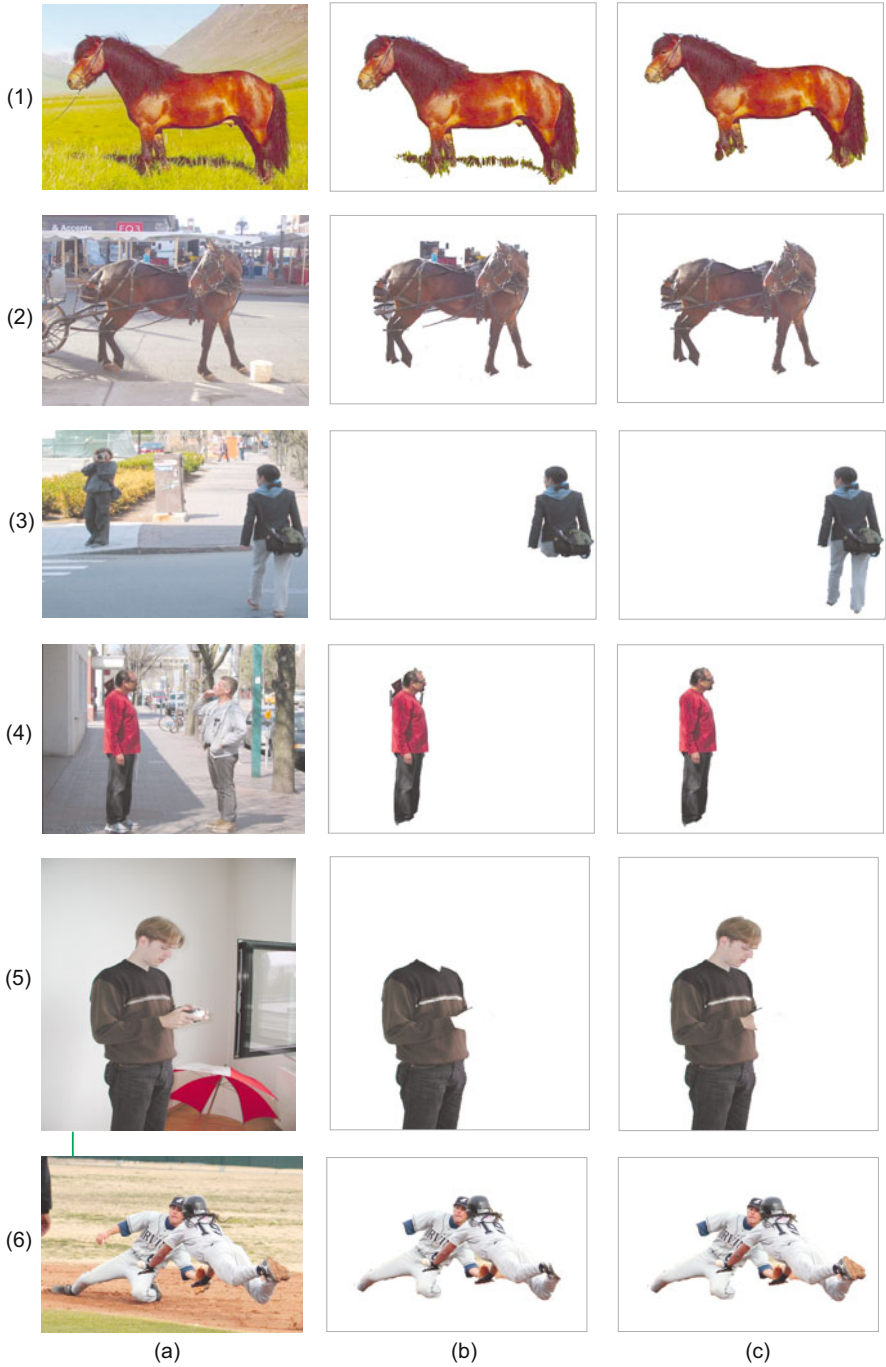
Segmentation can be achieved by minimizing the energy in the Markov random field formulated by an image [4,8]. However, energy minimization by searching for a solution over a restricted parameterized 2D manifold is computationally expensive, especially when we consider all of the possible parameters.

We segment an object based on the result of Eq. (14). Different segmentation results can be computed by varying the smoothness parameters  $\theta_e$  and  $\theta_d$ . Since we are designing an interactive image segmentation,  $\theta_e$  and  $\theta_d$  can be set interactively. We do not need to try all the parameters and evaluate the results based on the energy of MRF. The user can evaluate segmentation results. In practice,  $\theta_e$  and  $\theta_d$  are set according to the size of an input image. Assuming  $w$  and  $h$  are the width and height of the image, default values of  $\theta_e$  and  $\theta_d$  are calculated using  $\theta_d = \frac{\min(w,h)}{80}$  and  $\theta_e = \frac{\min(w,h)}{80}$ .

The segmentation are found directly from the filtering results. Although it seems that this approach is not as theoretically sound as the energy minimization framework, it works well in practice.

## 4 Results

We have implemented the proposed algorithm and tested it on many images and video sequences. The prior can also be provided by the users, although this is tedious and time consuming. We advocate computing shape priors automatically, as this leads to less work on the part of the user.



**Fig. 1.** Examples for image segmentation. (a) The input images. (b) Segmentation results using the grabcut method. (c) Segmentation results using the proposed method.

## 4.1 Interactive Image Segmentation

We segment images based on scribbles provided by the user. In our application, we compute the likelihood images using these scribbles. The regions in the scribbles are assumed to targets for segmentation. We are particularly interested in segmenting objects in difficult images which contain objects with a similar appearance exist. The segmentation results are shown in Fig. 1. We compare our method with the grab-cut method in [24]. We initialize the segmentation by providing a bounding box for each object, which is same to the initialization approach in [24]. We believe such initialization is less cumbersome than other manual interactions. We compute likelihood ratios based on the input using kernel estimation method. Then, we search for a shape prior in a specific prior set. We provide priors by collecting well-segmented objects. The distributions of the priors are learned using a method similar to [25]. The first example involves segmenting a horse. Segmentation using the approach in [24] labels the shadow regions under the horse as foreground. To get the result shown in the first example in Fig. 1, we run 5 iterations of their algorithm on the first example (In fact, the segmentation does not have much improvement after the second iteration). This problem is dealt with by using our method using the shape embedded geodesic distance transform. The proposed segmentation algorithm correctly segment the horse out by using the shape prior computed from the prior set. In the third example, we try to segment a girl. The lower body of the girl has similar appearance with the background. The algorithm from [24] can not give good results even after 7 iterations. The proposed method segments the girl well, thanks to the shape prior embedded distance transform. In other examples (but the last example) in Fig. 1, our segmentation results are better than those using the approach in [24]. The advantage of our algorithm is evident in the first five examples. However, in the last example in Fig. 1, we can not find a appropriate prior from the prior set. Therefore, we stop using shape priors in our segmentation. The segmentation result of our approach is even a little worse than the method in [24].

**Table 1.** Quantitative comparisons of image segmentation results

Images	1	2	3	4	5	6	All
Our method	7.2	6.6	10.7	7.9	11.3	15.1	9.7
Grab-cut [24]	13.4	12.3	38.5	11.6	24.5	14.6	13.5

For quantitative comparison, we compute the error rates as the percentage of mislabeled pixels inside the bounding boxes. Table 1 shows segmentation errors with respect to ground truth for the test images in Fig. 1. We observed that the segmentation errors of our method are apparently lower than the results by the grab-cut in the first five examples. However, the error rates of the two methods are similar in the 6th example because the foreground objects do not have appropriate shape prior in the prior set. This is one of the limitations of our method.



Besides the examples shown in Fig. 1, we compare the performance on 50 images, the average performance comparison of all the 56 examples is shown in the last column in Table 1.

## 5 Conclusions

We presented a novel geodesic segmentation algorithm that incorporates shape priors. This shape prior knowledge is helpful for computing a correct labeling with realistic shapes. The proposed method exhibits the desired properties of image segmentation. We also extended it for video segmentation. It can be applied in interactive image segmentation or semi-automatic video segmentation.

Prior knowledge is important to improve the performance of segmentation. This work learns shape priors in a batch mode. We are interested in developing a system that incrementally learns prior knowledge online. We believe such a system has a similar mechanism that may be at work with human-beings.

## References

1. Levin, A., Lischinski, D., Weiss, Y.: Colorization using optimization. *ACM Transactions on Graphics* 8, 11–19 (2004)
2. Protiere, A., Sapiro, G.: Interactive image segmentation via adaptive weighted distances. *IEEE Trans. on Image Processing* 16(4), 1046–1057 (2007)
3. Bai, X., Sapiro, G.: A geodesic framework for fast interactive image and video segmentation and matting. In: *Proc. of ICCV*, pp. 1–8 (2007)
4. Criminisi, A., Sharp, T., Blake, A.: GeoS: Geodesic image segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I. LNCS*, vol. 5302, pp. 99–112. Springer, Heidelberg (2008)
5. Wang, J., Yagi, Y.: Integrating color and shape-texture features for adaptive real-time tracking. *IEEE Trans. on Image Processing* 17(2), 235–240 (2008)
6. Wang, J., Yagi, Y.: Integrating shape and color features for adaptive real-time object tracking. In: *Proc. of Conf. on Robotics and Biomimetics*, pp. 1–6 (2006)
7. Cremers, D., Kohlberger, T., Schnorr, C.: Shape statistics in kernel space for variational image segmentation. *Pattern Recognition* 36, 1929–1943 (2003)
8. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: *Proc. of ICCV*, pp. 105–112 (2001)
9. Freedman, D., Zhang, T.: Interactive graph cut based segmentation with shape priors. In: *Proc. of CVPR*, pp. 755–762 (2004)
10. Bray, M., Kohli, P., Torr, P.: POSE CUT: Simultaneous segmentation and 3D pose estimation of humans using dynamic graph-cuts. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006. LNCS*, vol. 3952, pp. 642–655. Springer, Heidelberg (2006)
11. Wang, J., Makihara, Y., Yagi, Y.: Human tracking and segmentation supported by silhouette-based gait recognition. In: *Proc. of IEEE Int. Conf. on Robotics and Automation* (2008)
12. Besbes, A., Komodakis, N., Langs, G., Paragios, N.: Shape priors and discrete mrfs for knowledge-based segmentation. In: *Proc. CVPR*, pp. 1295–1302 (2009)

13. Wang, J., Yagi, Y., Makihara, Y.: People tracking and segmentation using efficient shape sequences matching. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009. LNCS, vol. 5995, pp. 204–213. Springer, Heidelberg (2010)
14. Leventon, M., Grimson, W., Faugeras, O.: Statistical shape influence in geodesic active contours. In: Proc. CVPR, pp. 316–323 (2000)
15. Chen, Y., Thiruvenkadam, S., Tagare, H., Huang, F., Wilson, D., Geiser, E.: On the incorporation of shape priors into geometric active contours. In: Proc. of IEEE Workshop on Variational and Level Set Methods, pp. 145–152 (2001)
16. Nguyen, H., Ji, Q.: Improved watershed segmentation using water diffusion and local shape priors. In: Proc. CVPR, pp. 985–992 (2006)
17. Teboul, O., Simon, L., Koutsourakis, P., Paragios, N.: Segmentation of building facades using procedural shape priors. In: Proc. CVPR (2010)
18. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.* 42(5), 577–685 (1989)
19. Cremers, D., Tischhäuser, F., Weickert, J., Schnörr, C.: Diffusion snakes: Introducing statistical shape knowledge into the mumford-shah functional. *International Journal of Computer Vision* 50(3), 295–313 (2002)
20. Trobin, W., Pock, T., Cremers, D., Bischof, H.: An unbiased second-order prior for high-accuracy motion estimation. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 396–405. Springer, Heidelberg (2008)
21. Werman, M., Weinshall, D.: Similarity and affine invariant distances between 2d point sets. *IEEE Trans. on Patt. Anal. and Mach. Intell.* 17(8), 810–814 (1995)
22. Ho, J., Peter, A., Ranganranjan, A., Yang, M.H.: An algebraic approach to affine registration of point sets. In: Proc. of ICCV, pp. 1335–1340 (2009)
23. Serra, J.: Image analysis and mathematical morphology. Academic Press, London (1982)
24. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: interactive foreground extraction using iterated graph cuts. *ACM Trans. on Graphics* 23(3), 309–314 (2004)
25. Toyama, K., Blake, A.: Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision* 48(1), 9–19 (2001)

# Shortest Path Based Planar Graph Cuts for Bi-layer Segmentation of Binocular Stereo Video

Xiangsheng Huang<sup>1</sup> and Lujin Gong<sup>2</sup>

<sup>1</sup> Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China  
xiangsheng.huang@ia.ac.cn

<sup>2</sup> State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, AMSS, CAS, Beijing 100190, China

**Abstract.** Separating a foreground layer from stereo video in real-time is used in many applications such as live background substitution. Conventional separating models using stereo, contrast or color alone are usually not accurate enough to be satisfactory. Furthermore, the powerful tool of graph cut which is well suited for segmentation is known to be not efficient enough especially for high resolution images. In this paper, we conquer these difficulties by fusing stereo with color and contrast to model the segmentation problem as an minimum cut problem of a planar graph and solving it by a specialized algorithm, parametric shortest paths [8] with a dynamic tree structure, in  $O(n \log n)$  time. Experimental results demonstrate the high accuracy and efficiency of the algorithm.

## 1 Introduction

Separating a foreground layer from stereo video is widely used in many applications, such as live background substitution, medical imaging, machine vision, and face recognition. The challenge is that both high quality and efficiency of the segmentation are required.

Many researches on obtaining a high quality segmentation have been conducted [3,10,19]. Stereo algorithms [7,9,12] that compute depth or occlusion can be used to get good results of layer extraction by the fact that the foreground and background should have different depths. However, most stereo algorithms can not do well in textureless regions. Researches on color and contrast based segmentation techniques [4,14] have been very active recently, which are very effective, even on low-texture images. But these are usually over segmentation and interactive methods, which are semiautomatic.

One of the most powerful techniques for making binary classification decisions is the graph cut method. Segmentation is one application of these types of decisions. Most implementations solve the equivalent problem of maximum flow, but traditional maximum flow algorithms, which are designed for general graphs, are too slow for live background substitution, especially in the cases that the images are of high resolution.

In [11], the authors presented two models for fusion of stereo with color and contrast, one of which solved the segmentation problem by ternary graph cut, that is,  $\alpha$ -expansion algorithm [6] with three labels. But this graph cut technique for general graphs is also not suitable for high resolution images in real-time applications, since as the scale of the graph increases, the running time of the graph cut procedure increases rapidly. Hence, one way to solve the segmentation problem for high resolution images in real-time is to slow down the increase rate of the running time of graph cut.

In this paper, we propose a novel solution, which fuses stereo with color, contrast, and a prior for intra-layer spatial coherence. Fusion of a variety of cues can improve the accuracy of the segmentation evidently compared to contrast, color, or stereo alone methods. Furthermore, in the consideration of efficiency, we model the segmentation problem with a variety of cues as a minimum cut problem in a planar graph, which can be solved much faster than the classical graph cuts by taking the advantage of the planar nature of the graph. The high efficiency is obtained by reformulating the maximum flow problem in the original graph as a parametric shortest path problem in the dual graph [8] and using standard dynamic tree data structures, which makes it possible to solve the minimum cut problem in  $O(n \log n)$  time. Experimental results imply that our algorithm is much faster than the classical graph cut methods, and the fusion of variety cues indeed works that the segmentation quality is better than using each of the cues alone.

This paper is organized as follows: In section 2, we model the segmentation problem as a planar graph and construct the energy function on it with a variety of cues. In section 3, we present our shortest path based planar graph cut algorithm in detail, which has a running time of  $O(n \log n)$ , by using dynamic tree structures. Experimental results measuring the accuracy and efficiency of the algorithm are given in section 4. Finally, section 5 provides conclusions and future work.

## 2 Construct Energy Function on Planar Graph

Separation of layers using color/contrast information or stereo information alone is known to be error-prone. Hence we fuse color, contrast and stereo matching information to infer layers accurately. The basic idea of our model is to minimize an energy function that consists of three items:

$$E = \alpha_c L + \alpha_c C + \alpha_s S \quad (1)$$

where  $\alpha$  is a vector of weights,  $L$ ,  $C$ ,  $S$  are the contrast, color and stereo item respectively.

The energy function  $E$  is defined on a set of edges between the pixels in the image. Different sets give different values of  $E$ . Given some pixels as the source and some other pixels as the sink, our aim is to search the sets  $\Omega$  whose edges can separate the source and the sink completely for an optimal one that has minimum energy value, which is

$$\min_{\Omega} E(\Omega) = \alpha_c L(\Omega) + \alpha_c C(\Omega) + \alpha_s S(\Omega) \quad (2)$$

Thus, we could convert the energy minimization problem into an equivalent minimum cut problem of the graph, whose vertices are the pixels and edges are the edges between the pixels, with some objective pixels as the source and some background pixels as the sink. The source and sink pixels are learned from the previous frames. The graph we have constructed is a planar graph.

## 2.1 Contrast Energy on Planar Graph

The contrast energy consists of three edge features: the Laplacian zero-crossing, gradient magnitude, and gradient direction as in Mortensen and Barrett's [13]. Thus we can obtain the cost of the edge between pixel  $p$  and its neighbor  $q$  as:

$$l(p, q) = \omega_Z \cdot f_Z(q) + \omega_G \cdot f_G(q) + \omega_D \cdot f_D(p, q) \quad (3)$$

where  $\omega$  is a vector of weights.

$f_Z$  represents the Laplacian feature, which is an approximation to the second derivative of the image. This feature is zero when the gradient of the intensity arrives a maximizer. Let  $I_L(p)$  be the Laplacian of pixel  $p$  in image  $I$ , then  $f_Z$  is given by:

$$f_Z(p) = \begin{cases} 0 & \text{if } I_L(p) = 0 \\ 1 & \text{if } I_L(p) \neq 0 \end{cases} \quad (4)$$

The gradient magnitude feature is computed by

$$G = \sqrt{I_x^2 + I_y^2} \quad (5)$$

where  $I_x, I_y$  denote the discrete horizontal and vertical derivatives respectively. This value is then scaled and inverted, such that high gradient values between pixels yield low cost edges, and vice versa:

$$f_G(p) = 1 - \frac{G}{\max(G)} \quad (6)$$

Finally, the gradient direction feature is computed by:

$$f_D(p, q) = \frac{2}{3\pi} \{ \arccos[d_p(p, q)] + \arccos[d_q(p, q)] \} \quad (7)$$

$$d_p(p, q) = D(p) \cdot T(p, q) \quad (8)$$

$$d_q(p, q) = T(p, q) \cdot D(q) \quad (9)$$

$$T(p, q) = \begin{cases} q - p & \text{if } D(p) \cdot (q - p) \geq 0 \\ p - q & \text{if } D(p) \cdot (q - p) < 0 \end{cases} \quad (10)$$

where  $D(p)$  is often given by:

$$D(p) = (I_y(p), -I_x(p)) \quad (11)$$

In essence,  $f_D$  assigns a high cost to edges between pixels whose gradient directions are similar but perpendicular to the link direction and a low cost to edges between pixels whose gradient directions are similar and parallel to the link direction.

Thus, we have:

$$L(\Omega) = \sum_{(p,q) \in \Omega} l(p, q) \quad (12)$$

## 2.2 Color-Likelihood Energy on Planar Graph

Like [4,14], we use Gaussian mixtures learned from the previous image frames that have been labeled by our algorithm to model the color likelihoods for foreground and background. Let  $P_F$ ,  $P_B$  denote the Gaussian mixtures of foreground and background respectively, and  $P$  denotes the Gaussian density, learned by pixelwise background maintenance [15,17], for each of the background pixels.  $P_B$  and  $P$  are combined when the stability flag  $s_k \in \{0, 1\}$  takes value 1, which indicates that there has been stasis over a sufficient number of previous frames. The color feature of an edge is then given by:

$$c^*(p, q) = \sqrt{(U(p, F) - U(q, F))^2 + (U(p, B) - U(q, B))^2} \quad (13)$$

which is some kind of gradient in the probability space.  $U(p, F)$  and  $U(p, B)$  are defined as:

$$U(p, F) = -\log P_F(p) \quad (14)$$

$$U(p, B) = -\log\left[\left(1 - \frac{s_k}{2}\right)P_B(p) + \frac{s_k}{2}P(p)\right] \quad (15)$$

$c^*$  is then scaled and inverted, such that high gradient values between pixels yield low cost edges:

$$c(p, q) = 1 - \frac{c^*(p, q)}{\max(c^*(p, q))} \quad (16)$$

Therefore, the color item is specified as:

$$C(\Omega) = \sum_{(p,q) \in \Omega} c(p, q) \quad (17)$$

## 2.3 Stereo Coherence Energy on Planar Graph

We use Gaussian mixtures  $P_{SF}$  and  $P_{SB}$ , which are learned from earlier image frames to model the likelihoods for disparity of the foreground and the background respectively.

Then we introduce SSD (sum-squared difference), which is  $L^2$ -norm of difference between image patches  $L(p)$ ,  $R(r)$  surrounding hypothetically matching pixels  $p$  and  $r$ . In the consideration of robustness, we use normalized SSD as follows:

$$N(L(p), R(r)) = \frac{1}{2} \frac{\|L(p) - R(r)\|^2}{\|L(p) - \bar{L}(p)\|^2 + \|R(r) - \bar{R}(r)\|^2} \in [0, 1] \quad (18)$$

We can give the stereo feature of an edge as:

$$s^*(p, q) = \sqrt{(M(p, F) - M(q, F))^2 + (M(p, B) - M(q, B))^2} \quad (19)$$

where  $M(p, F)$ ,  $M(p, B)$  are defined as:

$$M(p, F) = -\log\left[\sum_d P_{SF}(d) \exp(-\lambda N(L(p), R(p+d)))\right] \quad (20)$$

$$M(p, B) = -\log\left[\sum_d P_{SB}(d) \exp(-\lambda N(L(p), R(p+d)))\right] \quad (21)$$

$s^*$  can be viewed as some kind of gradient, which should be scaled and inverted, such that high gradient values between pixels yield low cost edges:

$$s(p, q) = 1 - \frac{s^*(p, q)}{\max(s^*(p, q))} \quad (22)$$

Then our stereo item can be formulated as:

$$S(\Omega) = \sum_{(p, q) \in \Omega} s(p, q) \quad (23)$$

### 3 Shortest Path Based Planar Graph Cuts

As mentioned in the above section, we have modeled segmentation as a minimum cut problem on a planar graph. In this section, we adopt parametric shortest paths graph cut with dynamic tree to minimize energy in  $O(n \log n)$  time.

#### 3.1 Parametric Shortest Paths

We solve the equivalent maximum flow problem of the graph to get minimum cut. The maximum flow problem of a planar graph can be reformulated as a parametric shortest path problem in the dual graph. For any value of the parameter  $\lambda$ , the shortest path distances in the dual graph define a flow with value  $\lambda$  in the original network, which builds up a connection between the problems in the original and dual graphs. Using dynamic tree data structure and special structure of parameterization, the algorithm runs in  $O(n \log n)$  time. However, when implementing the algorithm, we do not need to maintain the parameter  $\lambda$  explicitly.

### 3.2 Shortest Path Based Planar Graph Cuts

The detailed algorithm is as follows:

Shortest Path Based Planar Graph Cuts (SPPGC):

1. Initialization:

Fix an arbitrary dual vertex  $o$  (called the origin) in the dual graph. Compute the shortest path from the origin to each vertex  $p$ , let  $dist(p)$  denote the shortest path distance in the dual graph from  $o$  to  $p$ ,  $c(e)$  denote the weight of edge  $e$  in the original graph. Then we can define the *slack* of each dual edge  $e^*$  as follows:

$$slack(e^*) := dist(tail(e^*)) - dist(head(e^*)) + c(e) \quad (24)$$

Let  $T$  denote the single-source shortest path tree in the dual graph rooted at  $o$ . The edges in  $T$  are directed away from  $o$ . Thus, every dual vertex  $p \neq o$  has exactly one incoming edge in  $T$ , from its parent vertex, which we denote  $pred(p)$ . A dual edge  $e^*$  is called *tense* if  $slack(e^*) = 0$  and a primal edge  $e$  is called *loose* if neither its dual  $e^*$  nor its reversed dual  $rev(e^*)$  is tense. Let  $L$  be the subgraph of all loose edges, then  $L$  is a spanning tree.

2. Iterate:

While the source  $s$  and the sink  $t$  are in the same component of  $L$ :

$LP \leftarrow$  the path in  $L$  from  $s$  to  $t$

$p \rightarrow q \leftarrow$  the edge in  $P^*$  with minimum slack

$\Delta \leftarrow slack(p \rightarrow q)$

for every edge  $e$  in  $LP$

$slack(e^*) \leftarrow slack(e^*) - \Delta$

$slack(rev(e^*)) \leftarrow slack(rev(e^*)) + \Delta$

delete  $(p \rightarrow q)^*$  from  $L$

if  $q \neq o$

insert  $(pred(q) \rightarrow q)^*$  into  $L$

$pred(q) \leftarrow p$

3. Output:

For each edge  $e$

$\phi(e) \leftarrow c(e) - slack(e^*)$

Return  $\phi$

### 3.3 Dynamic Tree

To implement the above algorithm in  $O(n \log n)$  time, we need a special data structure, which is called a dynamic tree structure [1,2,16,18] to maintain the spanning tree  $L$  and the dual slacks. This data structure can support the following operations in  $O(\log n)$  amortized time: determine whether two nodes are in the same component, *expose* a path between two specified nodes, find the edge on the exposed path with minimum value, add some amount to all values on the exposed path, remove an edge, and insert an edge.



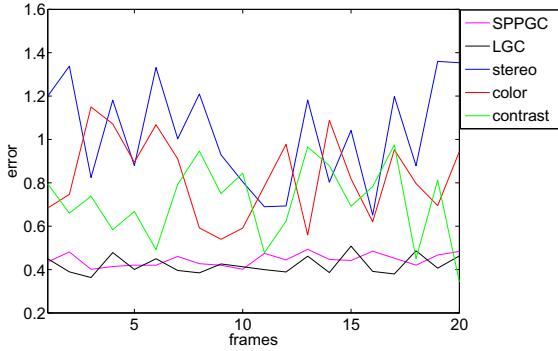


Fig. 1. Comparison of accuracy using different models

## 4 Experiments

We measure the shortest path based planar graph cut algorithm (SPPGC) in two aspects: accuracy and efficiency.

The parameters we used in this paper are selected manually now and we will search for a proper adaptive way to set the values of the parameters in the future.

### 4.1 Measure Accuracy of Segmentation

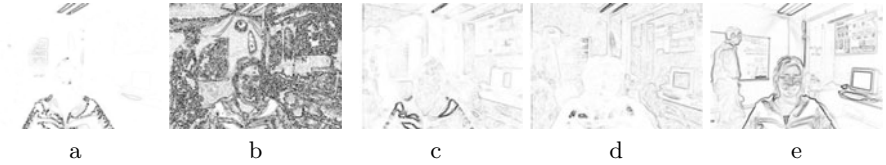
The accuracy of segmentation is evaluated by running the algorithm SPPGC on a stereo sequence of 20 frames, the ground truth of which is labeled manually. The pixels in the ground truth data is labeled by foreground, background and unknown. Error is measured as percentage of misclassified pixels, ignoring unknown pixels, which is used to mark the pixels along the boundary. For comparison, contrast, color and stereo alone algorithms are tested as well. These algorithms are simply obtained from SPPGC by keeping the contrast, color and stereo item alone in the cost function. The algorithm LGC described in [11] is also implemented and compared. As we can see in (Fig. 1), SPPGC and LGC have the similar performances, which are better than contrast, color and stereo alone algorithms in accuracy of segmentation. Thus, modeling the segmentation problem as a minimum cut problem of a planar graph with the weights of the edges given by a combination of contrast, color and stereo is reasonable. An example is given in (Fig. 2)-(Fig. 5).



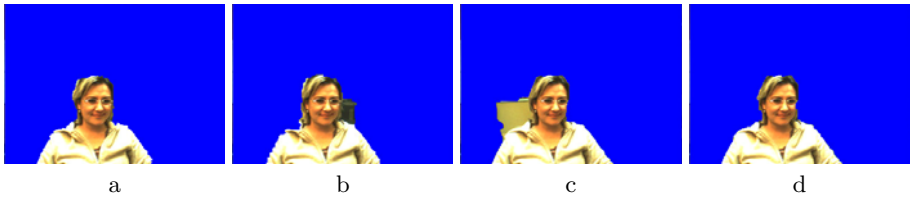
Fig. 2. Original left and right images



**Fig. 3.** Probability graphs for: (a) color of the foreground, (b) color of the background, (c) stereo of the foreground and (d) stereo of the background



**Fig. 4.** (a) and (b) are color energies of the foreground and background calculated by (16), (c) and (d) are stereo energies of the foreground and background calculated by (22), (e) is contrast energy calculated by (3)



**Fig. 5.** (a), (b) and (c) are the segmentation results obtained by using color, stereo and contrast alone, with (a) calculated by Fig. 4a and Fig. 4b; (b) by Fig. 4c and Fig. 4d, (c) by Fig. 4e; (d) is the result of SPPGC

## 4.2 Measure Efficiency of Segmentation

The algorithm SPPGC has been proved to have a running time of  $O(n \log n)$ , which is much faster than the general graph cut methods in the worst case theoretically. We now test SPPGC and the method of Boykov and Kolmogorov [5] (BK) with the same cost function on a set of images that have different resolutions to see the numerical performance of SPPGC. We use an image that consists of approximately 3 Megapixels and scale it down to different resolution images, which are used as the testing data with lower resolution. We can observe from (Fig. 6) that, two algorithms have performed similar on smaller images, while SPPGC has outperformed BK on larger images. As the resolution increases, the running times of the two methods increase at different rates, and the speed-up factor of SPPGC to BK becomes larger, which means that the spread of runtime of SPPGC is slower. This result shows that SPPGC is more suitable for handling high resolution images.

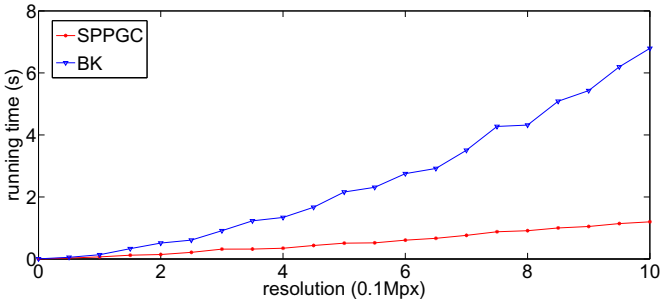


Fig. 6. Comparison of efficiency between BK and SPPGC

## 5 Conclusions

In this paper, we have modeled the segmentation problem as an energy minimization problem fusing stereo with contrast and color. Considering variety cues can enhance the quality of segmentation compared to stereo, contrast and color alone models. The energy minimization problem can then be converted into an equivalent maximum flow problem of a planar graph, which can be solved in  $O(n \log n)$  time using our SPPGC algorithm. SPPGC outperforms BK in terms of worst-case complexity, in terms of actual runtime in our experiment, and in terms of the observed spread of the runtime. The advantage of our solution makes it able to do an excellent job in the applications which require high accuracy and efficiency, especially in live background substitution. It is also well suited for applications with high resolution. In the near future, we will replace the current energy function with more complex and reasonable form and do more experiments to further improve the accuracy, efficiency and robustness of the algorithm.

## References

1. Acar, U.A., Blemloch, G.E., Harper, R., Vittes, J.L., Woo, S.L.M.: Dynamizing static algorithms, with applications to dynamic trees and history independence. In: Proc. 15th Ann. ACM-SIAM Symp. Discrete Algorithms, pp. 531–540 (2004)
2. Alstrup, S., Holm, J., De Lichtenberg, K., Thorup, M.: Maintaining information in fully dynamic trees with top trees. ACM Trans. Algorithms 1(2), 243–264 (2005)
3. Baker, S., Szeliski, R., Anandan, P.: A layered approach to stereo reconstruction. In: Proc. CVPR, Santa Barbara, pp. 434–441 (1998)
4. Boykov, Y., Jolly, M.-P.: Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In: Proc. Int. Conf. on Computer Vision (2001)
5. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE PAMI 26(9), 1124–1137 (2004)
6. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. on PAMI 23(11) (2001)

7. Criminisi, A., Shotton, J., Blake, A., Torr, P.H.S.: Gaze manipulation for one to one teleconferencing. In: Proc. ICCV (2003)
8. Erickson, J.: Maximum flows and parametric shortest paths in planar graphs. In: Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (2010)
9. Geiger, D., Ladendorf, B., Yuille, A.: Occlusions and binocular stereo. *Int. J. Computer Vision* 14, 211–226 (1995)
10. Jojic, N., Frey, B.: Learning flexible sprites in video layers. In: Proc. CVPR, Hawaii (2001)
11. Kolmogorov, V., Criminisi, A., Blake, A., Cross, G., Rother, C.: Probabilistic fusion of stereo with color and contrast for bilayer segmentation. *PAMI* 28(8) (2006)
12. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: Proc. ECCV, Copenhagen, Denmark (2002)
13. Mortensen, E.N., Barrett, W.A.: Intelligent Scissors for Image Composition. In: *Computer Graphics (SIGGRAPH 1995)*, Los Angeles, California, pp. 191–198 (1995)
14. Rother, C., Kolmogorov, V., Blake, A.: GrabCut—Interactive foreground extraction using iterated graph cuts. In: Proc. ACM Siggraph (2004)
15. Rowe, S.M., Blake, A.: Statistical mosaics for tracking. *J. Image and Vision Computing* 14, 549–564 (1996)
16. Sleator, D.D., Tarjan, R.E.: A data structure for dynamic trees. In: *JCSS*, pp. 362–391 (1981)
17. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proc. CVPR, pp. 246–252 (1999)
18. Tarjan, R.E., Werneck, R.F.: Self-adjusting top trees. In: Proc. 16th Ann. ACM-SIAM Symp. Discrete Algorithms, pp. 813–822 (2005)
19. Torr, P.H.S., Szeliski, R., Anandan, P.: An integrated Bayesian approach to layer extraction from image sequences. *PAMI* (2001)

# Color Information Presentation for Color Vision Defective by Using a Projector Camera System

Atsushi Yamashita, Rie Miyaki, and Toru Kaneko

Shizuoka University, Japan

**Abstract.** There are individual differences in color vision. It is difficult for a person with defective cones in the retina to recognize the difference of specific colors. We propose a presentation method of color information by using a projector camera system. The system projects border lines or color names on real object surfaces when they have specific color combinations. Effectiveness of the proposed method is verified through experiments.

## 1 Introduction

In this paper, we propose a color information presentation system for color vision defective by using a projector camera system.

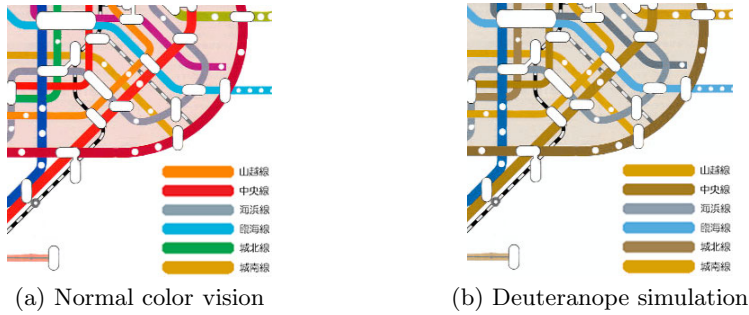
There are individual differences in the color vision. Human eye has cone cells that can sense colors. Cone cells are divided into three types by a difference of the spectral sensitivity; the long-wavelength-sensitive (L) cone, the middle-wavelength-sensitive (M) cone, and the short-wavelength-sensitive (S) cone. The individual difference in color vision comes from the lack or low sensitivity of three cone cells. The condition of the cone of all types without loss is called normal, the condition with a loss of the L cone is called protanopia, the condition with a loss of the M cone is called deuteranopia, the condition with a loss of the S cone is called tritanopia, and the condition of the cone of two kinds with loss is called cone monochromatism, respectively. It is difficult for a person with defective cones in the retina to recognize the difference of specific colors. For example, a person who lacks L cone has low sensitivity in red color. In this case, he or she may feel inconvenience in everyday life.

Figure 1 shows a route map in which the difference of colors indicates different routes. Figures 1(a) and (b) show a normal color vision and a color simulation result that a color vision defective (deuteranopia) senses 1, respectively. In Fig. 1(b), green and orange routes are difficult to distinguish with each other. Therefore, support system for visually impaired people is very significant.

In color universal design and color barrier-free approaches, color combinations that any color vision person is easy to distinguish should be used. However, such concepts do not spread in present day. Barriers are also left in several environments and situations such as route maps, signboards, posters and so on.

---

<sup>1</sup> In this paper, color simulation results are generated by using “UDing simulator” (Toyo Ink Mfg. Co., Ltd.).



**Fig. 1.** Example of individual difference in color vision

Therefore, there are studies of the use of image processing for supporting the visually impaired people [1, 2, 3], and especially for the color vision defective [4, 5, 6]. The main purpose of these studies for the color vision defective is constructing color conversion algorithms, and there are few studies that deal with real applications such as web page browsing [7] and a head mount display (HMD) [8]. A color modification method for web pages [7] does not treat with real objects. An HMD system [8] detects colors that are difficult to distinguish in acquired images by using a camera, and then displays boundary lines of color edges for users by using the HMD. However, registration between real objects and images that are displayed in HMD is not considered. In other words, color information is only in computer in these studies.

As to the presentation of color information, one of the most fruitful merits of augmented reality (AR) and mixed reality (MR) technologies is that we can recognize displayed color information in the same manner as real objects.

In some situation, an HMD is enough for a user to get visual information with using AR technology like AR tool kit [9]. The advantage of an HMD is that it is unaffected by environment light and does not prevent multi-user situations because it does not change environment [10]. However, an HMD is not suitable for a long time use because it gives a user a feeling of constraint. On the other hand, the method using not an HMD but a projector is also proposed [11]. A projector can easily add information over real objects. A projector is suitable for a long time use compared with an HMD. Therefore, we consider that the place that the system can use is not limited for living environments.

## 2 Purpose and Outline of Color Information Presentation

In this paper, we propose a presentation method of color information by using a projector camera system. The proposed system is adaptive to the individual and the place that the system works is not limited. The camera acquires color information and the projector presents color information (Fig. 2).

The system projects not only border lines [8], but also draws color names or overlays (paints) alternate colors on real object surfaces when they have specific

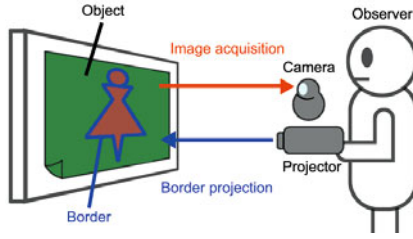


Fig. 2. Proposed projector camera system

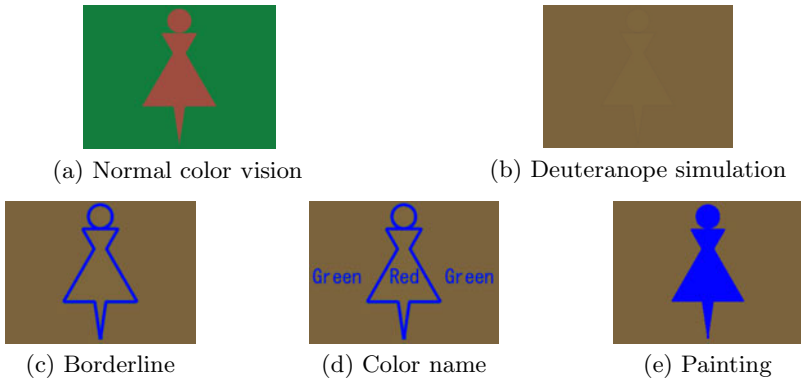


Fig. 3. Color combination difficult to distinguish and color information presentation

color combinations. This is a practical AR/MR application trying to improve a user’s color perception.

In our system, the three dimensional (3D) relationship between the projector and the camera is fixed. However, in mobile applications, the 3D relationship between the projector camera system and objects changes. Therefore, registration of projected images and real objects is realized by using projected markers.

In Fig. 3(a), a red picture is drawn on a green background. A deuteranope can hardly distinguish the difference of colors because the difference between green and red is not recognized like Fig. 3(b). Therefore, a color camera detects image regions that may appear ambiguous to a viewer. Then the projector overlays color information with three modes; by displaying boundary lines (Fig. 3(c)), by displaying color names (Fig. 3(d)), and by painting in another color (Fig. 3(e)).

The processing flow is shown in Fig. 4. At first, the system projects markers on real objects by using the projector. The camera acquires image and detect color(s) or color combination(s) that are difficult to distinguish. If there are color(s) or color combination(s) that are difficult to distinguish, the system generates an image that is projected on real objects. In this step, a projected image is registered with real objects by using projected markers. The system repeats the above procedure. If the 3D relationship between the system and real objects changes, the system detects motion and reprojects a new image.

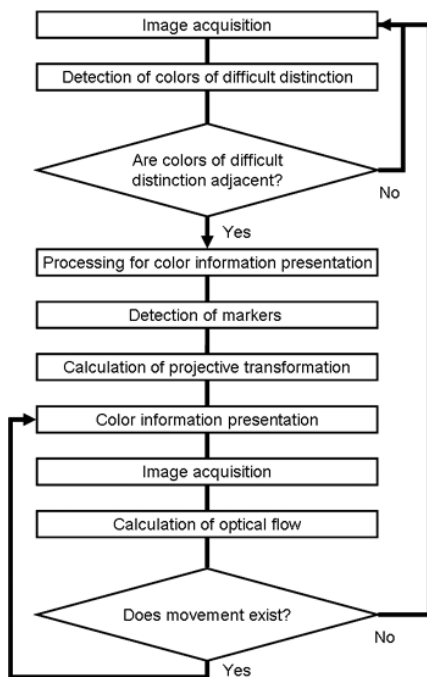


Fig. 4. Processing flow

### 3 Color Image Processing

The system examines color combinations that are difficult for the user to distinguish. At first, the system judges whether the acquired image has colors of difficult distinction.

In our system, color combination that the normal color vision feels similar is not extracted as difficult distinction color. For example, dark green and green are not a color combination of difficult distinction in Fig. 5(a), because the normal color vision people feels they are similar. On the other hand, dark green and red in Fig. 5(b) is judged as difficult distinction, because the color vision defective cannot judge them although they are different color.

In order to examine colors of difficult distinction, color confusion lines are used [12,13]. A color confusion line is a straight line radiated from the center of confusion (copunctal point) on the CIE1931 x-y chromaticity diagram (Fig. 6). The center of confusion is given by the type of the color vision [12]. Colors on a color confusion line are difficult to distinguish. Our method calculates a color confusion line that is linked from the center of confusion to a color of a pixel in an acquired image.

When the angle which two colors of color confusion lines make is small and two color points are away on the x-y chromaticity diagram, the system determines that two colors are colors of difficult distinction.



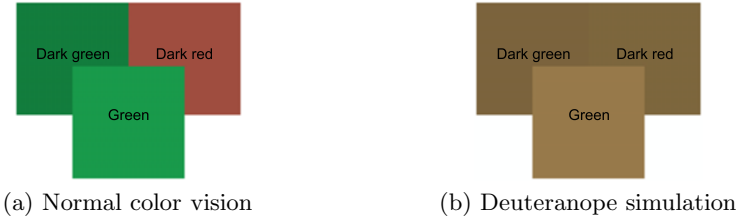


Fig. 5. Example of color combination

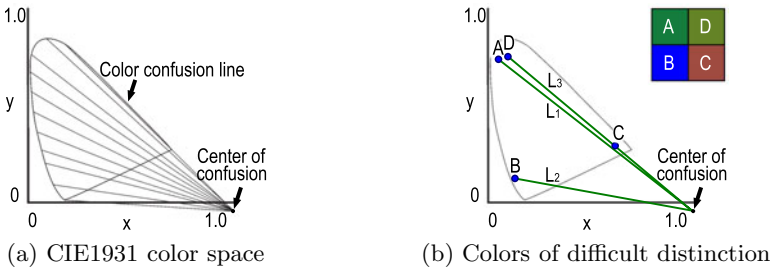


Fig. 6. Color confusion line

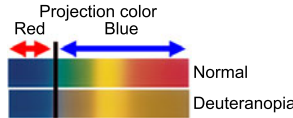
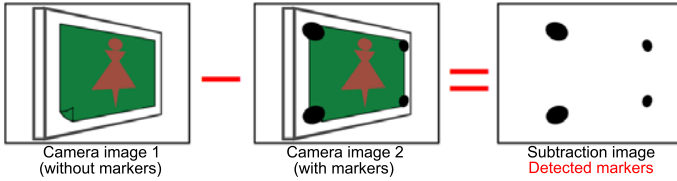


Fig. 7. Projection color

For example, when the color of the pixel in the acquired image is given as point A (dark green) of Fig. 6(b), the color confusion line becomes line  $L_1$ . In the same way, color confusion lines of B (blue), C (dark red), and D (green) are  $L_2$ ,  $L_3$ , and  $L_4$ , respectively. The combination of color A and color B is not judged as difficult distinction color, because the angle between line  $L_1$  and line  $L_2$  is large. The angle between line  $L_1$  and line  $L_3$  is small, and the distance between point A and point D is small. Therefore, the combination of color A and color D is not difficult distinction color. On the other hand, the distance between point A and point C is large. Therefore, the combination of color A and color C is difficult distinction colors.

The color that is projected is decided by considering the user’s color vision characteristics<sup>2</sup> (Fig. 7).

<sup>2</sup> The characteristics of the scene are not considered. The color of the surface affects the color of projected lights. In future work, the color for painting should be decided more carefully to show the information with appropriate color to the user.



**Fig. 8.** Marker detection by subtraction

The system provides three ways of color information presentation.

- (1) **boundary line presentation:** the system projects boundary lines on the place of the real object corresponding to pixels judged to be adjacent.
- (2) **color names presentation:** the system detects the area of each color and determines the presentation point of color names. The system projects color names on the presentation point of the real object.
- (3) **painting presentation:** the system projects colors on places of the real object corresponding to detected areas.

## 4 Registration of Projected Image and Real Object

In our proposed method, objects are assumed to be planar such as a signboard and a bulletin board. The system performs registration of a projected image and real objects by using projected markers.

First, the system projects markers on real objects. Next, the system takes an image of real objects with using a camera and detects markers in an acquired image. The system calculates a projective transformation matrix from detected markers. Finally, the system transforms a projection image by using calculated projective transformation matrix.

Our proposed method assumes that a user holds the system and uses it. Relative position and posture between the system and objects may always change. Therefore, the system projects and detects markers to make the registration of the projected image and the real objects every time at the image acquisition.

Markers are detected by using subtraction (Fig. 8). The system subtracts an acquired image without projecting markers from that with projecting markers, and detects marker positions. This processing is repeated and the marker positions are updated.

The system calculates the homography matrix  $\mathbf{H}$  from relations between measured marker positions in the acquired image and marker positions in the projected image. The projection image is transformed by using  $\mathbf{H}$  for registration of projected images and real objects.

Our system detects whether there is a movement between the system and real objects by using optical flow. If there is a movement,  $\mathbf{H}$  is recalculated.

## 5 Experiment

The experimental device consists of a projector, a camera (logitech web camera), and a computer (CPU: Intel Core 2 Duo 3.0GHz, Memory: 4GB) (Fig. 9). The

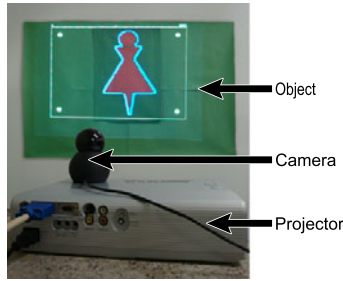


Fig. 9. Experimental equipment

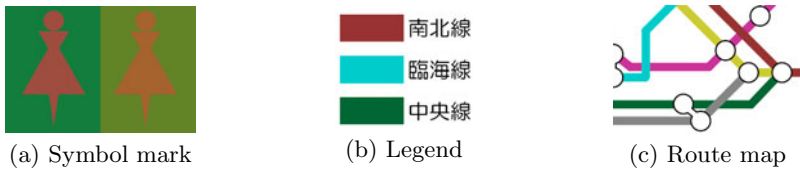


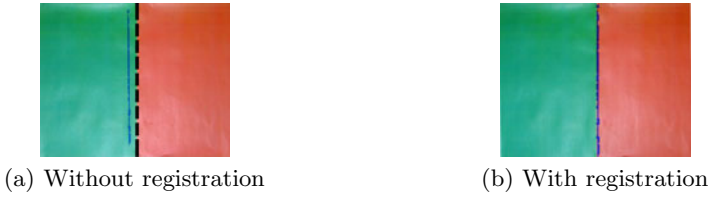
Fig. 10. Objects in experiment

experiment was performed in a room. The resolutions of acquired images were  $640 \times 480$  pixel. A threshold of an angle between two color confusion lines was decided in advance by trial and error.

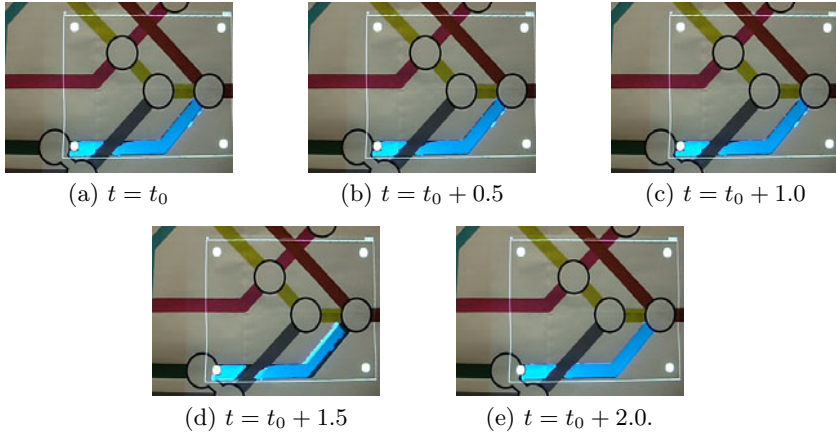
Figure 10 shows planar objects to use for experiment. Figure 10(a) shows printed emblems. In Fig. 10(a) from the left, the dark red emblem is drawn on a dark green background, the orange emblem is drawn on a yellow green background, and the red emblem is drawn on a white background. Figure 10(b) shows a legend of a map. In Fig. 10(b) from the top, a dark red quadrangle is drawn, a light blue quadrangle is drawn, and a dark green quadrangle is drawn. Figure 10(c) shows printed a route map. In Fig. 10(c), a red and a green route are drawn. Those objects were used for presentation experiment of boundary lines, color names, and painting, respectively.

Figure 11 shows a registration result of a projection image and a real object. In these figures, blue lines are projected boundary lines, and black dotted lines are boundaries of green and red regions (ground truth of boundary lines). Without registration (Fig. 11(a)), blue and dotted lines do not coincide with each other. On the other hand, they coincide with each other with registration (Fig. 11(b)).

Figure 12 shows the border projection result while the position of the projector camera system was changing. The computation time was about 4fps on an average. The system and the object were not moving between Fig. 12(a) and Fig. 12(c), while the relationship between the system and the object changes between Fig. 12(c) and Fig. 12(e) because the system moved. The movement was not detected from Fig. 12(a) to Fig. 12(c), and the system continued to project the same image and stable projection of color information was realized in Figs. 12(b) and (c). On the other hand, the movement was detected between



**Fig. 11.** Registration of projection image and real object



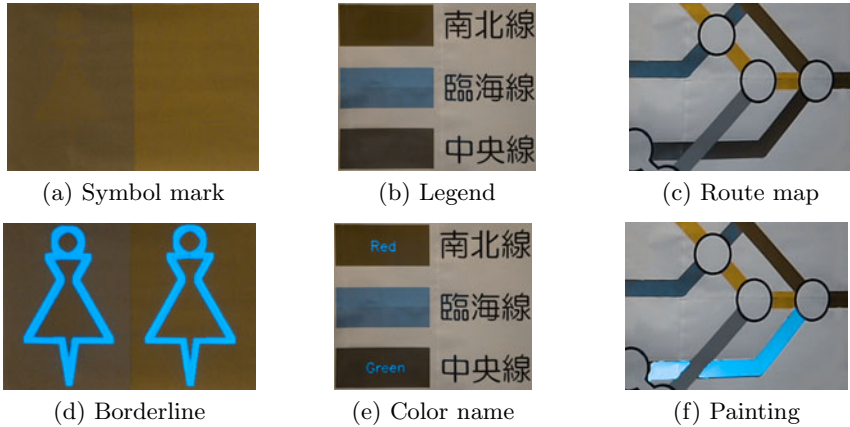
**Fig. 12.** Result of continuous border projection

Fig. 12(c) and Fig. 12(d), the system executed registration from the information of Fig. 12(d), and projected another image in Fig. 12(e). In Fig. 12(e), gap between the projected image and the real object was resolved. In this way, the system could successfully detect the motion of the system, and the projected images that coincided with the targets.

Figure 13 shows a result of the boundary line presentation, the color name presentation, and the painting presentation, respectively.

In Fig. 13(a), the angle between the dark red of color confusion line and the dark green of color confusion line is small with less than 1 degree. The distance between two colors is small. Therefore, boundary lines are projected on the border between the dark red and the dark green. Similarly, boundary lines are projected on the border between the orange and the yellow green. On the other hand, the angle between the dark green of color confusion line and the yellow green of color confusion line is as small as 2 degree. However, the distance between the two colors is not small. Therefore, boundary lines are not projected on the border between dark green and the yellow green.

In Fig. 13(d), boundary lines are projected on the border between the dark red and the dark green, and between the orange and the yellow green. Therefore, the method makes it easy to see emblems that are hard to see in Fig. 13(a).



**Fig. 13.** Results of projection

In Fig. 13(e), letters of “Red” are projected on a red part and letters of “Green” are projected on a green part. Whereas it is hard to distinguish the top quadrangle and the bottom quadrangle in Fig. 13(b), the method makes it easy to distinguish the top quadrangle and the bottom quadrangle in Fig. 13(e).

In Fig. 13(f), green line is painted with blue. Whereas it is hard to distinguish the red line and the green line in Fig. 13(c), the method makes it easy to distinguish the red line and the green line in Fig. 13(f).

The effectiveness of the method is shown by these results of boundary lines, color names, and painting presentation.

## 6 Conclusion

We propose a presentation method of color information with a projector camera system based on registration of projection image and real object using projected markers. We confirmed the effectiveness of the method by experimental results. The solution in this paper is simple yet effective. A color camera detects image regions that may appear ambiguous to a viewer. The projector then overlays lines, regions, and text to assist the viewer.

In future work, color calibration of the camera should be done in adapting to lighting condition change. Color information presentation must be considered when an object has a color gradation.

This system can be developed for tourists visiting other countries. Reading a subway map in Tokyo (Japan) is very difficult for non Japanese speaking/reading individuals, because the subway map in Tokyo is very complicated. There is a potential in our work for such applications, *e.g.* character translation which deserves to be explored.

## Acknowledgment

This research was partially supported by KAKENHI, Grant-in-Aid for Scientific Research (C), 21500164, and JGC-S Scholarship Foundation.

## References

1. Molton, N., Se, S., Lee, D., Probert, P., Brady, M.: Robotic Sensing for the Guidance of the Visually Impaired. In: Proceedings of the International Conference on Field and Service Robotics (FSR 1997), pp. 236–243 (1997)
2. Jacques, D.J., Rodrigo, R., McIsaac, K.A., Samarabandu, J.: An Object Tracking and Visual Servoing System for the Visually Impaired. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005), pp. 3510–3515 (2005)
3. Cardin, S., Thalmann, D., Vexo, F.: Wearable System for Mobility Improvement of Visually Impaired People. *Visual Computer Journal* 23, 109–118 (2006)
4. Ichikawa, M., Tanaka, K., Kondo, S., Hiroshima, K., Ichikawa, K., Tanabe, S., Fukami, K.: Preliminary Study on Color Modification for Still Images to Realize Barrier-Free Color Vision. In: Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics (SMC 2004), vol. 1, pp. 36–41 (2004)
5. Rasche, K., Geist, R., Westall, J.: Detail Preserving Reproduction of Color Images for Monochromats and Dichromats. *IEEE Computer Graphics and Application* 25(3), 22–30 (2005)
6. Yokota, S., Hashimoto, H., Sasaki, A., Takeda, D., Ohyama, Y.: Supporting Technologies for Weak-Eyesight Persons - Automatic Color Conversion and Training Vision. In: Proceedings of the SICE Annual Conference 2007, pp. 3064–3068 (2007)
7. Ichikawa, M., Tanaka, K., Kondo, S., Hiroshima, K., Ichikawa, K., Tanabe, S., Fukami, K.: Web-Page Color Modification for Barrier-Free Color Vision with Genetic Algorithm. *LNCS*, vol. 2724, pp. 2134–2146. Springer, Heidelberg (2003)
8. Tsutsui, T., Aoki, K.: A Wearable Type Color Barrier Free System. In: Proceedings of Workshop on Dynamic Image Processing for Real Application 2008 (DIA 2008), pp. 282–285 (2008) (in Japanese)
9. Kato, H., Billinghurst, M.: Marker Tracking and HMD Calibration for a video-based Augmented Reality Conferencing System. In: Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR 1999), pp. 85–94 (1993)
10. Karitsuka, T., Sato, K.: A Wearable Mixed Reality with an On-board Projector. In: Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR 2003), pp. 321–322 (2003)
11. Amano, T., Kato, H.: Shape Disparity Inspection of The Textured Object and Its Notification by Overlay Projection. In: Shumaker, R. (ed.) *VMR 2009*. LNCS, vol. 5622, pp. 405–412. Springer, Heidelberg (2009)
12. Grand, Y.L.: *Light, Colour and Vision*, 2nd revised edn. Chapman and Hall, Boca Raton (1968)
13. Fry, G.A.: Confusion Lines of Dichromats. *Color Research & Application* 17(6), 379–383 (1992)

# Simulating Artworks through Filter Blending

Crystal Valente and Reinhard Klette

*.enpeda.* Group, The University of Auckland, New Zealand

**Abstract.** This paper looks at a method of blending different artistic filters together to create a range of artistic effects. Instead of using a single painterly rendering technique, the methods used in three different filters can be blended together in a user defined way. The filters are arranged in a triangular structure where the user defines their chosen painting style by choosing a point in the triangle. This allows users to effectively create their own painting style and experiment with a range of different artistic effects.

The field of painterly rendering looks at methods of creating a simulated artwork from a source photograph. Artistic filters are inspired by methods used by real artists. Most of these filters look at a particular aspect of a real painting process and design an algorithm to simulate this process. Our contribution takes inspiration, not from one aspect of a painting process, but from the variety that is found in painting methods. We look at ways to blend aspects of different filters together to create a unique artistic effect that is chosen by the user.

We use three different artistic filters in this paper. The first is Aaron Hertzmann's painting algorithm described in [2]; it uses layers of curved brush strokes. Next is a pointillistic filter roughly based on [5] that attempts to simulate the works of Georges Seurat. The last is Papari and Petkov's method for creating impressionist paintings using Glass patterns as described in [3].

Sections [1], [2] and [3] look at the methods used in each of the different artistic filters. Section [4] described our methods for blending these filters together. Section [5] shows the results of this blending technique and Section [6] discusses our conclusions and ideas for future work.

## 1 Curved Brush Strokes

This section describes an algorithm presented by Aaron Hertzmann that simulates a layered painting style with brush strokes of varying sizes. Hertzmann's algorithm uses a layered approach; it starts with a rough approximation of the image and builds up more detail at each layer with steadily smaller brush strokes. These are curved strokes that follow object contours, as favored by many artists. The results of this process can be seen in Fig. [1].

**Layering of Strokes.** We start with a blank canvas and a reference image. The algorithm takes an array of difference brush sizes as a parameter. Each brush size defines a layer in the painting; starting with the largest brush and working down to smallest, defined by a minimum brush size  $b_{\min}$  and maximum size  $b_{\max}$  and equidistant values in-between.



**Fig. 1.** *Left:* Scene from Fiji. Original photograph by Marian Arnold. *Right:* The results of the curved strokes algorithm.

For each brush size we divide the image into a grid where the size of each cell is proportional to the current brush size. At each grid point we determine the total error of each pixel contained in this grid area by comparing the color of the canvas at this point to the color of the reference image. If the total error is above a threshold  $T$ , we add a new stroke to the canvas at the point in the neighborhood with the maximum error. Threshold  $T$  determines how closely the finished painting approximates the reference image and therefore how loose our painting style will be. We do not want the grid structure to make our strokes look too uniform, so strokes need to be painted in a random order. This layering process can be applied to strokes of any shape and orientation.

**Curved Brush Strokes.** We describe the creation of the curved brush strokes. Hertzmann’s algorithm for placing strokes is as follows. The process takes as input a brush radius  $R$ , a starting point  $(x_0, y_0)$ , and a maximum length  $L$ . We start at the point  $(x_0, y_0)$  and find the color value  $C$  at this point in the reference image.  $C$  gives us the color of the stroke that remains constant. We paint a circle of radius  $R$  and color  $C$  at point  $(x_0, y_0)$  to the image canvas. The next point in the stroke is computed by finding the normal to gradient at this point. We find the direction of the gradient  $\theta$  by determining the convolution of the Sobel operator with the luminance of the reference image in the x and y directions. The next point in our spline  $(x_1, y_1)$  is placed distance  $R$  from point  $(x_0, y_0)$  in direction  $\theta + \frac{\pi}{2}$ . This point is also a circle with radius  $R$  and color  $C$ . This process is repeated until all the control points in the stroke have been painted to the canvas. The process terminates when (a) the user defined maximum stroke length  $L$  is reached, or (b) the color of the stroke differs from the color of the reference image at the last control point more than it differs from the image canvas at that point.

## 2 Pointillism

The next filter used in our application implements a pointillistic style based on the works of Georges Seurat. Seurat was greatly influenced by the color theories of M. E. Chevreul and this is reflected in his work. Our filter incorporate several





**Fig. 2.** *Left:* Scene from Wharariki Beach, Golden Bay, NZ. Original photograph by Jenna Bowden. *Right:* The results of the pointillistic filter.

color distortions that attempt to emulate the color effects used by Seurat. This filter is roughly based on the work done by Yang and Yang in [5].

The problem of emulating Seurat’s pointillistic painting style is broken up into three layers. Each layer distributes points in a different way and has its own color distortions. A result of this process can be seen in Fig. 2.

**Color.** We implement three different color distortions based on our observations of Seurat’s colors. We refer to these as color restriction, saturation distortion, and divisionism. Color restriction restricts the image to a specific palette of colors. This palette is chosen to reflect the colors that Seurat is thought to have used. Chevreul’s color theory is based on hue, so this is the component that we pay the most attention to. When a color is initially chosen at each layer, the hue is set to closest color in the palette.

Saturation distortion implements Seurat’s bias toward bright colors by increasing the saturation when certain conditions are met. If a color has low brightness and low saturation, then its saturation is increased. This is done by breaking the colors up into sections depending on their brightness value. Each pixel has a probability of having its saturation changed that is determined by a random value. Many points still retain their original saturation no matter what the color value is at this point. For each point to be distorted, we determine which section the color  $c$  falls within based on its brightness value and change the saturation to  $S(c)$  as follows:

$$S(c) = \begin{cases} s & : s \geq M_i(c) \\ M_i(c) & : s < M_i(c) \end{cases} \quad (1)$$

where  $M_i(c) = 10(b_i - v)(b_i - b_{i-1})(m_{i-1} - m_i) + m_i$  defines the minimum saturation required for color  $c$ , where  $c$  falls into section  $i$  based on its brightness value  $v$ . The boundary of section  $i$  is defined by  $b_i$ , where  $c$  falls into section  $i$  if  $b_{i-1} < v \leq b_i$ . The basic minimum saturation for the section is  $m_i$ . This value is altered depending on how close  $v$  is to the boundary; the base case is  $M_1(c) = m_1$ . This method enhances the saturation of muddy colors while preserving the saturation of colors that are close to white or already have a high saturation value. This gives smooth transitions across section boundaries.

Finally, to include Seurat's concept of divisionism, once the values for a pixel are set, the hue can be changed based on a random variable to one of its nearest neighbors in the color wheel. This random paint method has a 0.5 chance of picking the color itself, and a 0.25 chance of picking the closest neighbor on each side of this color in the color wheel.

**Layering.** The first layer is the background layer; it is designed to fill up the image and set up our base colors. To fully color our canvas, the canvas is initially set to the original image. This is to avoid having white bits of canvas show through in the final image, what is acceptable in lighter areas but visually distracting in darker parts of the image. Once the original image has been copied to the canvas, we determine the placement of the points that are painted on this canvas. The distribution of dots is achieved using *Poisson disks*; for a point radius  $R_1$ , the minimum distance between sampled points is set to  $d = 2 \cdot R_1$ .

For each point chosen by the Poisson algorithm, a circle of radius  $R_1$  is painted with its center at this location. The color of this circle is the color of the original image at this point which is then color restricted. No further color distortion is performed for this layer.

For the middle layer, stroke placement is roughly based on the algorithm as described in Section 1 but instead of basing our error measure on color difference, we base our stroke placement on the difference in color intensity. The color of each point is again determined by the color of the original image at this point. For this layer however, we use color restriction, saturation distortion, and divisionism.

If only these first two layers are used, the filter has a tendency to swamp smaller details with points from the background, so we can also use a final layer to perform some edge detection to bring these details back into the picture.

### 3 Impressionism

Glass patterns emulate impressionism; they are based on a geometric transformation. First we create a vector field and a randomized image. Once these two components have been generated, the actual Glass pattern is created. The transformation for creating the Glass pattern is applied to the original image, see Papari and Petkov in [3] and Fig. 3 for a result.

**Vector field.** The vector field determines what kind of movement the brush strokes have. The impressionist paintings we are attempting to mimic contain swirling patterns based around the contours of the objects within in the scene. We create a vector field  $\mathbf{v}(\mathbf{r})$  that approximates this type of movement where  $\mathbf{r} = (x, y)$ . We start by computing the convolution  $I_\sigma = I \star \Delta_{x,y} G_\sigma$  of the original image with the gradient of the Gauss function, where  $I$  is the original image after smoothing and  $G_\sigma$  is the Gauss function with standard deviation  $\sigma$ . For image smoothing we used a simple median filter.

The area sampled over needs to be quite large in order to take into account some image data irregularity which might be "far away" from the current pixel. We have used a kernel size of  $31 \times 31$  for the convolution but have only sampled every 5th pixel within this area. This covers a fairly large area of the image



**Fig. 3.** *Left:* Scene from Oslo, Norway. Original photograph by Angela Palmer. *Right:* The results of the Glass patterns filter.

at each pixel without increasing the processing time too much. This gives us  $I_\sigma = [I_{\sigma x}, I_{\sigma y}]^T$  which is the gradient of each color channel in  $x$  and  $y$  direction.

The next step is to compute  $\theta_\sigma(x, y)$ , the angle of the "color gradient of the nearest edge". Let  $k$  be the number of color channels. At a pixel, calculate

$$E = \left( \sum_{i=1}^k I_{\sigma x}^{(i)} \right)^2, \quad F = \sum_{i=1}^k I_{\sigma x}^{(i)} \sum_{i=0}^k I_{\sigma y}^{(i)}, \quad G = \left( \sum_{i=1}^k I_{\sigma y}^{(i)} \right)^2 \quad (2)$$

Angle  $\theta_\sigma(x, y)$  is defined by the direction of the eigenvector associated with the maximum eigenvalue of the matrix

$$K_\sigma(x, y) = \begin{bmatrix} E & F \\ F & G \end{bmatrix} \quad (3)$$

There are two possible values of  $\theta_\sigma(x, y)$ :

$$\theta_+ = \frac{1}{2} \arctan \left( \frac{2F}{E - G} \right), \quad \theta_- = \theta_+ \pm \frac{\pi}{2} \quad (4)$$

We choose that value for  $\theta_\sigma$  where

$$M(\theta_\sigma) = \frac{1}{2} (E + G + \cos 2\theta_\sigma (E - G) + 2F \sin 2\theta_\sigma) \quad (5)$$

is maximum. This value expresses the "strength" of the gradient in direction  $\theta_\sigma$ . The vectors in the vector field  $\mathbf{v}(x, y)$  are of a fixed length  $a$ , which can be set by the user, and make a constant angle  $\theta_0$  with the color gradient  $\theta_\sigma$ . The vector field is defined as follows:

$$\mathbf{v}(x, y) = a [\cos(\theta_\sigma(x, y) + \theta_0), \sin(\theta_\sigma(x, y) + \theta_0)]^T \quad (6)$$

If the eigenvalues of  $K_\sigma(x, y)$  are equal then the value of  $\theta_\sigma$  are undefined; in this case we set  $\mathbf{v}(x, y) = \mathbf{0}$ .

**Random Image.** Our random image  $z(x, y)$  is created by generating white Gaussian noise which is then smoothed using a simple Gaussian filter. The white

noise is generated based on the Central Limit Theorem (i.e., the noise generated is not true white Gaussian noise if the random number generator used is not completely random).

**Continuous Glass Pattern.** Next we look at how to create a continuous Glass pattern given a random image  $z(\mathbf{r})$  and a vector field  $\mathbf{v}(\mathbf{r})$  where  $\mathbf{r} = (x, y)$ . For brevity's sake we focus on the most important formulas here. For a full mathematical discussion refer to [3]. As in [3] we first consider the differential equation

$$\frac{d\mathbf{r}}{dt} = \mathbf{v}(\mathbf{r}) \quad (7)$$

The solution, the trajectory  $\Phi_{\mathbf{v}}$ , is a map from  $\mathbb{R}^2$  to  $\mathbb{R}^2$  with  $\Phi_{\mathbf{v}}(\mathbf{r}, 0) = \mathbf{r}$ . In other words,  $\Phi_{\mathbf{v}}(\mathbf{r}, t)$  describes an arc from the current pixel  $\mathbf{r}$  to some new location. Each location along the arc is defined by the value of  $t$ .

For a continuous Glass pattern, instead of a point set we use the random image  $z(\mathbf{r})$ . This extension means that instead of taking the maximum over the binary case, we take the maximum of the gray values at each location. A continuous Glass pattern is a new image and defined as follows:

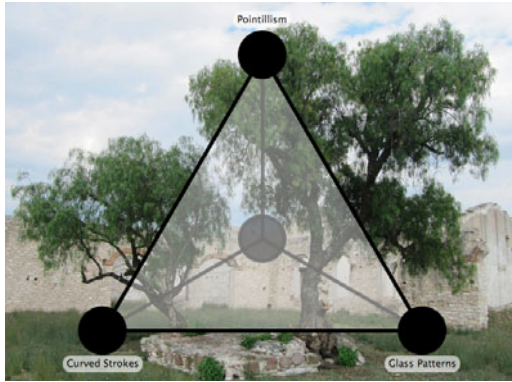
$$G_{\mathbf{v}}(\mathbf{r}) = \max_{t \in [0,1]} \{z[\Phi_{\mathbf{v}}(\mathbf{r}, t)]\} \quad (8)$$

At each point in this image, we take the pixel value at  $\mathbf{r}$  to be the maximum of the pixel values in  $z(\mathbf{r})$  that lie over the arc defined by  $\Phi_{\mathbf{v}}(\mathbf{r}, t)$  for every  $t \in [0, 1]$ . The first step in computing this is to integrate Eq. (7). This is done numerically using the Euler method; it gives a discrete approximation of the arc  $\Phi_{\mathbf{v}}(\mathbf{r}, t)$ . We then take the point  $\mathbf{p}$  in this arc where  $z(\mathbf{r})$  is at its maximum. The final result is a point set  $\mathbf{p}(\mathbf{r})$  which defines our continuous Glass pattern.

**Translating to the Image.** At this stage, for each point  $\mathbf{r}$  in the image we have an associated point  $\mathbf{p}$ . This describes a geometric transformation between the two points. Papari and Petkov describe two different ways that the structure of this continuous Glass pattern can be translated to the original image. We use their method of using the Glass pattern to translate the pixels of the original image. We have the original image  $I$  and the transformed image  $T$ . At each point  $\mathbf{r} \in I$  we find the point  $\mathbf{p}$  that is mapped to by the Glass pattern at  $\mathbf{r}$ . We record the pixel value of  $I(\mathbf{p})$  and set this to be the pixel value of  $T(\mathbf{p})$ . This process transfers the geometric structure of the continuous Glass pattern to the image itself. Areas with flat color remain the same. We can get around this by adding some noise to the original image before we process it.

## 4 Filter Blending

Our contribution to the field of painterly rendering is to combine the filters that we have examined (and partially modified) together to create a range of artistic effects. The filters that are used to create an image and the strength of these filters can be tailored to suite the subject of the image and the artistic intention of the user. For each filter combination, the strength of each of the filters  $f_1$  and  $f_2$  is determined by user defined influence parameters  $I(f_1)$  and  $I(f_2)$  where



**Fig. 4.** Screenshot of the user interface of the filter blending application. Scene from Mineral de Pozos, Mexico. Original photograph by Reinhard Klette.

$I(f_1) = 1 - I(f_2)$  and  $I(f_1), I(f_2) \in [0, 1]$ . The method used to produce the filter combination  $C(f_1, f_2)$  is different depending on the filters  $f_1$  and  $f_2$ .

**User Interface.** Our application has a triangular interface where each corner of the triangle represents a different filter and the center of the triangle represents the original image; for a screenshot of this concept see Fig. 4. The style we want to apply to our image is chosen by clicking somewhere within this triangle. Points along the edges of the triangle have the strongest filters.

**Curved Brush Strokes and Pointillism.** The curved brush strokes and pointillistic filters go for quite a different look but a lot of their underlying concepts are the same. The differences are the way that a new stroke position is determined and the color and shape of the stroke.

Our combined filter uses a three-layered approach. For our stroke radius, we take a stroke size between the strokes sizes of the two filters at each layer. To give this value a random appearance, we determine this size by getting a normally distributed random number  $z$  where the mean and standard deviation vary according to the influence parameter. For brush sizes  $b_p$  and  $b_c$  at the current layer where  $z$  is generated using mean  $m$  and standard deviation  $\sigma$ , we determine our final stroke radius  $b_{final}$  by  $b_{final} = b_p + z(b_c - b_p)$ .

For determining the position of the brush strokes at each layer, the image is divided into a grid of size  $b_{final}$ . At each grid point we determine two positions  $p_p = (x_p, y_p)$  and  $p_c = (x_c, y_c)$  which approximate the points chosen by the pointillistic and curved strokes filter respectively. Our final point  $p_{final}$  is a point between  $p_p$  and  $p_c$  where the influence parameters determine how close  $p_{final}$  is to each of the points.

The maximum stroke length  $l_{final}$  is determined in a similar way to the stroke radius. Given a user defined stroke length  $l_{max}$ ,  $l_{final}$  equals  $l_{final} = 1 + z(l_{max} - 1)$  where  $z$  is again a normally distributed random number. The amount of color distortion also varies according to the influence parameters.

**Curved Brush Strokes and Glass Patterns.** Those two filters have quite different approaches to the way they alter the image so there is no obvious scale

between them like with the curved brush strokes and pointillism. Instead we find a way to mix the concepts of the two filters together.

For the point halfway between the two filters we follow the usual method of the curved brush strokes filter, but instead of strokes following the normal of the image gradient, our brush strokes follow the vector field  $\mathbf{v}$  as defined in Section 3. As the image gets closer to curved strokes the filter gradually stops following this vector field and instead reverts to following the normal of the image gradient. We calculate the results of both methods for determining a new stroke control point and take the appropriate point between them based on the influence parameters. For a normal of the image gradient  $\mathbf{g} = (x_g, y_g)$  and a value of the vector field  $\mathbf{v} = (x_v, y_v)$ , the next control point is placed at  $\mathbf{p} = (x, y)$  where the distance of  $\mathbf{p}$  from each point is determined by the influence parameters. We also alter the maximum length  $l$  of the brush strokes based on the Glass patterns length parameter  $a$ . As the filter gets closer to impressionism,  $l$  tends toward  $a$ .

As the filter gets closer to impressionism, we start to decrease the influence of the curved brush strokes filter. This is done by gradually reducing the strokes radius parameters  $R_{max}$  and  $R_{min}$ , and the threshold  $T$  in proportion to the influence parameters. We also add noise to the reference image as the influence of impressionism increases to make the impressionist whirls more visible. We generate a small random number  $c$  as detailed in Section 3. In this case however we alter  $c$  in proportion to the impressionist influence parameter. When  $R_{min} \leq 1$  we discard the curved strokes algorithm and simply run the Glass patterns algorithm with the noise level set as above.

**Pointillism and Glass Patterns.** Our filter combination follows the basic method of the pointillistic filter. The background layer paints points as usual, but the other two layers mix their point placement with the Glass patterns method. For each point that is placed by the pointillistic part of the filter, we take the color of this point and use the Euler algorithm to paint more points of this color along the arc of a streamline defined by the vector field  $\mathbf{v}$  as defined in Section 3. We want the combined filter to retain the look of being made up of points, so these extra points that are generated have a probability of not being painted to the canvas to prevent the filter having the look of smooth strokes. This probability is proportional to the influences parameters so there are less points as the filter tends toward pointillism.

As the filter gets closer to impressionism we need to take a slightly different approach. We decrease the radius of the points so that we tend toward manipulating pixels rather than larger areas. We also gradually get rid of the color distortion of the pointillistic filter after a point.

## 5 Results

Since we are trying to approximate art, any evaluation is of course very subjective. Each combination creates a nice artistic effect that looks inspired by but distinct from the filters it is made up of. We look at each filter combination separately to give a more thorough evaluation. Figure 5 shows a combination of the curved strokes and pointillistic filters. This gives us an interesting effect





**Fig. 5.** *Left:* Scene from Gilleleje, Denmark. Original photograph by Angela Palmer. *Right:* The results of the curved strokes and pointillism combination.



**Fig. 6.** *Left:* Scene from Amantani Island, Lake Titicaca, Peru. Original photograph by Xnena Vitali Jaensch. *Right:* The results of the pointillism and Glass patterns combination.

with a variety of brush shapes and sizes. With the influence parameter at this level, less of the smaller points can be seen, but pointillism's interesting color distortions prove to be very effective when used with larger strokes as well as small points. The variety of stroke sizes and color distortions makes for an effect that is much more 'artistic', and for many images more visually pleasing.

Figure 6 shows a combination of the pointillistic and Glass patterns filters. The effect of color distortion is particularly striking in images like this where the saturation distortion brings out colors that are not normally noticeable in the scene. The color distortion mixes well with the geometric distortion as both implement different ideas of impressionism, departing from realism to give a nice 'impression' of the scene.

Figure 7 shows a combination of the curved strokes and Glass pattern filters. This combination creates a nice artistic effect that uses aspects of both filters to create an image that is perhaps more analogous to the painting process than any of the standalone filters. It incorporates the layers of the curved strokes filters, and develops a painting out of distinct strokes. It also uses the vector field from the Glass pattern filter however, which adds the idea of motion to the image.



**Fig. 7.** *Left:* Scene with a flower in Bangkok. Original photograph by Marian Arnold. *Right:* The results of the curved strokes and Glass patterns combination.

The result is an image with a variety of stroke sizes and nice layered effect but with strokes that have a nice flowing movement around object contours.

## 6 Conclusions and Future Work

We developed methods of blending filters together to create artistic effects that simulate a mixture of artistic styles, and our results illustrate this filter blending process (see also [4]). There are improvements that could be made to the individual filters but the blending process itself is extremely effective.

A possible extension of the work done here would be to combine the application with a camera and printer in order to create 'instant' portraits. Portraits are more difficult to render artistically than some other scenes as users can be sensitive about the amount of abstraction that is applied to faces, but extensions could be added using other fields such as face detection to tailor the results toward producing pleasing portraits. There is room here for further study into what features would best tailor the algorithm toward effects that users want to see in a portrait piece.

## References

1. Chevreul, M.E.: The Principles of Harmony and Contrast of Colors and Their Applications to the Arts, based on the first English edition of 1854 (introduction by Faber Birren). Reinhold Publishing Corporation, New York (1967)
2. Hertzmann, A.: Painterly rendering with curved brush strokes of multiple sizes. In: Proc. SIGGRAPH, pp. 453–460 (1998)
3. Papari, G., Petkov, N.: Continuous glass patterns for painterly rendering. IEEE Trans. Image Processing 18, 652–664 (2009)
4. Valente, C., Klette, R.: Aritstic emulation. Video on YouTube, uploaded (October 2010)
5. Yang, C.-K., Yang, H.-L.: Realization of Seurat's pointillism via non-photorealistic rendering. The Visual Computer 24, 303–322 (2008)



# A Data-Driven Approach to Understanding Skill in Photographic Composition

Todd S. Sachs<sup>1</sup>, Ramakrishna Kakarala<sup>2,\*</sup>, Shannon L. Castleman<sup>2</sup>,  
and Deepu Rajan<sup>2</sup>

<sup>1</sup> Aptina Imaging, San Jose, CA, USA

<sup>2</sup> Nanyang Technological University, Singapore 639798  
ramakrishna@ntu.edu.sg

**Abstract.** Photography requires not only equipment but also skill to reliably produce aesthetically-pleasing results. It can be argued that, for photography, skill is apparent even without sophisticated equipment. However, no scientific tests have been carried out to confirm that supposition. For that matter, there has been little scientific study on whether skill is apparent, whether it can be discerned by judges in blind tests. We report results of an experiment in which 33 subjects were asked to use identical cameras to photograph each of 7 pre-determined scenes, including a portrait, landscapes, and several man-made objects. Each photograph was then rated in a double-blind manner by 8 judges. Of those judges, 3 are professional photographic experts, and 5 are imaging researchers. The results show that expert judges are able to discern photographic skill to a statistically significant level, but that the enthusiasts, who are more akin to the general public, are not. We also analyse the photos using computer vision methods published in the literature, and find that there is no correlation between human judgements and the previously-published machine learning methods.

## 1 Introduction

Photography, like cooking, can be carried out by just about anyone with a minimum of equipment. However, in order to reliably produce aesthetically-pleasing results, skill is also required. While certain aspects of photography depend strongly on equipment, such as colour, focus, and exposure, there is at least one aspect that is a matter of skill: composition. Composition is taught in Arts faculties and is the subject of many books and papers, but few have analysed it from a scientific perspective. That may be because of the difficulty in framing scientific or engineering problems in studying composition. However, there is reason to believe that photographic composition is amenable to both scientific and engineering inquiry, since it depends on spatial arrangements of features and objects.

This paper is devoted to studying skill in photographic composition using a data-driven approach. We report results of an experiment when a group of

---

\* Corresponding author.

photographers, ranging in skill from novices to experts, took photographs of a set of predefined scenes using identical point-and-shoot cameras using identical settings (in full “auto” mode). The photographs were then rated numerically in terms of composition strength by an independent group of judges. We show that a subset of the judges, who are themselves professional photographers, are able to discern skill in a statistically-significant manner.

In order to put our work in context, we review relevant research in the field of photograph aesthetics. It is important to distinguish photographic quality from photographic appeal. As Savakis, Etz & Loui [1] point out, photo quality (sharpness, noise level, dynamic range) has been studied scientifically for decades but, in contrast, photo appeal has been studied very little from a scientific perspective. In [1], the authors experimentally determined the attributes that observers feel are important to deciding which pictures deserve emphasis in a photo album, and found that the most important is composition. In particular, Savakis *et al* found that composition was much more important (by at least a factor of 3) than either colourfulness or sharpness, two traditional measures of image quality. The photos used in [1] are from ordinary consumers, i.e., there was no segregation into those from professional photographers and amateurs. In contrast, Tong *et al* [2] explore the distinction of skill, and attempt to classify photographs into those taken by professionals and amateurs using computer vision techniques. Their methods rely on features extracted from the images such as sharpness, colourfulness, contrast, and saliency. The classifier that they develop correlates well (coefficient of 0.85) with rankings given by a group of 16 human observers. However, they do not consider composition as an attribute, nor possible equipment differences between professionals and amateurs. Ke, Tang, & Jing [3] also examine the choice of attributes that distinguish between experts and amateurs, and argue that the “bag of low-level features” approach taken by [2] is not as effective as using high level semantic features. Specifically, Ke *et al* propose that expert photos are distinguished from amateur shots by the attributes of “simplicity”, which they measure by spatial distribution of edges, “colourfulness” measured by color histograms and hue count, “sharpness” measured from the spatial frequency content, and two low-level features measuring contrast and brightness. Ke *et al* test their classifier on photos obtained from the website [dpchallenge.net](http://dpchallenge.net), and find that the sharpness attribute is the most discriminative in distinguishing between the top 10% most highly-rated photographs from the bottom 10% in their test set. However, their study does not consider composition as an attribute.

Datta, Joshi, Li, and Wang [4] propose a machine learning approach to rating aesthetic appeal of photographs. Like the previously-mentioned studies, Datta *et al* use the attributes of colourfulness, sharpness (depth of field), but, in a novel step, include consideration for composition by using the “rule of thirds”<sup>1</sup>, texture, and familiarity (measured by similarity to a group of standard images).

---

<sup>1</sup> A well-known maxim in photographic composition is that objects should be placed not in the center of the image, but at one-third or two-third the height or width to draw the user’s attention into the scene.

Their system is perhaps the first to explicitly consider composition in the colour and texture distribution measures. Datta *et al* compare their system with ratings obtained for photographs on [photo.net](#), and show good correspondence. More importantly for our paper, Datta & Wang [5] make their rating method, named ACQUINE (Aesthetic Quality Inference Engine) available online on the site [acquine.alipr.com](#).

Composition as an attribute is also considered by Luo & Tang [6], who, like previous researchers, develop methods for classifying expert and amateur photographs, but provide the novel step of extracting subjects from the background using sharpness as a cue. Specifically, they measure composition geometry by distance of the subject centroid to the rule-of-thirds points. Their method outperforms that of Ke *et al* [3] on the same data set obtained from [dpcchallenge.net](#).

It is interesting that the computer vision literature shows considerable interest in using sharpness and colourfulness as attributes of photograph aesthetics, whereas the study of Savakis *et al* shows that composition is far more important. Obviously, an image can be appealing even without being sharp or colourful; for example, the black-and-white photographs of Henri Cartier-Bresson are often slightly defocused and lack contrast, but are nevertheless powerful due to their composition [7].

No previous study that we are aware of has considered whether composition skill is measurable, the subject of our paper.

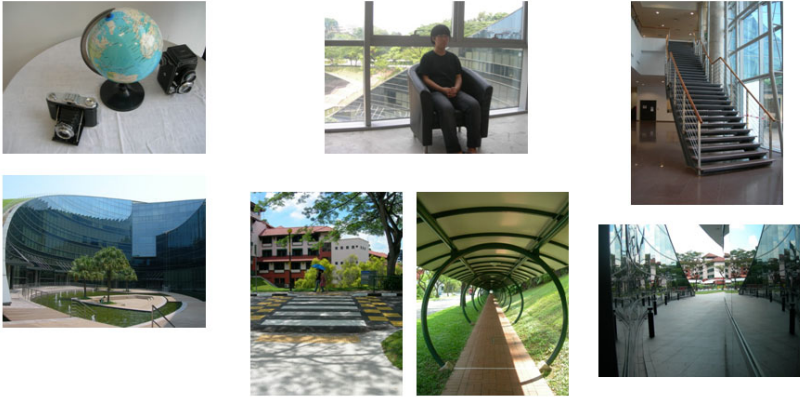
## 2 Experimental Methods

Our experiment is designed to test whether skill in photographic composition can be identified in a double-blind study. We recruited 33 unpaid volunteers to take part in the experiment. The group consisted of university students, staff, as well as professional photographers. Several of the students are undergraduate majors in photography, a point we return to below. The subjects were asked to identify their own skill level by answering a questionnaire, in which they were asked to estimate how many pictures they took each year (10, 100, 1000, or too many to count), whether they shared those pictures with others through photo sharing sites such as Flickr© or Picasaweb©, whether they received any formal training in photography or in other arts, and whether they have published or exhibited photographs. Through their answers, we assigned each photographer to one of four categories, in increasing level of skill: Novices, who rarely take photographs or use a camera; Amateurs, who frequently take photographs, may share them with friends and family, but do not have formal training in photography nor invest in equipment such as a SLR; Enthusiasts, who invest much time and resources in photographic aesthetics, possibly have formal training in photography (including the photo majors mentioned above); and Experts, who are professional photographers with published or exhibited work. The line between Amateur and Enthusiast is admittedly hard to draw; we expect some of our amateurs are better classified as enthusiasts and vice versa. Table 1 shows the number in each group. As could have been predicted from the nature of such an

**Table 1.** Distribution of the 33 photographers by skill level

Novices	Amateurs	Enthusiasts	Experts
4	15	12	2

experiment, few novices volunteered. The number of true Experts is admittedly few, but within the Enthusiast class there are 3 photography majors each with at least 3 years of formal training. The photo majors are arguably a group with skill somewhere in between the formally untrained Enthusiasts and the Experts, a point we explore quantitatively below. The subjects were asked to submit one photograph of each of 7 scenes. The scenes, which were chosen by two professional photographers, are as follows: (1) a still life of manmade objects; (2) a portrait of a person seated in a chair, whom the subjects were not allowed to ask to pose; (3) a two-story long indoor staircase; (4) an outdoor fountain; (5) a striped road crossing with traffic; (6) a covered walkway; and (7) a corner of a building with reflecting glass on both sides. The subjects were allowed to take as many photos as they liked of each scene, but then were required to select and give us their best one. Figure 1 shows examples of each of the 7 shots for illustrative purposes.

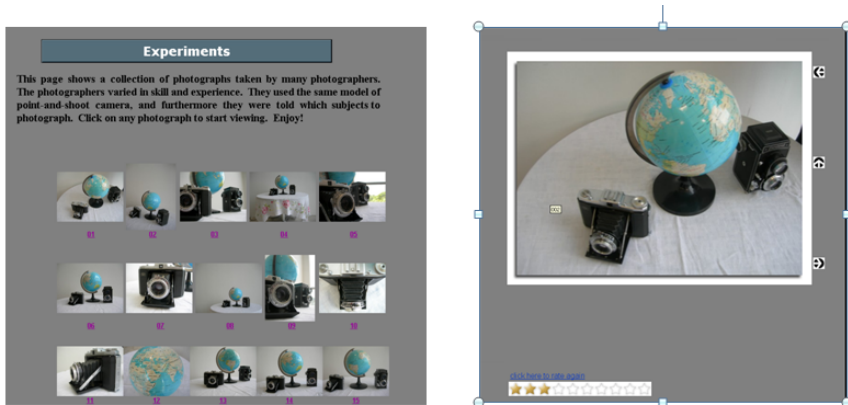


**Fig. 1.** Example photos of the seven scenes are shown. From left to right, top to bottom: still life; portrait; staircase; fountain; road; walkway; glass corner. Each photo was taken by a different photographer.

To make sure that the choice of equipment did not affect image quality, the subjects were given the same model of consumer point-and-shoot camera with 7 megapixel resolution. The cameras were set in full automatic mode, to ensure that picture quality was not affected by the mode chosen. For each of the 7 scenes, the subjects were allowed to move around within a demarcated area. The area was marked off by two professional photographers, with the objectives of allowing

**Table 2.** Guide given to judges for scoring

Score	Criterion
0 – 2	poor composition
3 – 4	some consideration and use of composition
5 – 6	average/ acceptable composition
7 – 8	some skill and use of composition
9 – 10	wonderful composition/ nice image



**Fig. 2.** Two screenshots of the web-based rating process used by the judges for collecting ratings from the panel. The left hand image shows a thumbnail view of some of the photos taken of the “still life”, and the right hand side shows the 10-star rating system.

both freedom of composition and also comparable photos. The movement area was marked on the ground by masking tape. Due to the freedom already provided by the movement area, we disabled the zoom on the cameras by taping over the zoom button. The combination of both zoom and movement area would lead to incomparable photos. The experiment was carried out over two separate days, between the hours of 11am-2pm to ensure similar light conditions outdoors. Of the  $33 \times 7 = 231$  photographs that were submitted, we eliminated 10 from consideration for violation of announced rules. For example, photos were removed if the subjects used zoom (despite the zoom button being taped over), or asked the portrait model to pose, or focused on bystanders rather the assigned scenes.

The remaining 221 photos were then judged in a double-blind manner by a panel of 8 judges, who were a separate group from the photographers. The panel included 3 professional photographers, and 5 imaging science and technology researchers from a leading digital image sensor company. The researchers can be considered to have similar characteristics to the Enthusiasts in the subject group. The judging was done online by selecting a rating based on 10 stars. Figure 2 illustrates the rating process. The judges were allowed to save their ratings, login

and out in order to break up the rating process according to their convenience, and also to revise their ratings if necessary. The judges were instructed to rate the photograph only on the composition, rather than on focus, lighting or colour. They viewed images of resolution  $600 \times 450$ . They were given the following guide for scoring

For comparison, we obtained the ACQUINE rating [5] by uploading all 221 images in our collection to [acquine.alipr.com](http://acquine.alipr.com).

### 3 Results

In this section, we compare the ratings between the two types of judges (professionals and researchers), the ACQUINE rating, and the identified skill of the photographers from the questionnaire. As a simple measure of statistical relationship, we use the Spearman rank correlation [8, pg 206]. The Spearman coefficient is computed by converting raw scores  $X_i$ ,  $Y_i$ , into ranks  $x_i$ ,  $y_i$ , and, in the event of tied ranks, using the standard Pearson correlation on ranks

$$\rho = \frac{\sum_i (x_i - \bar{x}) \sum_j (y_j - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_j (y_j - \bar{y})^2}}. \quad (1)$$

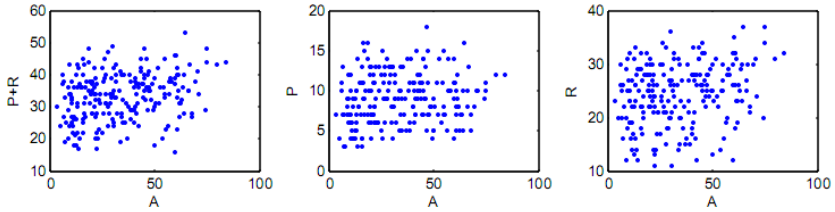
For example, raw scores 1.0, 1.1, 3.4, 10.0 are converted to ranks 1, 2, 3, and 4. Unlike the Pearson correlation, the Spearman coefficient  $\rho$  is nonparametric and indicates the degree to which one variable is a monotonic (not necessarily linear) function of another. We found no significant difference in our results between the Spearman and the Kendall  $\tau$  correlation, another widely used nonparametric measure.

We evaluated the consistency between all 9 judges: the 8 human judges and ACQUINE as follows. A  $9 \times 9$  Spearman rank coefficient matrix shows that none of the correlations are significantly negative, which indicates that no pair of judges have opposite views on what constitutes strength of composition. Table 3 summarizes the results, and shows that each group is consistent to the same degree among themselves, but there is less consistency between groups. None of the correlations are strong (i.e.,  $\rho > 0.7$ ), indicating that the judges are fairly independent. In fact, the correlations between the human judges and ACQUINE are weak (not more than 0.27). To demonstrate that visually, Figure 3 shows scatter plots of the mean rating of the human judges against the rating from ACQUINE. It is worth noting that ACQUINE takes into account many factors in assessing photo aesthetics, including but not limited to composition, whereas the human judges were instructed to pay attention only to composition. Therefore, there is no reason to expect good correlation between our human judges and ACQUINE; the point of the Figure 3 is to illustrate quantitatively that our judges are indeed rating images independently from the methods of ACQUINE.

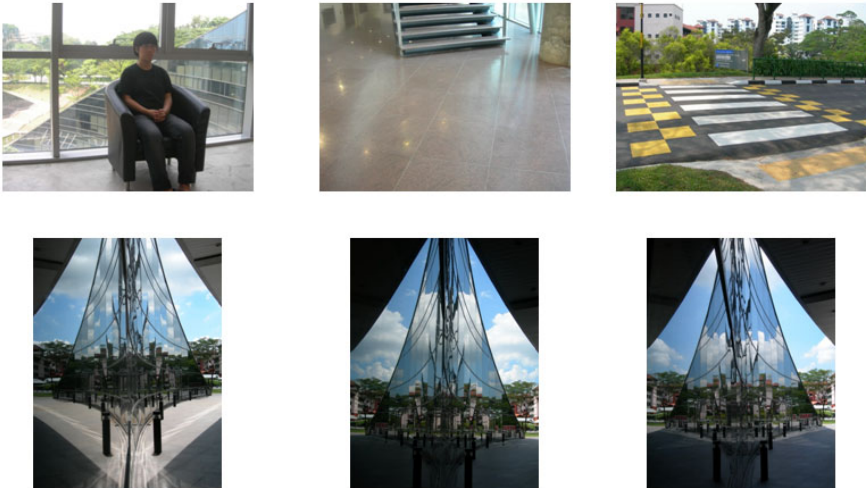
Although the groups had little correlation with each other, remarkably they found similar results for best images. Figure 4 shows that, for each group, the symmetry of the building corner was appealing. Interestingly, each group chose a slightly different picture of the corner as their best one.

**Table 3.** Minimum \ Maximum Spearman correlation in judge groups

Professionals (P)	Researchers (R)	Mixed (P vs R)	P vs Acquine (A)	R vs A
0.25\0.43	0.21\0.46	0.02\0.38	0.01\0.22	0.02\0.27



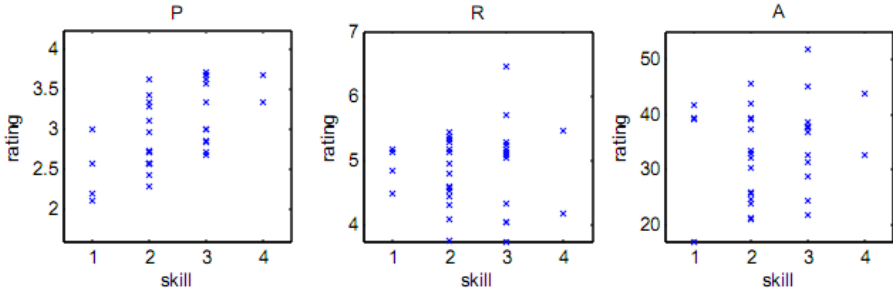
**Fig. 3.** The leftmost scatter plot compares all human judges (sum of scores) to ACQUINE (A), the middle plot compares professionals (P) to A, and the rightmost plot compares researchers (R) to A. The scatter plots confirm visually that there is low correlation between human judges and the automated rating system used by ACQUINE.



**Fig. 4.** From left to right, the top row are respectively the worst (lowest rated) pictures as selected by the professionals (P), researchers (R), and ACQUINE (A). The bottom row is the best (highest rated) for the respective groups. Remarkably, all three groups chose the building corner with its inherent symmetry as the best.

**Table 4.** Spearman rank correlation between photographer skill level and group rating. The top row uses skill rating 1-4, and the bottom row uses the modified 5-point skill rating discussed below, where photo majors are placed in between Enthusiasts and Experts.

Professionals (P)	Researchers (R)	P+R	ACQUINE (A)
0.59	0.05	0.25	0.09
0.60	0.03	0.21	0.08



**Fig. 5.** From left to right, the three graphs show ratings assigned to each of the skill levels by professional judges (P), camera researchers (R), and ACQUINE (A). An upward trend is apparent in the P graph, though not in the other graphs.

Next, we assessed whether any of the groups are able to determine photographic skill in their ratings. We measured the correlation between the skill level (on a scale of 1-4, with 1 for Novices, 2 for Amateurs, 3 for Enthusiasts, 4 for Experts) determined from the questionnaire, and the sum of ratings given to each photographer by each group. The results are summarized in the top row of Table 4. We see that the professional (P) ratings are consistent with skill to a much higher degree than either the researchers, or ACQUINE. Moreover, the correlation between P ratings and skill is statistically significant ( $p < 0.001$ ).

To illustrate visually the correlation numbers, Figure 5 shows a scatter plot of ratings assigned to each skill level. It is apparent that the left most graph, representing the professional judges, has an upward trend with skill, unlike the other graphs.

Since there were only 2 members of our Experts group, we next considered the scores given to the 3 photo majors with 3 years of formal training in photography. The majors were included in the Enthusiasts group for purposes of compiling the top row of Table 4. However, it is reasonable to place their skill somewhere between the untrained (formally) Enthusiasts and the accomplished Experts. Therefore, we created a new skill rating, with values of 1 and 2 as before for Novices and Amateurs, 3 for the untrained Enthusiasts, 4 for the photo majors, and 5 for the Experts. For this new skill rating, the correlations are shown in



the bottom row Table 4. We see little change in the results, indicating that it makes little difference whether we place the majors in between the Enthusiast and Expert groups.

## 4 Discussion

Our results show little correlation between human judges and ACQUINE. That may be expected given that the judges were instructed to rate only on composition, whereas ACQUINE considers other factors in addition to composition. But then how important is composition in ACQUINE’s internal weighting? From our results, we can conclude that either the weighting is low, or that the simple measure used by ACQUINE based on the rule of thirds is not a good predictor of the evaluation given by humans, or both.

The fact that all three groups (P, R, and A) found as their highest-rated images a shot of the building corner is interesting, and is likely to be connected with the ease of composing that shot. The subjects were restricted to stay in a relatively small box when composing, and within that box it is not difficult to find the symmetry of the building corner. The data also show that of the seven shots, the building corner received the highest mean score, 4.8, from the human judges, whereas the portrait received the lowest mean score, 3.6.

The computer-vision based aesthetic systems [3] [4] [6] aim to match ratings on popular photo-sharing sites. One can argue that they tend to rate “safe” techniques that do not challenge the viewer more highly, i.e., they favour colourful, sharp, and simple images. In contrast, the professional (P) judges in our study are aware of a greater variety of techniques, including those used by the Expert photographers to be distinctive. Therefore it is not surprising that our P group rates Experts highest, and also shows reasonably good correlation with skill level. Interestingly, the ratings given by our researcher (R) group have little correlation with skill. There may be at least two factors at work behind that result. First, the judges were instructed to rate a photo on composition only, but that is not easy to do in general. Second, the techniques that Experts (and photo majors) use may not be appreciated without formal training, as is the case with other art forms such as painting or music.

## 5 Conclusions and Future Work

This paper shows that skill in photographic composition is detectable to human judges from a collection of photographs. It shows that there are clear difference between ratings given by professional photographers, and those given by imaging researchers who are clearly interested in photography, but are not practicing photographers. We also see, as might be expected from the criteria used to rate the photographs, that there are clear differences between human judges and computer vision systems.

One interesting aspect of our database, that we have not explored fully, is the vantage point used for taking the photo. Experts are likely to use unusual

vantage points to make their photos stand out. Software tools such as Photo Tourism [9] and Photo Synth (photosynth.net) allow users to combine shots of a common subject, and to explore vantage points. We plan to investigate how vantage point varies with skill in future work.

**Acknowledgement.** We thank Dr. Shahidul Alam for valuable help in experimental design. We also thank Rajesh Somavarapu, Dawson Mao and Guo Yixiu for help in carrying out the experiments, and the subjects and judges for their time. This work was funded by the Institute for Media Innovation at NTU by a Seed Grant.

## References

1. Savakis, A., Etz, S., Loui, A.: Evaluation of image appeal in consumer photography. In: SPIE Human Vision and Electronic Imaging V (2000)
2. Tong, H., Li, M., Zhang, H.J., He, J., Zhang, C.: Classification of digital photos taken by photographers or home users. In: Proceedings of Pacific Rim Conference on Multimedia, pp. 198–205. Springer, Heidelberg (2004)
3. Ke, Y., Tang, X., Jing, F.: The design of high-level features for photo quality assessment. In: CVPR (1), pp. 419–426. IEEE Computer Society, Los Alamitos (2006)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006)
5. Datta, R., Wang, J.Z.: Acquine: aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In: Wang, J.Z., Boujemaa, N., Ramirez, N.O., Natsev, A. (eds.) Multimedia Information Retrieval, pp. 421–424. ACM, New York (2010)
6. Luo, Y., Tang, X.: Photo and video quality evaluation: Focusing on the subject. In: Forsyth, D., Torr, P.H.S., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 386–399. Springer, Heidelberg (2008)
7. Zakia, R.: Perception and imaging: photography—a way of seeing, 3rd edn. Elsevier Science Ltd., Cambridge (2007)
8. Maritz, J.S.: Distribution-free statistical methods. Chapman and Hall, London (1991)
9. Snaveley, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. In: SIGGRAPH Conference Proceedings, pp. 835–846. ACM Press, New York (2006)

# Image Information in Digital Photography

Jaume Rigau, Miquel Feixas, and Mateu Sbert

Graphics and Imaging Laboratory, University of Girona, Spain

**Abstract.** Image formation is the process of computing or refining an image from both raw sensor data and prior information. A basic task of image formation is the extraction of the information contained in the sensor data. The information theory provides a mathematical framework to develop measures and algorithms in that process. Based on an information channel between the luminosity and composition of an image, we present three measures to quantify the saliency, specific information, and entanglement of this image associated with its luminance values and regions. The evaluation of these measures could be potentially used as a criterion to achieve more aesthetic or enhanced images.

## 1 Introduction

The human visual system is able to reduce the amount of incoming visual data to a small but relevant amount of information for higher-level cognitive processing. Different computational models, most of them based on information theory, have been proposed to interpret the selective visual attention [8,13,2,6,7]. The biologically-inspired model of bottom-up attention of Itti et al. [8] permits us to understand our ability to interpret complex scenes in real time. The selection of a subset of available sensory information before further processing appears to be implemented in the form of a spatially circumscribed region of the visual field, called *focus of attention*, while some information outside the focus of attention is suppressed. This selection process is controlled by a *saliency map* which is a topographic representation of the instantaneous saliency of the visual scene and shows what humans find interesting in visual scenes.

On the other hand, *image formation* is the process of computing or refining an image from both raw sensor data and prior information about that image [9]. The main task of image formation is to extract the *information* contained in the raw sensor data to estimate the image. Information theory plays a basic role in this process: providing a theoretic framework, defining measures of optimality, developing algorithms, quantifying statistical quality, etc. It can be considered that image formation corresponds to our common concept of *photography*.

Instead of analyzing image information from a biologic perspective [8,13,2,6,7], in this paper we propose a mathematical approach based on an information channel between the luminosity and composition of that image — two basic features in photography. From this channel, saliency, specific information, and entanglement can be computed using different information-theoretic measures defined in the field of neural systems [5,3,1].

This paper is organized as follows. In Section 2, we review some basic information-theoretic measures. In Section 3, we present the information channel and the splitting algorithm used to analyze the information of an image. In Section 4, we describe three different measures of information associated to the luminance values and regions of an image. In Section 5, we show and discuss the obtained results. Finally, we present the conclusions.

## 2 Information-Theoretic Concepts

Information theory [4] deals with the transmission, storage and processing of information, and is used in fields such as physics, computer science, statistics, biology, image processing, learning, etc.

Let  $\mathcal{X}$  be a finite set, let  $X$  be a random variable taking values  $x$  in  $\mathcal{X}$  with distribution  $p(x) = Pr[X = x]$ . Likewise, let  $Y$  be a random variable taking values  $y$  in  $\mathcal{Y}$ . An information channel  $X \rightarrow Y$  between two random variables (input  $X$  and output  $Y$ ) is characterized by a *probability transition matrix* (composed of conditional probabilities) which determines the output distribution given the input.

The *Shannon entropy*  $H(X)$  of a random variable  $X$  is defined by

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x). \quad (1)$$

It measures the average *uncertainty* of a random variable  $X$ . All logarithms are base 2 and entropy is expressed in bits. The convention that  $0 \log 0 = 0$  is used. The *conditional entropy* is defined by

$$H(Y|X) = - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) = \sum_{x \in \mathcal{X}} p(x) H(Y|x), \quad (2)$$

where  $p(y|x) = Pr[Y = y|X = x]$  is the conditional probability and  $H(Y|x) = - \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x)$  is the entropy of  $Y$  given  $x$ . The conditional entropy  $H(Y|X)$  measures the average uncertainty associated with  $Y$  if we know the outcome of  $X$ .  $H(X) \geq H(X|Y) \geq 0$  and, in general,  $H(Y|X) \neq H(X|Y)$ .

The *mutual information* (MI) between  $X$  and  $Y$  is defined by

$$I(X; Y) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{p(y)}. \quad (3)$$

It is a measure of the *shared information* between  $X$  and  $Y$ . It can be seen that  $I(X; Y) = I(Y; X) \geq 0$ .

The *relative entropy* or *Kullback-Leibler distance* between two probability distributions  $p = \{p(x)\}$  and  $q = \{q(x)\}$  defined over  $\mathcal{X}$  is given by

$$KL(p|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}, \quad (4)$$

where, from continuity, we use the convention that  $0 \log 0 = 0$ ,  $p(x) \log \frac{p(x)}{0} = \infty$  if  $p(x) > 0$ , and  $0 \log \frac{0}{0} = 0$ . The relative entropy  $KL(p|q)$  is a divergence measure between the *true* probability distribution  $p$  and the *target* probability distribution  $q$ . It can be proved that  $KL(p|q) \geq 0$ .

### 3 Compositional Information Channel

Two of the most basic elements of a photograph are composition and luminosity. In order to analyze their correlation, we use an information channel between the luminance histogram and the regions of the image. This channel permits us to investigate, from an information theory perspective, the shared information between them and, more particularly, the saliency, the specific information, and the entanglement associated to each luminance value and region (see Sec. 4).

In this section, we review the information channel between the color (in our case, luminance) histogram and the regions of an image, introduced by Rigau et al. [10], and then we describe the partitioning algorithm which progressively splits the image by extracting the maximum information at each step. The information channel  $C \rightarrow R$  is defined between the random variables  $C$  (input) and  $R$  (output), which represent respectively the set of bins ( $\mathcal{C}$ ) of the color histogram and the set of regions ( $\mathcal{R}$ ) of the image. Given an image  $\mathcal{I}$  of  $N$  pixels, where  $N_c$  is the frequency of bin  $c$  ( $N = \sum_{c \in \mathcal{C}} N_c$ ) and  $N_r$  is the number of pixels of region  $r$  ( $N = \sum_{r \in \mathcal{R}} N_r$ ), the three basic elements of this channel are:

- The conditional probability matrix  $p(R|C)$ , which represents the transition probabilities from each bin of the histogram to the different regions of the image, is defined by  $p(r|c) = \frac{N_{c,r}}{N_c}$ , where  $N_{c,r}$  is the frequency of bin  $c$  into the region  $r$ . Conditional probabilities fulfill  $\forall c \in \mathcal{C}. \sum_{r \in \mathcal{R}} p(r|c) = 1$ .
- The input distribution  $p(C)$ , which represents the probability of selecting each intensity bin  $c$ , is defined by  $p(c) = \frac{N_c}{N}$ .
- The output distribution  $p(R)$ , which represents the normalized area of each region  $r$ , is given by  $p(r) = \frac{N_r}{N} = \sum_{c \in \mathcal{C}} p(c)p(r|c)$ .

According to (3), the MI between  $C$  and  $R$  is given by

$$I(C; R) = \sum_{c \in \mathcal{C}} p(c) \sum_{r \in \mathcal{R}} p(r|c) \log \frac{p(r|c)}{p(r)} \quad (5)$$

and represents the *shared information* or *correlation* between  $C$  and  $R$ .

We now describe a greedy mutual-information-based algorithm [10] which splits the image in quasi-homogeneous regions. This procedure takes the full image as the unique initial partition and progressively subdivides it in a binary space partition according to the maximum MI gain for each partitioning step. The algorithm generates a partitioning tree for a given ratio of MI gain  $I(C; R)/H(C)$ , or a predefined number of regions.

This partitioning process can also be visualized from

$$H(C) = I(C; R) + H(C|R), \quad (6)$$

where  $R$  is the random variable which represents the set of regions of the image that varies after each new partition. The acquisition of information increases  $I(C; R)$  and decreases  $H(C|R)$ , producing a reduction of uncertainty due to the equalization of the regions. The maximum MI that can be achieved is  $H(C)$ . The more complex the image the further down the regions we have to go to achieve a given level of information. The rate of the information extraction will depend on the degree of order in the image. Fig. 1*b.i* and Fig. 2*b.i* show decompositions obtained using a MI ratio of 1/3. Observe that the number of regions is much bigger in the second image because this contains more detailed and contrasted areas.

## 4 Image Information Measures

In this section, we study how information is distributed in the image by computing three different information measures associated with each luminance value and region. As we have seen in Sec. 3, the MI between  $C$  and  $R$  expresses the degree of correlation or the information transfer between the set of luminance bins and the regions of the image. This interpretation can be extended to consider the information associated to a single luminance value, that is, the information gained on  $R$  by the observation of a intensity value  $c$ , and vice versa. To obtain this information, MI can be decomposed in different alternative ways [5,3,1]. Although many definitions of information are plausible, we present here the three most “natural” decompositions of  $I(C; R)$ .

### 4.1 Saliency

From (3), the MI between color and regions can be expressed as

$$I(C; R) = \sum_{c \in \mathcal{C}} p(c) \sum_{r \in \mathcal{R}} p(r|c) \log \frac{p(r|c)}{p(r)} = \sum_{c \in \mathcal{C}} p(c) I_1(c; R), \quad (7)$$

where we define

$$I_1(c; R) = \sum_{r \in \mathcal{R}} p(r|c) \log \frac{p(r|c)}{p(r)} \quad (8)$$

as the *surprise* associated with the color  $c$  and can be interpreted as a measure of its *saliency*. Itti and Baldi [7] provide experimental evidence that Bayesian surprise best characterizes what attracts human gaze. According to Bruce and Tsotsos [2], certain visual events such as a bright flash of light will almost result in an observer’s gaze being redirected.

High values of  $I_1(c; R)$  express a high surprise and identify the most salient colors. It is important to observe that (8) is as a Kullback-Leibler distance,  $I_1(c; R) = KL(p(R|c)|p(R))$ , where  $p(R|c)$  is the conditional probability distribution between  $c$  and the image regions, and  $p(R)$  corresponds to the distribution of region areas. It can be shown that  $I_1$  is the only positive decomposition of MI [5].

Similarly, the surprise associated with a region can be defined from the reversed channel  $R \rightarrow C$ , so that  $R$  is the input and  $C$  the output. From the Bayes' theorem,  $p(c, r) = p(c)p(r|c) = p(r)p(c|r)$ , the MI (7) can be rewritten as

$$I(R; C) = \sum_{r \in \mathcal{R}} p(r) \sum_{c \in \mathcal{C}} p(c|r) \log \frac{p(c|r)}{p(c)} = \sum_{r \in \mathcal{R}} p(r) I_1(r; C), \quad (9)$$

where we define

$$I_1(r; C) = \sum_{c \in \mathcal{C}} p(c|r) \log \frac{p(c|r)}{p(c)} \quad (10)$$

as the surprise associated with region  $r$  and can be interpreted as its saliency. Analogous to  $I_1(c; R)$ , high values of  $I_1(r; C)$  correspond to the most salient regions. Measures  $I_1$  have been previously used to quantify the color and region information in Van Gogh's paintings [11]. The measure  $I_1(r; C)$  has been also used to evaluate the saliency of a painting, comparing well with Itti-Koch model [12].

## 4.2 Specific Information

The definition of specific information  $I_2$  was proposed by DeWeese and Meister [5]. From (5), mutual information can be expressed as

$$I(C; R) = H(R) - H(R|C) = \sum_{c \in \mathcal{C}} p(c)[H(R) - H(R|c)] = \sum_{c \in \mathcal{C}} p(c) I_2(c; R), \quad (11)$$

where

$$I_2(c; R) = H(R) - H(R|c) = - \sum_{r \in \mathcal{R}} p(r) \log p(r) + \sum_{r \in \mathcal{R}} p(r|c) \log p(r|c) \quad (12)$$

is the *specific information* of  $c$  and expresses the change in uncertainty about  $R$  when  $c$  is observed. A large value of  $I_2(c; R)$  means that we can easily predict a region given the color  $c$ .

Following a similar process for the reversed channel  $R \rightarrow C$ , the specific information associated with region  $r$  is given by

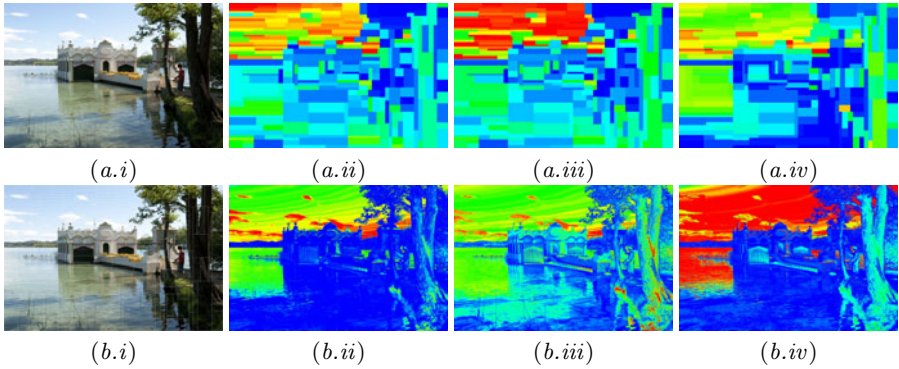
$$I_2(r; C) = H(C) - H(C|r) = - \sum_{c \in \mathcal{C}} p(c) \log p(c) + \sum_{c \in \mathcal{C}} p(c|r) \log p(c|r) \quad (13)$$

and expresses the predictability of a color known the region. Note that  $I_2(c; R)$  and  $I_2(r; C)$  can take negative values [5].

## 4.3 Entanglement

Butts [3] proposed another decomposition of MI based on the *stimulus specific information*  $I_3$ . In our framework, this measure, which we call *entanglement*, is defined by

$$I_3(c; R) = \sum_{r \in \mathcal{R}} p(r|c) I_2(r; C). \quad (14)$$



**Fig. 1.** (a.i) Banyoles Lake, Spain [f/16, 1/80, ISO200]. (b.i) Image decomposition (328 regions,  $H(C) = 7.890$ ,  $I = 2.623$ ) of (a.i). (a.ii-iv)  $I_1$ ,  $I_2$ , and  $I_3$  maps from the channel  $R \rightarrow C$ . (b.ii-iv)  $I_1$ ,  $I_2$ , and  $I_3$  maps from the channel  $C \rightarrow R$ .

A large value of  $I_3(c; R)$  means that the specific information  $I_2(r; C)$  of the regions that contain the color  $c$  are very informative.

Following a similar process for the reversed channel  $R \rightarrow C$ , the entanglement associated with each region is given by

$$I_3(r; C) = \sum_{c \in \mathcal{C}} p(c|r) I_2(c; R). \quad (15)$$

Similarly to  $I_3(c; R)$ , a large value of  $I_3(r; C)$  means that the specific information  $I_2(c; R)$  of the colors contained in a region  $r$  are very informative.

These measures emphasize a univocal relationship between color and regions, and can be interpreted as the correlation between specific regions and colors. For instance, a particular color can have a high value because is identified by a characteristic region. In the same way, a region with a single color that doesn't appear in other regions will show a high  $I_3$  value. An example could be the background of an image.

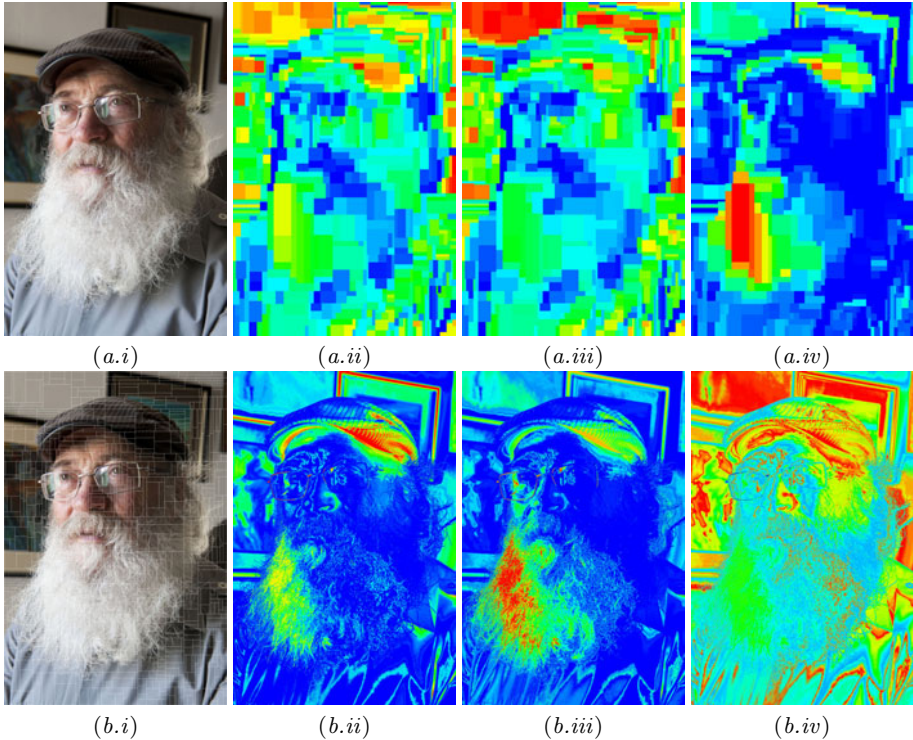
In conclusion,  $I_1$ ,  $I_2$ , and  $I_3$  represent three different ways of quantifying the information associated to a luminance value  $c$  and to a region  $r$ . While  $I_1$  is always positive and non-additive,  $I_2$  can take negative values but is additive, and  $I_3$  can take negative values and is non additive [5,3,11].

## 5 Results and Discussion


In this section we analyze the behavior of the  $I_1$ ,  $I_2$ , and  $I_3$  measures, with the following considerations for each image:

- The color RGB is filtered by the luminance function  $Y_{709} = 0.2126R + 0.7152G + 0.0722B$ .
- The luminance histogram has 256 bins.
- The information channel is based on a MI ratio of  $\frac{1}{3}$ .

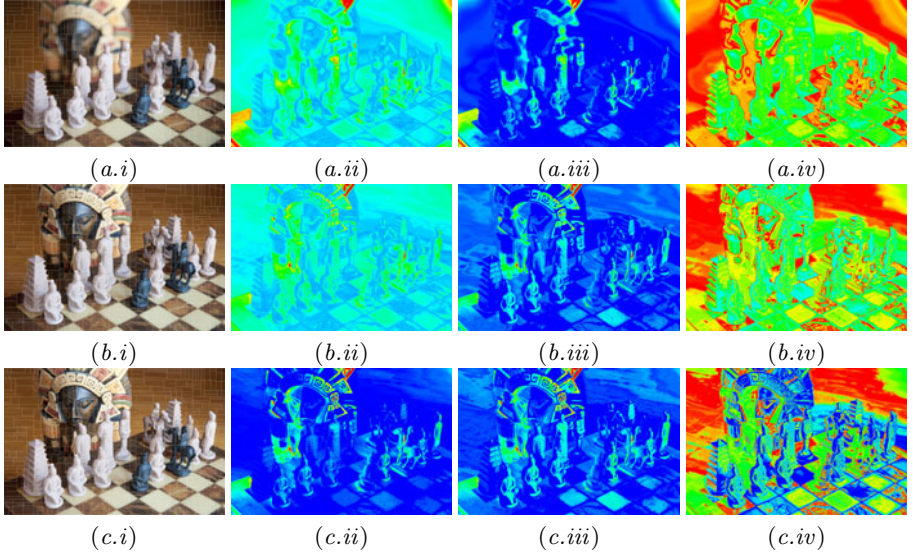




**Fig. 2.** (a.i) Cadaqués' man [f/8, 1/15, ISO400]. (b.i) Image decomposition (854 regions,  $H(C) = 7.839$ ,  $I = 2.613$ ) of (a.i). (a.ii-iv)  $I_1$ ,  $I_2$ , and  $I_3$  maps from the channel  $R \rightarrow C$ . (b.ii-iv)  $I_1$ ,  $I_2$ , and  $I_3$  maps from the channel  $C \rightarrow R$ .

- The color and region information maps are shown using a thermal-scale  (i.e., the lowest intensity corresponds to the blue and the highest to the red).
- The outliers are defined outside  $\mu \pm 3\sigma$  (i.e., three-sigma rule: for a normal distribution, nearly all values, 99.7%, lie within 3 standard deviations of the mean).

We can observe the behaviour of our measures in Fig. 2. The channel  $R \rightarrow C$  is shown in the first row with (a.ii) saliency  $I_1$ , (a.iii) specific information  $I_2$ , and (a.iv) entanglement  $I_3$ . In the second row we have the channel  $C \rightarrow R$  with the same measures. In the region saliency map (a.ii), regions with a higher measure value are the ones with an average color far away from the average color in the image. These are regions with a low color probability, and hence a high saliency. More salient parts in the color saliency map (b.ii) are, by order, the clouds, sky, illuminated water, and foreground tree trunk. The border is clearly defined between illuminated and non-illuminated water. The region specific information map (a.iii) represents the predictability of a color given a region. Thus, sky colors are the most predictive ones, followed by illuminated water colors. The

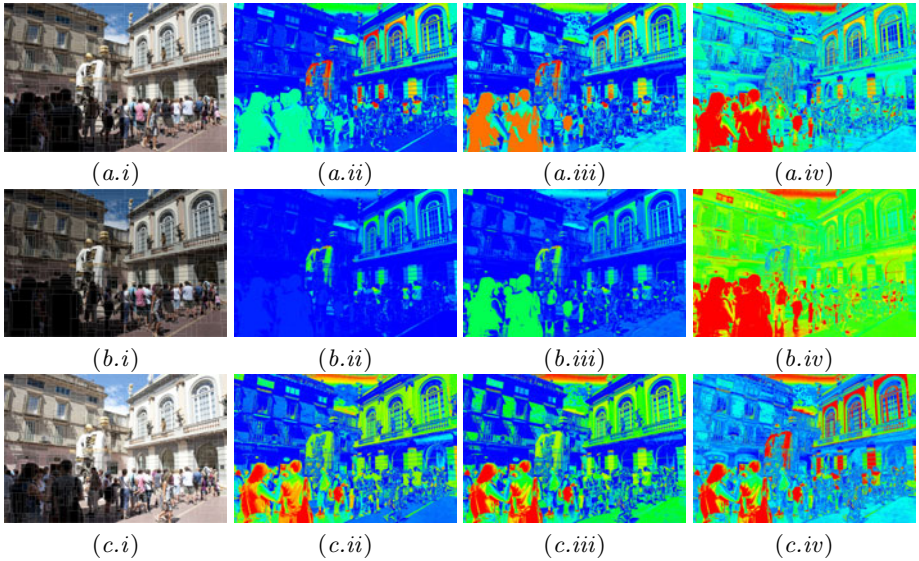


**Fig. 3.** Chess. (a.i) Image decomposition (489 regions,  $H(C) = 7.674$ ,  $I = 2.558$ ) [f/2.8, 1/160, ISO400]. (b.i) Image decomposition (657 regions,  $H(C) = 7.638$ ,  $I = 2.546$ ) [f/8, 1/20]. (c.i) Image decomposition (830 regions,  $H(C) = 7.683$ ,  $I = 2.560$ ) [f/16, 1/5]. (a-c.ii-iv)  $I_1$ ,  $I_2$ , and  $I_3$  maps from the channel  $C \rightarrow R$  in (a-c.i), respectively.

color specific information map (b.iii) gives us a detailed account of the image showing a more balanced range of values than the corresponding saliency map. Finally, observe that the behaviour of the entanglement is similar for regions and color (a-b.iv) showing a high correlation in the sky and the illuminated water with their corresponding colors, and a medium correlation in the tree trunk (most difference is caused by mapping the range to termic scale). In Fig. 2, we show another set of maps illustrating the behavior of the measures in a portrait.

We use Fig. 3 to comment the relationship of depth of field (DOF) with the information channel. In general, with high values of DOF, we need a higher number of regions to extract the same level of information because the image becomes more clear and sharper, i.e., it contains more information. On the contrary, with low values of DOF, the image is more blurred, and in general the number of region decreases due to the fact that there is less information to extract. This results in more defined and contrasted information maps for a higher DOF.

The interaction of exposure with our three measures is illustrated in Fig. 4. By overexposing a dark zone, we can uncover hidden information and the number of regions of the MI decomposition would increase. Otherwise, underexposing a burned zone, new information might appear and the number of regions would also increase. In an ideal case of exposure, under or overexposure might hide details and the number of regions would decrease. In all the cases, our measures reflect the changes in the exposure. We show in (a-c.ii) the color saliency maps for  $C \rightarrow R$  where the salient areas for different exposures can be compared. The color



**Fig. 4.** *Queuing in Dali's Museum, Figueres, Spain.* (a.i) Image decomposition (1,543 regions,  $H(C) = 7.594$ ,  $I = 2.531$ ) [f/8, 1/400, ISO200]. (b.i) Image decomposition (1,262 regions,  $H(C) = 7.060$ ,  $I = 2.353$ ) [underexposure 1/800]. (c.i) Image decomposition (1,880 regions,  $H(C) = 7.752$ ,  $I = 2.584$ ) [overexposure 1/200]. (a-c.ii-iv)  $I_1$ ,  $I_2$ , and  $I_3$  maps from the channel  $C \rightarrow R$  in (a-c.i), respectively.

specific information (a-c.iii) and entanglement maps (a-c.iv) are also depicted. Note for example how a lot of details appear in the left-bottom (people) and right-top (window) of the image when we overexpose the image (c.ii-iii), while the details of these areas disappear when we underexpose them (b.ii-iv). In the entanglement map, high correlations are shown in the case of overexposure (c.iv).

## 6 Conclusions

We have presented here three information-theoretic measures for saliency, specific information, and entanglement of luminance values and regions in an image. These measures extend previous work done on the study of artistic style in paintings, and are based on the information channel between colors and regions in the image, quantifying the correlation between color and compositional characteristics of the image. We have also shown how the information channel reflects changes in DOF and exposure. At this stage, we have only evaluated qualitative visual 2D results in order to show the behavior of these new measures and specially the informativeness associated to each color of the image. We believe that our measures represent an improvement in the understanding of the information contained in an image, and can have potential applications in several areas, as artistic style classification and image enhancement. Future work will be addressed to statistically analyze the results for a wide range of images and

different levels of image decomposition. Further work will be done to identify which approach ( $I_1$ ,  $I_2$ , or  $I_3$ ) is the most appropriate in any particular case or how the different results might be combined.

**Acknowledgments.** This work has been funded in part by grant number TIN2010-21089-C03-01 of the Ministry of Science and Technology (Spanish Government) and grant number 2009-SGR-643 of the *Generalitat de Catalunya* (Catalan Government).

## References

1. Bezzi, M.: Quantifying the information transmitted in a single stimulus. *Biosystems* 89(1–3), 4–9 (2007), selected Papers presented at the 6th International Workshop on Neural Coding
2. Bruce, N.D., Tsotsos, J.K.: Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision* 9(3), 1–24 (2009)
3. Butts, D.A.: How much information is associated with a particular stimulus? *Network: Computation in Neural Systems* 14, 177–187 (2003)
4. Cover, T.M., Thomas, J.A.: *Elements of Information Theory*. Wiley Series in Telecommunications (1991)
5. De Weese, M.R., Meister, M.: How to measure the information gained from one symbol. *Network: Computation in Neural Systems* 10, 325–340 (1999)
6. Gao, D., Han, S., Vasconcelos, N.: Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(6), 989–1005 (2009)
7. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. *Vision Research* 49(10), 1295–1306 (2009)
8. Itti, L., Koch, C.: Computational modelling of visual attention. *Nature Reviews: Neuroscience* 2(3), 194–203 (2001)
9. O’Sullivan, J.A., Blahut, R.E., Snyder, D.L.: Information-theoretic image formation. In: *Information Theory: 50 Years of Discovery*, pp. 50–79. IEEE Press, Los Alamitos (2000)
10. Rigau, J., Feixas, M., Sbert, M.: An information theoretic framework for image segmentation. In: *IEEE International Conference on Image Processing (ICIP 2004)*, vol. 2, pp. 1193–1196. IEEE Press, Los Alamitos (2004)
11. Rigau, J., Feixas, M., Sbert, M.: Informational dialogue with Van Gogh’s paintings. In: Brown, P., Cunningham, D.W., Interrante, V., McCormack, J. (eds.) *Computational Aesthetics 2008*. Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging, pp. 115–122. Eurographics Association (June 2008)
12. Wallraven, C., Cunningham, D., Rigau, J., Feixas, M., Sbert, M.: Aesthetic appraisal of art — from eye movements to computers. In: Deussen, O., Hall, P., Gibson, S., Hushlack, G., Shaw, J. (eds.) *Computational Aesthetics 2009*. Eurographics Workshop on Computational Aesthetics in Graphics, Visualization and Imaging, pp. 137–144. Eurographics Association (May 2009)
13. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision* 8(7), 1–20 (2008)

# Automatically Detecting Protruding Objects When Shooting Environmental Portraits

Pei-Yu Lo, Sheng-Wen Shih, Jen-Chang Liu, and Jen-Shin Hong

Department of Computer Science and Information Engineering,  
National Chi Nan University,  
No. 1, University Rd., Puli, Nantou, 54561, Taiwan  
hisopenny@gmail.com, {swshih, jcliu, jshong}@ncnu.edu.tw

**Abstract.** This study proposes techniques for detecting unintentional protruding objects from a subject’s head in environmental portraits. The protruding objects are determined based on the color and edge information of the background regions adjacent to the head regions in an image sequence. The proposed algorithm consists of watershed segmentation and KLT feature tracking model for extracting foreground regions, a ROI (Region of Interest) extracting model based on face detection results, and a protruding object detection model based on the color clusters and edges of the background regions inside the ROI. Experimental evaluations using four test videos with different backgrounds, lighting conditions, and head ornaments show that the average detection rate and false detection rate of the proposed algorithm are 87.40% and 12.11% respectively.

**Keywords:** Photo Composition, Protruding Object, Computational Photography.

## 1 Introduction

Beyond the lighting and chromatic aspects, it is well known that the composition, i.e., the arrangement of visual elements in the image frame, is also an essential aspect in the creation of quality photos. Although there are no absolute rules exist that ensure good composition in every context, there are various heuristic rules-of-thumb, such as “rule of third”, that help to achieve an aesthetic appealing photo composition when applied properly. Such rules are routinely applied as guidelines likely to increase the aesthetic appreciation of photographs.

Aiming to develop advanced intelligent digital cameras, there have been commercial interests these years to develop digital still camera with “composition advising” functions (e.g., [2][3]). A number of research studies have also devoted towards this goal. For example, [14] developed an intelligent system that positions the features of interest in an automatic robot camera using the rule of thirds. [1] developed computational models of photographic aesthetics and a system that aids the user to select the optimal composition of a given scene. [13] developed computational models for visual balance/symmetry for photos overlaid with texts. Overall speaking,



incorporating composition advising functions in a digital camera often require sophisticated visual object recognition and aesthetic computing techniques.

Beyond those widely-applied photo composition rules, there are also certain common photographic “mistakes” that may degrade the aesthetic appeal of photos, including tilted horizon, unintentional dissection lines, unintentional amputation, protruding objects from a subject’s head, unwanted distracting objects in a scene, etc. Avoiding these mistakes is particularly critical when taking environmental portraits, which often focus both on the main subject and on their surroundings backgrounds that provide more character to the subject. A protruding object in an environmental portrait usually refers to objects such as trees, street lights, windows frames or steeples which protrude abruptly from the head regions. Examples of environmental portraits with unintentional protruding objects are shown in Fig. 1. Certainly, it is preferred if the camera can automatically cut down the harshness of the protruding objects by using a smaller depth of field to lessen its impact, or simply provide warning messages to advice the photographer to eliminate the object by moving around the camera.



**Fig. 1.** Examples of an unintentional protruding object in an environmental portrait

In line with [4], aiming to develop intelligent composition-advising functions to avoid common photography compositional mistakes, this study aims to develop algorithm for the detection of protruding objects in environmental portraits. This study focuses on applications where the camera is mounted a tripod and the vibration of human hand is not concerned. The rest of this paper is organized as follows. Section 2 describes algorithms for automatically detecting protruding objects. Section 3 describes experimental results conducted to evaluate the performance of the proposed algorithms. Conclusions are given in Section 4.

## 2 Algorithm for Protruding Object Detection

From a single image frame, it is difficult to develop algorithms to distinguish between a protruding object which always stays in the background and head ornaments moving along with the subject’s head (e.g., the hat shown in Fig. 1(b)). Therefore, it is favorable to develop such algorithms based on sequence of images. This study assumed that the photographer previews and records an image sequence using a video camera when taking photos. The camera is assumed to be stationary (as is hold in a

tripod) such that the background region changes merely slightly and slowly as compared to the subject’s motions in the scene. In this way, information of the subject’s motion can be conveniently applied to segment the foreground regions which refer to the subject together any objects, such as hat or adornments, moving alongside with the subject’s head. After the foreground regions and the ROI (Region of Interest) is extracted, the protruding objects can be determined based on the color and edge information of the background regions inside the ROI.

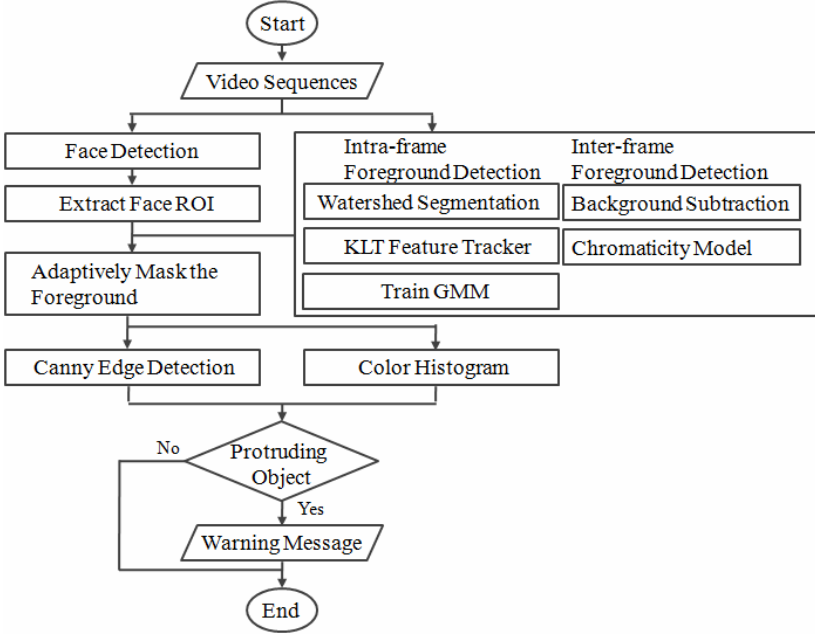


Fig. 2. Flowchart of the protruding object detection algorithm

The proposed protruding object detection system consists of three major modules:

- 1) Foreground region detection module: for extracting candidate foreground regions in the images.
- 2) Face detection module: for extracting ROI based on the face region detected by face recognition techniques.
- 3) Protruding object classification module: for classifying protruding objects based on the edge numbers and color clusters of the background regions inside the ROI.

The flowchart of the algorithm is shown in Fig. 2 and will be elaborated in the following subsections.

### 2.1 Foreground Region Extraction

Overall, an intra-frame and an inter-frame foreground region detection processes are developed and integrated for reliably detecting foreground regions. Assuming a

stationary camera setting, the Watershed segmentation algorithm [6][7] and KLT (Kanade-Lucas-Tomasi) feature tracking algorithm [8][9] can be applied to detect moving image blocks and thereby determine the background regions and the intra-frame foreground regions. In addition, an adaptive background subtraction method [10] is incorporated to further improve the detection rate. The background subtraction method can automatically develop a self-update reference background model to determine the inter-frame foreground regions. Since the background subtraction uses more than one image frames to determine the foreground regions, it is referred to as the “inter-frame” foreground detection process. Details of these detection procedures are elaborated in the following.

### 2.1.1 Intra-frame Foreground Region Detection

The intra-frame foreground detection process consists of the following three procedures.

1) Watershed algorithm is first applied to segment images. A pre-processing Gaussian filter is used to smooth images in order to avoid over-segmentation caused by watershed algorithm. In addition, a mathematical morphology filter is applied for post-processing segmented images to deal with cluttered scenes. The segmented regions with the same watershed label are drawn in the same color as shown in Fig. 3(a).

2) KLT feature tracking algorithm is applied to detect favorable feature points at corners or edges of objects in the image. Motions of features in an image stream are calculated based on these feature points. We extract pixels with 5×5 masks centering the feature points with the same watershed label as the candidate foreground regions. Examples of KLT features and the candidate foreground regions are shown in Fig. 3(b) and 3(c) respectively.

3) The foreground colors are modeled using a Gaussian Mixture Model (GMM) with five Gaussian components. The mean and the standard deviation parameters  $[\mu_{gr}^i \mu_{gg}^i \mu_{gb}^i \sigma_{gr}^i \sigma_{gg}^i \sigma_{gb}^i]$  of the foreground colors in Gaussian component  $i$  for each R, G and B channel are estimated. The foreground color model is constantly updated in real time by using simple recursive updates [10]. In practice, since computing the GMM probability of every pixel in the image is rather time consuming, a simplified method is adopted in this work to speed the foreground detection process. In the simplified method, the foreground color model is used to segment the intra-frame foreground regions based on the following criterion.

$$F_{intra}(x) = 1, \text{ if } |x_c - \mu_{gc}^i| < 3 \cdot \min(\sigma_{gc}^i, \sigma_{cam}), c \in \{r, g, b\}, \quad (1)$$

where  $x$  is the current pixel to be compared to the model,  $i$  is the index of a component of the GMM, and  $\sigma_{cam}$  is the variance of the camera noise. If any color channel of a pixel fits either one of the components of the GMM, it is regarded as a foreground pixel. In practice, for a stationary camera setting, it is not necessary to train GMMs frame by frame because typically there is no significant change between adjacent image frames when taking environmental portraits.





**Fig. 3.** (a) Watershed segmentation result; (b) KLT feature points; (c) extracted foreground regions following watershed segmentation and KLT feature point extraction

**2.1.2 Inter-frame Foreground Region Detection**

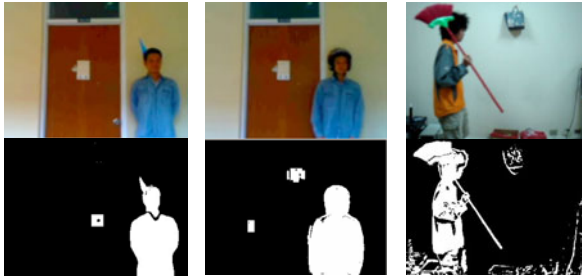
Since the above intra-frame foreground detection algorithm may often falsely include various static pixels in the background regions, an adaptive background subtraction method [10] is applied in this study to further improve the detection accuracy. The adaptive background subtraction model applies both RGB color model and chromaticity model. Given the means and variances of the RGB model denoted by  $[\mu_r \ \mu_g \ \mu_b \ \sigma_r \ \sigma_g \ \sigma_b]$ , and means and standard deviations of the chromaticity model denoted by  $[\mu_{r_c} \ \mu_{g_c} \ \sigma_{r_c} \ \sigma_{g_c}]$ , the adaptive background subtraction model is calculated according to the following:

$$F_{rgb}(x) = 1, \text{ if } |x_c - \mu_c| > 3 \cdot \max(\sigma_c, \sigma_{cam}), c \in \{r, g, b\}, \tag{2}$$

$$F_{chroma}(x) = 1, \text{ if } |x_c - \mu_c| > 3 \cdot \max(\sigma_c, \sigma_{cam}), c \in \{r_c, g_c\}, \tag{3}$$

where the chromatic values are computed as  $r_c = \frac{r}{r+g+b}$  and  $g_c = \frac{g}{r+g+b}$ , respectively.

Using  $F_{rgb}$  and  $F_{chroma}$  to remove falsely detected static pixels in  $F_{intra}$  and patch pixels with obvious chromaticity change, the final fused foreground region  $F$  is calculated by  $F = F_{intra} \cap (F_{rgb} \cup F_{chroma})$ . Results of the proposed foreground detection algorithm to drive complex silhouettes are shown in Fig. 4. Notably, a sudden change of environmental illumination may lead to false detections as shown in the second column of Fig. 4. However, integrating the intra-frame and inter-frame foreground detection results can reduce the false detections.



**Fig. 4.** First row: original images. Second row: foreground detection results.

### 2.2 Estimating the Region of Interest (ROI)

In principle, the ROI for detecting objects protruding across the head region of the subject can be approximated based on the subject’s face region. Viola-Jones face detector [11] is applied in this study to estimate the face region. The output of the face detector is a list of rectangular regions circumscribing the detected potential face regions. The ROI for protruding objects is specified as a rectangular region slightly larger than the face region (as shown in Fig. 5a). As such, potential protruding objects adjacent to both top and side regions around the head are properly attended to. An example of the face detection result and the corresponding ROI are shown in Fig. 5.



**Fig. 5.** (a) ROI template; (b) the face detection result is circumscribed by a green rectangular; (c) the corresponding ROI for protruding object detection is circumscribed by a yellow rectangle

### 2.3 Estimating the Protruding Objects

In general, a ROI of an image frame with protruding objects should have more edges and color clusters than ROIs of adjacent frames without any protruding objects. Therefore, after the foreground regions and the ROI are extracted, whether there is a protruding object is determined based on the color and edge information of the background regions inside the ROI. To get the background regions inside the ROI, the foreground region is first masked on the ROI. Further, mimicking the typical shape of a face, an elliptic mask which is centered on the detected face region is used to compensate possible fragile foregrounds computed.

For obtaining the edge features, Canny edge detector [12] is applied to detect the edges inside the ROI. Define  $E_{pixs}$  as the number of edge pixels in the current frame  $t$ , and  $E_{pre\_pixs}$  as the number of edge pixels in the previous frame ( $t-1$ ). For each image frame, an adaptive base value of the number of edge pixels, denote by  $E_{base}$ , is used for estimating the likelihood of existence of a protruding object. Heuristically,  $E_{base}$  is constantly updated every 5 frames as follows:

$$E_{base} = \begin{cases} E_{pixs}, & \text{if } E_{base} = 0 \\ \min(E_{pixs}, E_{base}), & \text{if } E_{pixs} - E_{pre\_pixs} > 0 \\ (E_{pixs} + \max(E_{pixs}, E_{base}))/2, & \text{otherwise} \end{cases} \quad (4)$$

The value of  $E_{base}$  represents an estimate of the number of edge pixels inside the face region. When the number of edge pixels inside the ROI increased, it signifies that the face region of the subject may be approaching a protruding object. Therefore, an

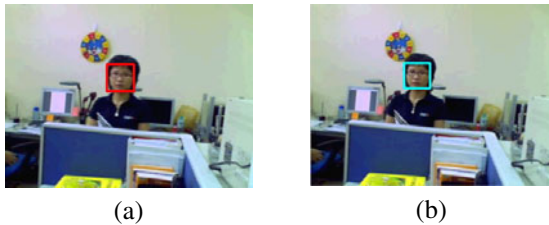
overestimate of  $E_{base}$  will increase the protruding object detection rate. At the beginning,  $E_{base}$  is initialized as zero so that it will be updated as  $E_{pixs}$  of the first image frame according to equation (4). When the number of edge pixels is increasing (i.e.,  $E_{pixs} - E_{pre\_pixs} > 0$ ) and  $E_{base}$  is greater than  $E_{pixs}$ , it means that the number of edge pixels in the ROI is overestimated and should be reduced. Likewise, when  $E_{pixs}$  is decreasing and  $E_{base}$  is greater than  $E_{pixs}$ ,  $E_{base}$  has to be reduced too. However, if  $E_{pixs}$  is decreasing and  $E_{base}$  is smaller than  $E_{pixs}$ , then it may indicate an underestimate of  $E_{base}$ . Therefore, when  $E_{pixs}$  is decreasing, update equation is defined as shown in equation (4).

For obtaining the color features, the color ROI image is quantized into 16 bins to be able to effectively compute the number of colors in subsequent processes. Given the number of ROI pixels other than the foreground and elliptic mask as  $color_{all}$ , and the number of pixels in each bin as  $color_i$ ,  $i \in [1, 16]$ , we increase the number of color clusters  $cluster$  by one if  $(color_i / color_{all} > 0.2)$  is true.

The system then determines whether the object is a protruding object or not according to the number of edge pixels and color clusters based on the following criterion:

$$\text{protruding object} = \text{true, if } (cluster \geq 2) \cap \left( \frac{E_{pixs}}{E_{base}} > th \right). \quad (5)$$

The heuristic threshold value  $th$  was set to 1.2 with which the obtained average detection rates in our initial evaluation experiments appeared to be satisfactory. Example frames with detected protruding objects are shown in Fig. 6.



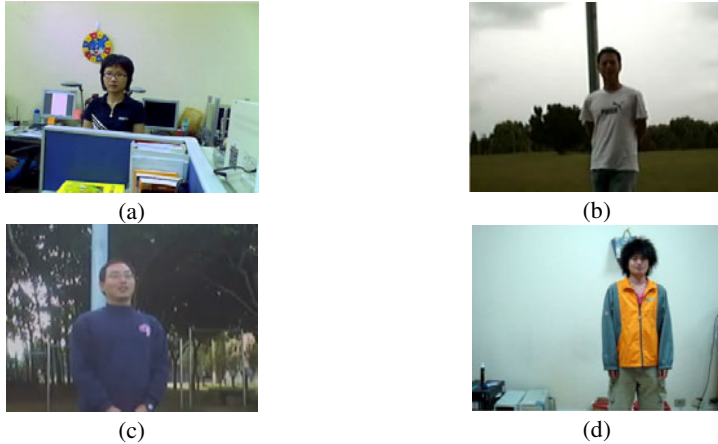
**Fig. 6.** Two example video frames showing detected protruding objects. The face region is circumscribed by a rectangular.

### 3 System Evaluations and Analysis

The proposed protruding object detection algorithm was tested on four video sequences shot with cameras mounted on a tripod. The background in each video changes slowly relative to the motions of the main subject in the scene. Section 3.1 describes the testing data set. Evaluation experiments and effectiveness analysis are presented in Section 3.2 and Section 3.3 respectively.

### 3.1 Data Set

We recorded four video sequences in a variety of scenes as shown in Fig. 7. The videos are designed to evaluate the performance of the proposed algorithm in scenes with different backgrounds, lighting conditions, and large-size head ornaments.



**Fig. 7.** Four different test scenes: (a) indoor; (b) outdoor with a pure background; (c) outdoor with a cluttered background; (d) subject with afro hair.

### 3.2 Evaluations

The performance of the proposed protruding object detection algorithm was evaluated by comparing the system outputs with the ground-truth data using the four performance measurements listed in Table 1.

**Table 1.** Four performance measurements applied in the system evaluation experiments

	Image Frames with Protruding objects	Image Frames w/o Protruding objects
Detected	True Positive ( <i>TP</i> )	False Positive ( <i>FP</i> )
Non-detected	False Negative ( <i>FN</i> )	True Negative ( <i>TN</i> )

The detection rate (*DR*) and the false detection rate (*FDR*) are calculated respectively based on the following formulas:

$$DR = \frac{TP}{TP + FN}, \quad FDR = \frac{FP}{TP + TN}. \quad (6)$$

Since the protruding object detection is based on the face detection results, ( $TP+TN$ ) is the number of frames with face detected in the video sequence. The average detection rate and false detection rate of these four different test scenes are denoted as  $DR_{all}$  and  $FDR_{all}$ . Evaluation results are presented in Table 2. The results show that, the detection rate and false detection rate for the four test video ranges from 74.00% to 95.24% and from 0% to 27.59%, respectively.

**Table 2.** Evaluation results for different videos

	Indoor scene	Outdoor scene with simple background	Outdoor scene with cluttered background	Subject with Afro hair
<i>DR</i>	90.91%	74.00%	95.24%	89.47%
<i>FDR</i>	17.95%	0.00%	2.90%	27.59%
<i>DR<sub>all</sub></i>			87.40%	
<i>FDR<sub>all</sub></i>			12.11%	

## 4 Conclusions and Future Works

This study applies computer vision and image processing techniques to develop an intelligent composition-advising function for automatic detection of protruding objects when shooting environmental portraits. The protruding object detection system consists of a foreground region detection module, a face detection module and a protruding object classifier. Experimental evaluations show that the detection rate and false detection rate of protruding objects in the test videos are around 88% and 12%, respectively. Ongoing works are currently underway to improve the current techniques with further concerns on the vibrations of hand-held cameras.

**Acknowledgements.** This study is supported by the National Science Council of Taiwan, under Grant No. NSC 99-2410-H-260-052 and NSC 99-2221-E-260-028.

## References

1. Liu, L., Chen, R., Wolf, L., Cohen, D.: Optimizing Photo Composition. *Computer Graphics Forum* 29-2, 469–478 (2010)
2. Miyake, T., Soga, T.: Digital Still Camera with Composition Advising Function, and Method of Controlling Operation of Same. Fujifilm Corporation, United States Patent, Patent Number: 7317458 (2008)
3. Suarez, L.A.F.: Picture Composition Guidance System. Sony Corporation, Sony Electronics Inc., United States Patent, Patent Number: 5873007 (1999)
4. Shen, C.T., Liu, J.C., Shih, S.W., Hong, J.S.: Towards Intelligent Photo Composition-Automatic Detection of Unintentional Dissection Lines in Environmental Portrait Photos. *Expert Systems with Applications* 36, 9024–9030 (2009)
5. Cavalcanti, C., Gomes, H., Meireles, R., Guerra, W.: Towards Automating Photographic Composition of People. In: *IASTED International Conference on Visualization, Imaging, and Image Processing*, pp. 25–30 (2006)
6. Vincent, L., Soille, P.: Watersheds in Digital Spaces: An Efficient Algorithm Based on Immersion Simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 583–598 (1991)
7. Meyer, F.: Color Image Segmentation. In: *International Conference on Image Processing and its Applications*, pp. 303–306 (2002)
8. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features. Carnegie Mellon University, Technical Report CMU-CS-91-132 (1991)
9. Shi, J., Tomasi, C.: Good Features to Track. In: *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 593–600 (1994)

10. McKenna, S.J., Jabri, S., Duric, Z., Wechsler, H., Rosenfeld, A.: Tracking Groups of People. *Computer Vision and Image Understanding* 80, 42–56 (2000)
11. Viola, P., Jones, M.: Rapid Objects Detection using a Boosted Cascade of Simple Features. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 511–518 (2001)
12. Canny, F.J.: A Computational Approach to Edge Detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 8, 679–698 (1986)
13. Lai, C.Y., Chen, P.H., Shih, S.W., Liu, Y., Hong, J.S.: Computational Models and Experimental Investigations of Effects of Balance and Symmetry on the Aesthetics of Text-Overlaid Images. *International Journal of Human Computer Studies* 68(1-2), 41–56 (2010)
14. Byers, Z., Dixon, M., Smart, W.D., Grimm, C.: Say Cheese! Experiences with a Robot Photographer. *AI Magazine* 25(3), 37–46 (2004)

# Artist-Led Suggestions towards an Approach in Content Aware 3D Non-photorealistic Rendering

Martin Constable

School of Art Design and Media, Nanyang Technological University, Singapore

**Abstract.** Referencing practice in traditional drawing, the author attempts to expand upon the knowledge landscape informing current approaches in 3D NPR rendering and thereby to indicate possible areas for fruitful enquiry. The author presents three examples of drawing practice: incomplete perimeters, lines that suggest form and lines that suggest color. Each case is accompanied by examples of drawings from modern or pre-modern artists. A need for a content-aware approach to rendering is indicated. Informing this enquiry is the fact that the author taught drawing for 20 years before working with computer engineers on 3D NPR rendering.

## 1 Introduction

In his ongoing collaboration with the computer graphics engineers of Nanyang Technological University the author has been fascinated and humbled by their examination of drawing from a fundamental point of view.

Drawings are objects that have been left behind by the process of their manufacture and they do not lend themselves easily to being reverse engineered. Learning the rules of drawing by examining these artifacts is a bit like learning chess by staring at a chess board.

Artists can help with this investigation. They can illuminate a drawing with insights into the process behind its manufacture. Some aspects of drawing practice are natural such that even a child could grasp them. Others are entirely unnatural in that they can only be known if they have been taught. The biggest thing that an artist who has been trained in the western classical manner has to learn is how to move the form of the drawing from the flat of the paper into the illusional volume of the picture space.

In this paper some of these formal aspects are examined: incomplete perimeters, lines that suggest form and lines that suggest color are examined. These three aspects of drawing practice have been chosen because they all require an understanding of drawing informed by a 3D spatial context.

## 2 Lines That Describe Incomplete Perimeters

The profile or 'outline' of an object is a formal attribute that we first encounter as children when we are moving from what G. H. Luquet describes as the scribbling

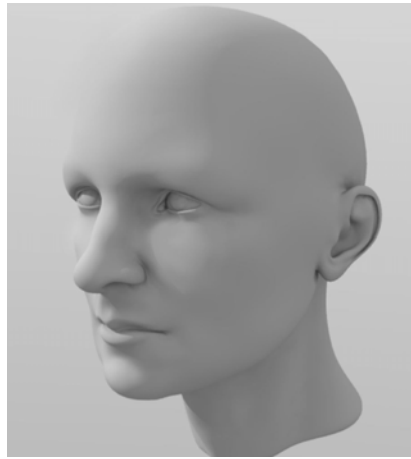


Scribbling stage (age 1 yr)



Schematic stage (age 6 yrs)

**Fig. 1.** From the Scribbling stage to the Schematic stage: the development of an awareness of perimeter in children's drawings as outlined by the work of G H Luquet



**Fig. 2.** 3D rendering of a head showing an invisible edge where the foreground and background are of the same lightness value

stage of drawing to the schematic (Fig. 1). Having discovered the perimeter as children, novice artists are very reluctant to let go of it and will draw it even when it is clearly not there.

Consider the model of the head in Fig. 2: because the forehead and cheek areas of the head are the same lightness value as the wall against which they are situated, their profile is invisible. However, the author has yet to encounter a drawing student who does not complete the profile whether it is apparent or not.

An experienced artist will often break or soften a profile to the point where it is very faint or altogether invisible. In Fig. 3 the far side of the sitter's face is barely discernable. Besides being (perhaps) a true response to the light conditions at the time of the portrait sitting, this strategy also serves an aesthetic purpose. Broken lines within a drawing's perimeter can create a sense of air flowing through the drawing and avoid the impression of flatness that a drawing with a complete perimeter often has.





**Fig. 3.** *'Portrait of the Fox Madox Brown'* (detail), Rossetti Dante, 1860 (showing barely discernible perimeter line of far side of face)

Another case is presented in Fig. 4 where the artist has drawn into the background that immediately borders the face. He has done this so that the relative lightness of the face to the background is apparent. Though this drawing looks very different from Fig. 3, we can see the same loss of perimeter where the darkness under the nose meets the darkness of the background. This loss of perimeter forces the volume of the head into the background and thereby increasing the sense of form. As a comparison Fig. 5 present what that portion of the drawing would look like were the perimeter complete.

## 2.1 Proposal: A Light Aware NPR

A 3D NPR rendering algorithm draws a perimeter line round an object that is invariably derived from a straightforward relationship between the camera and the geometry of the object. However, as has been demonstrated, this perimeter is not sacrosanct to drawing and its loss can be advantageous to a spatial reading of the form.

In the paper *'Coherent Stylized Silhouettes'* [1] Robert D. Kalnins and Philip L. Davidson et. al. describe a way to render stylized silhouettes. Using this approach broken perimeter lines are possible.

However, unlike the broken perimeters shown in figures Fig. 3 and Fig. 4 their result does not derive from the position of the light, nor the light value



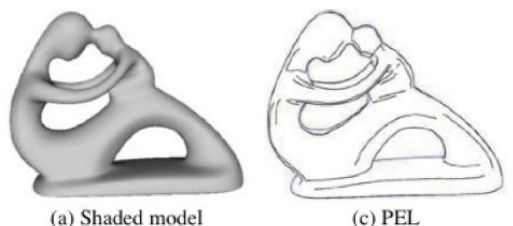
**Fig. 4.** Detail of *'The Architect Andr -Marie Chatillon, Jean-Auguste-Dominique Ingres, 1860*



**Fig. 5.** Fig. 4 with the perimeter of the nose completed by the author

relationship between the background and the foreground. The lines that they produce are a stylistic veneer placed on top of the form geometry, though no less interesting for it.

The author proposes a 3D NPR rendering algorithm that take a lead from the artist's awareness of relative lightness values between the foreground and the background. A prime need of such a shader is that it must be aware of where the light is. In their paper *'An Effective Illustrative Visualization Framework Based on Photic Extremum Lines (PELs)'* [2] Xuexiang Xie, Ying He et. al. have done productive work on this subject (Fig. 6). However, their rendering algorithm is still done in ignorance of the relative light value relationships between the object and its environment.

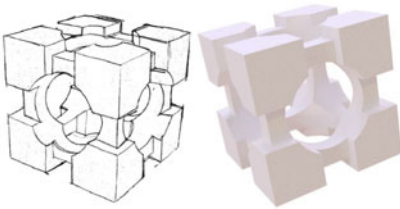


**Fig. 6.** Showing lines rendered in response to the lighting conditions using the procedure outlined by Xuexiang Xie, Ying He et. al. in *'An Effective Illustrative Visualization Framework Based on Photic Extremum Lines (PELs)'*. Notice how the lines are following the form of the shadows, not just the geometry.

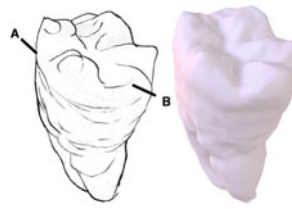
## 2.2 Lines That Suggest Form

In their study *'Where Do People Draw Lines?'* [3] Forrester Cole et. al. showed that there was a reasonably straightforward correspondence between the lines that a person draws when depicting a piece of regular, hard-edged geometry and the exterior and interior occluding contours of that geometry (Fig. 7).

However, this linear relationship is not sustained when the geometry is less regular and more organic. In such instances some of the lines were suggestive of the form rather than descriptive (Fig. 8). In drawing and painting practice suggestion is usually used when the detail of a form is too complex to depict. In *'Programmable Rendering of Line Drawing from 3D Scenes'* by Stephane Grabli and Emmanuel Turquin this quality is called indication and is described as resulting from a pruning of detail.



**Fig. 7.** Drawing (left) and 3D model geometric style source shape (right) from *Where Do People Draw Lines?*, Forrester Cole et. al.



**Fig. 8.** Drawing (left) and 3D model organic style source shape (right) from *Where Do People Draw Lines?*, Forrester Cole et. al. showing depiction in line (A) and suggestion in line (B).

In the drawing *'The Rocks'* (detail Fig. 9) by Van Gogh the profile of the foreground against the sky is clearly visible. As a descriptor of the form it functions efficiently. However, the marks within the drawing that represent rocky ground (Fig. 10) have a more suggestive relationship with form and look almost abstract up close.

**Proposal: A Suggestive Line Multi-Object NPR.** Suggestive lines are not new in NPRs. In *'Suggestive Contours for Conveying Shape'* [4] M. A. Kowalski et. al. describe a way to combine contours and suggestive contours to enable an NPR render to draw complex undulations in form.

However, its success is limited to the rendering of single objects. It would therefore only work if the many rocks and plants depicted in the Van Gogh drawing detailed in Fig. 10 were a single object.

The author proposes a 3D NPR rendering algorithm that can be applied to many objects yet react as a single thing.

In the case of a small drawing of a large crowd of people where each person is a separate piece of geometry (Fig. 11), the figures are small enough in relation to the size of the drawing and are numerous enough to appear as a single mass.



**Fig. 9.** Detail of 'The Rocks', Vincent Van Gogh, 1888. Note the clear perimeter lines.



**Fig. 10.** Detail of 'The Rocks', Vincent Van Gogh, 1888. Note the suggestive lines.



**Fig. 11.** Simulated result from a Suggestive Line Multi-Object NPR showing complete perimeter round general mass of figures and suggestive lines within the perimeter. Also showing complete perimeter around an isolated figure.

The proposed algorithm would draw a collection of lines inside of this mass that suggest the people without depicting them individually. However, the perimeter of the mass would be drawn as a complete line. If a single figure walked away from the crowd, then the algorithm would change strategy to draw the entirety of this particular figure's perimeter.

### 2.3 Lines That Suggest Color

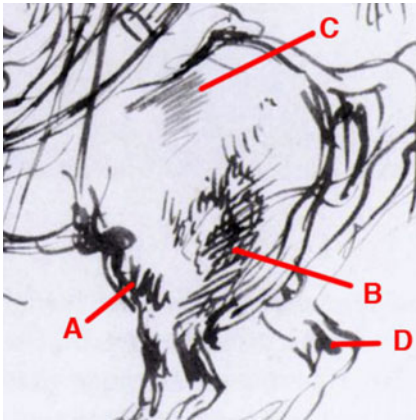
The texture of a line can be varied in many ways depending on the medium involved. The lines left behind by a pencil can be changed by dragging the pencil sideways, using it fast, using it slow, pressing it down hard, using it blunt, using it sharp etc. However, the lines of novice artists are usually simple and 'binary' in their nature (i.e. line/not line) and do not vary much in their lightness or texture.

To address this issue the author would set his students the task of drawing from imagination the difference between two objects that are identical in every way except their hue.

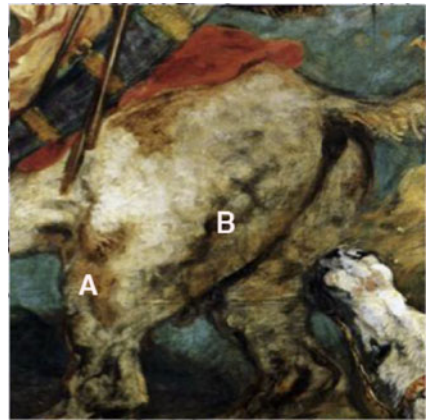
In order to successfully complete this exercise the lines must be able to convey values of difference other than just lightness.

To help them he would show them the drawings of great colorists like Bonnard or Delacroix whose drawings both display a clear correspondence between line and color. In the study (Fig. 12) for the mural 'Attila, Followed by the Barbarian Hordes, Trample on Italy and the Arts' Delacroix has varied the lines around the back of the horse. This variation is more than just tonal. They vary in the following ways:

- Pen pressing down on the paper hard and vertically<sup>1</sup> to produce a wide mass, with a wet and heavily pigmented ink load
- Sharp, diagonal masses of lines with a quite dry and very heavily pigmented ink load
- Diagonal lines<sup>2</sup> with a faint watery ink load
- Single marks, wet load and low pigmentation



**Fig. 12.** Study for *Attila, Followed by the Barbarian Hordes, Trample on Italy and the Arts*, Eugene Delacroix, 1843-1847.

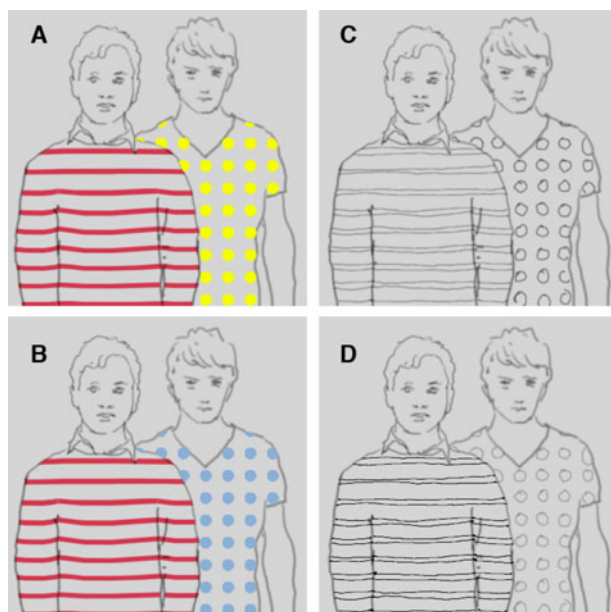


**Fig. 13.** Annotated detail of *Attila, Followed by the Barbarian Hordes, Trample on Italy and the Arts*, Eugene Delacroix, 1843-1847: with A being a warm area of color and B being a cool area.

If we examine a detail of the painting (Fig. 13) we can see that the warm, burned umber of flank area A has been contrasted with the cold, raw umber of flank area B. This instituting of a contrast along the warm/cool axis of a painting

<sup>1</sup> A vertical pen line produces a wide mark.

<sup>2</sup> A diagonal pen line produces a thin mark.

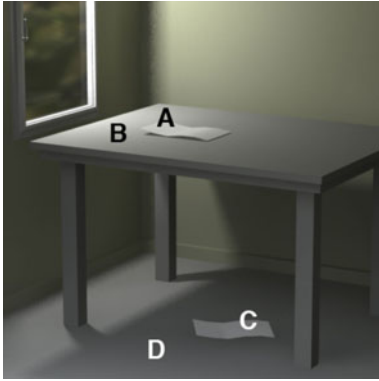


**Fig. 14.** The red stripes in images A and B are identical in HSL value yet they register differently in the relative hierarchies of the images. The drawings C and D clearly preserves this difference.

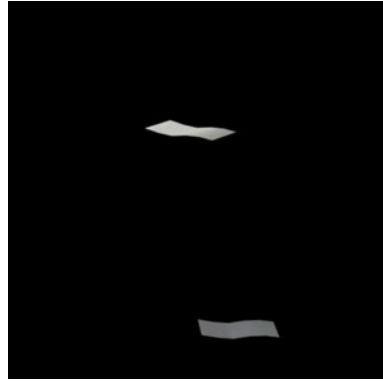
is a classic ploy to that has been used by artists for hundreds of years. There is a clear correspondence between these areas in the painting and the line masses A and B in the preparatory drawing. In both the drawing and the painting these areas are approximately the same lightness value and hold their difference through a value other than lightness. In the drawing it is the sharpness of the lines that distinguish them from each other, in the painting it is the temperature of the color.

**Proposal: Lines for Colors.** In 'Programmable Rendering of Line Drawing from 3D Scenes' a [5] technique is described that enables the automatic calculation of lines that are the same color as that of the material. However, this is a literal 'like for like' approach that does not take advantage of drawing's strength as a vehicle of symbolic signification.

At its simplest a color aware 3D NPR rendering algorithm might render lines as simple linear equivalents for color (e.g. red = sharp and thin line). However, a more complex approach would be for it to be 'aware' about color within a drawing as a set of relative relationships that express the three values: hue, saturation and lightness in a triangular and dynamic relationship to each other. Such a 3D NPR render would not depict an absolute relationship between a color and a drawing mark but a relative one. In Fig. [14] the red stripes in A and B are identical. However, their relative position in the color hierarchies of



**Fig. 15.** The relative white of the piece of paper A against the piece of paper B.



**Fig. 16.** The sheets of paper from Fig. 15 isolated against black.

the drawings are different. This difference is preserved in the drawings C and D, where the two reds are drawn in different ways.

### 3 Conclusion

Though light, detail and color have been covered separately in this paper, the one thing that unites them all is that artists consider these values not as relative but absolute. This relativity is often a complex and nested thing as illustrated in Fig. 15:

- The sheet of paper A is lighter than the tabletop B (local relation)
- The sheet of paper C is lighter than the floor D (local relation)
- The area beneath the table is darker than the area on top of the table (global relation)

A consequence of this last fact is that the sheet of paper C is darker than the sheet of paper A. This can be seen more clearly if the papers are separated as in Fig. 16. However, a novice is almost certain to depict both sheets of paper as pure white.

A notable point needs to be made to the novice student at this point: that all the values in a painting or a drawing exist relative to each other and that this relationship is complex. Furthermore, everything to which a value can be affixed is subject to this simple principle including, hue, texture, depth, detail, apparent movement, narrative elements etc. When teaching drawing, this primacy of considering all values relatively is an important enough principle to bring up on the first day of teaching, and to be re-iterating on the last.

A consideration of relative values can not be made without a consideration of the scene. In fact, a scene can be defined as being a set of interconnected

relative values and the 3D NPR strategies proposed in this paper all need to be aware of these interconnected values in order to function. It is proposed that for its similarity to the way that artists formulate their drawings, a content-aware approach presents an avenue for fruitful future realms of enquiry into 3D NPR rendering.

## References

1. Kalnins, R., Davidson, P., Markosian, L., Finkelstein, A.: Coherent stylized silhouettes. *ACM Transactions on Graphics (TOG)*, 133–138 (2003)
2. Xie, X., Ying, H., Feng, T., Seah, H.-S., Gu, X., Hong, Q.: An effective illustrative visualization framework based on photic extremum lines (PELs). *IEEE Trans. Visualization and Computer Graphics* 13, 1328–1335 (2007)
3. Cole, F., Golovinskiy, A., Limpaecher, A., Stoddart-Barros, H., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S.: Where do people draw lines? In: *International Conference on Computer Graphics and Interactive Techniques*, Article No. 88, pp. 185–196 (2008)
4. DeCarlo, D., Finkelstein, A., Rusinkiewicz, S., Santella, A.: Suggestive contours for conveying shape. *ACM Trans. Graphics (TOG)* 22, 848–855 (2003)
5. Stephane, G., Turquin, E., Durand, F., Sillion, F.: Programmable rendering of line drawing from 3d scenes. *ACM Trans. Graphics (TOG)* 29, 1–20 (2010)



# Ground Truth Evaluation of Stereo Algorithms for Real World Applications

Sandino Morales and Reinhard Klette

.*eneda.* group, Dept. Computer Science, University of Auckland, New Zealand

**Abstract.** Current stereo algorithms are capable to calculate accurate (as defined, e.g., by needs in vision-based driver assistance) dense disparity maps in real time. They have become the source of three-dimensional data for several indoor and outdoor applications. However, ground truth-based evaluation of such algorithms has been typically limited to data sets generated indoors in laboratories. In this paper we present a new approach to evaluate stereo algorithms using ground-truth over real world data sets. Ground truth is generated using range measurements acquired with a high-end laser range-finder. For evaluating as many points as possible in a given disparity map, we use two evaluation approaches: A direct comparison for those pixels with available range data, and a confidence measure for the remaining pixels.

**Keywords:** Performance evaluation, stereo algorithms, laser range finder.

## 1 Introduction

Vision-based stereo algorithms are designed to generate three-dimensional (3D) information from two-dimensional (2D) data recorded with two or more video cameras. State-of-the-art stereo algorithms are capable to perform in real-time “accurate” disparities for almost all the points visible in the input images. Current applications for stereo algorithms, among many others, are vehicle navigation (robots [17], forklifts [21], wheelchairs [20], and so forth) or industrial safety equipment [1].

We are interested in the evaluation of stereo algorithms in the context of vision-based Driver Assistance Systems (DAS) [11] for improving those techniques. DAS requires that the detection of depth is accurate on every road, under all kinds of weather conditions, and in any traffic context. Therefore, stereo algorithms need to be evaluated in the real-world, and not only on data representing a few seconds of recording but hours or days.

The evaluation of stereo algorithms is either based on ground truth data, allowing direct comparisons between true disparity values and those obtained with the algorithms; or it is performed in the absence of ground truth using various ideas for still ensuring some kind of objective testing. For real-world video data it is the ultimate goal to provide ground truth as well. Synthetic

---

<sup>1</sup> <http://www.pilz.com/products/sensors/camera/f/safeteye/>

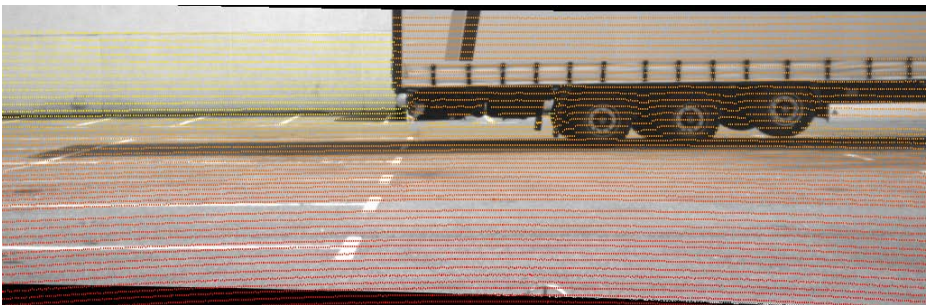
(i.e., computer generated stereo pairs) or engineered (i.e., images captured under highly controlled conditions, using structured light for generating ground truth) data do have their own characteristics [8], and do not cover the “challenges” as occurring in real-world data.

Real-world data do not come (typically) with ground truth. Therefore, diverse methods have been proposed to evaluate the algorithms even in absence of ground truth. In [1], the evaluation was done by measuring the number of successfully matched pixels using a left-right consistency check [9]. Some authors used an extra image (e.g., prediction error in [23]) or a third video sequence (see the third view in [15]) as ground truth. Confidence measures are another example of evaluation in the absence of ground truth [6,16]. The idea is to measure the reliability of the calculated disparity value for each pixel. Techniques, specifically designed for DAS, were proposed in [14,22]; these evaluation schemes can only be applied if some conditions are satisfied in the recorded scenes.

We generate ground truth using precise depth measurements acquired with a laser range-finder (LRF). The generation of ground truth (or of accurate 3D models) using LRF’s has been investigated before [2,10,17]. However, those publications do not report about the evaluation of stereo algorithms using laser range data. Stereo algorithms are discussed together with laser range data in [19] at selected feature areas.

The evaluation scheme in this paper analyzes stereo algorithms on recorded video sequences based on available ‘sparse’ (but uniformly distributed) ground truth and also applying a confidence measure for dealing with the ‘gaps’. We use Velodyne’s HDL-64E S2 range-finder [24]. For the distance interval of interest (about 5 to 120 m), the available accuracy is defined by possible errors of less than 10 cm (the producer even sees the error at 1.5 cm at most in 5 to 120 m).

The obtained range data are insufficient for evaluating an entire dense disparity map, e.g. a VGA image has  $640 \times 480 = 307,200$  pixels, and the used LRF generates up to 24,000 points in the field of view of the reference camera in our stereo set up; see Fig. 1. Thus, we combine two approaches for the evaluation. If ground truth data are available at a specific pixel, we perform a direct



**Fig. 1.** Sample image showing combined laser range-finder and image data. Ground truth points (i.e., points acquired with the laser range-finder) are color encoded from red (for close) to green (for further away).

comparison between the calculated disparity value and the ground truth. For the remaining points, we use a geometrical approach using “close” range readings to generate a confidence measure. This approach allows us to evaluate stereo algorithms for outdoor real-world data based on true measurements. Data sets can be recorded in all kinds of weather where the LRF will work in, or road conditions.

The main contributions of this paper are the measures proposed to evaluate dense algorithms against sparse (less than 10%) ground truth. The data provided contain sub-pixel accurate ground truth for real-world scenes, and this was not available prior to the use of a laser range-finder. This data set has been made publicly available for future research considerations, see [4].

The structure of this paper is as follows. In Section 2 we present the proposed approach. We continue with experiments in Section 3, and finalize with conclusions in Section 4.

## 2 Approach

We generate sparse ground truth disparity maps with the LRF, and perform the evaluation by fusing a direct comparison approach (where true values from the LRF were available) and a confidence measure (for the remaining points). See Fig. 2 for a flow chart of the proposed approach.

**Ground Truth Disparity Map Generation.** We record range data of the surrounding environment of the *ego-vehicle* (i.e. the vehicle carrying the stereo camera and the LRF) using a high-end LRF [24]. The provided accuracy data (precision of 1.5 centimeters within a range from one to 120 metres) needs to be slightly corrected, and 10 cm can be used as an upper bound in our experiments.

The rotational architecture of the LRF allows us to obtain readings from 64 lasers in a full 360° rotation. Its optimum resolution (depending on the rotational speed) is of 0.09° (horizontal) times 0.4° (vertical). The vertical field of view of 26.8° provides sufficient information for modeling the road and the objects that would be of interest in a driving scene.

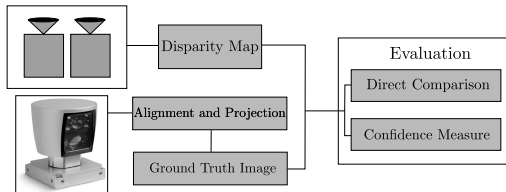


Fig. 2. Flow chart of the used approach

Assume for now that the coordinate systems of the LRF and the stereo camera have been calibrated and aligned. Then, we are able to project the output of the LRF (a set of 3D points) onto a 2D image  $G$  using the (internal) parameters of the stereo camera. The ground truth disparity value  $G(\mathbf{x})$  of a pixel  $\mathbf{x} \in G$  is defined by

$$G(\mathbf{x}) = \frac{f \cdot b}{Z(\mathbf{x})} \quad (1)$$

where  $f$  denotes the focal length of the stereo camera,  $Z$  the distance from the camera (at pixel  $\mathbf{x}$ ) in the depth direction, and  $b$  is the distance between the optical centers of the cameras (the length of the baseline). For pixels where there is no distance measure available, a distinctive negative value is assigned (as disparity values are strictly positive). For the images that we use for our experiments (i.e.,  $1024 \times 334 \approx 342,000$  pixels), we are able to obtain ground truth values for almost 7% (about 24,000) of the pixels. These are the only points we are able to perform a direct comparison.

In the context of DAS, the final goal is to analyze the performance of stereo (or any) algorithms in outdoor dynamic environments. Thus, it does not make sense to scan the same scene multiple times to get more range readings. Instead, we use the available measurements to generate a confidence measure to evaluate the remaining points.

**Direct Comparison.** Where range data is available we use the percentage of badly calculated pixels (BCP) as quality metric. Let  $D$  be a disparity map obtained with a given stereo algorithm, and  $G$  the generated ground truth image. Let  $\Omega$  denote the set of pixels in  $G$  and  $D$  such that  $G(\mathbf{x}) > 0$  (i.e., pixels with a valid measurement from the LRF) and  $D(\mathbf{x}) > 0$  (i.e. pixels with invalid disparities were also identified with a negative value). Let  $T$  be a predefined tolerance threshold, and

$$\delta(\mathbf{x}) = \begin{cases} 1, & \text{if } |G(\mathbf{x}) - D(\mathbf{x})| \geq T \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Then, the BCP of  $D$  is as follows:

$$B = \frac{100\%}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \delta(\mathbf{x}) \quad (3)$$

where  $|\cdot|$  denotes the cardinality of a set.

**Confidence Measure.** To complement the direct comparison (i.e., to evaluate also points where no range data are available), we use a simplified version of the approach presented in [3]. In that paper, the authors used a probabilistic scheme to deal with non organized point clouds generated by a LRF of small objects under controlled conditions (i.e., indoor scenes).

Given three “close” pixels in the ground truth image  $G$ , we define a patch  $P_G \subset G$  and its 3D version  $\overline{P_G}$  by back projecting the three pixels into the 3D space. Using the corresponding pixels in the disparity map  $D$ , we generate the respective patches  $P_D$  and  $\overline{P_D}$ . The evaluation is then made by comparing the geometric properties of the 3D patches.

The selection of the three “close” pixels is as follows. Given a pixel  $\mathbf{x} \in G \cap \Omega$ , its closest neighbors are the points generated by the same laser beam  $L_{\mathbf{x}}$  (recall that the horizontal resolution of the LRF is  $0.09^\circ$ ) in the previous or in the

next shot, followed by the points generated by one laser beam below or above  $L_{\mathbf{x}}$  (there are 64 lasers in the LRF). Thus, we choose to generate the patches  $P_G$  using two pixels from the same laser and one either from the laser above or below (creating a triangle). A patch is only defined if the disparity value of all the selected pixels is within a predefined range. If the selected pixels are also elements of  $\Omega \cap D$ , we generate the corresponding patch  $P_D$ . This patch also contains the pixels in  $D$  within the triangle defined by the three selected pixels. Once both patches have been defined, we analyze the geometric properties of their respective back projections (i.e 3D sets),  $\overline{P_G}$  and  $\overline{P_D}$ .

Let  $\overline{P} \subset \mathbb{R}^3$  be one of this patches, the *centroid*

$$c(\overline{P}) = \frac{1}{|\overline{P}|} \sum_{\overline{\mathbf{x}} \in \overline{P}} \overline{\mathbf{x}} \quad (4)$$

is calculated, as well as the *deviation* of the points in  $\overline{P}$  with respect to  $c(\overline{P})$ :

$$\text{Dev}(\overline{P}) = \sqrt{\frac{1}{|\overline{P}| - 1} \sum_{\overline{\mathbf{x}} \in \overline{P}} (\overline{\mathbf{x}} - c(\overline{P}))^2} \quad (5)$$

Note that  $\overline{\mathbf{x}} \in \mathbb{R}^3$ . Now, let  $P_G$  and  $P_D$  be corresponding patches in  $G$  and  $D$ , respectively. The confidence measure is calculated based in the distance between the centroid of the back projected patches,  $\overline{P_G}$  and  $\overline{P_D}$ , and the ratio of their respective deviations. Let  $\Delta_P$  be the Euclidean distance between  $c(\overline{P_G})$  and  $c(\overline{P_D})$ , and

$$\rho = \frac{\text{Dev}(\overline{P_G})}{\text{Dev}(\overline{P_D})} \quad (6)$$

Then, the confidence measure index for  $P_D$  is calculated as

$$CM(P_D) = \frac{2\rho}{\rho^2 + 1} \left( 1 - \frac{\Delta_P}{\Delta_{\max}} \right) \quad (7)$$

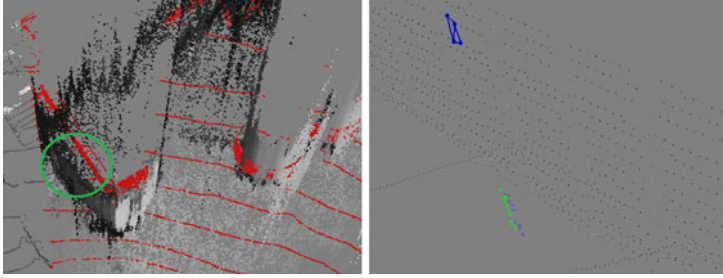
where  $\Delta_{\max}$  is the maximum possible Euclidean distance between the centroids.

The range of  $CM$  is  $[0, 1] \subset \mathbb{R}$ ; where a value close to one indicates that both patches are geometrically alike, and thus that the disparity results are reliable. Low values imply a low confidence in the calculated disparity values. To obtain a high confidence value (i.e., a value close to one), it is necessary that the centroids of both patches are close to each other and the ratio  $\rho$  of the variances is close to one.

The first factor in Eq. (7) penalizes the index more if  $\rho < 1$ , as it is expected that  $\overline{P_G}$  would be a more homogeneous set than  $\overline{P_D}$ . See Fig. 3 for an example of two pairs of analyzed patches in a sample 3D scene as viewed from above.

### 3 Experiments

Our experimental data set was captured using the LRF and two grey-scale (12 bits per pixel) cameras, all mounted in the same ego-vehicle. The cameras were



**Fig. 3.** A 3D test scene from the containers sequence (see Section 3) as viewed from above. *Left:* The red dots are the points returned with the LRF within the field of view of the reference camera. The grey points are the back projected points from a sample disparity map. *Right:* The gray points in here represent the LRF points. The projection of the highlighted blue points define two patches in the ground truth image; while the green and purple dots are the back projection of the two corresponding patches in the sample disparity map.

placed behind the windshield, while the LRF was attached to a rack on the roof. The coordinate system from the LRF was calibrated according to the external parameters of the reference camera (the left camera) of the stereo set up using the method proposed in [13], where a closed-form solution of the Perspective-n-Point problem was presented. We use the internal parameters of the stereo camera to project the 3D points from the LRF to generate the ground truth image.

For defining the patches, we use a disparity threshold of one, so that they were generated with points that are really close to each other. The threshold for the BCP quality metric was set to one.

**Data Set.** We illustrate the presented approach by using three sequences recorded in “simple” environments. The objective of using these sequences is to “grow” a first experience using this approach and to validate if there is a good correlation between the direct comparison and the confidence measure’s indexes.

The size of the images is of  $1024 \times 334$  pixels, reduced to  $930 \times 289$  due to the rectification procedure for stereo analysis. Range data were recorded using the five revolutions per second configuration of the LRF, in order to obtain the maximum number of measurements (around 24,000 pixels with positive value in the ground truth image). All the sequences are stop-and-go ones, in order to minimize synchronization issues between the camera (set to 20 frames per second) and the LRF. Developing and approach to generate ground truth in dynamic scenes is out of the scope of this paper. Sample frames of each sequence are shown in Fig. 1 and 4.

*Wall sequence.* Recorded while driving towards a wall that covers the entire field of view of the cameras. In the lower right corner of the images there is a small car and a trailer. Both objects are only present in the first part of the sequence.

*Wall-trailer sequence.* Recorded while driving towards the same wall as in the wall sequence. In this case there is a trailer that covers almost half of the reference and match image. This sequence turned out to be a good example for miscalculated disparities, as the trailer’s cover has areas with no texture at all. There are also two areas below the trailer where it is possible to see the road behind the trailer.

*Container sequence.* In this scene two different kinds of containers are present, one with a square base and two with a circular one. There is also a small part of a building with an intensity that it is very similar to the intensity of the curved containers so we are expecting “not so good” results from the stereo algorithms for this sequence. There are also two staircases with thin handrails that even the LRF had problems to detect.

**Results.** The stereo algorithms used in this work are briefly identified below. We use a local standard *dynamic programming* (DP) stereo algorithm [18]. Two global algorithms: *belief propagation stereo* (BP), with a coarse-to-fine approach [5] and a quadratic cost function [7], and a graph cut (GC) [12] algorithm. Finally, a *semi-global matching* (SGM) approach with mutual information as the cost function [9] was also used.

The algorithms are tested with respect to pixel accuracy. But, the approach presented here, as well as the data set, are well suited to test sub-pixel accuracy disparities. We are not aware of an existing real-world data set that can evaluate the performance of sub pixel accurate algorithms. See Table II for a summary of the results for all algorithms and both sequences.

*Wall sequence.* For this sequence, we expect the disparity values to get better as the ego-vehicle approaches the wall. This is due to the inverse proportionality of distance to disparity, thus small errors in disparity have a large effect at large distances. The algorithms behave as expected with respect to the BCP index; the percentage of badly calculated pixels decreases as the ego-vehicle gets closer to the wall. SGM had a high peak among the last five frames, where it has a poor performance on the road area. For CM, an average of 14,500 patches were analyzed (so above 50% of of the points in the disparity maps were considered for the evaluation). For GC and SGM, the CM index showed a consistent behavior with BCP, even the same peak for SGM in the last five frames can be identified here. The DP algorithm showed a relatively constant CM index. But, there is a low peak in the last five frames (the same set of frames that made SGM have a high BCP peak). In these frames the disparities obtained for the wall are not as homogeneous as it is expected, this can be barley detected with the BCP



**Fig. 4.** Sample images of the *wall* (left) and *container* (right) sequences. For a sample image of the *wall-trailer* sequence see Fig. III.

**Table 1.** Summarized results for the three sequences with both quality metrics. The results for the confidence measure (CM) are presented as the average over the entire sequence (first column), and the percentage of the number of patches with an index below 0.5 (second respective column), again over the entire sequence. For BCP is only shown the average percentage over each one of the sequences.

Alg.	Wall				Wall-Trailer				Containers			
	CM		BCP		CM		BCP		CM		BCP	
	Avg.	"> 0.9" "< 0.5"			Avg.	"> 0.9" "< 0.5"			Avg.	"> 0.9" "< 0.5"		
BP	0.43	4.2	64.6	28.5	0.43	6.0	63.2	33.5	0.44	4.8	64.1	29.5
DP	0.49	7.1	56.8	14.8	0.47	10.1	56.5	22.1	0.51	13.9	54.1	16.3
GC	0.29	0.5	85.8	35.1	0.34	0.6	82.2	38.7	0.28	0.7	88.2	42.6
SGM	0.37	3.2	67.9	35.3	0.38	2.5	75.9	50.2	0.36	2.4	76.6	59.5

index. For BP, the CM index decreases over the sequence, in contrast with the behavior of the BCP index, as this indicates that there are less miscalculated points (according to the low BCP score) but the miscalculations are larger.

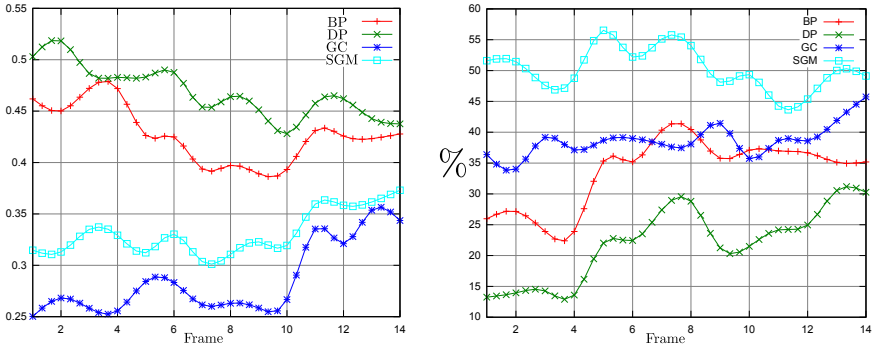
*Wall-trailer sequence.* As expected, most of the algorithms have problems with the trailer’s cover, as it is almost textureless. With respect to BCP, the algorithms had a similar performance, showing the worst results at the end of the sequence, when the trailer occupied almost the half of the stereo images. The exception was SGM. SGM handles this area better than the other algorithms. Its BCP index showed an improvement on its performance in the last part of the sequence. However, this algorithm had a poor performance in the road area making it the worst performing algorithm.

The results for CM show a good correlation with BCP. The confidence index decreases for DP and BP as the trailer is getting closer to the cameras, but increases for SGM. The GC algorithms did not follow the same pattern as with BCP; the last frames are the ones with highest CM value (but still very low). This can be explained as in the first half of the sequence, where the two areas below the trailer are visible; as this sequence goes forward, one of these zones goes out of the field of view of the cameras. The GC algorithms had more trouble detecting those background zones than the other algorithms. This can be detected with the CM index. However, the BCP index keeps going higher indicating that the disparity maps are still affected by the trailer’s cover, but that the accuracy of the disparity values are better. The average number of patches calculated for this sequence was 14,300 (almost 50% of the points).

*Container sequence:* While both staircases and the building on the right side are present in the stereo images, the results for all the metrics for DP, GC and SGM show a failure. They all have problems detecting the thin structures from the staircases and the almost equal intensities of the circular containers on the right and the building next to it.

The BP algorithm behaved differently, but consistently for the two metrics. Its best performance is on the first part of the sequence, and starts decreasing





**Fig. 5.** Results for the wall-trailer sequence. *Left:* Results for CM, a value close to one indicates a high confidence in the disparity map. *Right:* Plot for the BCP, larger values implies a larger number in the miscalculated points.

from frame five. It looks like it had more trouble than the others with an almost saturated background area that grows as the sequence goes forward.

The GC and SGM algorithms swap their ranking under different metrics, see Table 1. For BCP and SGM, the GC algorithm had a better performance than SGM. This does not represent a drawback for our approach as one metric counts the number of pixels that were miscalculated (BCP) while the other one focuses on how accurate the disparity values are (CM). The average number of analyzed patches for this sequence was 14,600 implying that there were evaluated more than 50% of the pixels in the disparity map.

## 4 Conclusions and Future Work

In this work we present a ground truth-based approach to evaluate stereo algorithms over real-world sequences. We evaluate the algorithms by comparing the calculated disparity maps against ground truth images generated using a high-end LRF. As the ground truth images are not dense enough to evaluate all the pixels in the disparity maps, we follow two evaluation criteria: Where ground-truth data are available, we use a well-known quality metric to evaluate the corresponding disparity values. For the remaining points, we use a confidence measure that compares the geometric properties of corresponding point sets in the ground truth images and in the disparity maps. We also include a few experiments to show the effectiveness of the presented approach. In the experiments we noticed a good correlation between the measures used.

Using the direct comparison approach, we were capable to evaluate around 7% of the pixels in a disparity image. However, when we also use the confidence measure, we could evaluate the majority of the points. The exact number depends on the scene.

The obtained evaluation results need to be addressed in work aiming at improvements of stereo matching algorithms. We have a lot more experimental

data, and those accumulated data will help further to identify weakness and strength of particular matching strategies, cost functions, or further algorithmic “ingredients” of stereo matching.

*Acknowledgements.* The first author thanks Dr. Uwe Franke for the opportunity to be a part of his research team at Daimler A.G. for six months, and Dr. Stefan Gehrig for valuable guidance.

## References

1. Banks, J., Corke, P.: Quantitative evaluation of matching methods and validity measures for stereo vision. *Int. J. Robotics Research* 20, 512–532 (2001)
2. Eid, A., Farag, A.: A unified framework for performance evaluation of 3-d reconstruction techniques. In: *Proc. Comp. Vision Pattern Recognition Workshop*, vol. 3, pp. 33–41 (2004)
3. Egnal, G., Mintz, M., Wildes, R.: A stereo confidence metric using single view imagery with comparison to five alternative approaches. *Image Vision Computing* 22, 943–957 (2004)
4. .enpeda. Group, University of Auckland: EISATS (.enpeda. sequence analysis test site) (2010), <http://www.mi.auckland.ac.nz/EISATS>
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. *Int. J. Comp. Vision* 70, 41–54 (2006)
6. Gherardi, R.: Confidence-based cost modulation for stereo matching. In: *Proc. ICPR*, pp. 1–4 (2008)
7. Guan, S., Klette, R., Woo, Y.W.: Belief propagation for stereo analysis of night-vision sequences. In: *PSIVT 2009. LNCS*, vol. 5414, pp. 932–943 (2009)
8. Haeusler, R., Klette, R.: Benchmarking stereo data (Not the matching algorithms). In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) *Pattern Recognition. LNCS*, vol. 6376, pp. 383–392. Springer, Heidelberg (2010)
9. Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: *Proc. CVPR*, pp. 807–814 (2005)
10. Huang, F., Klette, R., Scheibe, K.: *Panoramic Imaging: Sensor-Line Cameras and Laser Range-Finders*. Wiley, Chichester (2008)
11. Klette, R., Vaudrey, T., Wiest, J., Haeusler, R., Jiang, R., Morales, S.: Current challenges in vision-based driver assistance. In: *Progress in Combinat. Image Analysis, Research Publ. Services, Singapore* (2010)
12. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) *ECCV 2002. LNCS*, vol. 2352, pp. 82–96. Springer, Heidelberg (2002)
13. Lepetit, V., Moreno-Noguer, F., Fua, P.: Accurate non-iterative  $o(n)$  solution to the PNP problem. In: *Proc. ICCV*, pp. 874–885 (2007)
14. Liu, Z., Klette, R.: Approximate ground truth for stereo and motion analysis on real-world sequences. In: *PSIVT 2009. LNCS*, vol. 5414, pp. 874–885 (2009)
15. Morales, S., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. In: Jiang, X., Petkov, N. (eds.) *CAIP 2009. LNCS*, vol. 5702, pp. 1078–1086. Springer, Heidelberg (2009)
16. Mordohai, P.: The self-aware matching measure for stereo. In: *Proc. ICCV*, pp. 1841–1848 (2009)

17. Murray, D., Little, J.J.: Using real-time stereo vision for mobile robot navigation. *Aut. Robots* 8, 161–171 (2000)
18. Ohta, Y., Kanade, T.: Stereo by two-level dynamic programming. In: *Proc. Int. Joint Conf. Artificial Int.*, pp. 1120–1126 (1985)
19. Reulke, R., Lubert, A., Haberjahn, M., Piltz, B.: Validierung von mobilen Stereokamerasystemen in einem 3D-Testfeld. In: *Proc. 3D-NordOst* (2009)
20. Satoh, Y., Sakaue, K.: An omnidirectional stereo vision-based smart wheelchair. *J. Image Video Processing* 2007, 1–11 (2007)
21. Seelinger, M., Yoder, J.D.: Automatic pallet engagement by a vision guided forklift. In: *Proc. IEEE Int. Conf. Robotics Automation*, pp. 4068–4073 (2005)
22. Steingrube, P., Gehrig, S., Franke, U.: Performance evaluation of stereo algorithms for automotive applications. In: *Proc. Int. Conf. Comp. Vision Systems*, pp. 285–394 (2009)
23. Szeliski, R.: Prediction error as a quality metric for motion and stereo. In: *Proc. ICCV*, pp. 781–788 (1999)
24. Velodyne Lidar Inc.: Velodyne’s HDL-64E S2 user manual, <http://www.velodyne.com/lidar/hdlproducts/hdl64e.aspx>

# Vehicle Ego-Localization by Matching In-Vehicle Camera Images to an Aerial Image

Masafumi Noda<sup>1,\*</sup>, Tomokazu Takahashi<sup>1,2</sup>, Daisuke Deguchi<sup>1</sup>,  
Ichiro Ide<sup>1</sup>, Hiroshi Murase<sup>1</sup>, Yoshiko Kojima<sup>3</sup>, and Takashi Naito<sup>3</sup>

<sup>1</sup> Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601, Japan

<sup>2</sup> Gifu Shotoku Gakuen University, Nakauzura 1-38, Gifu, 500-8288, Japan

<sup>3</sup> Toyota Central Research & Development Laboratories, Inc., 41-1 Aza Yokomichi,  
Oaza Nagakute, Nagakute, Aichi, 480-1192, Japan

`mnoda@murase.m.is.nagoya-u.ac.jp`

**Abstract.** Obtaining an accurate vehicle position is important for intelligent vehicles in supporting driver safety and comfort. This paper proposes an accurate ego-localization method by matching in-vehicle camera images to an aerial image. There are two major problems in performing an accurate matching: (1) image difference between the aerial image and the in-vehicle camera image due to view-point and illumination conditions, and (2) occlusions in the in-vehicle camera image. To solve the first problem, we use the SURF image descriptor, which achieves robust feature-point matching for the various image differences. Additionally, we extract appropriate feature-points from each road-marking region on the road plane in both images. For the second problem, we utilize sequential multiple in-vehicle camera frames in the matching. The experimental results demonstrate that the proposed method improves both ego-localization accuracy and stability.

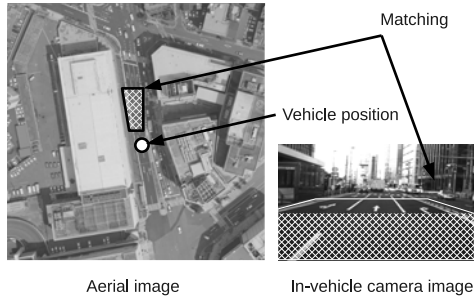
## 1 Introduction

The vehicle ego-localization task is one of the most important technologies for Intelligent Transport Systems (ITS). Obtaining an accurate vehicle position is the first-step to supporting driver safety and comfort. In particular, ego-localization near intersections is important for avoiding traffic accidents. Recently, in-vehicle cameras for the ego-localization have been put to practical use. Meanwhile, aerial images have become readily available, for example from Google Maps [1]. In light of the above, we propose a method for accurate ego-localization by matching the shared region taken in in-vehicle camera images to an aerial image.

A global positioning system (GPS) is generally used to estimate a global vehicle position. However, standard GPSs for a vehicle navigation system have an estimation error within about 30–100 meters in an urban area. Therefore, a relatively accurate position is estimated by matching information, such as a geo-location and an image taken from a vehicle, to a map. Among them, map-matching [2] is one of the most prevalent methods. This method estimates a

---

\* Corresponding author.



**Fig. 1.** Vehicle ego-localization by matching in-vehicle camera image to an aerial image: Shaded regions in both images correspond

vehicle position by matching a vehicle’s driving trajectory calculated from rough estimations using GPS to a topological road map. Recently, in-vehicle cameras have been widely used; therefore, vehicle ego-localization using cameras has been proposed [3,4,5]. This camera-based vehicle ego-localization matches in-vehicle camera images to a map, which is also constructed from in-vehicle camera images. In many cases, the map is constructed by averaging in-vehicle camera images with less-accurate geo-locations. Therefore, it is difficult to construct a globally consistent map.

In contrast, aerial images that covers a wide region and with a highly accurate geo-location have also become easily available, and we can collect them at low-cost. There are some methods that ego-localize an aircraft by matching aerial images [6,7]. However, the proposed method estimates a vehicle position. The proposed method matching the shared road-region of in-vehicle camera images and an aerial image is shown in Figure 1. Pink et al. [8] have also proposed an ego-localization method based on this idea. They estimate a vehicle position by matching feature-points extracted from an aerial image and an in-vehicle camera image. An Iterative Closest Point (ICP) method is used for this matching. As feature-points, the centroids of road markings, which are traffic symbols printed on roads, are used. This method, however, has a weakness in that a matching error occurs in the case where the images differ due to illumination conditions and/or occlusion. This decreases ego-localization accuracy.

There are two main problems to be solved to achieve accurate ego-localization using in-vehicle camera images and an aerial image. We describe these problems and our approaches to solve them.

- 1) **Image difference between the aerial image and the in-vehicle camera image:** The aerial image and the in-vehicle camera image have large difference due to viewpoints, illumination conditions and so on. This causes difficulty in feature-point matching. Therefore, we use the Speed Up Robust Feature (SURF) image descriptor [9]. The SURF image descriptor is robust for such differences of view and illumination. Additionally, since the road-plane region in the images has a simple texture, the feature-points extracted by a general method tend to be too few and inappropriate for the matching.



**Fig. 2.** Feature-point map: White dots represent feature-points

Therefore, we extract feature-points appropriate for the matching from each road-marking region.

- 2) **Occlusion in the in-vehicle camera image:** In a real traffic environment, forward vehicles often exist. They occlude the road-markings in the in-vehicle camera image, and thus matching to an aerial image fails. However, even if the feature-points are occluded in some frames, they may be visible in other frames. Therefore, we integrate multiple in-vehicle camera frames to extract feature-points, including even those occluded in specific frames.

Based on the above approaches, we propose a method for vehicle ego-localization by matching in-vehicle camera images to an aerial image. The proposed method consists of two stages. The first stage constructs a map by extracting feature-points from an aerial image, which is performed offline. The second stage ego-localizes by matching in-vehicle camera images to the map.

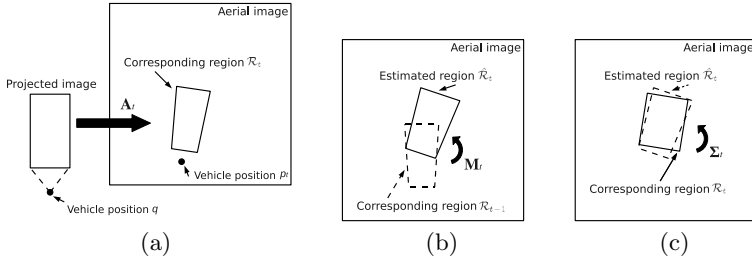
This paper is organized as follows: Section 2 proposes a method of map construction from an aerial image, and Section 3 proposes a method of ego-localization by matching in-vehicle camera images to the map, in real time. Experimental results are presented in Section 4, and discussed in Section 5. Section 6 summarizes this paper.

## 2 Construction of Feature-Points Map for Ego-Localization

A feature-points map is constructed from an aerial image for the ego-localization. To adequately extract the applicable feature-points, we first extract road-marking regions and then extract the unique feature-points from each region. We then construct a map for the ego-localization using SURF descriptors [9], which are robust against the image difference between the aerial image and the in-vehicle camera image. Figure 2 shows a feature-point map constructed from the aerial image. In this paper, the road region of the intended sequences is manually extracted in advance to evaluate the proposed method. We will automatically extract the region by a segmentation method in future work.

The map construction process is divided into the following steps:

1. Emphasize road markings by binarizing an aerial image, then split it into multiple regions by a labeling method.



**Fig. 3.** Overview of the proposed method: (a) Correspondence of a projected image and the region in aerial image. (b) Estimation of the current corresponding region. (c) Estimation of an accurate corresponding region.

2. Eliminate the regions considering appropriate road-marking size.
3. Extract feature-points  $\mathbf{x}_n (n = 1, \dots, N)$  from the road-marking regions in the binary image by Harris corner detector.
4. Calculate the SURF descriptor  $\mathbf{f}_n$  around  $\mathbf{x}_n$  from the aerial image.

The feature-point map is represented as the pairs of the position and the SURF descriptor  $\{(\mathbf{x}_1, \mathbf{f}_1), \dots, (\mathbf{x}_N, \mathbf{f}_N)\}$ . In this paper, we treat objects on the road such as vehicles and trees as well as road markings, though the detection of these objects is required in a fully developed system.

### 3 Ego-Localization by Matching the In-Vehicle Camera Images to the Map

#### 3.1 Overview

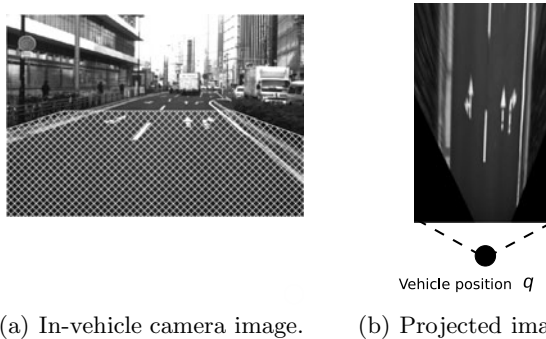
Vehicle ego-localization is achieved by sequentially matching in-vehicle camera images to a map constructed from an aerial image. The proposed method ego-localizes a vehicle at time step  $t$  (frame) by the following steps:

1. Transformation of an in-vehicle camera image to a projected image
2. Sequential matching between projected images
3. Matching of the projected image to the map using multiple frames
4. Estimation of the vehicle position

The proposed method first transforms the in-vehicle camera image to a projected image to simplify the matching process. Then, the proposed method finds a region  $\mathcal{R}_t$  in the map that corresponds to the in-vehicle camera image as shown in Figure 3(a). The homography matrix  $\mathbf{A}_t$  in this figure transforms the projected image on  $\mathcal{R}_t$ . Then, we estimate the vehicle position  $\mathbf{p}_t$  as

$$\mathbf{p}_t = \mathbf{A}_t \mathbf{q}, \tag{1}$$

where  $\mathbf{q}$  is the vehicle position in the projected image, as shown in Figure 4(b) and Figure 3(a), obtained from the in-vehicle camera parameters.



**Fig. 4.** Transformation of an in-vehicle camera image to a projected image: the shaded region in (a) is transformed to the projected image (b)

The proposed method updates  $\mathbf{A}_t$  by the two-step estimation shown in Figure 3(b) and Figure 3(c).  $\mathbf{A}_t$  is then updated as

$$\mathbf{A}_t = \Sigma_t \mathbf{A}_{t-1} \mathbf{M}_t. \quad (2)$$

$\mathbf{M}_t$  and  $\Sigma$  are the homography matrices.  $\mathbf{M}_t$  transforms the projected image to the estimated corresponding region  $\hat{\mathcal{R}}_t$  from the previous frame as shown in Figure 3(b). Then,  $\mathbf{M}_t$  is estimated by the sequential matching between projected images. The estimated region, however, contains some error due to the matching error  $\Sigma_t$ , which transforms the estimated region to an accurate corresponding region  $\mathcal{R}_t$  as shown in Figure 3(c). Therefore,  $\Sigma_t$  is estimated by the matching of the projected image to the map. In this matching, multiple in-vehicle camera frames are used to improve the matching accuracy. This aims to increase the number of feature-points and to perform accurate matching in a situation where part of the road markings are occluded in the in-vehicle camera images. We detail the ego-localization process below.

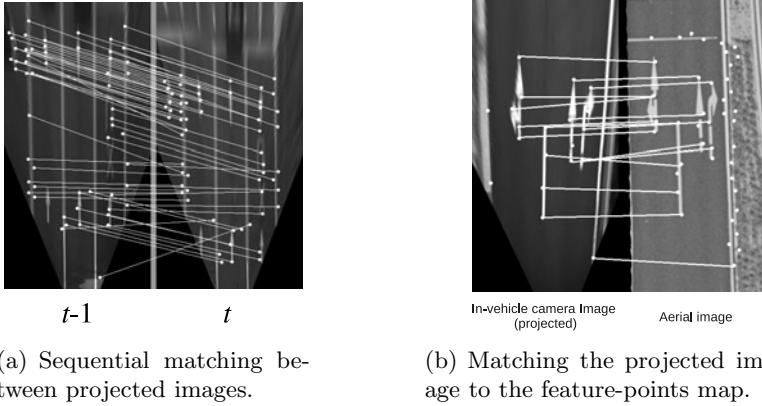
### 3.2 Transformation of an In-Vehicle Camera Image to a Projected Image

An in-vehicle camera image is transformed to a projected image as shown in Figure 4. To transform the projected image, a  $3 \times 3$  homography matrix is used. The matrix is calculated in advance from the in-vehicle camera parameters: installed position, depression angle and focal length. The vehicle position  $\mathbf{q}$  in a projected image is also obtained using the matrix.

### 3.3 Sequential Matching between Projected Images

To estimate  $\hat{\mathcal{R}}_t$ , the proposed method performs the matching between sequential projected images. The projected image at  $t$  is represented as  $I_t$ .  $\mathbf{M}_t$ , shown in Figure 3(b), is obtained by matching between the feature-points in  $I_{t-1}$  and  $I_t$ .





**Fig. 5.** Two step matching (Corresponding feature-point pairs in the projected images: The dots represent the feature-point in each image and the lines show their correspondence)

The feature-points are extracted by Harris corner detector, then matched by Lucas-Kanade’s method. Figure 5(a) shows the initial correspondence between the feature-points.  $\mathbf{M}_t$  is calculated by minimizing the LMedS criterion by selecting the correspondences.  $\hat{\mathcal{R}}_t$  is calculated from  $\mathbf{M}_t$  and  $\mathbf{A}_{t-1}$ .

### 3.4 Matching of the Projected Image to the Feature-Points Map Using Multiple Frames

$\hat{\mathcal{R}}_t$  contains some error, which is represented as a homography matrix  $\Sigma_t$  shown in Figure 3(c). We calculate  $\Sigma_t$  by matching the projected image to the map to obtain the accurate corresponding region  $\mathcal{R}_t$ . In this matching, in order to improve the accuracy and stability in a situation where occlusions occur in the in-vehicle camera image, multiple in-vehicle camera frames are used. We first explain a matching method the only uses a single frame, and then how to extend it to that uses multiple frames.

**Matching using a Single Frame.** We extract the feature-points from the projected images in the same manner as described in Section 2. The position of a feature-point extracted from  $I_t$  is represented as  $\mathbf{y}_{t,l_t}$  ( $l_t = \{1, \dots, L_t\}$ ), where  $L_t$  is the number of feature-points. The SURF descriptor of  $\mathbf{y}_{t,l_t}$  is represented as  $\mathbf{g}_{t,l_t}$ . Thus, the feature-points could be represented as  $\{(\mathbf{y}_{t,1}, \mathbf{g}_{t,1}), \dots, (\mathbf{y}_{t,L_t}, \mathbf{g}_{t,L_t})\}$ .

For the matching, each feature-point position  $\mathbf{y}_{t,l_t}$  is transformed to  $\mathbf{y}'_{t,l_t}$  in the map as

$$\mathbf{y}'_{t,l_t} = \mathbf{A}_{t-1} \mathbf{M}_t \mathbf{y}_{t,l_t}. \tag{3}$$

Feature-point pairs are chosen so that they meet the following conditions:

$$\left\{ \begin{array}{l} \|\mathbf{y}'_{t,l_t} - \mathbf{x}_n\| < r \\ \min_{l_t} \|\mathbf{g}_{t,l_t} - \mathbf{f}_n\| \end{array} \right. , \tag{4}$$

**Table 1.** Dataset

Set No.	Length (m)	Aerial image	In-vehicle camera image	
		Occlusion	Occlusion	Time
1	85	small	small	day
2	100	small	large	night
3	100	large	small	day
4	75	large	large	day

where  $r$  is the detection radius. Figure 5(b) shows the feature-point pairs. Then,  $\Sigma_t$  is obtained by minimizing the LMedS criterion by selecting the correspondences.

**Matching using Multiple Frames.** To achieve accurate matching in a situation where occlusions occur in some in-vehicle camera images, we integrate the feature-points in the multiple in-vehicle camera frames. The feature-points at  $t'$  are represented as  $\mathcal{Y}_{t'} = \{\mathbf{y}_{t',1}, \dots, \mathbf{y}_{t',L_{t'}}\}$ . They are transformed to  $\mathcal{Y}'_{t'} = \{\mathbf{y}'_{t',1}, \dots, \mathbf{y}'_{t',L_{t'}}\}$  in the map coordinate.  $\mathbf{y}'_{t',1}$  is transformed as

$$\mathbf{y}'_{t',l_{t'}} = \begin{cases} \mathbf{A}_{t'-1} \mathbf{M}_{t'} \mathbf{y}_{t',l_{t'}} & t' \text{ is current frame} \\ \mathbf{A}_{t'} \mathbf{y}_{t',l_{t'}} & \text{otherwise} \end{cases} . \quad (5)$$

Then, the feature-points in the  $F$  multiple frames including the current frame are used for the matching. Then, we obtain  $\Sigma_t$  in the same manner as in the case of a single frame.

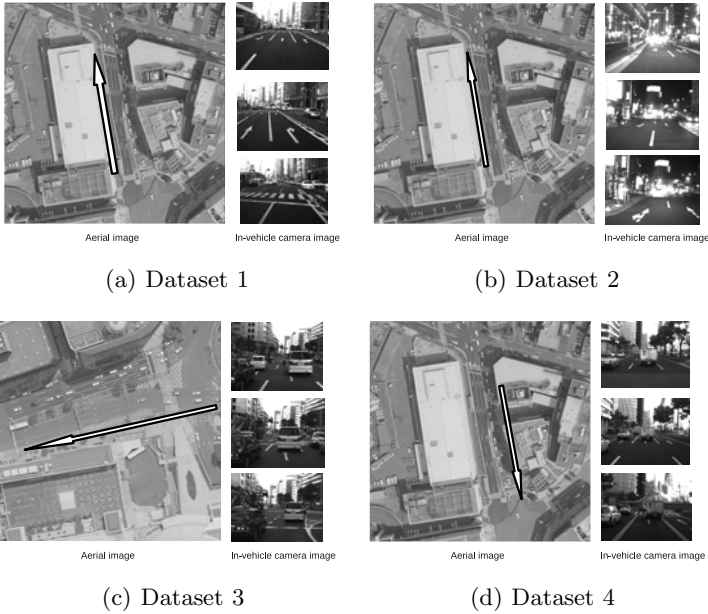
### 3.5 Estimation of the Vehicle Position

Finally,  $\mathbf{A}_t$  is calculated by Equation 2, and the vehicle position  $\mathbf{p}_t$  is estimated by Equation 1. As for the matrix  $\mathbf{A}_0$  at the initial frame, it is obtained by a global matching method in the map without the estimation of  $\hat{\mathcal{R}}_0$ .

## 4 Experiment

### 4.1 Setup

We mounted a camera, a standard GPS and a high accurate positioning system (Applanix, POSLV) [10] on a vehicle. The standard GPS contains an error of about 5–30 meters, which was used for the initial frame matching. The high-accuracy positioning system was used to obtain the reference values of vehicle positions. We used four sets of an aerial image and an in-vehicle camera image sequence with different capturing conditions. Table 1 shows the specification of the datasets and Figure 6 shows examples. The resolution of the aerial image was 0.15 meters per pixel. The resolution of the in-vehicle camera image was  $640 \times 480$  pixels, and its frame-rate was 10 fps. Occlusions in the aerial image occurred due to vehicles, trees and so on. Occlusions in the road regions in an



**Fig. 6.** Datasets: Four sets of an aerial image and an in-vehicle camera image sequences

aerial image occurred due to vehicles, trees and so on. We defined a road segment in an aerial image which was occluded less than 10% as a small occlusion, and that occluded more than 50% as a large occlusion by visual judgment. Occlusions in the in-vehicle camera images were due to forward vehicles.

## 4.2 Evaluation

We evaluated the ego-localization accuracy by the Estimation Error and the Possible Ratio defined by the following equations:

$$\text{Estimation error} = \frac{\text{The sum of estimation errors in available frames}}{\text{The number of available frames}}, \quad (6)$$

$$\text{Possible ratio} = \frac{\text{The number of available frames}}{\text{The number of all frames}}. \quad (7)$$

The Estimation Error is the average error between the estimated vehicle position and the reference value. On the other hand, the Possible Ratio represents the stability of the estimation. So, we use available frames in which the estimation was achieved successfully to calculate the Estimation Error. The available frames were checked by the size and twisting of the corresponding region, which was transformed from the projected image to the aerial image. When the Possible Ratio was less than 0.50, we did not calculate the Estimation Error.

In this experiment, we compared the ego-localization accuracy between the proposed method and a method based on [8]. The comparative method used

**Table 2.** Experimental result

Set No.	Proposed		Compared	
	Error (m)	Possible Ratio	Error (m)	Possible Ratio
1	0.60	1.00	0.72	0.83
2	0.70	1.00	0.75	0.90
3	0.98	0.73	N/A	0.30
4	N/A	0.12	N/A	0.04

only the center position of road markings as the feature-point, then performed the matching of these feature-points to the map using the ICP method. In this matching, the comparative method used only a single in-vehicle camera frame. On the other hand, the proposed method used five frames selected from frames for the previous five seconds with the same interval.

### 4.3 Initial Estimation

For the initial estimation, we performed matching between a projected image and a circular region in an aerial image with the radius of 30 meters around the location measured by a standard GPS. In cases where the estimation failed in the frame, we also performed this initial estimation in the next frame.

### 4.4 Experimental Result

Table 2 shows the ego-localization accuracy. Each row shows the Estimation Error and the Possible Ratio of each dataset. We confirmed from this result that the proposed method improved the accuracy for all datasets compared with the comparative method. In the case of Dataset 1 with small occlusion in both the in-vehicle camera image sequence and the aerial image, the Estimation Error was 0.60 meters by the proposed method. Furthermore, the Possible Ratio 1.00 was achieved by the proposed method, compared to 0.83 by the comparative method. Thus, we also confirmed the high stability of the proposed method. In the case of Dataset 2 with the in-vehicle camera image sequence taken at night, the Estimation Error and the Possible Ratio also improved.

In the case of Dataset 3 with a large occlusion in the in-vehicle camera image sequence, an Estimation Error of 0.98 and Possible Ratio of 0.73 were achieved by the proposed method. In contrast, a Possible Ratio of only 0.30 was achieved by the comparative method, and the Estimation Error was not available because the possible rate was less than 0.50. Finally, in the case of Dataset 4, there was a large occlusion in the aerial image, and ego-localization by both methods was not available in most frames due to mismatching of the feature-points.

The estimation of the proposed method consumed about 0.6 (sec) per frame when we used a computer whose CPU was Intel(R) Core(TM) i7 860 2.80GHz.

## 5 Discussion

- 1) **Image Difference between the Aerial Image and the In-vehicle Camera Image:** For matching the in-vehicle camera image to the aerial image, we extracted unique feature-points from road markings, and used the SURF descriptor. From the results of Datasets 1 and 2, the proposed method improved the Estimation Error and the Possible Ratio. The results demonstrated that the proposed method could make the matching robust for the image difference between the images.
- 2) **Occlusion in the In-vehicle Camera Image:** The feature-points extracted from the in-vehicle camera image were occluded in some frames. However, they were not occluded in other frames. From the result of Dataset 3, we confirmed that the matching using the multiple frames in the proposed method worked well in such situations. In this experiment, we fixed the number of frames used for the matching. We consider that adapting the number to the changes of occlusions could further improve the performance.
- 3) **Limitation of the Proposed Method:** From the result of Dataset 4, the proposed method could not estimate accurately the vehicle position when a large occlusion existed in the aerial image. To solve this problem, we need to construct a map without occlusions. In future work, we will detect the occluded regions and interpolate them by using in-vehicle camera images.

## 6 Conclusion

We proposed a vehicle ego-localization method using in-vehicle camera images and an aerial image. There are two major problems in performing accurate matching: the image difference between the aerial image and the in-vehicle camera image due to view-points and illumination conditions; and occlusions in the in-vehicle camera image. To solve these problems, we improved the feature-point detector and the image descriptor. Additionally, we extracted appropriate feature-points from each road marking region on the road plane in both images, and utilized sequential multiple in-vehicle camera frames in the matching. The experimental results demonstrated that the proposed method improves both the ego-localization accuracy and the stability. Future work includes construction of a feature-points map without occlusions by using in-vehicle camera images.

## Acknowledgement

Parts of this research were supported by JST CREST and MEXT, Grant-in-Aid for Scientific Research. This work was developed based on the MIST library (<http://mist.murase.m.is.nagoya-u.ac.jp/>).

## References

1. Google Inc.: Google Maps (2005), <http://maps.google.com/>
2. Brakatsoulas, S., Pfoser, D., Salas, R., Wenk, C.: On map-matching vehicle tracking data. In: Proc. 32nd Conf. on Very Large Data Bases, pp. 853–864 (2005)
3. Kawasaki, H., Miyamoto, A., Ohsawa, Y., Ono, S., Ikeuchi, K.: Multiple video camera calibration using EPI for city modeling. In: Proc. 6th Asian Conf. on Computer Vision, vol. 1, pp. 569–574 (2004)
4. Ono, S., Mikami, T., Kawasaki, H., Ikeuchi, K.: Space-time analysis of spherical projection image. In: Proc. 18th Int. Conf. on Pattern Recognition, pp. 975–979 (2006)
5. Uchiyama, H., Deguchi, D., Takahashi, T., Ide, I., Murase, H.: Ego-localization using streetscape image sequences from in-vehicle cameras. In: Proc. Intelligent Vehicle Symp. 2009, pp. 185–190 (2009)
6. Lin, Y., Yu, Q., Medioni, G.: Map-enhanced UAV image sequence registraton. In: Proc. 8th Workshop on Applications of Computer Vision, pp. 15–20 (2007)
7. Caballero, F., Luis Merino, J.F., Ollero, A.: Homography based Kalman filter for mosaic building. applications to UAV position estimation. In: Proc. Int. Conf. on Robotics and Automation, pp. 2004–2009 (2007)
8. Pink, O., Moosmann, F., Bachmann, A.: Visual features for vehicle localization and ego-motion estimation. In: Proc. Intelligent Vehicle Symp. 2009, pp. 254–260 (2009)
9. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: SURF: Speeded up robust features. Computer Vision and Image Understanding 110, 346–359 (2008)
10. Applanix Corp.: POS LV (2009), <http://www.applanix.com/products/land/pos-lv.html>

# A Comparative Study of Two Vertical Road Modelling Techniques

Konstantin Schauwecker and Reinhard Klette

Computer Science Department, The University of Auckland  
Private Bag 92019, Auckland 1142, New Zealand

**Abstract.** Binocular vision combined with stereo matching algorithms can be used in vehicles to gather data of the spatial proximity. To utilize this data we propose a new method for modeling the vertical road profile from a disparity map. This method is based on a region-growing technique, which iteratively performs a least-squares fit of a B-spline curve to a region of selected points. We compare this technique to two variants of the  $v$ -disparity method using either an envelope function or a planarity assumption. Our findings are that the proposed road-modeling technique outperforms both variants of the  $v$ -disparity technique, for which the planarity assumption is slightly better than the envelope version.

## 1 Introduction

Vehicles have become more and more intelligent over the past few years. Nowadays, drivers are supported by a range of helpful *driver assistant systems* (DAS). Some DAS, such as advanced automatic cruise control or parking assistants, only work well if information about the spatial environment is available. This data is gathered, for example, by using radar sensors. The problem with radar is that it only provides a measurement in the vehicle surroundings along a given direction.

It can be expected that future DAS will be more intelligent and thus require a more detailed model of the spatial proximity. This data can be gathered in principle using binocular vision (the human visual system may be cited as a proof). Top-performing *stereo matching* algorithms provide a dense measure for the disparity of most visible pixels. From the resulting *disparity map* we can reconstruct the 3D origin of each pixel and thus receive a detailed representation of the vehicle environment. An intelligent car not only has to gather this data but is also required to “understand” it. In particular, it is important that it recognizes properties of the road such as its geometry in 3D space, surface properties, speed bumps, obstacles on the road, and so forth. Our study is focussing on the road profile, modeled by a geometric manifold.

An accurate road profile helps to identify other vehicles and objects on the road by comparing the height of matched points with the road model: Points that are significantly above the road must belong to obstacles; of course, a road may also be elevated with visible objects next to the road that are below road level, but those objects would not be *on the road*.

Different manifold models may be selected to be fitted to the disparity data obtained for the road profile. Ideally, a road model should precisely match the vertical road profile. The common planarity assumption does not support that.

In this paper we present a new technique for creating a vertical road model, which is based on *B-spline curves* and a *region-growing* process. We evaluate the performance of this method and compare it to the widely used *v*-disparity approach. We use a common *belief propagation* stereo algorithm for the stereo matching part, but this is not crucial for the processing pipeline or the comparison (because we use it uniformly for both techniques), and it could be replaced by another stereo algorithm.

## 2 Related Work

The simplest way to model the vertical road profile is to assume that the road is planar and its normal perpendicular to the horizon, as done by Weber et al. [1]. This assumption is known to be inaccurate, and more advanced methods have been proposed.

The method introduced by Labayrade et al. [2] is based on *v-disparity images*. In this method, the disparity map is first transformed into a new virtual image, by counting the occurrences of each disparity value in each image row and plotting the result. The disparities corresponding to the road surface are likely to be incident with a curve. In [2] this curve is modeled as a piecewise linear curve, and its segments are detected through a standard *Hough transform*, which delivers a set of best matching straight lines. Those lines are mapped into one polygonal chain by either calculating the upper or lower envelope.

The approach proposed in [3] is based on approximating the road by a three-dimensional quadratic model. The first step in this procedure is to convert the disparity map into a digital elevation map. The quadratic model is then fitted using a region-growing method. First, a small region close to the *ego-vehicle* (i.e., the car where the system is operating in) is selected and used for fitting the first version of the model. This region is then iteratively extended by including matching adjacent pixels; the model is continuously refitted.

In [4], Nedeveschi et al. perform an approximation of the road surface by fitting a *clothoid*, which is a polynomial of degree three. The approximation process works by first reconstructing a *lateral view* of the scene from the disparity map. In a next step, a polar histogram is created that counts the number of points near a range of selected polar lines. The angle of the polar line with the maximum of surrounding points will be selected as being the pitch angle for the clothoid curve. The curvature of the clothoid is then detected using a similar histogram.

Wedel et al. introduce an approximation of the road surface in [5] that uses a B-spline curve with equidistant nodes. The control points of this curve are found with a least-squares method, which is not solved directly but embedded into a special *Kalman filter*. The curve is fitted to a region that is believed to match the road and was detected before with a *free-space estimation* algorithm. To improve accuracy, constraints are introduced which require that the height



and the gradient of the curve equals zero at the camera position. Furthermore, solutions with high gradients and curvatures are penalized.

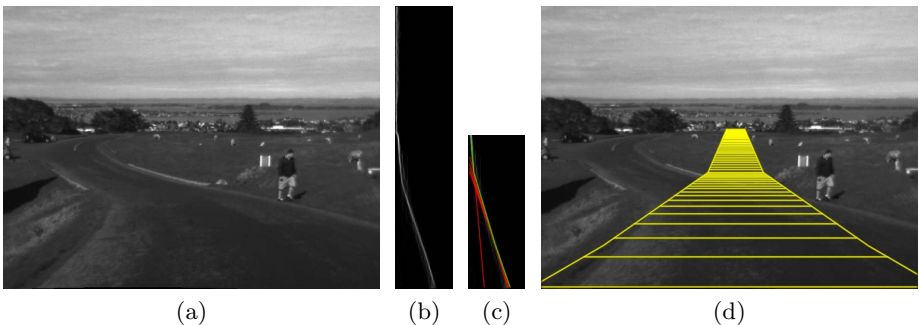
### 3 V-Disparity Images

The previously mentioned  $v$ -disparity method appears to be the most popular approach within the set of introduced vertical road-modeling techniques. This might be due to its simplicity. However, [3] criticizes the  $v$ -disparity approach by stating that it requires the road to occupy most of the image, and that it is sensitive to changes in roll-angle. Furthermore, the usage of  $v$ -disparity images for performing a piecewise linear approximation of the road surface, as done in [2], is not as accurate as other modeling approaches that rely on higher order curves [5].

Nevertheless, the  $v$ -disparity method can provide good results on predominantly straight roads without large curvatures [2,5]. We therefore chose to implement it as a reference system for evaluating our own technique.

Figure 1b shows the  $v$ -disparity image we obtained with our implementation for an example of a stereo pair; one image of the pair is shown in Fig. 1a. The  $v$ -disparity image has the same height as the input image and its width is equal to the number of possible disparity values. The intensity of a pixel  $(u, v)$  in this image represents how often the disparity  $u$  occurs in image row  $v$ .

Figure 1c shows the best matching lines (red) found with the Hough transform, and their upper envelope (green). We clipped the image to avoid false matches from sections above the road. The decision on whether to use the upper or lower envelope is done with the same method as in [2], which is by comparing the intensity sum of all pixels along both possible envelopes. In Fig. 1d a perspective projection of the resulting envelope function has been overlaid on one image of the input stereo pair.



**Fig. 1.** (a) Image of a stereo pair. (b) Corresponding  $v$ -disparity image. (c) Lines found by the Hough transform. (d) Perspective projection of the road profile.

## 4 B-Spline Road-Modeling

Using B-spline curves to model the vertical road profile, as done in [5], allows to model road profiles whose curvature changes its sign. None of the other approaches we discussed is capable of doing this. Thus, if we encounter such a road, those techniques will become largely inaccurate. We have developed a system for approximating the road by a B-spline curve, which uses a different approach from the one presented in [5].

A B-spline curve is fitted in [5] to a set of points extracted by a free-space estimation algorithm. This means that the accuracy of the created road model strongly depends on the used free-space algorithm; its results may be inaccurate if there are difficulties in detecting road boundaries. We propose an alternative approach that is based on a region-growing technique. In general, our method may potentially be more accurate if clear road boundaries are missing.

### 4.1 Region-Growing

Our method works on the set of 3D-points we obtain, when we reconstruct the 3D-location for all pixels of the disparity map. For initializing the region growing process, we select a small region of points close to the ego-vehicle, for which we have a high confidence that they are part of the road. We increase our confidence by only selecting points that do not deviate much from the model of the previous frame. A B-spline curve is then fitted to those points and used for finding further road points. This will increase the set of points we “understand” to be part of the road, and we use the enlarged set to fit a new and presumably more accurate curve. The selection of points and fitting of a new curve is repeated, either for a predefined number of iterations, or until a termination criterion is met.

Our region-growing method differs from the one proposed in [3], in that we do not require an elevation map. Furthermore, we allow the selection of new points that are not adjacent to already selected ones in the disparity map. This drastically reduces the number of iterations required, as more points can be selected in a single step compared to [3]. For the tested stereo sequences, less than twenty iterations were necessary for all stereo pairs.

### 4.2 Least-Squares Fitting

Once a region of identified road points has been selected, we use these points for fitting a uniform B-spline curve. For this task we use the method of least-squares, which has also been used in [3,5] for model fitting. This means that we try to minimize the error

$$E = \sum_{k=0}^m (B(z) - P_k)^2 = \sum_{k=0}^m \left( \sum_{j=0}^n N_j(t_k) Q_j - P_k \right)^2 \quad (1)$$

where  $P_k$  is in the set of selected road points,  $B(z)$  is the wanted B-spline curve,  $N_j$  a B-spline basis function, and  $Q_j$  is in the set of B-spline control points.

In [5], a solution is found by feeding the least-square equations as a measurement into a special Kalman filter. We cannot use this method because we develop our solution by an iterative process. Applying a Kalman filter to inaccurate intermediate estimates would disrupt the filter state.

We thus determine our solution the ordinary way, using the method of linear least-squares. The subject of fitting B-spline curves is discussed in [6]. To find a solution, we need to formulate our problem as a system of linear equations. In the ideal case, all measurement points lie exactly on the B-spline curve and can thus be expressed in terms of the control points and a matrix  $A$ , containing the B-spline basis functions, with

$$P = AQ \quad \text{and} \quad A = \begin{bmatrix} N_0(t_0) & N_1(t_0) & \cdots & N_n(t_0) \\ N_0(t_1) & N_1(t_1) & \cdots & N_n(t_1) \\ \vdots & \vdots & \ddots & \vdots \\ N_0(t_m) & N_1(t_m) & \cdots & N_n(t_m) \end{bmatrix} \quad (2)$$

In the case of linear least-squares, we can transpose the above equation to obtain the *normal equation*

$$A^T A Q = A^T P \quad (3)$$

of our linear system. The above linear system is invertible and can thus be solved by matrix inversion; we have the solution

$$Q = (A^T A)^{-1} A^T P \quad (4)$$

### 4.3 Error Model

Whether a point should be included in the enlarged region or not, is decided by its vertical distance to the previously fitted curve. For evaluating a new candidate point we need an error model that tells us the maximum distance that is still acceptable. In [3], the error in  $z$ - and  $y$ -direction is calculated by the formulas<sup>1</sup>

$$z_{err} = \left| \frac{z^2 \cdot d_{err}}{b \cdot f - z \cdot d_{err}} \right| \quad \text{and} \quad y_{err} = \left| \frac{y \cdot z_{err}}{z} \right| \quad (5)$$

where  $d_{err}$  is the disparity error,  $b$  is the baseline and  $f$  is the focal length.

With those two equations we can determine the maximum error from triangulation in  $y$ -direction. Experiments with this error model showed that it is not sufficient for our region-growing approach. It is missing the error introduced by the curve fitting, which can cause a displacement of the curve along the  $z$ -axis. This displacement can be as large as the the error in  $z$ -direction  $z_{err}$ .

We have to take this error into account when we decide whether a given point with the  $z$ -coordinate  $z$  should be considered to be part of the road or not. To do this, we not only examine the curve at position  $z$  but also at  $z + z_{err}$  and  $z - z_{err}$ . If the vertical distance of a candidate point to any of the three selected curve points is less than  $y_{err}$  plus a tolerance threshold  $s$ , then the point will be considered to be part of the road and added to the selected region.

<sup>1</sup> The original equations contain the camera height, as the origin is assumed to be on the road. We use the camera position as origin and thus do not require this variable.

#### 4.4 Region-Reduction

Because we allow a large number of pixels to be selected in one iteration, it is “very likely” that pixels outside the road are falsely selected; then those pixel distort the region-growing process. To cope with this problem we exclude some particular pixels from the selected region, even though they meet the criteria of our error model. We call this step *region-reduction* as it counteracts to the region-growing step. For performing region-reduction, we use a set of independent techniques:

**Z-Distance Limit.** The fitted B-spline curve is accurate for points that are not far from the selected region. The accuracy greatly declines the farther the curve is extrapolated. We thus limit the evaluation of a curve to be not continued beyond a distance  $d$  to the farthest point in the current region. This also limits the disconnectivity in  $z$ -direction: The distance between a new point and its closest selected neighbor can never exceed  $d$ .

**Connectivity Constraint.** We enforce connectivity in the  $xy$ -plane by treating our selection masks as a binary image, and use a flood-fill algorithm to extract a connected subset. The seed is selected as any point from the initial region. Points that are not in the extracted subset will be removed.

**Density Constraint.** The flood-fill algorithm does not remove erroneous regions if they share just a single pixel with the current road region. Therefore, we eliminate such connecting pixels by evaluating the number of selected pixels in the neighborhood of a pivot pixel. Pixels that have less than the required number of neighbors, will be removed. A rectangular neighborhood is evaluated in constant time based on the use of *integral images* [7].

**SSR Threshold.** If some erroneous points are selected, those points will have a small disruptive impact on the curve fitting. This may cause a selection of more and more erroneous points in subsequent iterations. In such cases we need to stop the iteration earlier. We perform this decision by calculating the *sum of squared residuals* (SSR). If the SSR per selected pixel exceeds a given threshold value, the iteration will be discontinued.

## 5 Smoothness Constraint

A smoothness constraint is used in [5] that penalizes high gradients and curvatures because we expect the road to be never extremely steep or curved. For this purpose, two penalizing quantities are introduced in [5], which are based on the squared first and second order derivative of the B-spline curve. The penalizing quantities are embedded in the update equation of the employed Kalman filter.

As we do not make use of a Kalman filter, we cannot apply the same method. We therefore suggest an alternative approach, which embeds the constraint equations into the least-squares linear system. To achieve this, we have to introduce two different penalizing quantities based on the absolute derivatives

$$w_1 \int |B'(z)| dz \quad \text{and} \quad w_2 \int |B''(z)| dz \quad (6)$$

where  $w_1$  and  $w_2$  are two factors controlling the influence of the constraint.

If we insert the B-spline equation into those quantities we obtain

$$w_1 \int |B'(z)| dz = w_1 \int \sum_{j=0}^n |N'_j(t_k) \cdot Q_j| dt_k = w_1 \sum_{j=0}^n |Q_j| \underbrace{\int |N'_j(t_k)| dt_k}_{I_1} \quad (7)$$

$$w_2 \int |B''(z)| dz = w_2 \int \sum_{j=0}^n |N''_j(t_k) \cdot Q_j| dt_k = w_2 \sum_{j=0}^n |Q_j| \underbrace{\int |N''_j(t_k)| dt_k}_{I_2} \quad (8)$$

Because we are using uniform B-splines, all basis functions  $N_j$  are shifted copies of each other. This means that the integrals over their absolute gradient and curvature  $I_1$  and  $I_2$  will both be constant. If we replace the products of the integrals and weighting factors  $I_1 w_1$  and  $I_2 w_2$  with a new weighting factor  $w_s$ , we can unify both equations in one single formula. Further, we can eliminate taking the absolute values if we shift the curve along the positive  $y$ -axis, such that all points are positive. We thus receive a simplified penalizer

$$w_s \sum_{j=0}^n Q_j \quad (9)$$

We want to find a solution where this sum becomes a small value. This is equivalent to finding a small value for the squared entity

$$\left( \sum_{j=0}^n w_s Q_j - 0 \right)^2 \quad (10)$$

If we compare this expression with Eq. (II), we realize that it has the same structure as the inner term. This inner term calculates the contribution of one control point to the overall error. If we interpret Eq. (10) in this context then 0 would be a measurement point and  $w_s$  the value of all B-spline basis functions. We can thus incorporate the smoothness constraint into the least-squares system by using a new matrix  $\hat{A}$  and point vector  $\hat{P}$ , which both contain one additional row, where

$$\hat{A}_{m+1} = [w_s \ w_s \ \dots \ w_s] \quad \text{and} \quad \hat{P}_{m+1} = 0$$

## 6 Gradient Constraint

We introduce another constraint, which corresponds to the gradient constraint used in [5], and penalizes solutions for which the first derivative at the origin is nonzero. As the ego-vehicle is standing flat on the road surface, we assume that the gradient of the road equals 0 at the camera position.

The derivative of a B-spline curve can be calculated by replacing the individual basis functions with their derivatives. We can thus implement the new constraint

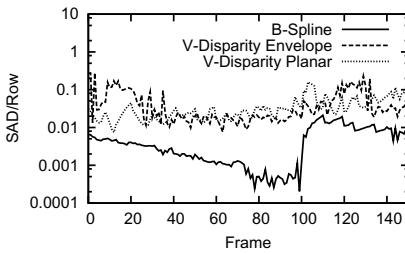
by adding another row to the system matrix  $\hat{A}$  that contains the value of the derived basis functions at position 0, multiplied with a weighting factor  $w_g$  to control the influence of the constraint. Furthermore, we need to add a new point with a value of 0 (the desired gradient) to the point vector. The new row of the resulting matrix  $\tilde{A}$ , and the new point of the point vector  $\tilde{P}$ , are thus as follows:

$$\tilde{A}_{m+2} = [w_g N'_0(0) \ w_g N'_1(0) \ \cdots \ w_g N'_n(0)] \quad \text{and} \quad \tilde{P}_{m+2} = 0 \quad (11)$$

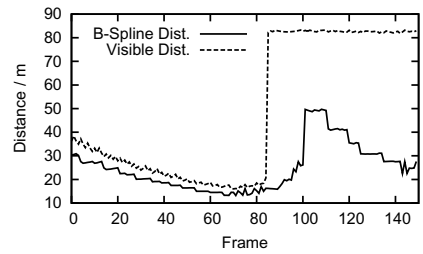
## 7 Results

To judge the performance of our new road-modeling algorithm we performed a comparative evaluation with two versions of the  $v$ -disparity method. The first version matches the method discussed in [2] and creates a polygonal chain, while the second one only selects the best matching line and thus creates a planar model. We tested both variants and our algorithm on the second synthetic driving sequence in Set 2 of EISATS [8]. The used disparities are the provided ground-truth values rounded to the nearest integer, which is the best result we could expect from a stereo matching algorithm which is not aiming at subpixel precision.

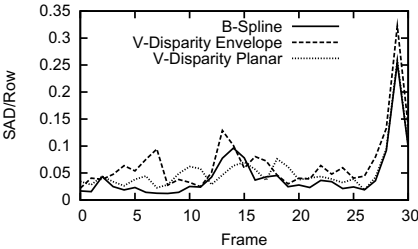
We were able to extract the road pixels from the provided ground-truth for the first 150 frames, which gives us a precise measure of the road profile at each image row. For a quantitative evaluation we calculate the *sum of absolute differences* (SAD) per image row between estimated and ground-truth road profile. We



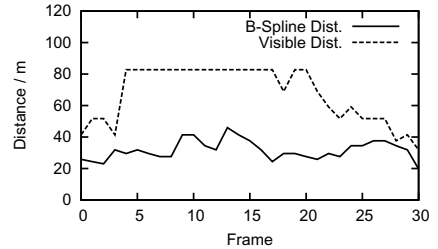
(a) Error on synthetic sequence



(b) Distances in synthetic sequence

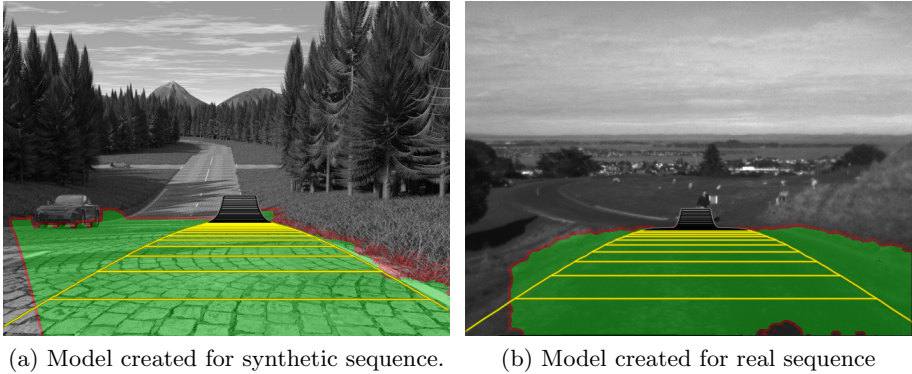


(c) Error on real sequence



(d) Distances in real sequence

**Fig. 2.** Quantitative comparison of B-spline and  $v$ -disparity road-modeling



**Fig. 3.** Examples for created B-spline road model curves

perform this evaluation only up to the last image row selected by our region-growing algorithm. This is in favor of the  $v$ -disparity approach, which does not set a distance boundary and is more inaccurate with increase in distance.

Figure 2a shows the results we receive for the three cases on a logarithmic scale. Our approach clearly outperforms both variants of the  $v$ -disparity method. The envelope-based  $v$ -disparity version does not perform any better than the simpler planar version. Figure 2b compares the fitted distance to the maximal distance at which the road is still observable with a minimum disparity. The sudden jump in visible distance is caused by driving over a hill that occludes the road in the beginning of the sequence. This is in favor of the  $v$ -disparity approach, which does not set a distance boundary and would otherwise be evaluated until the maximum visible distance.

In a second evaluation we compared the performance of all methods on a real world sequence. The used stereo sequence has been recorded on a hilly and windy road without prominent road boundaries, and should thus present a difficult challenge for any algorithm. We manually extracted the road from 30 frames and used the median disparity for the road pixels in each image row to obtain an estimate of the road profile. Figure 2c shows the comparison of this road profile to the results of the tested road-modeling techniques. Our approach still performs predominantly better than both  $v$ -disparity variants, but with a much smaller margin and not for all frames. It appears that the planar  $v$ -disparity variant performs better than the envelope version for most of the tested frames. The sudden increase in the modeling error at the end of the sequence can be explained by the much worse performance of the used stereo matching algorithm during this section. Figure 2d compares the visible and fitted distances for the tested sequence.

## 8 Conclusions

In this research we have proposed a new method for modeling the vertical road profile using B-spline curves. The method does not require a free-space estimation and has proven to work on scenes where the road is not constrained by any prominent boundaries. Examples for the performance of this method on a synthetic and real world scene are shown in Figs. 3a and 3b. In our experiments, the new method performed better than both tested versions of the popular  $v$ -disparity technique. The advance was major on the tested synthetic sequence but minor on the real world sequence. We suspect that this gap is caused by the lower accuracy of the disparity map for the real world sequence. Using better stereo matching algorithms could thus improve the results of our road-modeling technique.

Furthermore, we have found that using the envelope of best matching straight lines for the  $v$ -disparity method does not produce any better results. On the real world sequence, the performance of the envelope based  $v$ -disparity implementation produced the worst results for most of the frames, while performing roughly equal to the planar  $v$ -disparity version on the synthetic sequence. Our road-modeling method has proven to be competitive to both tested  $v$ -disparity approaches. Nevertheless, more research is required to further improve the accuracy. The method could particularly benefit from taking features of the intensity image into account and introducing a temporal filter. This could, however, distort the comparison if the  $v$ -disparity results are not filtered as well.

## References

1. Weber, J., Koller, D., Luong, Q.T., Malik, J.: Integrated stereo-based approach to automatic vehicle guidance. In: IEEE International Conference on Computer Vision, pp. 52–57 (1995)
2. Labayrade, R., Aubert, D., Tarel, J.P.: Real time obstacle detection in stereovision on non flat road geometry through “ $v$ -disparity” representation. In: IEEE Intelligent Vehicle Symposium, pp. 646–651 (2002)
3. Oniga, F., Nedevschi, S., Marc, M., Thanh, B.: Road surface and obstacle detection based on elevation maps from dense stereo. In: IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 859–865 (2007)
4. Nedevschi, S., Danescu, R., Frentiu, D.: High accuracy stereovision approach for obstacle detection on non-planar roads. In: IEEE Intelligent Engineering Systems (INES), pp. 292–297 (2004)
5. Wedel, A., Badino, H., Rabe, C., Loose, H., Franke, W., Cremers, D.: B-spline modeling of road surfaces with an application to free-space estimation. IEEE Transactions on Intelligent Transportation Systems 10, 572–583 (2009)
6. Eberly, D.: Least-Square Fitting of Data with B-Spline Curves. Geometric Tools, LLC (2008)
7. Viola, P., Jones, M.: Robust real-time object detection. International Journal of Computer Vision 57, 137–154 (2002)
8. The University of Auckland: Multimedia Imaging Technology Portal – EISATS (2010), <http://www.mi.auckland.ac.nz/EISATS> (viewed April 19, 2010)



# The Six Point Algorithm Revisited

Akihiko Torii<sup>1</sup>, Zuzana Kukelova<sup>2</sup>, Martin Bujnak<sup>3</sup>, and Tomas Pajdla<sup>2</sup>

<sup>1</sup> Tokyo Institute of Technology, Tokyo, Japan  
torii@ctrl.titech.ac.jp

<sup>2</sup> CMP, Czech Technical University in Prague, Prague, Czech Republic  
{kukelova,pajdla}@cmp.felk.cvut.cz

<sup>3</sup> Bzovicka 24, 85107, Bratislava, Slovakia

**Abstract.** This paper presents an algorithm for estimating camera focal length from tentative matches in a pair of images, which works robustly in practical situations such as automatic computation of structure and camera motion from unknown photographs, e.g. from the web or from various instruments mounted on a vehicle. We extend the standard 6-pt algorithm based on the observations: (i) the quality of the estimation of this algorithm is strongly correlated with the ratio of the singular values of the essential matrix computed from inliers, which is calibrated by using the estimated focal length, returned by RANSAC and (ii) the reprojection error of the affine camera model, fit to the inliers, predicts the uncertainty in the estimated focal length. Furthermore, for scenes with dominant plane we propose a novel algorithm calculating relative orientation and unknown focal length given a plane homography and a single off the plane point correspondence. The performance of the proposed algorithm is demonstrated on a set of real images having different focal lengths.

## 1 Introduction

Several systems for automatic structure from motion computation have been recently published, developed and made available [1,3,4,6]. All these systems need to know internal camera calibration to recover camera poses. Without exception, they all use the 5-pt algorithm [7,8,9] to compute the relative camera poses from 5 image matches by RANSACing [10] tentative image matches [11,12,13,14].

It turns out that with modern digital cameras, two out of the five internal calibration parameters [16, p.157], the skew and the pixel aspect ratio, can always be safely set to 0 and 1, respectively. The remaining three parameters, the principal point and the focal length would, however, need to be autocalibrated [16] in order to allow working with images from completely unknown sources, e.g. web, or allowing free image scaling and cropping. Despite the vast body of literature on the autocalibration, all the above mentioned systems avoid it since it is an ill conditioned process in general. Instead, they adopt more practical approach by assuming that the principal point is in the center of the image and the focal length is correctly stored in image EXIF.

Somewhat surprisingly, the above systems do not even autocalbrate the focal length which would be very practical since zooming is one of the most common camera control. The generalization of the calibrated 5-pt algorithm for cameras with unknown (but

same) focal length, the 6-pt algorithm, is well known since [17] and has been further simplified and enhanced [9]. So, why is it not used?

The main problem with the 6-pt algorithm is that it fails or returns rather imprecise results in many real situations due to presence of critical motions [19]. Critical motions for a camera pair with unknown focal length are quite common since they appear, e.g., for camera pure translation or revolute motions which keep camera optical axes intersecting. It can be shown that setting focal lengths incorrectly skews the reconstruction which means that it is necessary to initialize  $f$ 's sufficiently close to correct values, which is in all mentioned situations impossible. The importance of calibration priors has been made clear in [20].

The bundler [1] can often reconstruct the scene sufficiently well even if there are some images without calibration priors. This is true especially when the scene is captured by a large number of images and therefore there is a high chance of finding an initial seed reconstruction from some image pair with focal length close to the prior expected. Then, other focal lengths can be estimated by a direct linear transfer method and further improved by bundle adjustment on top of the good initial seed reconstruction. In contrast, focal length autocalibration becomes really necessary when reconstructing a scene from a small number of images captured by very different instruments.

In this paper we analyze limits of the 6-pt algorithm performance and develop its robust version for RANSAC “which works”. We demonstrate that (i) the quality of the estimation of this algorithm is strongly correlated with the ratio of the singular values of the fundamental matrix computed from inliers returned by RANSAC and (ii) the reprojection error of the affine camera model, fit to the inliers, predicts the uncertainty in the estimated focal length. Based on our observations we develop several criteria and extend the existing 6-pt algorithm. Furthermore, for scenes with dominant plane we propose an algorithm calculating relative orientation and focal length given a plane homography and a single off the plane point correspondence. The performance of our algorithm is demonstrated on a set of real images having three different focal lengths.

## 2 Limits of the 6-pt Algorithm

It is known that the 6-pt algorithm [17] is rarely used in structure from motion pipelines due to the several problems it has. These problems can be divided into the following categories: 1. Problems with critical motions e.g. when optical axes of the cameras are parallel or intersecting [19]. 2. Planar scenes. 3. Camera pairs with different focal lengths. 4. Cameras with large focal lengths.

In the first two situations it is not possible to estimate reasonable epipolar geometry (EG) and focal length because there exist several Euclidean interpretations of the given structure. The third situation can't be handled using standard 6-pt algorithm, since this algorithm is dedicated to cameras with unknown but same focal length. The last situation is close to the critical configuration with planar scene resp. to the affine camera case. In practice, any camera configuration that is close to the critical one can cause problems and give inaccurate results. Unfortunately, most of these configurations are common in real situations e.g. when taking photos from a moving car or moving around an object.

### 3 Selection Criteria

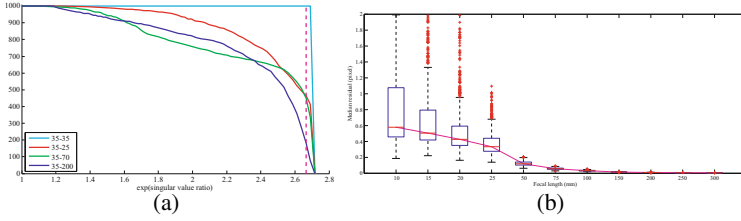
*Singular value ratio.* We have found that the quality of the estimation of the 6-pt algorithm is strongly correlated with the ratio of the two largest singular values of the fundamental matrix computed from inliers returned by RANSAC. This singular value ratio is close to one for “good” image pairs, like the pairs with the same or closely the same focal lengths in a general configuration. On the other hand this value is usually rather small for “bad” image pairs, like image pairs with large focal lengths, pairs with different focals, or critical configurations.

Figure 1(a) shows the cumulative graph of singular value ratios computed between pairs with the same and different focal lengths. The cumulative graph is generated by counting the number of image pairs having the singular value ratio greater than the value on x-axis. A general 3D scene consisting of 100 points uniformly distributed in a sphere of 500 mm radius has been generated. Two images were simulated as if taken by one camera with fixed  $f = 35$  mm lens and the others varied  $f = 25, 35, 70,$  and  $200$  mm. Gaussian noise are used as image noise with standard deviation  $\sigma =$  half a pixel in 1000 resolution. The camera centers are approximately 1700 mm away from the scene. The relative camera motions were created by 1000 random motions with motion size 500 mm. The cyan line which corresponds to the singular value ratio computed from pair of images with the same focal lengths is clearly distinct from the singular value ratios computed by pair of different focal lengths images.

*Affine residual.* When objects in a scene are distant from an observer and images are taken by a standard perspective camera with a very large focal length, the image projection model is adequately expressed by the orthographic projection. This is because all rays incident to the image plane are nearly parallel and then the perspective effects induced by the central projection become weak. In such a case, the relative camera motion will be better fit by affine epipolar geometry [16] and therefore it makes difficult to estimate correct EG and focal length. Although the estimation of EG cannot be improved, it is still possible to detect such pair of images with a large focal length by computing affine epipolar geometry from all supports of the EG and by evaluating the residuals.

Figure 1(b) shows boxplot of the median residuals w.r.t. affine epipolar geometry computed from a pair of images with focal lengths 10 to 300 mm. The geometric configurations of scene and cameras are exactly same as the setting used in the singular value ratio. The graph clearly shows that the larger the focal lengths are, the smaller the median residuals w.r.t. the affine epipolar geometry.

*Planarity test.* Using six point correspondences the fundamental matrix  $F$  can be parameterized as a linear combination of a basis  $F_1, F_2, F_3$  of the space of all compatible fundamental matrices. We can write  $F = xF_1 + yF_2 + F_3$ . It is known [21] that for six points on the plane, all matrices  $F_1, F_2, F_3$  in this space have rank 2. Therefore, we can detect planarity by testing whether arbitrary linear combination of  $F_1, F_2, F_3$  has rank 2. In case of planarity detected, the fundamental matrix and the focal length can be estimated by using the plane+parallax algorithm, if there exists at least one point out of the plane, as proposed in the following section.



**Fig. 1.** Criteria of selecting focal lengths. (a) Cumulative graph of showing singular value ratios computed between pairs of 35-35, 35-25, 35-70, 35-200 mm focal length images. The magenta dashed line indicates the threshold which we used in real experiment in Section 6. (b) Median residuals of the affine fundamental matrix fit between a pair of perspective images with focal length 10 to 300 mm.

## 4 Plane+Parallax for Cameras with Unknown Focal Length

### 4.1 Problem Formulation

In this section we formulate the problem of estimating epipolar geometry i.e. the essential matrix  $E$  and the unknown focal length from images of five points, four of which are coplanar and one is off the plane.

The images of four coplanar points define the homography  $H$ . For all 3D points  $X$  lying on the plane defined by these four points holds  $x' = Hx$ , where  $x'$  and  $x$  are projections of point  $X$  in the first and second view. The important property holds for points off this plane. The image  $x'$  of some off the plane 3D point  $X$  in the second view and the point  $\tilde{x}' = Hx$ , mapped by the homography  $H$ , lie on the epipolar line of  $x$ , since both are images of points on the ray through  $x$ . Therefore  $l'_x = x' \times Hx$  is an epipolar line in the second view. Thus we can write

$$e' = x' + sv, \tag{1}$$

where  $v$  is the normalized vector of the epipolar line  $l'_x$  and  $s$  is an unknown parameter. Another possibility which holds also for epipoles at infinity is to write

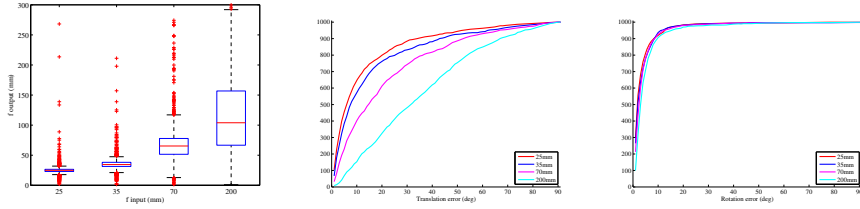
$$e'^T (x' \times Hx) = 0, \tag{2}$$

which means that the epipole  $e'$  lies on the epipolar line  $x' \times Hx$ . The epipoles at infinity appear in the case of special motion, where the translation is parallel to the image plane, and the rotation axis is perpendicular to the image plane. Since this motion is critical for the 6-pt algorithm in the following we will use the parameterization  $\square$ .

It is known  $\square$  that for the fundamental matrix  $F$  holds

$$F = [e']_{\times} H. \tag{3}$$

Therefore once the epipole  $e'$  and the homography  $H$  induced by any plane are estimated, the fundamental matrix can be computed uniquely. Note that this fundamental matrix is already singular.



**Fig. 2.** Performance evaluation of the 4 + 1 plane+parallax algorithm. (a) is the boxplot of estimated focal lengths. (b) is the translation error  $\angle(\mathbf{t}_{estimate}, \mathbf{t}_{true})$ . (c) is rotation error computed from the angle of the rotation of  $\angle(\mathbf{R}_{estimate}^{-1} \mathbf{R}_{true})$ .

The well known algorithm [16] for computing  $F$  given the homography induced by a plane uses images of six points, four of which are coplanar and two are off the plane. The images of four coplanar points define the homography, and the remaining two points off the plane define two epipolar lines which intersect at the epipole  $e'$ . A focal length can be computed from a given fundamental matrix [16] but it is known as a complicated problem because the estimated focal lengths often become complex. In contrast, this is not a problem for our algorithm since the focal length is calculated together with the essential matrix. In our case we have only one point off the plane. Therefore we can only parameterize  $e'$  with one unknown parameter  $s$  using Equation 1.

We assume that both cameras are calibrated up to an unknown common focal length  $f$ . Then for the essential matrix  $E$  holds

$$E = K^T FK, \tag{4}$$

where  $K \simeq \text{diag}([f \ f \ 1])$  is a diagonal calibration matrix. It is known [16] that two singular values of the essential matrix  $E$  are equal and the third is zero. This can be written as

$$2EE^T E - \text{trace}(EE^T)E = 0. \tag{5}$$

Matrix equation 5 gives nine equations in elements of  $E$  from which three are algebraically independent. If we express  $E$  using equations 1, 3 and 4 we obtain nine fifth degree equations in two unknowns  $s$  and  $f$  or in  $w = 1/f^2$ , from which six are linearly independent. From these equations only two are algebraically independent since  $E$  is already singular. Therefore we have six equations in two unknowns which result in five solutions for  $f$  and  $s$ . Next we show how these equations can be solved.

### 4.2 The 4 + 1 Plane+Parallax Solver

The problem formulation from Section 4.1 can be easily solved using Gröbner basis method for solving systems of polynomial equations. This method was recently successfully used to solve many minimal problems in computer vision [17,22,8,23,24,25].

The Gröbner basis method is an algebraic method based on polynomial ideal theory which generates special bases of ideals, called Gröbner bases [27]. Gröbner bases have

the same solutions as the initial system of polynomial equations but are often easier to solve. Using these bases, special matrices, also called action matrices can be constructed. These matrices have a nice property, that solutions to a system of polynomial equations can be easily obtained from their eigenvalues and eigenvectors. Therefore these matrices can be viewed as a generalization of well known companion matrices used for solving one polynomial equation in one unknown. More details of the Gröbner basis method and its applications in computer vision can be found in [26,27,23,28]. Also, automatic generator of polynomial equation solvers based on this Gröbner basis method is proposed and demonstrated in [29].

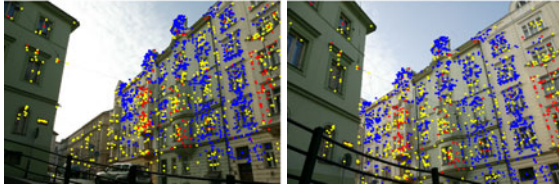
For our problem formulation with two unknowns results this method to the solver which consists of one Gauss-Jordan (G-J) elimination of the  $10 \times 15$  matrix and computation of eigenvectors of the  $5 \times 5$  action matrix. The  $10 \times 15$  is obtained by adding monomial multiples of initial six fifth degree polynomial equations up to total degree six and then removing unnecessary polynomials using method from [24]. Eigenvectors of the  $5 \times 5$  action matrix give us up to five real solutions for  $f$  and  $s$ , from which we compute  $e'$  and the essential matrix  $E$ .

Figure 2 demonstrates performance of the  $4 + 1$  plane+parallax algorithm. 1000 samples of 5 coplanar points + one out of the plane are generated in a 3D scene. A homography is computed from 5 tuple and focal length and EG is estimated by the P+P algorithm. Two images were simulated as if taken by a pair of camera with  $f = 25, 35, 70,$  and  $200$  mm lens. For each focal length, 1000 pairs of images are generated with randomly generated camera motion but keeping the baseline 500 mm. Gaussian noise is used as image noise with standard deviation  $\sigma =$  half a pixel in 1000 resolution. Figure 2(a) shows the estimated focal lengths. Figures 2(b) and (c) show the cumulative histogram of translation error  $\angle(\mathbf{t}_{estimate}, \mathbf{t}_{true})$  evaluated as the angle between the estimated and the true translation direction and rotation error computed from the angle of the rotation of  $\angle(\mathbf{R}_{estimate}^{-1} \mathbf{R}_{true})$ , which is ideally zero [16]. The translation tends to be estimated less accurately compared to the focal length and rotation. The precision of translation direction is strongly correlated with the precision of homography estimation so that translation is less accurately estimated in the plane+parallax. We strongly recommend to refine EG by using the 5pt algorithm using points calibrated by the focal length estimated by the plane+parallax.

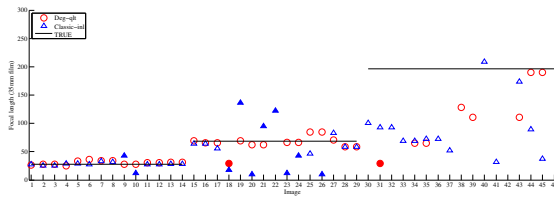
## 5 The Pipeline of Finding Focal Length

For the robust estimation of focal length estimation from a pair of images, we use DEGENSAC [21] which checks the degeneracy of samples and refines the hypothesis by using the plane-and-parallax algorithm when there exists homography supported by many matches. The main modification of using DEGENSAC in our 6 point case is the H-degeneracy detection, which is the detection of coplanar samples, and the plane-and-parallax algorithm. For the H-degeneracy detection of 6pt case, it is sufficient to check if there exists any 5 tuple consistent to a plane homography. Figure 3 shows an example of epipolar geometry and focal length estimation by using DEGENSAC with the  $4 + 1$  plane+parallax solver.

For the selection of pairs of images having the same focal length, we first generate the image similarity matrix using the visual vocabulary technique [30] and select 10



**Fig. 3.** Example of epipolar geometry and focal length estimation. The estimated focal length is 30.1 mm and the EXIF focal is 27.5mm. The best model is found by the plane+parallax algorithm in DEGENSAC . Blue dots are the supports of H. Red dots are new supports of Fcomputed by the plane+parallax algorithm starting from H. Yellow dots are the tentative matches generated by matching SURF image features [14].



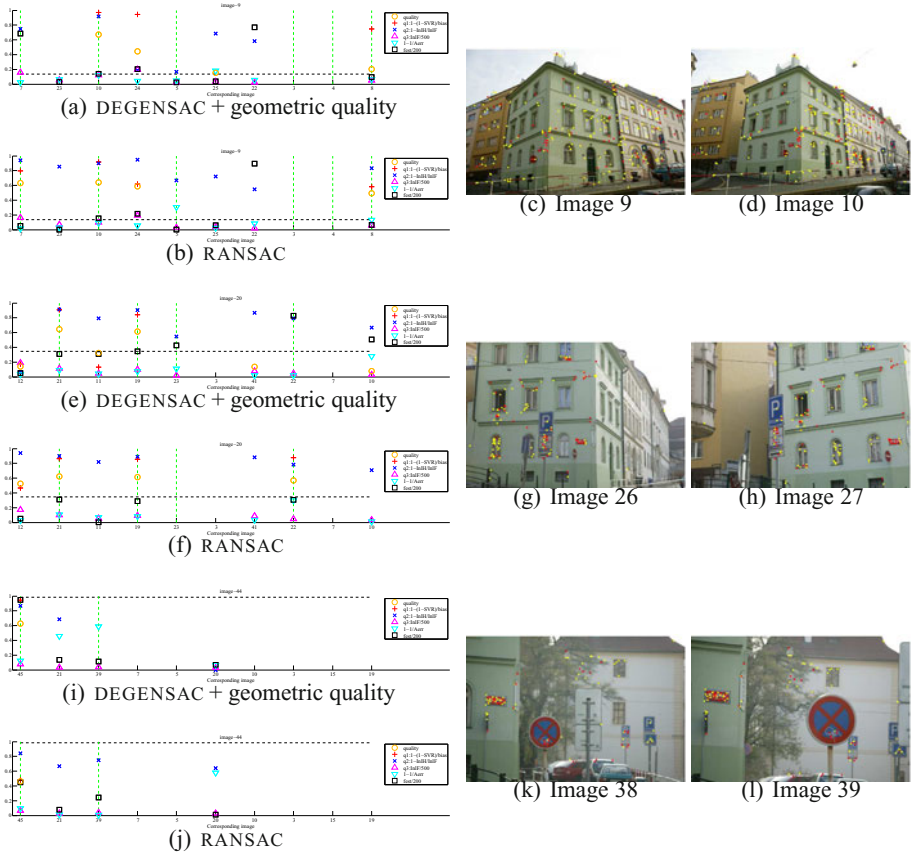
**Fig. 4.** Performance evaluation of focal length selection. The red  $\circ$  is the focal length estimated and selected by our pipeline. The blue  $\triangle$  is the focal length computed by classic RANSAC . The filled markers indicate the focal length are computed by a pair of images with different focal lengths (false selection).

most similar images for every target image. The camera motion  $E$ , the focal length  $f$ , the matches  $\mathbf{m}$  supporting  $E$  and  $f$ , and matches  $\mathbf{m}^H$  supporting homography  $H$  are computed by running DEGENSAC for each pair of images. Then, the ratio  $s$  of singular values of  $E^S$  computed by 8-pt algorithm (least squares) and the median  $r$  of reprojection errors w.r.t. the affine fundamental matrix are computed. We reject such a pair of images if it satisfies any one of criteria that the singular value ratio  $s < 0.98$ , the median affine reprojection error  $r < 0.5$  pixel, and the number of supports  $|\mathbf{m}|$  of  $E$  is less than  $|\mathbf{m}^H|$  of  $H$ . These criteria reject pairs of images likely degenerated or computed from image pairs having different focal lengths. Then, the best pair of images are selected by computing the confidence of the estimated model such that  $q = (q_1 + q_2 + q_3)/3$  where  $q_1 = 1 - (1 - s)/(1 - 0.98)$ ,  $q_2 = |\mathbf{m}|/500$ ,  $q_3 = 1 - |\mathbf{m}^H|/|\mathbf{m}|$ . In this experiment, the focal lengths for all pairs of images are computed but in practice it is possible to test only similar pairs using the image similarity matrix as already described.

## 6 Experiments

We demonstrate our pipeline on real images of a city scene. The dataset involves three different focal lengths images: 14 images of 27.5 mm, 15 images of 68.8 mm, and 17 images of 196.5 mm focal length of a unit in standard film size.





**Fig. 5.** Quality scores for (c) 'Image 9', (g) 'Image 26' and (k) 'Image 38' having the focal lengths 27.5 mm, 68.8 mm and 196.5 mm, respectively. (a) and (b) show the quality scores computed between Image 9 and the other 10 images selected using the image similarity. Orange 'o' is the quality score computed from red '+', blue '\*', and magenta 'Δ' defined in Section. Cyan '∇' shows the median affine residual. Black '□' is the estimated focal length. The horizontal black dashed line is the ground truth value of focal length and the vertical green dashed line indicates the pair has the same focal length. (c) is the image 9. (d) is the selected image 10 as the best pair by our pipeline. The yellow and red dots are the tentative matches and the supports of EG, resp. In the same manner, the quality scores for 'Image 26' and 'Image 38' are shown in (e)-(f), and (i)-(j). (h) and (l) are again the pairs selected by our method.

Figure 5 shows results of the quality scores computed by our pipeline and the corresponding images giving the best focal length estimate. The quality scores computed w.r.t. 'Image 9', which has 27.5 mm focal length, are shown in Figure 5(a). The total quality  $q$  is in orange 'o' and the components  $q_1$ ,  $q_2$ , and  $q_3$  are shown in red '+', blue '\*', and magenta 'Δ'. All scores are normalized to fit in the graph and especially  $q_1$  is divided by a certain constant to enhance the differences. The median affine residual  $r$  used for rejecting likely degenerated image pairs is in cyan '∇'. The estimated focal



length w.r.t. each pair is in black '□'. The horizontal black dashed line is the ground truth value of focal length and the vertical green dashed line indicates the pair has the same focal length. (c) is the reference image 9 and (d) is the selected image 10 as the best pair. The yellow and red dots are the tentative matches and the supports of EG in (c) and (d). In Figure 4, the red  $\circ$  is the focal length estimated by our pipeline. The blue  $\triangle$  is the focal length computed by classic RANSAC. The pair having the largest number of support is simply selected as the best pair. The filled markers indicate the focal length are computed by a pair of images with different focal lengths (false selection). The red  $\circ$  is not printed if all images to the target image are rejected by our criteria. Our proposed pipeline successfully recovers all except only one 27.5 mm and 68.8 mm focal lengths selecting correct pair of images. It is very difficult to estimate large focal length 196.5 mm but two of them are correctly estimated. Note that most of inaccurate estimates are rejected by our criteria so that there are fewer red  $\circ$  markers in 196.5 mm images.

## 7 Conclusions

An algorithm for robustly estimating camera focal length in a pair of images is presented based on the geometric analysis of limits on the standard 6-pt algorithm. Through the experiments on synthetic and real data, we showed the quality of the estimation of epipolar geometry and focal length is strongly correlated with the ratio of the singular values of the essential matrix computed from inliers returned by RANSAC. Also, the reprojection error of the affine camera model, fit to the inliers, predicts the uncertainty in the estimated focal length. The extension of the existing 6-pt algorithm with the 4 + 1 plane+parallax solver improved the quality of the estimation.

## Acknowledgements

This research was supported by EC project FP7-SPACE 218814 PRoVisG and by Czech Government under the research program MSM6840770038.

## References

1. Snavely, N., Seitz, S.M., Szeliski, R.S.: Photo Tourism: Exploring image collections in 3D. In: SIGGRAPH, pp. 835–846. Implementation at, <http://phototour.cs.washington.edu/bundler/>
2. Goesele, M., Snavely, N., Curless, B., Hoppe, H., Seitz, S.M.: Multi-view stereo for community photo collections. In: ICCV (2007)
3. Snavely, N., Seitz, S., Szeliski, R.: Skeletal graphs for efficient structure from motion. In: CVPR (2008)
4. Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J.M.: Modeling and recognition of landmark image collections using iconic scene graphs. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 427–440. Springer, Heidelberg (2008)
5. Photosynth, <http://photosynth.net>

6. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building Rome in a day. In: ICCV (2009)
7. Nister, D.: An efficient solution to the five-point relative pose. *IEEE PAMI* 26(6), 756–770 (2004)
8. Stewénius, H., Engels, C., Nister, D.: Recent developments on direct relative orientation. *ISPRS J. of Photogrammetry and Remote Sensing* 60, 284–294 (2006)
9. Kukulova, Z., Bujnak, M., Pajdla, T.: Polynomial eigenvalue solutions to the 5-pt and 6-pt relative pose problems. In: *BMVC 2008* (2008)
10. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM* 24(6), 381–395 (1981)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. *IJCV* 60(2), 91–110 (2004)
12. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: *BMVC 2002*, pp. 384–393 (2002)
13. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. *IJCV* 60(1), 63–86 (2004)
14. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). *CVIU* 110(3), 346–359 (2008)
15. Hartley, R.: In defence of the 8-point algorithm. In: *CVPR 1995*, pp. 1064–1070 (1995)
16. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge (2003)
17. Stewenius, H., Nister, D., Kahl, F., Schaffalitzky, F.: A minimal solution for relative pose with unknown focal length. In: *CVPR 2005*, pp. 789–794 (2005)
18. Li, H.: A simple solution to the six-point two-view focal-length problem. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*. LNCS, vol. 3954, pp. 200–213. Springer, Heidelberg (2006)
19. Kahl, F., Triggs, B.: Critical motions in Euclidean structure from motion. In: *CVPR 1999*, pp. 366–372 (1999)
20. Fitzgibbon, A., Robertson, D., Criminisi, A., Ramalingam, S., Blake, A.: Learning priors for calibrating families of stereo cameras. In: *ICCV 2007* (2007)
21. Chum, O., Werner, T., Matas, J.: Two-view geometry estimation unaffected by a dominant plane. In: *CVPR 2005*, pp. 772–779 (2005)
22. Stewenius, H., Nister, D., Oskarsson, M., Astrom, K.: Solutions to minimal generalized relative pose problems. In: *OMNIVIS 2005* (2005)
23. Kukulova, Z., Pajdla, T.: A minimal solution to the autocalibration of radial distortion. In: *CVPR 2007* (2007)
24. Bujnak, M., Kukulova, Z., Pajdla, T.: A general solution to the P4P problem for camera with unknown focal length. In: *CVPR 2008* (2008)
25. Byröd, M., Kukulova, Z., Josephson, K., Pajdla, T., Åström, K.: Fast and robust numerical solutions to minimal problems for cameras with radial distortion. In: *CVPR 2008* (2008)
26. Cox, D., Little, J., O’Shea, D.: *Using Algebraic Geometry*, 2nd edn., vol. 185. Springer, Heidelberg (2005)
27. Cox, D., Little, J., O’Shea, D.: *Ideals, Varieties, and Algorithms*. Springer, Heidelberg (2007)
28. Byröd, M., Josephson, K., Åström, K.: Improving numerical accuracy of Gröbner basis polynomial equation solver. In: *ICCV 2007* (2007)
29. Kukulova, Z., Bujnak, M., Pajdla, T.: Automatic generator of minimal problem solvers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008*, Part III. LNCS, vol. 5304, pp. 302–315. Springer, Heidelberg (2008)
30. Sivic, J., Zisserman, A.: Video google: Efficient visual search of videos. In: *CLOR 2006*, pp. 127–144 (2006)

# Multi-body Segmentation and Motion Number Estimation via Over-Segmentation Detection

Guodong Pan and Kwan-Yee Kenneth Wong

Department of Computer Science, The University of Hong Kong, Hong Kong

**Abstract.** This paper studies the problem of multi-body segmentation and motion number estimation. It is well known that motion number plays a critical role in the success of multi-body segmentation. Most of the existing methods exploit only motion affinity to segment and determine the number of motions. Motion number estimated in this way is often seriously affected by noise. In this paper, we recast the problem of multi-body segmentation and motion number estimation into an over-segmentation detection problem, and introduce three measures, namely loss of spatial locality (LSL), split ratio (SR) and cluster distance (CD), for over-segmentation detection. A hierarchical clustering method based on motion affinity is applied to split the motion clusters recursively until over-segmentation occurs. Over-segmentation is detected by Kernel Support Vector Machines trained under supervised learning using the above three measures. We leverage on Hopkins155 database to test our method and, with the same motion affinity measure, our method outperforms another state-of-the-art method. To the best of our knowledge, this paper is the first to tackle the problem of multi-body segmentation and motion number estimation from the perspective of over-segmentation detection.

## 1 Introduction

To reconstruct or understand a dynamic scene consisting of multiple moving objects observed by a static or moving camera, the trajectories of image features are often segmented using their motion affinity. Estimation of the motion number is critical to such a multi-body segmentation, and its failure often leads to a high error rate in the motion segmentation. In this paper, we refer to *motion number* as the number of independently moving objects in a scene.

Most of the existing works, if not all, exploit only motion affinity to segment and determine the number of motions. In the factorization method presented by Costeira and Kanade [1], the motion number was determined by sorting the shape interaction matrix and detecting blocks via minimizing the Frobenius norm of the shape interaction matrix subject to some physical constraints. This detection method suffers a lot from noisy data, especially when the noise level is high. Gear [2] converted the data matrix into an echelon form, and features of the same motion shared the same zero positions in the synthetic case. The motion number was then given by the number of different configurations. He also provided a bipartite graph model for real data with noise, and tried to explain

it with probabilistic models. Nevertheless, he admitted that real data was too complex to be explained by this model. Vidal et al. [3] presented the concept of multi-body fundamental matrix for the segmentation problem, and retrieved the motion number from the rank of the matrix of Veronese mapping of trajectories. It is a non-trivial problem to estimate the rank of a matrix with noise. This method also requires a minimum number of trajectories for each motion, which may not be practical. In [4], trajectories were clustered based on the distance of subspace using spectral clustering. In [5], the authors introduced the ordered residual metric, and clustered the trajectories also by spectral clustering. For the spectral clustering method in [6], the motion number was equivalent to the multiplicity of the zero eigenvalue of graph Laplacian, and the affinity matrix of trajectories was usually generated in such a manner as Normalized Cut [7]. The parameters of this model are quite influential, but are difficult to adjust for different applications. From the perspective of information theory, Ma et al. [8] modelled the problem via lossy data coding and compression, with the assumption that the mixed data were drawn from a mixture of Gaussian distributions. Given data to be compressed and a distortion criterion, the motion number and segmentation were obtained by minimizing the coding length. This method generalizes the problem but only considers data with mixtures of Gaussian distributions. [9] and [10] tackled the motion number estimation problem with a sampling method based on Torr's extension of Schwarz' BIC approximation [11]. Recent work [12] applied the Dirichlet Process Mixture Models to the motion hypotheses, and obtained the motion number when the process converged. However, with a median scale of disturbance and noise, the converged state was unsteady. [13] focused on the change of motion number in video and proposed a method based on an outlier detection approach. Most of the methods above determine the motion number only from the motion information, except [9] and [10] which used a local sampling scheme [14].

There is no doubt that motion affinity is a key factor for motion number estimation. However, this is by no means the only factor that matters. In this paper, we recast the problem of multi-body segmentation and motion number estimation into an over-segmentation detection problem, and introduce three measures, namely *loss of spatial locality* (LSL), *split ratio* (SR) and *cluster distance* (CD), to detect the occurrence of over-segmentation. A hierarchical clustering method based on an improved ordered residual metric is applied to split the motion clusters recursively until over-segmentation occurs. Supervised learning is employed to train Kernel Support Vector Machines using the above three measures motion affinity measure, our method outperforms another state-of-the-art method. To the best of our knowledge, this paper is the first to tackle the problem of multi-body segmentation and motion number estimation from the perspective of over-segmentation detection.

The rest of paper is organized as follows. Section 2 states our problem statement. Section 3 introduces the proposed measures for over-segmentation detection. The hierarchical clustering method and classifiers for over-segmentation

detection are described in Section 4. In Section 5, experiments and comparisons are presented. Finally conclusion and future work are discussed in the Section 6.

## 2 Problem Statement

Suppose several rigid objects are moving independently in a scene with different 3D motions, and a video camera is used to observe them. Feature points of the objects and the background are tracked through the video sequence. The problem of multi-body segmentation is to find the number of rigid motions and group the trajectories according to their motion affinity. Motion affinity refers to the degree to which motions share similar rotation and translation in 3D space. In this paper, we focus on objects in rigid motions and only consider the case when different moving objects have different 3D motions. We assume that all features are visible and tracked throughout the video sequence.

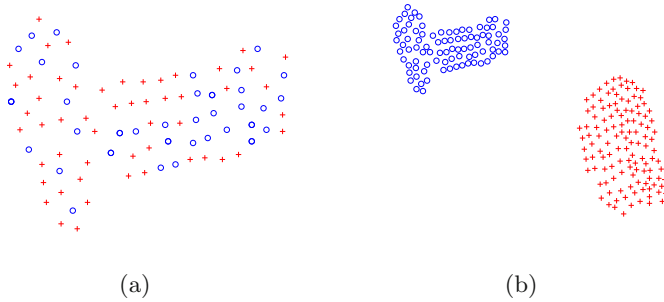
## 3 Measures for Over-Segmentation Detection

In this paper, the motion number is estimated by a recursive splitting approach. An initial motion cluster containing all the trajectories is recursively split into smaller clusters until over-segmentation occurs. When the recursion stops, the number of the resulting motion clusters simply gives the motion number. In the following subsections, we will introduce three measures for over-segmentation detection.

### 3.1 Loss of Spatial Locality

Assume that the moving objects are not transparent. Feature points of the same motion often scatter locally unless occlusion exists. Without occlusion, if two sets of features segmented into two different motions overlap, these features are likely being over-segmented. An example is shown in Fig. 1, where plus and circle marks denote features segmented into two different motions. The segmentation in Fig. 1(b) is more reasonable than that in Fig. 1(a) because there is no overlapping of the features, and hence shape integrity is not violated. Obviously, the overlapping of features in different motion clusters is a strong cue for over-segmentation.

Based on the above observation, we introduce a measure, namely *loss of spatial locality* (LSL), for over-segmentation detection. Given a motion affinity measure, a dataset can be divided into a number of motion clusters. For each element in a cluster, the number of its neighbors belonging to a different cluster is counted. LSL is defined as the total sum of such a number for all elements in all clusters, and it provides a measure for the degree of overlapping. If a feature set of the same motion is segmented into two motion clusters with a perfect motion affinity measure, every feature will have a probability of 0.5 to be selected into either



**Fig. 1.** Plus and circle marks denote features segmented into two different motions. (a) Overlapping of features segmented into different motions suggests the occurrence of over-segmentation. (b) There is no overlapping of the features and hence shape integrity is not violated.

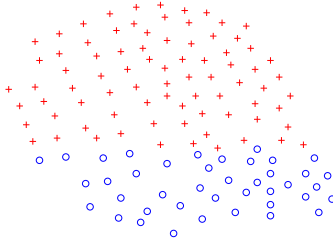
cluster. A high LSL score would therefore mean the clusters are highly overlapped and vice versa. For simplicity,  $K$ -Nearest Neighbor is used in determining the neighbors of a feature, and LSL is formulated as

$$LSL = \frac{1}{FN} \sum_{f=1}^F \sum_{i=1}^N G(x_{i,f}, k), \quad (1)$$

where  $F$  is the number of frames in the sequence,  $N$  is the number of feature points,  $x_{i,f}$  is the  $i$ -th feature point in the  $f$ -th frame,  $G(x, k)$  is the number of neighbor points belonging to a different cluster within the  $k$ -nearest point set of  $x_{i,f}$  in term of image distance.

### 3.2 Split Ratio and Cluster Distance

Over-segmentation can also occur when there is no overlapping of feature sets. This can happen when the motion affinity measure is too sensitive which segments features on a rigid object into non-overlapping but adjacent motion clusters (see Fig. 2). For example, consider a car translating and rotating at a road junction. Motion affinity between features in the front (at the back) of the car would often score higher than those between the front and the back of the car. Consequently, features in the front of the car would often be segmented into one motion, and those at the back would be segmented into another motion. Obviously, LSL cannot detect this type of over-segmentation. Nonetheless, human can perceive such features sharing one single motion because (1) these non-overlapping clusters are relatively close to each other, and (2) they share similar motions. Based on these observations, two further measures, namely *split ratio* (SR) and *cluster distance* (CD), are introduced. SR is defined as the ratio of the smallest image distance between features in separate clusters to the largest one. It provides a measure for the distance between two non-overlapping clusters with respect to their sizes. Over-segmentation would produce a low SR score.



**Fig. 2.** Over-segmentation can also occur when there is no overlapping of feature sets. This can happen when the motion affinity measure is too sensitive which segments features on a rigid object into two non-overlapping but adjacent motion clusters.

CD is defined as the distance between two cluster centers in the motion space. It measures how similar the motions of the two clusters are. Over-segmentation would produce a low CD score.

## 4 Hierarchical Clustering with Supervised Classifiers for Over-Segmentation Detection

As mentioned before, the problem of multi-body segmentation is recast into an over-segmentation detection problem. A hierarchical clustering approach is adopted to recursively split the motion clusters until over-segmentation occurs. Initially, all trajectories are considered as one single motion cluster. An improved Ordered Residual metric is employed to split each motion cluster in two smaller clusters. This corresponds to building a binary tree in which the root node contains all the trajectories. Each split will produce two child nodes, the union of which is their parent node. After each split, classifiers trained under supervised learning are used to detect the occurrence of over-segmentation based on the previously introduced measures, namely loss of spatiality locality (LSL), split ratio (SR) and cluster distance (CD). If over-segmentation is detected in the split at a particular motion cluster, its child nodes will be removed from the binary tree and further splitting of its child clusters will be prohibited. Alg. 1 summarizes the algorithm of the proposed hierarchical clustering method. The improved Ordered Residual metric used for clustering and the classifiers used for over-segmentation detection will be described in detail in the following subsections.

### 4.1 Dual Pass Ordered Residual Method

Several motion affinity measures have been mentioned in Section 1, such as shape interaction matrix [1], Local Subspace Affinity [4], and Ordered Residual [5]. Among these measures, the Ordered Residual method strongly interests us since it provides a more robust statistic estimation of motion affinity. In this paper, we propose an improved version of this method called *Dual Pass Ordered Residual method*, which is computational more efficient than the original method

**Algorithm 1.** Algorithm of the hierarchical clustering method.

---

```

Track image features to produce the trajectory data  $W$ ;
Estimate the motion affinity  $K$  between each trajectory using Dual Pass Ordered
Residual method;
Dimension reduction: Project  $K$  onto the 4-D subspace corresponding to the 4 largest
singular values and get a 4-D point set  $D$ ;
Create an empty queue  $Q$  and add a node  $R$  containing  $D$  to it;
Create an empty binary tree  $T$  and add  $R$  as the root node;
while  $Q$  not empty do
  Retrieve a node  $N$  from  $Q$ ;
  Split the point set in  $N$  into two clusters by K-means;
  Compute LSL, SR and CD for the two child clusters;
  Assign the values of LSL, SR and CD to  $N$ ;
  Use classifiers to decide if over-segmentation occurs;
  if over-segmentation not occurs then
    Add two new nodes containing the new clusters into  $Q$ ;
    Add the two new nodes as child nodes of  $N$  in  $T$ ;
  end if
end while
The number of clusters (motions) is given by the number of leaf nodes in  $T$ .

```

---

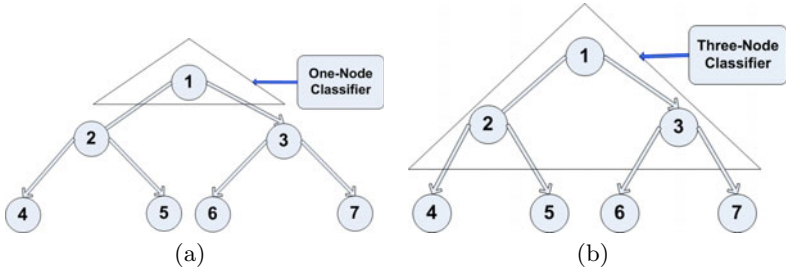
proposed in [5]. As its name suggests, the proposed method consists of two passes. In the first pass, we follow [5] in the way that a sufficient number of trajectory sets are randomly drawn to generate a hypothesis set, and the affinity matrix is computed. In the second pass, we fully exploit the information retrieved from the first pass by decomposition of the affinity matrix to obtain the nearest  $k$  neighbor of each trajectory in the motion space. For each trajectory, we obtain a refined hypothesis of the subspace by decomposition of the trajectories in the  $k$  neighbors instead of those selected randomly in the whole trajectory space. The number of hypotheses is independent of the size of the sampling, and we can obtain a satisfactory motion affinity matrix within two passes.

## 4.2 Classifiers for Over-Segmentation Detection

Although three measures for over-segmentation detection have been introduced in Section 3, it is still difficult to find a simple function relating them to make a decision on the occurrence of over-segmentation. Furthermore, over-segmentation is more or less a subjective perception, with different people giving different opinions. Hence, a machine learning approach is adopted in this paper to learn the decision function.

Each cluster node in the binary tree is associated with three features, namely *loss of spatial locality* (LSL), *split ratio* (SR) and *cluster distance* (CD), computed from its child nodes. A single-node structure and a triple-node structure are designed for classifying the split of a root node and a non-root node respectively. The single-node structure contains only one single node (see Fig. 3(a)), and is





**Fig. 3.** A single-node structure for the root node and a triple-node structure for the non-root node

used for determining whether to split the root node or not based its associated features (i.e., LSL, SR and CD). The triple-node structure contains three nodes, including the node under consideration, its parent node, and its sibling node (see Fig. 3(b)), and is used for determining whether to split a non-root node or not based on the features of all three nodes (i.e., nine values in total). A classifier is trained for each type of structures respectively.

In the training stage, Kernel Support Vector Machines (SVM) with radial basis function [15] are trained under supervision. Observations are the features of the structures associated with each node, and labels are the decisions of whether to split or not. Observation collection includes two stages: dataset selection from the database as a training set and feature extraction. For dataset selection, we exploit two methods of cross-validation, namely K-fold cross validation and Hold-out cross validation, to evaluate the performance as the volume of the training set decreases. K-fold cross validation partitions the database into  $k$  folds, and uses  $k - 1$  portions as the training set and the rest for testing. For Hold-out cross validation, a portion of data will be hold out for testing and the rest will be used as training data. Both methods are applied because we want to find out the least portion of data needed to train the classifiers while keeping the performance. For each validation method, we train several SVMs and select the classifier giving the best performance. Feature extraction is carried out by the hierarchical clustering method introduced in the previous subsection, but without over-segmentation detection. A decision is labelled when a new structure appears.

## 5 Experiments

We leveraged on the benchmark of Hopkins155 [16] for experiments. Our method was applied to various real dynamic scenes with two to five motions, with both rigid and articulated motions. There are 119 two-motion examples, 35 three-motion examples and 1 five-motion example. To demonstrate the effectiveness of the proposed measures for over-segmentation detection, we compared our method with the spectral clustering method presented in [5]. To ensure a fair comparison, both methods used the same motion affinity metric as described in [5]. We have not compared our method with some other methods such as

[9] and [10] because the performance of such methods heavily depends on the implementations.

To cluster the motions, we first computed the motion affinity as described in [5], centered the kernel matrix  $K$  and obtained its projection points  $P_4$  in the 4-D subspace by eigen-value decomposition  $K = RDR^T$  and  $P_4 = D_4^{\frac{1}{2}}R(:, 1 : 4)^T$ , where  $D_4$  is the  $4 \times 4$  diagonal block of  $D$  associated with the largest 4 eigenvalues, and  $R(:, 1 : 4)$  consists of the 4 columns of  $R$  associated with the largest 4 eigenvalues. K-Means method was then used to cluster  $P_4$  into two groups. This was done once in the testing stage but repeated eight times for the training stage to find the correct clustering. We computed LSL, SR and CD from the two groups for over-segmentation detection. The neighbor number for LSL was chosen to be 1 since we found any number within the range  $[1, \dots, 5]$  would give similar performance. For K-fold method, we trained  $k$  SVMs and selected the one with the best performance as our classifier. With Hold-out cross validation method, for each fixed portion, we repeated  $1/portion$  times, each time trained one SVM and selected the one with best performance.

**Table 1.** Error rates for K-fold cross-validation

<i>FoldNumber</i>	2	3	4	5	6	7	8
<i>Overall</i>	18.7%	16.1%	14.2%	17.4%	14.8%	15.5%	15.5%
<i>TwoMotion</i>	0%	0.8%	0%	0%	0%	0%	0%
<i>ThreeMotion</i>	77.1%	68.6%	60.0%	74.3%	62.3%	65.7%	65.7%
<i>FiveMotion</i>	100%	100%	100%	100%	100%	100%	100%

**Table 2.** Error rates for Hold-out cross-validation

<i>PortionForTest</i>	95%	90%	85%	80%	75%	70%	65%	60%	55%
<i>Overall</i>	21.3%	21.9%	21.3%	21.3%	21.3%	19.4%	18.1%	18.1%	15.5%
<i>TwoMotion</i>	0%	0%	0%	0.8%	0%	0%	0%	0%	0%
<i>ThreeMotion</i>	91.4%	94.3%	91.4%	88.6%	91.4%	82.9%	77.1%	77.1%	65.7%
<i>FiveMotion</i>	100%	100%	100%	100%	100%	100%	100%	100%	100%

The error rate with K-fold is listed in Table 1. Overall error rate is defined by the ratio of the number of erroneously estimated examples to the total number in the database. Error rate of each motion number is also listed for analysis. We summarize the results of Hold-out in Table 2. From the tables, we can see our method did well in two-motion case but was not satisfactory for the three-motion and five-motion cases for both cross-validation methods. We also notice that the error rate of two-motion case in Hold-out was not very sensitive to the number of training samples. For example, the error rate associated with the case using 5% of data for training is the same with those using more training data. However, the error rate of three-motion case decreases as training data increase from 5% to 45%, which may indicate that there may be an insufficient number of three-motion and five-motion samples in the training set.

Table 3 below copies the results shown in Table 2 of [5] for ease of reference.

**Table 3.** Error Rates for [5]

<i>Database</i>	Hopkins 155
<i>Overall</i>	36.63%
<i>TwoMotions</i>	32.63%
<i>ThreeMotions</i>	50.34%

With benefit from the features for over-segmentation detection, our method outperforms [5] in most cases. For two-motion case, our method can virtually achieve an error rate of 0%. For three-motion case, the result of [5] is a little better than ours. One possible reason for the poor performance of our method is that the number of SVM parameter for three-motion and five-motions case is larger than that of the two-motion case, while the number of samples for the former in the database is much less than that of the later. The database hence provides an insufficient training set for the more-motion case.

## 6 Conclusion and Future Work

In this paper, we recast the problem of multi-body segmentation and motion number estimation into an over-segmentation detection problem. The main contributions of our work are (1) the introduction of three measures, namely loss of spatial locality, split ratio and cluster distance, for over-segmentation detection; (2) the introduction of the Dual Pass Order Residual method for computing motion affinity; (3) the introduction of a hierarchical clustering method for multi-body segmentation with a supervised learning approach for over-segmentation detection. We leverage on Hopkins155 database to test our method and, with the same motion affinity metric, our method outperforms another state-of-the-art method. To the best of our knowledge, this paper is the first to tackle the problem of multi-body segmentation and motion number estimation from the perspective of over-segmentation detection. In the future, more exploration should be focused on the structures and features of over-segmentation that determine complex decision trees, such as a classifier structure for more than two motions.

## References

1. Costeira, J.P., Kanade, T.: A multibody factorization method for independently moving objects. *International Journal of Computer Vision* 29, 159–179 (1998)
2. Gear, C.: Multibody grouping from motion images. *International Journal of Computer Vision* 29, 133–150 (1998)
3. Vidal, R., Ma, Y., Soatto, S., Sastry, S.: Two-view multibody structure from motion. *International Journal of Computer Vision* 68, 7–25 (2006)
4. Yan, J., Pollefeys, M.: A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In: *European Conference on Computer Vision*, pp. 94–106 (2006)

5. Chin, T.J., Wang, H., Suter, D.: The ordered residual kernel for robust motion subspace clustering. In: *Neural Information Processing Systems* (2009)
6. Luxburg, U.V.: A tutorial on spectral clustering. Technical report, Max Planck Institute for Biological Cybernetics (2007)
7. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: *Computer Vision and Pattern Recognition*, pp. 395–416 (1997)
8. Ma, Y., Derksen, H., Hong, W., Wright, J.: Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1546–1562 (2007)
9. Schindler, K., Suter, D.: Two-view multibody structure-and-motion with outliers through model selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 983–995 (2006)
10. Schindler, K., Suter, D., Wang, H.: A model-selection framework for multibody structure-and-motion of image sequences. *International Journal of Computer Vision* 79, 159–177 (2008)
11. Bab-Hadiashar, A., Suter, D.: In: *Data Segmentation and Model Selection for Computer Vision* Ch.6, pp. 143–178. Springer, Heidelberg (2000)
12. Jian, Y.D., Chen, C.S.: Two-view motion segmentation with model selection and outlier removal by ransac-enhanced dirichlet process mixture models. *International Journal of Computer Vision* 88, 489–501 (2010)
13. Ozden, K.E., Schindler, K., Gool, L.V.: Multibody structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 1134–1141 (2010)
14. Schindler, K., Suter, D.: Two-view multibody structure-and-motion with outliers. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 676–683. IEEE Computer Society, Los Alamitos (2005)
15. Bishop, C.M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus (2006)
16. Tron, R., Vidal, R.: A benchmark for the comparison of 3-d motion segmentation algorithms. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2007)

# Improvement of a Traffic Sign Detector by Retrospective Gathering of Training Samples from In-Vehicle Camera Image Sequences

Daisuke Deguchi, Keisuke Doman, Ichiro Ide, and Hiroshi Murase

Graduate School of Information Science, Nagoya University  
Furo-cho Chikusa-ku Nagoya, Aichi 464-8601, Japan

**Abstract.** This paper proposes a method for constructing an accurate traffic sign detector by retrospectively obtaining training samples from in-vehicle camera image sequences. To detect distant traffic signs from in-vehicle camera images, training samples of distant traffic signs are needed. However, since their sizes are too small, it is difficult to obtain them either automatically or manually. When driving a vehicle in a real environment, the distance between a traffic sign and the vehicle shortens gradually, and proportionally, the size of the traffic sign becomes larger. A large traffic sign is comparatively easy to detect automatically. Therefore, the proposed method automatically detects a large traffic sign, and then small traffic signs (distant traffic signs) are obtained by retrospectively tracking it back in the image sequence. By also using the retrospectively obtained traffic sign images as training samples, the proposed method constructs an accurate traffic sign detector automatically. From experiments using in-vehicle camera images, we confirmed that the proposed method could construct an accurate traffic sign detector.

## 1 Introduction

In recent years, ITS (Intelligent Transport Systems) technologies have become widely available in our driving environment. In particular, understanding of the road environment in ITS is one of the most important technologies for a safe driving assistance system. Since traffic sign detection and recognition are key components for understanding the road environment, several methods have been proposed [1,2,3,4]. Bahlmann et al. proposed a method for detecting traffic signs from in-vehicle camera images [3]. They employed a cascaded AdaBoost classifier [5] for rapid detection, and color Haar-like feature is used for improving the accuracy of the detection. Although their method is accurate and fast enough, it requires a tremendous number of traffic sign images for training the AdaBoost classifier. Doman et al. solved this problem by generating training samples according to image degradation models [4]. Although this method can generate numerous training samples, it is still difficult to generate various appearances actually observed in the real environment as shown in Fig. 1. For constructing a traffic sign detector easily and accurately, it is necessary to obtain a large

number of training samples from real environment without manual intervention. Also, if a traffic sign detector is constructed before applying it to an unknown environment, it is required to reconstruct the detector by using new training samples obtained in the environment. Wöhler tried to solve these problems by constructing a pedestrian detector by obtaining training samples automatically from in-vehicle camera images [6]. In this method, pedestrians were detected by using a previously constructed detector, and training samples were obtained by tracking them forward in the time space. However, to exclude false positives from training samples, this method requires that an initial detector should be relatively accurate. Therefore, it still requires a large number of training samples for constructing the initial detector. To solve this problem, this paper introduces knowledge about appearance changes of traffic signs when driving a vehicle.

Training samples of distant traffic signs are required for constructing an accurate traffic sign detector that can detect distant traffic signs from in-vehicle camera images. However, since their sizes are too small in in-vehicle camera images, it is difficult to obtain them either automatically or manually. When driving a vehicle in a real environment, the distance between a traffic sign and the vehicle shortens gradually, and proportionally, the size of the traffic sign becomes larger. Therefore, if we can know the position of the large traffic sign, small traffic signs (distant traffic signs) can be obtained by tracking it back in the image sequence. Based on this idea, the proposed method greatly reduces the number of initial training samples, and then constructs an accurate traffic sign detector by gathering training samples retrospectively from in-vehicle camera image sequences. To use the traffic sign detector in a real environment, not only precision but also recall of the detector should be high. Therefore, the aim of the work presented in this paper is to construct a traffic sign detector having a high F-measure.

Section 2 describes the details of the proposed method. Then, experiments using in-vehicle camera images are shown in section 3. We discuss the results in section 4. Finally, we will conclude this paper in section 5.

## 2 Method

This paper proposes a method for constructing an accurate traffic sign detector by gathering training samples retrospectively from in-vehicle camera images. To construct an accurate traffic sign detector, traffic sign images for training should be gathered in various sizes from small (low resolution) through to large (high resolution). However, as shown in Fig. 2(a), it is difficult and time consuming to obtain numerous small traffic sign images (distant traffic signs) segmented accurately, since their sizes are small. On the other hand, large traffic sign images (close traffic signs) shown in Fig. 2(c) can be segmented accurately, and it is comparatively easy to recognize them automatically. Also, if the position of a large traffic sign is obtained, it is easy to track small traffic signs from it. Therefore, based on these ideas, the proposed method employs two strategies for gathering various traffic sign images: (1) find large traffic signs (high resolution),



Fig. 1. Examples of various appearances of traffic signs



Fig. 2. Appearances observed at distant, middle and close traffic signs from a vehicle

and (2) retrospective tracking from a large traffic sign to a small one. Then, the proposed method constructs a traffic sign detector by using samples obtained automatically. Figure 3 shows very common and important traffic signs when driving a vehicle in Japan. Therefore, we consider these traffic signs as our targets in this paper.

The proposed method consists of two parts: (1) retrospective gathering of traffic sign images from in-vehicle camera images, and (2) construction of a traffic sign detector by using them. The following sections describe details of these two parts.

### 2.1 Retrospective Gathering of Traffic Sign Images

Figure 4 shows a flowchart of our proposed method. The proposed method employs a nested cascade of a Real AdaBoost classifier for the detection of large traffic signs [11][12]. Then, retrospective tracking is used for gathering small traffic sign images automatically. The following sections describe details of these steps.



Fig. 3. Target traffic signs

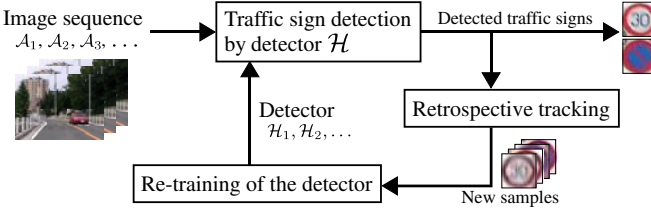


Fig. 4. Flowchart of the proposed method

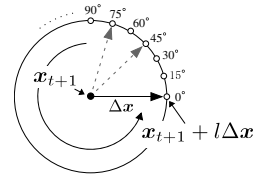


Fig. 5. Edge detection of a traffic sign

**Detection of a large traffic sign.** First, the proposed method searches traffic sign candidates from in-vehicle camera images by using a traffic sign detector  $\mathcal{H}$  based on a nested cascade of a Real AdaBoost classifier. The process of traffic sign detection is performed in the same manner as in [5]. Since this search process is performed by placing a detection window over the entire region of an image, in general, many candidates are obtained around a traffic sign. By using this characteristic, the proposed method merges the detected candidates according to the distance between them. Mean shift clustering [7] is used for this merge process. This step reduces the number of candidates by merging candidates detecting a same traffic sign. Then, false positives are removed by evaluating the number of the merged candidates. Finally, the positions of the detected candidates are used as the initial position of retrospective tracking described in the next section.

**Retrospective tracking of traffic signs.** This step extracts small (low resolution) traffic signs by tracking them back in the image sequence from a detection result of the previous step. This is formulated as a process that iteratively computes the center and the size of the  $(t - 1)$ -th traffic sign by using those of the  $t$ -th one.

First, the red component of an input image (each traffic sign has a red edge) is normalized by its intensity, and then an image  $\mathbf{F}$  is obtained by applying a Gaussian filter. The edge of a traffic sign is computed by evaluating

$$\nabla \mathbf{F}_t(\mathbf{x}_{t+1} + l\Delta \mathbf{x}) \cdot \Delta \mathbf{x} < 0, \quad (1)$$

where  $\nabla \mathbf{F}_t(\mathbf{x})$  is a gradient of an intensity at  $\mathbf{x}$ , and “ $\cdot$ ” is an inner product of vectors. In this process, the proposed method searches the edge pixel along the



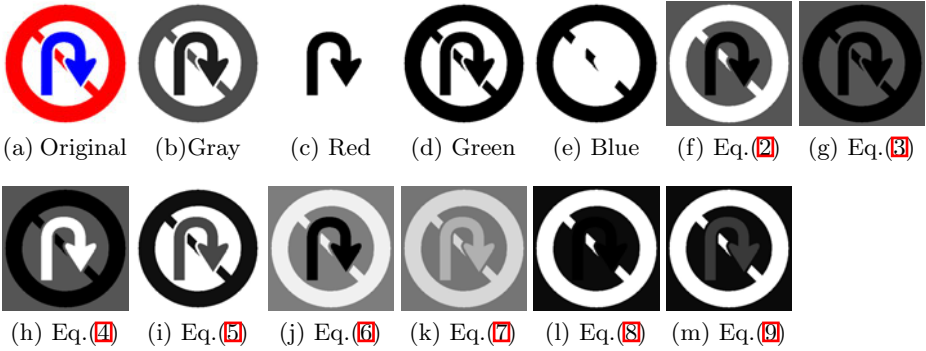


Fig. 6. Examples of color feature images for computing LRP features

direction  $\Delta \mathbf{x}$  from the center of previously detected traffic signs by increasing  $l$ , as shown in Fig. 5. Finally, the center and the size of a traffic sign are calculated by fitting a circle to the edge [8]. In this fitting process, we use RANSAC approach to avoid the effect of inappropriate edge detection results. The proposed method tracks traffic signs back in the image sequence by repeating this process by  $t \leftarrow t - 1$ .

### 2.2 Construction of a Traffic Sign Detector

Our traffic sign detector  $\mathcal{H}$  is constructed based on a nested cascade of a Real AdaBoost classifier [11,12]. The weak classifier for the Real AdaBoost classifier uses LRP (Local Rank Pattern) features [10], and these features are calculated from twelve types of color values. Color values used in this step consist of gray scale value ( $f_1$ ), RGB values ( $f_2 \sim f_4$ ), normalized RGB values ( $f_5 \sim f_7$ ), and opponent color values ( $f_8 \sim f_{12}$ ) [9]. Here,  $f_5 \sim f_{12}$  are calculated as

$$f_5(\mathbf{x}) = \frac{r(\mathbf{x})}{r(\mathbf{x}) + g(\mathbf{x}) + b(\mathbf{x})}, \tag{2}$$

$$f_6(\mathbf{x}) = \frac{g(\mathbf{x})}{r(\mathbf{x}) + g(\mathbf{x}) + b(\mathbf{x})}, \tag{3}$$

$$f_7(\mathbf{x}) = \frac{b(\mathbf{x})}{r(\mathbf{x}) + g(\mathbf{x}) + b(\mathbf{x})}, \tag{4}$$

$$f_8(\mathbf{x}) = 0.06 r(\mathbf{x}) + 0.63 g(\mathbf{x}) + 0.27 b(\mathbf{x}), \tag{5}$$

$$f_9(\mathbf{x}) = 0.30 r(\mathbf{x}) + 0.04 g(\mathbf{x}) - 0.35 b(\mathbf{x}), \tag{6}$$

$$f_{10}(\mathbf{x}) = 0.34 r(\mathbf{x}) - 0.60 g(\mathbf{x}) + 0.17 b(\mathbf{x}), \tag{7}$$

$$f_{11}(\mathbf{x}) = \frac{f_9(\mathbf{x})}{f_8(\mathbf{x})}, \tag{8}$$

$$f_{12}(\mathbf{x}) = \frac{f_{10}(\mathbf{x})}{f_8(\mathbf{x})}, \tag{9}$$

**Table 1.** Detection rate of the constructed detectors  $\mathcal{H}_0, \mathcal{H}_1, \dots, \mathcal{H}_4$ 

Detector	Precision	Recall	F-measure
$\mathcal{H}_0$	0.982	0.636	0.772
$\mathcal{H}_1$	0.978	0.878	0.925
$\mathcal{H}_2$	0.968	0.940	0.954
$\mathcal{H}_3$	0.956	0.955	0.955
$\mathcal{H}_4$	0.945	0.960	0.953

where  $r(\mathbf{x})$ ,  $g(\mathbf{x})$  and  $b(\mathbf{x})$  represent red, green and blue values at a pixel  $\mathbf{x}$ , respectively. Figure 6 shows examples of color values calculated by these equations.

In the training of the nested cascade of a Real AdaBoost classifier, traffic sign images gathered in the previous section are used as positive samples for training the classifier. Then, the trained classifier is used for gathering new traffic sign images in the next loop as shown in Fig. 4. By iterating these processes, the proposed method gathers training samples automatically, and constructs an accurate traffic sign detector iteratively.

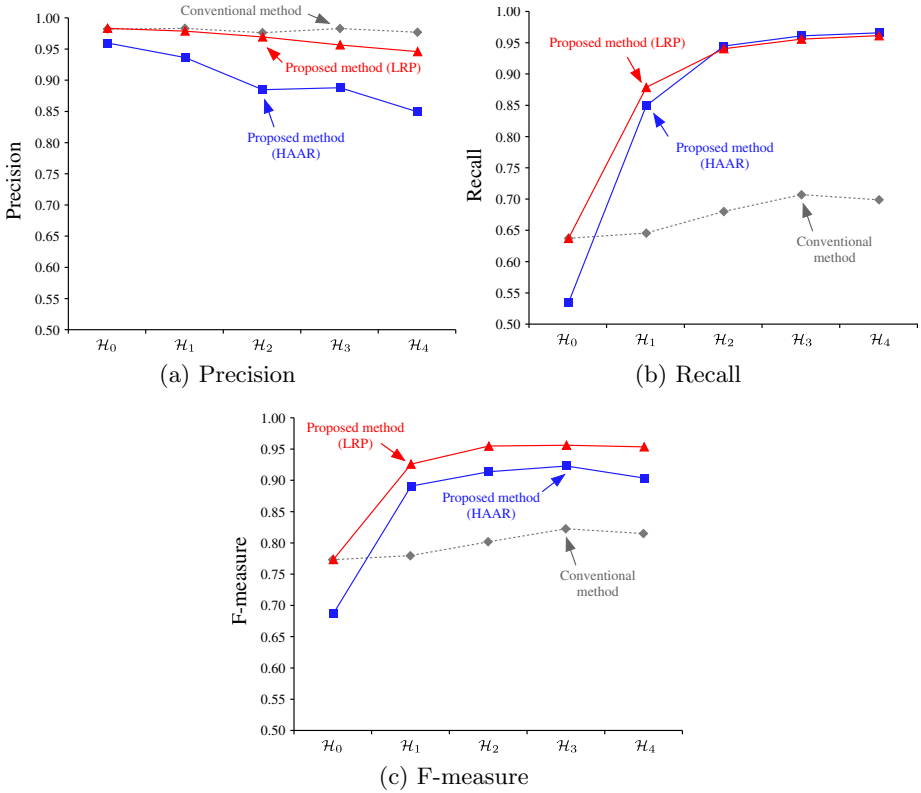
### 3 Experiment

Experiments using in-vehicle camera images were conducted for evaluating the effectiveness of the proposed method. We used SANYO Xacti DMX-HD2 as an in-vehicle camera, and the size of the captured images was  $640 \times 480$  pixels (30 fps). We prepared five image sequences ( $\mathcal{A}_0, \mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ , and  $\mathcal{A}_4$ ) containing 3,907 images in total for training. We also prepared 2,967 images for evaluation. Here, each image contains at least one traffic sign with a size between  $15 \times 15$  pixels and  $45 \times 45$  pixels. Negative samples were randomly selected from 180 in-vehicle camera images containing no traffic sign, and 2,500 negative samples were used for training in each stage of the cascade.

In this experiment, we constructed five traffic sign detectors by the following steps: At first, we manually selected thirteen large traffic signs from dataset  $\mathcal{A}_0$ , and 500 traffic sign images were generated by changing their clipping positions. Then, we constructed an initial detector  $\mathcal{H}_0$  by using these 500 images. Second, by applying the processes described in section 2.1, the proposed method gathers traffic sign images from dataset  $\mathcal{A}_1$  by using detector  $\mathcal{H}_0$ . Then, traffic sign images used in  $\mathcal{H}_0$  and traffic sign images gathered in the above step are used for constructing a second detector  $\mathcal{H}_1$ . Similarly,  $\mathcal{H}_2, \mathcal{H}_3$ , and  $\mathcal{H}_4$  are constructed by applying the same steps.

To evaluate the effectiveness of the retrospective gathering of training samples proposed in this paper, we compared the following three methods:

**Proposed method (LRP).** This method uses LRP features in section 2.2. Traffic sign detectors  $\mathcal{H}_0, \mathcal{H}_1, \dots, \mathcal{H}_4$  are constructed using training samples obtained by the proposed method.



**Fig. 7.** Results of detectors  $\mathcal{H}_0$  ‘  $\mathcal{H}_4$  constructed by the proposed method and the conventional method in precision, recall and F-measure

**Proposed method (HAAR).** This method uses Haar-like features instead of LRP features in section 2.2. Here, Haar-like features [5] are features based on intensity difference, and widely used for object detection methods, especially face detection. Other processes are same as the Proposed Method (LRP).

**Conventional method.** This method uses training samples generated from thirteen large traffic images by changing their clipping positions (X and Y coordinates of the top-left of the clipped image). These training images are same as ones used for training  $\mathcal{H}_0$  for the proposed method. In this method,  $\mathcal{H}_0, \mathcal{H}_1, \dots, \mathcal{H}_4$  are constructed by changing the number of images generated from the large traffic sign images.

Table 1 shows the results of the constructed detectors  $\mathcal{H}_0, \mathcal{H}_1, \dots, \mathcal{H}_4$  of the proposed method (LRP) in precision and recall rates with corresponding F-measures. Figure 7 shows the results of detectors  $\mathcal{H}_0$  ‘  $\mathcal{H}_4$  constructed by the proposed method (LRP), the proposed method (HAAR), and the conventional method in precision, recall and F-measure. Examples of the detection results by the proposed method (LRP) are shown in Fig. 8.



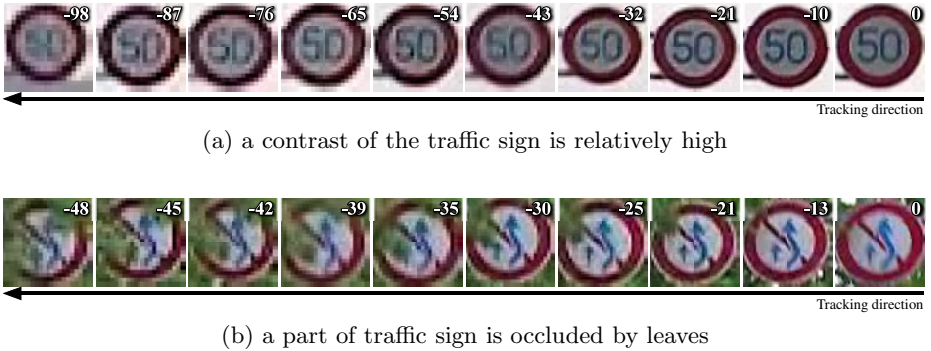
**Fig. 8.** Examples of detection results by the proposed method (LRP). (a) there is an object similar to the target traffic signs, which is located above the traffic signs but correctly not detected, and (b) although a traffic sign is occluded by a pole, the proposed method succeeded to detect it.

When using Intel Xeon W5590  $3.33 \text{ GHz} \times 2$ , the finally constructed detector required 0.122 sec. (8.2 fps) in average for detecting traffic signs from an image. This means that the proposed method can detect traffic signs every 2 meters when the vehicle moves at 60 km/h.

## 4 Discussions

As mentioned earlier, both precision and recall of a constructed traffic sign detector should be high. That is, it is required that the constructed detector should have high F-measure reflecting both precision and recall. From this point of view, as can be seen from Table 1, the proposed method could construct an accurate traffic sign detector (0.955 in F-measure) automatically by obtaining various traffic sign images from only thirteen large traffic sign images inputted manually. The accuracy of the constructed detector gradually improved by applying the proposed method iteratively. Also, as shown in Fig. 7, this can be observed from the comparison of the proposed method and the conventional method. Although the precision of the proposed method slightly degrades compared to that of the conventional method, the proposed method could obtain much higher recall rate. Therefore, F-measure was greatly improved by the proposed method. From these results, since only a small number of training samples is required as an input for the proposed method, this can greatly reduce the cost for constructing a detector. Therefore, the proposed method will be quite useful for improving the accuracy of a traffic sign detector without manual intervention.

To evaluate the effectiveness of the LRP features, we compared LRP features and Haar-like features in precision, recall, and F-measure, shown in Fig. 7. To construct an accurate traffic sign detector, training samples obtained by the method must be labeled correctly. In the case of the method using Haar-like



**Fig. 9.** Results of retrospective tracking of a traffic sign. Relative frame number is shown at the top right of each image.

features, some false positives are included in the training samples obtained automatically by the proposed method. Therefore, the precision of the constructed detector gradually decreased. On the other hand, in the case of using LRP features, since few false positives are included in the obtained training samples, the precision of the proposed method (LRP) is much higher than the proposed method (HAAR). However, the proposed method (LRP) still gathered a small number of false positives for training samples. We intend to improve the performance of automatic gathering of training samples in our future work.

Figure 9 shows examples of retrospective tracking of traffic signs proposed in this paper. As shown in Fig. 9(a), it can be confirmed that the proposed method could obtain traffic sign images in various resolutions from low to high. Although a part of a traffic sign in Fig. 9(b) is occluded by leaves, some edges of the traffic sign can still be observed. Since these edges were extracted, the proposed method was able to track it correctly. However, the method failed to track traffic signs when their resolution was too poor. We intend to deal with this problem in our future work.

## 5 Conclusions

This paper proposed a method for constructing an accurate traffic sign detector by automatic gathering of various traffic sign images based on retrospective tracking. First, the proposed method detects large (high resolution) traffic signs from in-vehicle camera images. Then, retrospective tracking is applied for obtaining small traffic sign images. By applying these steps, the proposed method allows us to automatically gather real traffic sign images in various sizes from a small one to a large one. Finally, a traffic sign detector is constructed by using the gathered traffic sign images. We evaluated the accuracy and the effectiveness of the proposed method by applying it to actual in-vehicle camera images. Experimental results showed that the proposed method could improve the accuracy

of the traffic sign detector satisfactorily. Future works include: (i) improvement of the tracking of small traffic signs, (ii) evaluation by applying the method to many more cases.

## Acknowledgement

Parts of this research were supported by a Grant-in-Aid for Young Scientists from MEXT, a Grant-In-Aid for Scientific Research from MEXT, and JST CREST. MIST library (<http://mist.murase.m.is.nagoya-u.ac.jp/>) was used for developing the proposed method.

## References

1. Maldonado-Bascón, S., Lafuente-Arroyo, S., Gil-Jiménez, P., Gómez-Moreno, H., López-Ferreras, F.: Road-sign detection and recognition based on support vector machines. *IEEE Transactions on Intelligent Transportation Systems* 8(2), 264–278 (2007)
2. Loy, G., Barnes, N.: Fast shape-based road sign detection for a driver assistance system. In: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1, pp. 70–75 (2004)
3. Bahlmann, C., Zhu, Y., Ramesh, V., Pellkofer, M., Koehler, T.: A system for traffic sign detection, tracking, and recognition using color, shape, and motion information. In: *Proceedings of IEEE Intelligent Vehicles Symposium*, pp. 255–260 (2005)
4. Doman, K., Deguchi, D., Takahashi, T., Mekada, Y., Ide, I., Murase, H.: Construction of cascaded traffic sign detector using generative learning. In: *Proceedings of International Conference on Innovative Computing Information and Control, ICICIC-2009-1362* (2009)
5. Viola, P., Jones, M.: Robust real-time face detection. *International Journal of Computer Vision* 57(2), 137–154 (2004)
6. Wöhler, C.: Autonomous in situ training of classification modules in real-time vision systems and its application to pedestrian recognition. *Pattern Recognition Letters* 23(11), 1263–1270 (2002)
7. Cheng, Y.: Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(8), 790–799 (2005)
8. Coope, I.D.: Circle fitting by linear and nonlinear least squares. *Journal of Optimization Theory and Applications* 76(2), 381–388 (1993)
9. Burghouts, G.J., Geusebroek, J.M.: Performance evaluation of local colour invariants. *Computer Vision and Image Understanding* 113(1), 48–62 (2009)
10. Hradis, M., Herout, A., Zemcik, P.: Local rank patterns – novel features for rapid object detection. In: Bolc, L., Kulikowski, J.L., Wojciechowski, K. (eds.) *ICCVG 2008. LNCS*, vol. 5337, pp. 239–248. Springer, Heidelberg (2009)
11. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. *Machine Learning* 37(3), 297–336 (1999)
12. Huang, C., Ai, H., Wu, B., Lao, S.: Boosting nested cascade detector for multi-view face detection. In: *Proceedings of the International Conference on Pattern Recognition*, vol. 2, pp. 415–418 (2004)

# Statistical Modeling of Long-Range Drift in Visual Odometry

Ruyi Jiang<sup>1</sup>, Reinhard Klette<sup>2</sup>, and Shigang Wang<sup>1</sup>

<sup>1</sup> Shanghai Jiao Tong University, Shanghai, China

<sup>2</sup> University of Auckland, Auckland, New Zealand

**Abstract.** An intrinsic problem of visual odometry is its drift in long-range navigation. The drift is caused by error accumulation, as visual odometry is based on relative measurements. The paper reviews algorithms that adopt various methods to minimize this drift. However, as far as we know, no work has been done to statistically model and analyze the intrinsic properties of this drift. Moreover, the quantification of drift using offset ratio has its drawbacks. This paper models the drift as a combination of wide-band noise and a first-order Gauss-Markov process, and analyzes it using Allan variance. The model's parameters are identified by a statistical method. A novel drift quantification method using Monte Carlo simulation is also provided.

## 1 Introduction

Visual odometry uses camera(s) to incrementally calculate a robot's motion between frames and to position the robot in all six degrees of freedom in a 3D world. Compared with other positioning sensors (e.g., odometry, GPS, IMU and so forth), visual odometry has its own characteristics. Classical odometry, installed on a robot's wheel axis, is usually deceived by wheel slippage, especially in an outdoor environment. GPS is not always available for navigation, due to signals being missing or jammed. A typical example is the successful application of visual odometry for NASA's MER missions [4]. Also, compared to GPS and IMU, cameras in visual odometry are relatively cheap.

Visual odometry has been widely applied in many fields, such as driver assistance or autonomous driving [11], simultaneous localization and mapping (SLAM), helicopter navigation [10], and underwater navigation [6]. Many algorithms have been tested to implement visual odometry using monocular [13] or stereo [10,14], perspective or omnidirectional [13] cameras. A popular framework for visual odometry is based on feature matching and tracking [11,14]. While considering that a feature-based method is sensitive to systematic errors due to intrinsic and extrinsic camera parameters, appearance-based visual odometry uses the appearance of the world to extract motion information (e.g., [13]). Recently, a direct method was also tested for visual odometry with very accurate results [5].

Among all these algorithms, one intrinsic problem of visual odometry is its drift in long-range navigation. The drift is caused by error accumulation, as

visual odometry is based on relative measurements. Relative motion matrices between frames are concatenated to produce the final position. Small errors in these matrices accumulate during this process to a large amount, and the distance measurement drifts from its real trajectory after some long-time navigation. For feature-based algorithms, the sources for these small errors are mainly uncertainties of feature localization and triangulation.

Section 2 reviews algorithms that adopted various methods to minimize this drift. Section 3 models the drift as a combination of wide-band noise and first-order Gauss-Markov process, and analyzes it using Allan variance, named after David W. Allan [1]. The identification of model parameters using statistical methods is also introduced. Experiments and discussions are provided in Sections 4 and 5.

## 2 State of the Art

Before proceeding to drift-minimization algorithms, we discuss at first a method to quantify drift. Currently, the *offset ratio* (OR), ratio of the final drift value to the traveled distance, is the common choice to measure the drift when running a visual odometry algorithm over some distance, from tens or hundreds of meters to several kilometers. (Drawbacks of OR, and a better quantification method are discussed later in this paper.)

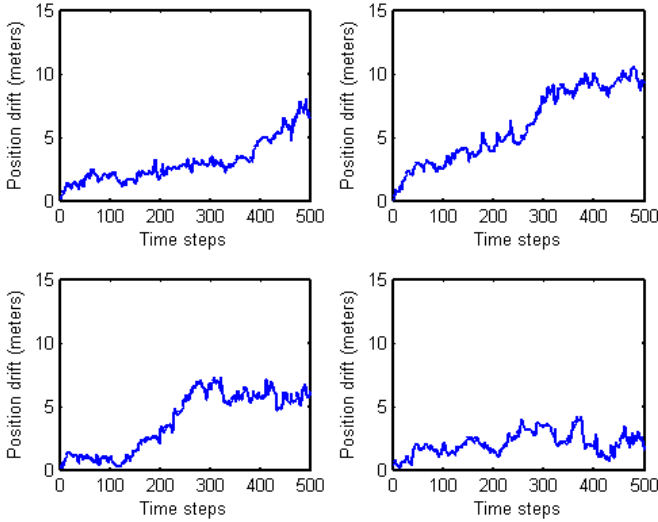
Note that the following review of algorithms is not for visual odometry, but for drift-minimizing methods adopted in these algorithms. It has been proved that integrating visual odometry with other positioning sensors, such as gyro or GPS, minimizes the drift. But this is not the problem to be discussed in this paper. As visual odometry alone has its practical and theoretical meaning, the paper analyzes drift without any help from other sensors.

A “fire wall” was inserted into sequences by Nistér [11] to act against error propagation. With fire walls, relative poses, estimated before the application of the fire wall, only affect the choice of the coordinate system for subsequent poses, and relative poses after the fire wall are estimated as if the system was starting again. It is supposed that fire walls suppress the propagation of gross errors and slow down the error buildup. From the provided experimental results, visual odometry had an accuracy [compared to the ground truth from a Differential Global Positioning System (DGPS)] of 1.07%, 4.86% and 1.63% for three outdoor runs with traveled distances of 185.88, 266.16, and 365.96 meters, respectively.

Bundle adjustment is another scheme that can be adopted to suppress error accumulation. It is widely used for solving the off-line structure and motion problem, and the SLAM problem. Full bundle adjustment is almost impossible for on-line long range navigation, as there is a huge number of poses and features to be optimized. A sliding-window sparse bundle adjustment was applied by Sünderhauf [14] for visual odometry. A subset of several images (the number is fixed, or adaptive to the motion vector) is continuously selected to perform bundle adjustment. Experiments on simulation show that sparse bundle adjustment slows down the drift.

Though many papers on visual odometry use various methods to suppress the drift, no work has been done to explicitly model the drift. Clark et al. [12]





**Fig. 1.** Position drifts after running a visual odometry algorithm (using ABSOLUTE algorithm, see the experiment section), with the same simulated motion vectors for the same time steps. Note that with the simulated motion vectors as the ground truth, camera’s pose can be estimated using visual odometry algorithm. It can be seen that drift values can be quite different. This results into incapability of the offset ratio.

analyze the contribution of position and orientation errors to the overall drift, and observed that the drift does not grow linearly in the distance traveled, but super-linearly. The growth was regarded as  $\mathcal{O}(dist^{\frac{3}{2}})$ , but no specific models and parameters were provided.

As a new and promising sensor, visual odometry needs a methodology for systematic and comparative analysis of its drift, in order to quantify the performance of various algorithms. For this purpose, OR has its drawbacks. First, drift does not increase linearly with the distance traveled, as stated by Clark et al. [12] and further proved in this paper. Thus, OR, when running algorithms on some distance, changes with the traveled different distances. Moreover, running the same algorithms on the same dataset repeatedly produces quite different ORs. The reason is that drift is a random process, and it does not always increase, but sometimes also decrease at some places, as errors in different motion vectors compensate to some extent. Thus, using end-point values (the final drift value, and the final traveled distance) is inappropriate to model the whole random process. An example is shown in Fig. 1. Considering these findings, a more accurate quantification method is introduced later.

### 3 Drift Model, Analysis and Quantification

In this paper, coordinate frame transformations are used to represent both poses and motions. Using general notations, a pose  $E$  is the transformation from the

world coordinate frame into that of the camera, and a motion  $M$  is a transformation of the coordinate frame of the camera at time  $t$  into that at time  $t + 1$ . Drift in orientation is limited to the range  $[-\pi, \pi]$ , and it contributes to the drift in position; this paper only considers positional drift. We consider positional drift in world coordinates along  $x$ -,  $y$ -, and  $z$ -axes separately, and use the  $x$ -axis as an example in the following modeling and analysis.

The concatenated camera pose at time  $t$  is  $E_t$ , and the estimated motion from time  $t$  to  $t + 1$  is  $M_t$ . Then  $E_{t+1} = E_t \cdot M_t$ . Note that the right multiplication with  $M_t$  is because the motion  $M_t$  is relative to the camera coordinate frame at  $t$ . The general structure of  $E$  and  $M$  is of the form  $[R_{3 \times 3} \ T_{3 \times 1}; 0 \ 0 \ 0 \ 1]_{4 \times 4}$ , where  $R_{3 \times 3}$  is a rotational matrix, and  $T_{3 \times 1} = [x, y, z]^T$  is a translational vector. Then the translational drift  $dx_{t+1}$  in  $x$ -direction is equal to

$$dx_{t+1} = x_{t+1} - \bar{x}_{t+1} \quad (1)$$

where  $x_{t+1}$  and  $\bar{x}_{t+1}$  are the estimated and the true position in  $x$ -direction at time  $t + 1$ , respectively.

**Drift Model.** As drift increases unboundedly for an assumed unlimited time, modeling and analysis of drift is confined within a limited time region. This matches the analysis of drift for inertial sensors. For a limited number of time steps, the drift  $\{dx_i, i = 1, 2, \dots, N\}$ , as established in Eq. (1), is modeled in discrete form as

$$dx_i = \omega_{n_i} + b_i \quad (2)$$

where  $\omega_n$  is zero-mean wide-band noise with variance  $\sigma_n^2$ , and  $b$  is a first-order Gauss-Markov process. This process is given by

$$b_i = \left(1 - \frac{1}{\tau}\right)b_{i-1} + \omega_a \quad (3)$$

where  $\tau$  is a constant called correlation time, and  $\omega_a$  is the driving noise modeled as zero-mean wide-band noise with variance  $\sigma_a^2$ . The variance of the Gauss-Markov process  $\sigma_b^2$  equals

$$\sigma_b^2 = \sigma_a^2 / \left(\frac{2}{\tau} + \frac{1}{\tau^2}\right) \quad (4)$$

The model of Eq. (2) has been widely used as a static stochastic error model in drift analysis for inertial sensors [15]. In this paper we show that this model can easily also be introduced into drift analysis for visual odometry; we reveal some important findings.

**Identification of Drift Model Parameters.** The parameters for the drift model of Eq. (2) are  $\sigma_n^2$ ,  $\tau$  and  $\sigma_b^2$ , and they can be estimated from experimental data using various identification techniques.

Parameter  $\sigma_n^2$  can be specified as the value of the Allan variance corresponding to 1 second averaging time [15]. Parameter  $\sigma_b^2$  is the variance of experiment data after removing the high frequency components.

The time constant  $\tau$  can be estimated from experimental autocorrelation data. This is because the first-order Markov process has an autocorrelation known as

$$R_b(T) = \sigma_b^2 e^{-T/\tau} \quad (5)$$

For the normalized autocorrelation  $\bar{R}_b(T)$  (normalization means  $\bar{R}_b(0) = 1$ ), we have that  $\tau = T$  when  $\bar{R}_b(T) = e^{-1}$ . In this way, the time constant  $\tau$  can be estimated as being the value of  $T$  corresponding to the normalized autocorrelation value 0.368 (i.e.,  $e^{-1}$ ). For an example of parameter identification, see the experimental section.

**Drift Quantification by  $\tau$  and  $\sigma_b^2$  Using Monte Carlo Simulation.** As  $\sigma_n$  is usually small and stable, the time constant  $\tau$  and the variance of the Gauss-Markov process  $\sigma_b^2$ , from the drift model as established by Eq. (2), are characterizing parameters of drift. Parameters  $\tau$  and  $\sigma_b^2$  of a drift sequence can be estimated as specified above. However, it can be expected that drift sequences, while running the same visual odometry algorithm, produce different  $\tau$  and  $\sigma_b^2$  values. In order to model drift properties of a specific visual odometry algorithm, Monte Carlo simulation is used.

Monte Carlo simulation is a technique that propagates uncertainties in input variables of a model into the output, depending on the given probability distributions. The procedure for parameter identification for a specific visual odometry algorithm is as follows:

1. Generate a random motion vector for every frame in visual odometry. Possible motion vectors are restricted by real situations.
2. Use the generated motion to simulate feature registration.
3. Estimate motion and concatenate for  $N$  frames.
4. Calculate  $\tau$  and  $\sigma_b^2$ .
5. Run Step 1 to 4 repeatedly (say,  $n$  times).

This procedure provides  $n$  possible  $\tau$  and  $\sigma_b^2$  values. An analysis of these values, using some statistical methods (e.g., histograms, accumulated statistics, confidence intervals, and so forth), finally illustrates the distribution of  $\tau$  and  $\sigma_b^2$  for a specific visual odometry algorithm.

**Drift Analysis Using Allan Variance.** The Allan variance is a method of analyzing a time sequence to pull out the intrinsic noise in the system as a function of the averaging time. It was originally developed to analysis the stability of clocks, but can also be adapted for any other type of outputs. Full details of the construction of Allan variance can be found here [15]. A typical application of Allan variance is the analysis of navigation errors for inertial sensors [15]. Here, we use Allan variance to analysis the drift and validate the drift model as established above.

A special capability of the Allan variance is that noise types can be identified by matching different curve slopes in an Allan variance chart. The different slopes on the plot indicate the unique time regions dominated by the sensor's

specific noise. As only wide-band noise and a first-order Gauss-Markov process are considered, the specific curve slopes for time regions dominated by these two kinds of noise are  $-1/2$  and  $1/2$  respectively. For the correspondence of other kinds of noise with some slope values, we refer to [15].

## 4 Experiments

Experiments are conducted to illustrate the validation of the established drift model, as well as the stability and robustness of the new drift quantification method. Moreover, some important facts of the drift in visual odometry are also revealed from the experimental results.

Simulated data is more preferable than real data for drift analysis, as simulation controls the source of error (e.g., feature location uncertainty for feature-based visual odometry) and removes outliers, which are common for real data. We simulate a sequence of stereo pairs. Only feature-based visual odometry algorithms are considered, as they are more general compared to appearance-based and direct visual odometry, and they are also easy to be controlled by the errors in the estimated motion matrix. No feature matching and tracking failures are considered in the simulation; thus, robust regression is not adopted to remove the outliers. We apply a similar scheme as used by Badino [2] for generating matched and tracked features.

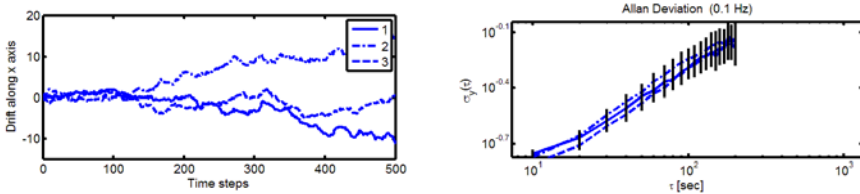
Two typical feature-based visual odometry algorithms were implemented, to illustrate the behavior of drift. The first one, named ABSOLUTE here, estimates the motion matrix between frames as an absolute orientation problem. Motion matrices are directly concatenated to estimate camera poses. Thus, drift is not suppressed, and is expected to be large. The second algorithm, named SBA here, uses a sliding window sparse bundle adjustment to optimize the motion matrices as estimated by the ABSOLUTE method. The whole implementation is similar to the one reported by Sünderhauf [14]. The number of features, tracked for both algorithms, is set to be 200, and the feature localization uncertainty is 0.5 pixel. For SBA, the number of stereo frames for bundle adjustment equals five. We assume that the stereo frames are taken one second apart in the following.

**Drift Analysis Using Allan Variance.** This experiment illustrates the use of Allan variance for visual odometry drift analysis. Some facts can be observed from the Allan variance.

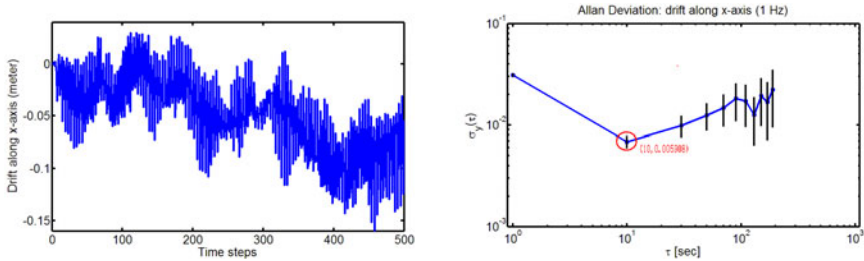
Both cameras move the same way. For every simulation instance of running the algorithms, different drift curves are obtained. Results for ABSOLUTE and SBA algorithms are shown in Figs. 2 and 3.

Though every running of ABSOLUTE algorithm gives a sequence of different drift, it produces a similar Allan variance chart. The main curve slope in Allan variance chart (Fig. 2, right) produced from ABSOLUTE algorithm is  $1/2$ , which means that the noise modeled as first-order Gauss-Markov process dominates the whole time region. In other words, the estimated camera trajectory will obviously drift from its real one since the first step.

For SBA there is an important point (see Fig. 3, circled point on the right), which separates the curve with slope  $-1/2$  from the curve with slope  $1/2$ . It means that the dominating noise in the first ten seconds (the  $x$ -coordinate of the point) is wide-band noise, while for the time after ten seconds, noise modeled as a first-order Gauss-Markov process will take over. In our situation (a one second image sampling interval, five frames bundle adjustment), it states that among ten frames, the drift behaves as wide-band white noise, while for more then ten frames, first-order Gauss-Markov noise dominates.

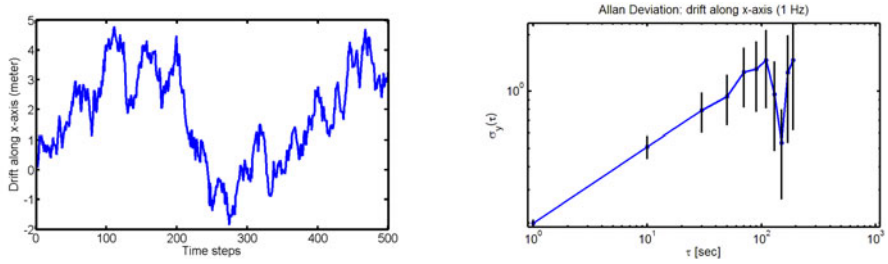


**Fig. 2.** Drift analysis using Allan variance. *Left:* Five drift instances from running the ABSOLUTE algorithm for 500 time steps. *Right:* Allan variance of these five drift instances. Note that the vertical bar is for the uncertainties.

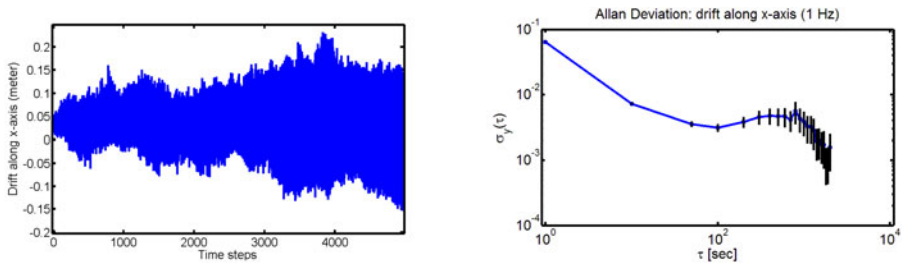


**Fig. 3.** Drift analysis using Allan variance. *Left:* A drift instance from running the SBA algorithm for 500 time steps. *Right:* Allan variance of the drift instance. The point circled by red is the point where the Allan variance curve changes slope from  $-1/2$  to  $1/2$ . The coordinate of this point is (10, 0.005908).

**Drift with Static Camera.** In this experiment we illustrate the behavior of visual odometry with cameras being static. The drift values (along the  $x$ -axis) from both the ABSOLUTE and SBA algorithm are shown in Figs. 4 and 5. It can be seen from Fig. 4 that ABSOLUTE has a large drift value, which is similar to the situation when the stereo pair is under motion. Also the drift error dominates the whole time region, which is illustrated by the Allan variance chart with a dominating slope of  $1/2$ . The SBA suppresses the drift to a small value even after a long time. The Allan variance in Fig. 5 reveals that the main error source in the drift value is random noise, with no obvious drift.



**Fig. 4.** Drift with static camera using ABSOLUTE visual odometry. *Left:* Drift value along  $x$ -axis for 500 time steps. *Right:* Allan variance of the drift value.



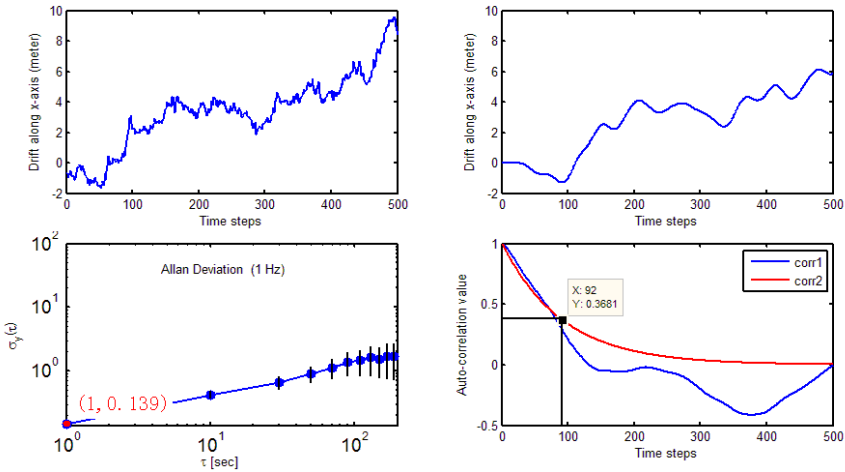
**Fig. 5.** Drift with static camera using SBA visual odometry. *Left:* Drift value along  $x$ -axis for 5,000 time steps. *Right:* Allan variance of the drift value. Note that the number of time steps here is larger than that in Fig. 4, to illustrate the behavior of drift for long time.

**Drift Model Parameter Identification.** Here we discuss a real example to illustrate the identification of drift model parameters  $\sigma_n^2$ ,  $\tau$  and  $\sigma_b^2$ . A drift sequence from the ABSOLUTE algorithm is used, as shown on the upper left of Fig. 6. The whole procedure is similar for drift value estimations for other visual odometry algorithms.

Parameter  $\sigma_n^2$  can be estimated from the Allan variance calculated from raw drift values. As shown in the lower left of Fig. 6, the Allan deviation  $\sigma_y$ , corresponding to one second, equals 0.139. Thus,  $\sigma_n^2 = 0.139^2 \approx 0.019$ .

After filtering the raw drift with a low-pass filter to remove the high frequency components caused by wide-band noise,  $\sigma_b^2$  can be estimated as the variance of the filtered drift. In this example, the filtered drift is shown on the upper right of Fig. 6, and  $\sigma_b^2 = 18.8$ .

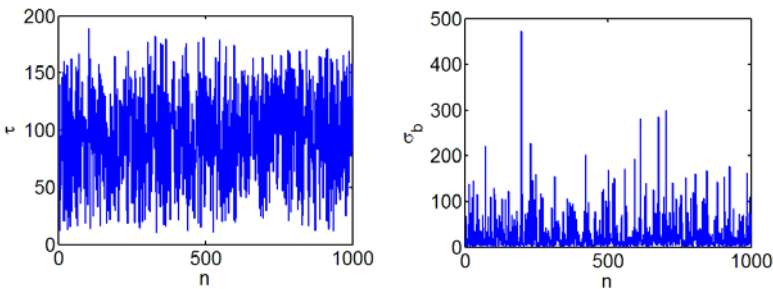
The time constant  $\tau$  is the time value corresponding to the normalized autocorrelation value 0.368. In this example, the autocorrelation curve using filtered drift is shown in blue; see lower left of Fig. 6. The model fitted curve by the expected autocorrelation function as Eq. (5) is shown in red; see also lower left of Fig. 6. Only the first 150 time steps are used to fit the model. As the time value for normalized autocorrelation 0.368 is 92 (the marked point; lower left of Fig. 6).



**Fig. 6.** An example of drift model parameter identification. *Upper left:* The raw drift value for 500 time steps. *Upper right:* Drift values after low-pass filter. *Lower left:* Allan variance. *Lower right:* Autocorrelation values. Note that the blue curve is the raw autocorrelation value using the drift values as shown in the upper right, while the red curve is the model-fitted autocorrelation for the first-order Gauss-Markov process.

**Drift Quantification Using Monte Carlo Simulation.** This experiment calculates the parameters  $\tau$  and  $\sigma_b^2$  for the ABSOLUTE algorithm. The number  $n$  of Monte Carlo iterations equals 1 000. The Monte Carlo simulation results are shown in Fig. 7.

The analysis of the Monte Carlo simulation results can be conducted using histogram, summary statistics, and so on. For the results in Fig. 7, the mean of  $\tau$  is about 94.8, while the mean of  $\sigma_b$  equals 27.6. It can be seen from this example that the parameters ( $\tau$  and  $\sigma_b$ ) to quantify the drift for a specific algorithm take much more factors into account than *offset ratio*. Moreover, *offset ratio* is only for running of a visual odometry algorithm once on a trajectory, which means it will always change for different trajectory with various length. While, the new quantification method using Monte Carlo simulation provides an overall evaluation of the algorithm.



**Fig. 7.** Monte Carlo simulation results for ABSOLUTE algorithm. (*Left:*  $\tau$ . *Right:*  $\sigma_b$ ). Note that the horizontal axis is the simulation number.

## 5 Discussion and Conclusions

Modeling and analyzing long-range drift in visual odometry is of practical and theoretical significance. This paper models drift as a random process, combining wide-band noise and a first-order Gauss-Markov process. Model parameters can be identified from experimental data. The Allan variance, offering the possibility to separate between these two sources of error, is adopted to analysis the drift. Experimental results reveal several important facts:

1. Modeling drift as a combination of wide-band noise and a first-order Gauss-Markov process is validated. This can be seen from the Allan variance of the drift.
2. Analyzing drift from various visual odometry algorithms using Allan variance is a feasible way to tell where drift, and not white noise, becomes dominating.
3. Quantifying drift from a specific algorithm by  $\tau$  and  $\sigma_b^2$  is a more accurate way than the usual offset ratio method.

## References

1. Allan, D.W.: Statistics of atomic frequency standards. *Proceedings of the IEEE* 54(2), 221–230 (1966)
2. Badino, H.: Binocular ego-motion estimation for automotive applications. PhD thesis. Frankfurt/Main University (2008)
3. Calvetti, D.: A stochastic roundoff error analysis for the convolution. *Mathematics of Computation* 59, 569–582 (1992)
4. Cheng, Y., Maimone, M.W., Matthies, L.: Visual odometry on the Mars exploration rovers. *IEEE Robotics Automation Magazine* 13(2), 54–62 (2006)
5. Comport, A.I., Malis, E., Rives, P.: Real-time quadrifocal visual odometry. *Int. J. Robotics Research* 29, 245–266 (2010)
6. Corke, P., Detwiler, C., Dunbabin, M., Hamilton, M., Rus, D., Vasilescu, L.: Experiments with underwater robot localization and tracking. In: *IEEE Int. Conf. Robotics Automation*, pp. 4556–4561 (2007)
7. Flenniken, W.S.: Modeling inertial measurement units and analyzing the effect of their errors in navigation applications. Master thesis, University of Auburn (2005)
8. IEEE standard specification format guide and test procedure for single-axis laser gyros. *IEEE Std 647<sup>TM</sup> – 2006* (2006)
9. Kelly, A.: Linearized error propagation in odometry. *The Int. J. of Robotics Research* 23, 179–218 (2004)
10. Kelly, J., Sukhatme, G.S.: An experimental study of aerial stereo visual odometry. In: *IFAC Symp. Intelligent Autonomous Vehicles* (2007)
11. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: *IEEE Conf. Computer Vision Pattern Recognition*, vol. 1, pp. 652–659 (2004)
12. Olson, C.F., Matthies, L.H., Schoppers, M., Maimone, M.W.: Stereo ego-motion improvements for robust rover navigation. In: *IEEE Int. Conf. Robotics Automation*, pp. 1099–1104 (2001)



13. Scaramuzza, D., Siegwart, R.: Appearance-guided monocular omnidirectional visual odometry for outdoor ground vehicles. *IEEE Trans. Robotics* 24, 1015–1026 (2008)
14. Sünderhauf, N., Protzel, P.: Towards using sparse bundle adjustment for robust stereo odometry in outdoor terrain. *Towards Autonomous Robotic Systems*, 206–213 (2006)
15. Wall, J.H., Bevly, D.M.: Characterization of inertial sensor measurements for navigation performance analysis. In: *Proc. of ION GNSS*, Long Beach, CA (2005)

# Object Discrimination and Tracking in the Surroundings of a Vehicle by a Combined Laser Scanner Stereo System

Mathias Haberjahn and Ralf Reulke

German Aerospace Center, Institute of Transportation Systems,  
Rutherfordstrasse 2, 12489 Berlin, Germany  
{mathias.haberjahn,ralf.reulke}@dlr.de

**Abstract.** The use of sensor data for observing the surrounding environment of a vehicle is becoming increasingly popular. Especially for detecting dangerous situations, which occur too fast for the human senses, sensor systems are needed. In the following paper such a sensor system consisting of a stereo camera and a multilayer laser scanner mounted in front of a test vehicle is introduced. Both sensors are used to detect and track obstacles and other traffic objects independent from each other for future data fusion. An overview of the complete process for the object discrimination including a novel approach for a sensor cross calibration and a new method for the object refinement and the object tracking is given. The effectiveness of the algorithms are tested with real road reference data, obtained through highly precise GPS data.

**Keywords:** stereo vision, laser scanner, segmentation, object discrimination, tracking, competitive data fusion.

## 1 Introduction

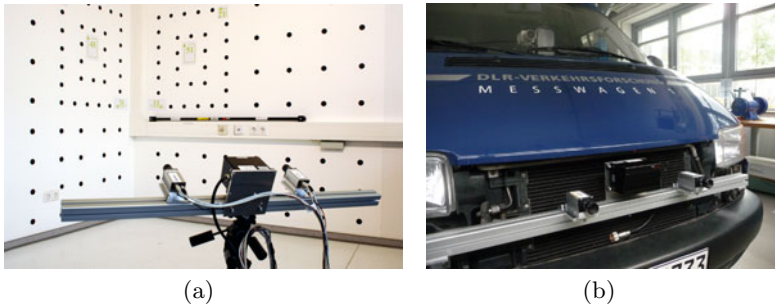
High accident rates still force the automobile industry and science to develop new methods to assist the human driver in dangerous situations. For this purpose, different sensors are tested or already in use. Sensor systems working with only one measurement method like mono or stereo camera systems ([9], [10]), radar or laser scanner [6] are commonly used.

Among the advantages every measurement method has a specific drawback resulting in false or missing alarms. To overcome these problems it is obvious to combine different sensor types and to fuse their data. So far, attempts were made to combine stereo vision with radar [4] or a laser scanner with a mono camera [3] or a stereo camera system ([12], [11]). All of these fusion systems define one main sensor and a second support sensor for weighting the object hypotheses. But in this constellation, not detected objects from the main sensor remain undiscovered or correct object hypotheses could be negated through false information from the second sensor. To avoid this loss of information, the stereo camera and the laser scanner used in this work are considered as equal sensors, able to work as competitive stand-alone systems.

The paper focuses on the process of object discrimination for both sensors. In section 2 the sensor setup, a new cross calibration procedure and a short summary of stereo vision are introduced. The segmentation of the raw data for the sensors and the object shaping is described more detailed in section 3. After the discussion of the object tracking including a smart approach for the object association in section 4 the algorithms are finally tested against ground truth data in section 5.

## 2 Sensor Setup

The setup consists of two monochrome *PicSight P141M Smart* cameras by *Leutron Vision* and a 4-layer laser scanner by *IBEO*. The devices are mounted on a profile for usage as a stand-alone system (Fig. 1(a)) or in front of the test vehicle (Fig. 1(b)).



**Fig. 1.** Sensor setup. (a) Stand alone setup in labor. (b) Setup mounted on test vehicle.

The cameras are able to acquire up to 20 images per second running with a maximum resolution of 1392 by 1040 pixel over a GigE connection. The laser scanner covers a region of 3.2 vertically and 110 horizontally with four beams and a maximum scan rate of 50Hz.

Both cameras are triggered simultaneously over a TTL pulse. Tests with an optical binary clock have shown that the latency is significantly below  $100\mu s$ . For the synchronization between the stereo system and the laser scanner the synchronization out signal from the the laser, which is emitted in the middle of the scan, is used to trigger the cameras at a frequency of 12.5Hz.

The embedded processing units of the cameras are used to synchronize the internal clock to a *sntp* network. That provides the possibility to append every acquired image a highly accurate ntp time stamp.

### 2.1 Calibration

**Stereo System.** Three successive steps are needed to calibrate the stereo camera system.

At first the *interior orientation* including the principle point, the principle distance and the correction for the lens distortion for both cameras have to be determined. For the modeling of the lens distortion a 7-parameter model by *Brown* [2] is used. The calibration process is performed with an optical target wall (see Fig. 1(a)).

The *relative orientation* is needed for the following rectification step of the stereo image pair and to obtain the *absolute orientation*. Through a set of corresponding image coordinates of a normalized stereo image pair the relative orientation can be calculated.

The absolute orientation describes the point transformation from a model coordinate system to a superordinate world coordinate system [8]. For the translation and rotation six parameters are needed and one parameter is used to scale the system. Only the scaling factor is needed which equals the length of the base vector between booth optical centers. Additionally, the deviation of the transformed model coordinates from the corresponding world coordinates is a measurement for the quality of all three consecutive calibration steps. The deviation is typically less than 1mm.

**Cross Calibration Laser Scanner.** There are well engineered calibration procedures for different kinds of laser scanner like terrestrial laser scanner [14] or profiler [15]. But these strategies cannot be applied to a multilayer laser scanner. The terrestrial scanner can be calibrated over perfectly scanned geometrical elements or the intensity information which are turning the scanner into a camera. Compared to that, the multilayer scanner has only four scan lines which are insufficient for that kind of calibration. In addition, the laser beams cannot be detected by a visual camera.

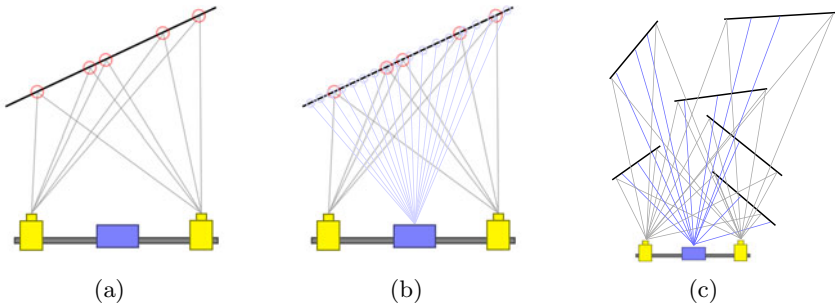
So a new approach was developed trying to determine the systematic measuring errors and the orientation to the stereo system in one step. To achieve this, the stereo system is used to deliver reference data in the form of planes. Over a spatial intersection, object points on a plane are collected and used to calculate the plane parameters, the normal vector  $N$  and a plane point  $P$ . The laser points are fitted onto their matching reference planes through an adjustment (see Fig. 2). To ensure that the stereo based reference planes are sufficiently accurate this process is executed under laboratory conditions with a specific range limit.

The functional model of the adjustment consists of an orientation part with a rotation  $R_L$  and a translation  $T_L$  and a fault model to cover the systematic error of the scanning device. The deviation between the measured  $d_M$  and the true point distance  $d$  is linearly described with a constant factor  $a$  and an offset value  $b$ . As the laser scanner generates polar coordinates a stretching factor  $s$  for the measured vertical angle  $\Phi_M$  is introduced. For simplicity the vertical measurement angle  $\Theta_M$  remains unaffected:

$$d = ad_M + b \quad (1)$$

$$\Phi = s\Phi_M \quad (2)$$

$$\Theta = \Theta_M \quad (3)$$



**Fig. 2.** Laser calibration process. (a) Determining object points on plane through image point triangulation. (b) Determining object points on plane by the laser scanner. (c) Repeating step (a) and (b) for several spatially well distributed planes.

Trough a conversion from a polar to Cartesian coordinate  $L$  it is possible to put all together into the plane equation to set up the functional model solved for the measured distance:

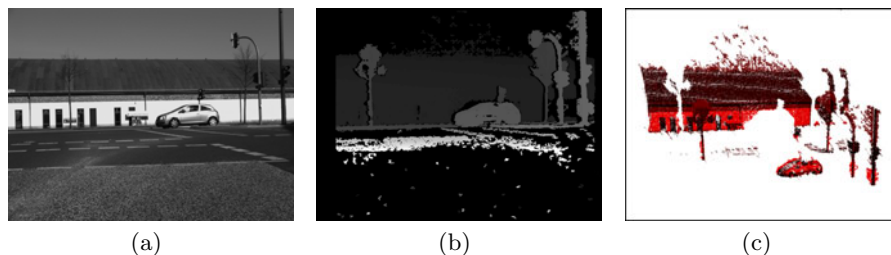
$$d_M = \frac{1}{a} \left( \frac{NP - NT}{NRW} \right) - b, \text{ with } L = dW, \text{ and } W = \begin{pmatrix} \cos \Theta \cos \Phi \\ \sin \Theta \cos \Phi \\ \sin \Phi \end{pmatrix} \quad (4)$$

Each measured laser point lying on a reference plane is used as an observation for the overdetermined system of nonlinear equations which is solved by a general least squares adjustment (for more details see [7]). As a result we get the correction of the laser raw data which is simultaneously transformed into the stereo coordinate system for further processing. After the calibration the RMS of the laser points referring to the reference planes is less than 2cm.

## 2.2 Stereo Processing

The process of obtaining 3D world coordinates out of corresponding 2D image points from two images observing the same scene is known as stereo vision. To do so, the interior parameters of the cameras, the relative orientation of the normalized images and the length of the base vector must be known. To simplify the correspondence search on both images called *stereo matching* it is helpful to project both images in one plane. In addition, the epipolar lines mapping the same world point must lie on the same image height and parallel to the base vector (stereo rectification). In this stereo normal configuration the world point can be determined trough the horizontal disparity of the corresponding image points and and the principle distance of the camera.

For illustration of this process, Fig. 3 shows the evolution from the image data trough the disparity image to the world points on a typical traffic scene (hereinafter referred to as *scene 1*). Here the *semi-global block matching* algorithm by *OpenCV* is used. The stereo points in the results are generated through a faster block matching algorithm.



**Fig. 3.** Stereo processing on scene 1. (a) Left stereo image. (b) Disparity image. (c) Resulting 3D world coordinates.

## 3 Object Segmentation

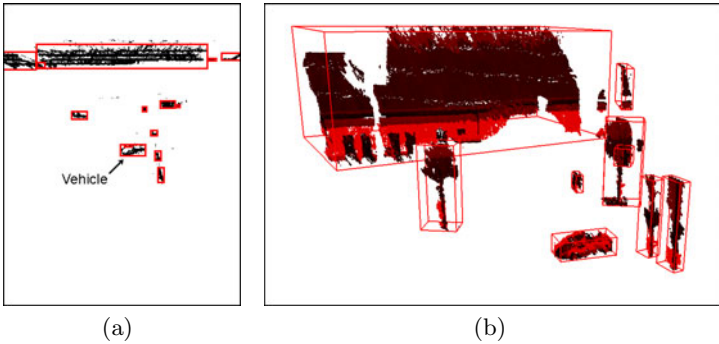
### 3.1 Stereo Camera

The object information like obstacles or traffic vehicles has to be extracted from the stereo raw data. To achieve this, the disparity image or the derived 3D points are used for the clustering. In [12] camera fronto-parallel objects are detected through their representation as vertical straight lines in a *v-disparity image*. In addition to the *v-disparity* [10] uses the *u-disparity image* and a region growing algorithm to find segments in the disparity image. Another way for the object grouping is to reduce the 3D information to a 2D occupancy grid as done in [11].

In this work the stereo segmentation is based on an approach by [9] which combines the whole informational content from mono image processing and stereo vision. In order to simplify and accelerate the process the segmentation is divided into two main steps.

At first the obtained and filtered 3D points are mapped on the horizontal plane in a predefined depth map which divides the ground in cells of constant size. In contrast to [9], here the depth map uses the disparity values as the ordinates and the lateral ranges as the abscissa. The disparity was chosen for the longitudinal range to avoid the scattering of the stereo data in the depth. Each cell has a constant width and a height of one disparity. The grey level of the cells corresponds to the accumulated mapped world points. After the creation of the depth map a binary depth map is derived by thresholding the grey levels of the cells. Regarding to the decreasing point density with the distance, this threshold depends on the distance, too. The binary map is used to extract the point clusters from the bird's-eye view with a region growing algorithm. As a result we get a set of 2D segments with a depth and a width description (Fig. 4(a)).

In the second step the depth and width information from the previously extracted 2D segments is used to define a region of interest on the original image enclosing the specific object. Those region of interest images called layers are used to refine the object size and to further segment the object. With the *canny edge detector* the object contour is extracted from the layer image and the complete 3D object bounding box can be established (Fig. 4(b)).



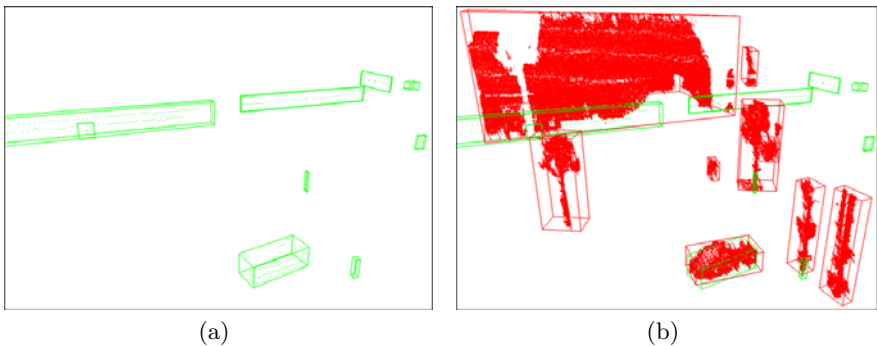
**Fig. 4.** Steps of stereo segmentation on scene 1. (a) Segmented binary depth map in bird's-eye view. (b) Complete stereo segmentation after step 2.

### 3.2 Laser Scanner

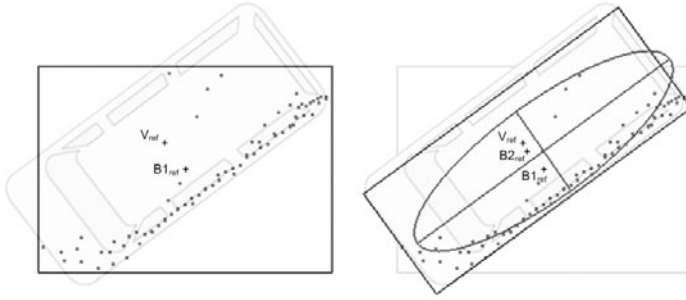
The clear arrangement and high accuracy of the laser point data allows a more straightforward segmentation solution compared to the stereo data. The euclidean distance of two points is taken into account to check if they belong to one segment. For each direction a constant distance threshold  $o$  is used to build an segmentation ellipsoid around every point. Considering the radial measuring principle of the scanner, a distance  $d_{lp}$  dependent value  $s$  is added to the constant distance threshold. Every Point  $P'$  is then connected to a Point  $P''$  if the following inequality holds true:

$$\frac{(P'_x - P''_x)^2}{(o_x + s_x * d_{lp})^2} + \frac{(P'_y - P''_y)^2}{(o_y + s_y * d_{lp})^2} + \frac{(P'_z - P''_z)^2}{(o_z + s_z * d_{lp})^2} \leq 1 \quad (5)$$

Fig. 5 shows the results of the laser segmentation on scene 1.



**Fig. 5.** (a) Segmentation of laser points in scene 1. (b) Overlapping object information from both sensors.



**Fig. 6.** The ellipse shaped box (right) approximates the real vehicle box more exactly than the unshaped bounding box (left)

### 3.3 Object Shaping

To achieve an accurate object tracking it is important to have an precise reference point for the tracking object. A simple but inaccurate way is to determine the center of gravity over all object points [6]. A better solution seems to be to use the center point of the bounding box as the reference point. For that, the bounding rectangular box with the minimal volume has to be shaped. For simple object shapes as received from a laser scanner a usual object box fitting [16] is suitable. But for large unstructured object point sets coming from a stereo system this fitting could be expensive.

A nice way to avoid this, is to fit an ellipse over the object points to get the optimal bounding box. For this purpose, the points are mapped on the horizontal plane into 2D space. An ellipse fitting approach by Fitzgibbon et al. [5] is used which allows a ellipse specific fitting of scattered data computationally efficient (Fig. 6). Ellipse specific fitting means that the algorithm returns always an ellipse solution independently of the input data.

The shaping is linked with a merging process where overlapping objects or objects which are lying too close to each other according to a proper threshold are grouped.

To obtain an even better approximation of the real object shape and the center point, the real object size is estimated and adapted during the tracking. In contrast to the *two line approach* [6] where only the length and the width of the object is tracked, here all three dimensions are updated.

## 4 Object Tracking

In order to derive a more comprehensive description of the surrounding traffic situation the extracted objects need to be tracked. Here, a multi object tracking approach for each sensor is used. The process of moving traffic objects is covered through a five parameter state model (6) also described in [1]. To describe the complete behavior of those moving objects like acceleration or cornering the



observation of the yaw angle ( $\psi$ ) and the yaw rate ( $\omega$ ) are needed in addition to the common parameters direction ( $x, y$ ) and velocity ( $v$ ). This leads to a non-linear process model which is linearized for the usage in an *extended Kalman filter*.

$$\hat{x}_k = (x_k \ y_k \ \psi_k \ v_k \ \omega_k)^T \quad (6)$$

Besides the usual measurement of the position coordinates, the measurement value for the yaw rate is extracted from the shaped object. Through this additional information the filter becomes more stable and reliable.

For the tracking innovation step an association between the set of predicted objects and the set of measured objects is needed. A sophisticated approach for the association was inspired by a work about image feature correspondences achieved through *singular value decomposition* (SVD) [13]. Feature points coming from two images are mapped under the constraint of proximity and similarity in a single SVD operation. In the case of associating objects the proximity property describes the euclidean distance of the object box centers. For a stereo based object the similarity property is determined by the average intensity of the including object points. In contrast, the average echo width of the including points from a laser object are used as the similarity description.

## 5 Results

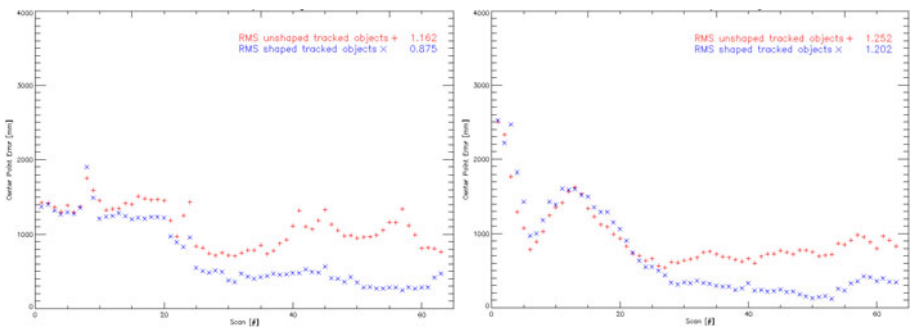
To obtain reference data for testing the object discrimination with real data a second test vehicle equipped with two GPS receivers was used. The two GPS receivers (*Trimble 5700*) together with a base station (*Trimble 750*) are able to work in the *Real Time Kinematic* mode. This enables a point measurement with an accuracy below 2cm with a frequency of 10Hz under normal conditions for receiving. The alignment of the two GPS antennas and the dimensions of the test vehicle were determined in advance to calculate the object box of the vehicle through the GPS positions.

During the measurement the first test vehicle equipped with the sensors stays in a fixed position to avoid additional errors. To transform the GPS positions into the coordinate system of the left camera, the exterior orientation of the left camera with respect to the GPS coordinate system has to be known. In order to achieve this, an arrangement of several optical targets in the measurement scene was used, which were measured by GPS (see Fig. 7).

Different routes were driven to cover the performance of the object discrimination as completely as possible. The results have shown that the center points of the tracked object boxes, which are derived from shaping are lying closer to the ground truth data than the center points of the tracked unshaped object boxes. The shaping process is especially effective during cornering where the additional angle information improves the tracking performance. In the following representative example the test vehicle turns right in front of the sensors. Figure 8 illustrates the distance from the calculated object box center to its reference center.



**Fig. 7.** Arrangement for exterior orientation of the camera (left) and gps setup (right)



**Fig. 8.** Accuracy of the tracked unshaped and shaped object boxes, derived from laser point data (left) and stereo data (right)

As you can see, after an initial phase for the tracking both measurements stabilize with a better accuracy for the shaped objects. In this example the center deviation for the stereo objects is partially smaller compared to the laser scanner objects. This could be caused by different starting conditions for the tracking or the speed of the object size adaption. In this example, the object tracking by the laser scanner performs better, indicated by a smaller RMS value.

## 6 Conclusion

The capability of extracting object hypotheses for a stereo and laser scanner system without interdependency was shown. Besides, a novel cross calibration procedure and a smart approach for refining the object shape over ellipse fitting were introduced and successfully tested. Up to a certain distance both sensors are able to deliver equivalent results. Otherwise the laser scanner performs better. In the next step the sensor data will be fused on different processing levels to verify the effects on object discrimination quality and the amount of false detections.

## References

1. Boehringer, F.: Gleisselektive Ortung von Schienenfahrzeugen mit bordautonomer Sensorik. Universitt Karlsruhe, PhD: 178 (2008)
2. Brown, D.C.: Close-range camera calibration. In: Photogrammetric Engineering, pp. 855–866 (1971)
3. Catala Prat, A., Reulke, R., Kstner, F.: Early Detection of hazards in driving situations trough multi sensor fusion. In: Proc. of FISITA World Automotive Congress, pp. 527–536 (2008)
4. Fang, Y., Masaki, I., Horn, B.: Depth-based target segmentation for intelligent vehicles:fusion of radar and binocular stereo. *IEEE Trans. Intell. Transp. Syst.* 3(3), 196–202 (2002)
5. Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct Least Square Fitting of Ellipses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(5), 476–480 (1999)
6. Fuerstenberg, K.C., Dietmayer, K.: Pedestrian Recognition and Tracking of Vehicles using a vehicle based Multilayer Laserscanner. In: 10th World Congress on Intelligent Transport Systems, Madrid, Spain (2003)
7. Haberjahn, M.: Cross-Kalibrierung eines Mehrzeilen-Laserscanner- und Stereokamera-Systems zur Fahrzeugumfelderfassung. 3D-Nordost 2009, Berlin, Germany (2009)
8. Horn, B.K.P.: Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Amer. A* 4(4), 629–642 (1987)
9. Huang, Y., Fu, S., Thompson, C.: Stereovision-Based Object Segmentation for Automotive Applications. *EURASIP Journal on Applied Signal Processing*, 2322–2329 (2005)
10. Kormann, B., Neve, A., Klinker, G., Stechele, W.: Stereo Vision Based Vehicle Detection. In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPH 2010. Angers, France, pp. 17–21 (2010)
11. Kumar, S., Gupta, D., Yadav, S.: Sensor Fusion of Laser & Stereo Vision Camera for Depth Estimation and Obstacle Avoidance. *International Journal of Computer Application* 1(26) (2010)
12. Perollaz, M., Labayrade, R., Royere, C., Hautière, N., Aubert, D.: Long Range Obstacles Detection Using Laser Scanner and Stereovision. In: *Procs. IEEE Intelligent Vehicles Symposium 2006*, Tokyo, Japan, pp. 182–187 (2006)
13. Pilu, M.: Uncalibrated Stereo Correspondence by Singular Value Decomposition. Technical Report HPL-97-96, Digital Media Department, HP Laboratories Bristol (1997)
14. Rietdorf, A.: Automatisierte Auswertung und Kalibrierung von scannenden Messsystemen mit tachymetrischem Messprinzip. PhD thesis, Technische Universitt Berlin (2005)
15. Santolaria, J., Pastor, J.J., Brosted, F.J., Aguilar, J.J.: A one-step intrinsic and extrinsic calibration method for laser line scanner operation in coordinate measuring machines. *Electronic Journals - Measurement Science and Technology* 20 (2009)
16. Wender, S., Dietmayer, K.: 3D Vehicle Detection using a Laser Scanner and a Video Camera. In: 6th European Congress on ITS, Aalborg (2007)

# Realistic Modeling of Water Droplets for Monocular Adherent Raindrop Recognition Using Bézier Curves

Martin Roser, Julian Kurz, and Andreas Geiger

Department of Measurement and Control  
Karlsruhe Institute of Technology (KIT)  
D-76131 Karlsruhe, Germany

**Abstract.** In this paper, we propose a novel raindrop shape model for the detection of view-disturbing, adherent raindrops on inclined surfaces. Whereas state-of-the-art techniques do not consider inclined surfaces because they assume the droplets as sphere sections with equal contact angles, our model incorporates cubic Bézier curves that provide a low dimensional and physically interpretable representation of a raindrop surface. The parameters are empirically deduced from numerous observations of different raindrop sizes and surface inclination angles. It can be easily integrated into a probabilistic framework for raindrop recognition, using geometrical optics to simulate the visual raindrop appearance. In comparison to a sphere section model, the proposed model yields an improved droplet surface accuracy up to three orders of magnitude.

## 1 Introduction

Outdoor navigation and surveillance demand for reliable and robust computer vision algorithms. They have to meet stringent conditions concerning disturbances caused by arbitrary weather conditions. In fact, there are various atmospheric influences which restrict the usability of these systems such as fog, rain or snow. Especially in rainy weather it is often the case that adherent waterdrops on the lens-protecting glass disturb the view of a camera. Although a lot of research has been pursued in robotics [16], computer vision [5,7,11] and for driver assistance [9,12,19], raindrop detection still remains a challenging task. This might be for several reasons: Water droplets on a glass surface exhibit a large variety in shape and size. Transparency makes their appearance highly dependent on the image background. Moreover, water droplets on the protecting glass of a camera are subject to severe out-of-focus blur which lowers their distinguishability from the scene background.

Recent work on raindrop detection [8,12] assumes a simple sphere section for modeling the droplet boundary. Especially on tilted planes where gravity causes an unidirectional droplet deformation, this assumption does not hold. While high-order polynomials are more adequate, a physical interpretation of the fitted parameters is hard.

In this paper we propose a novel raindrop shape model, that provides a physically interpretable parameter set of low dimensionality. Our main contribution is a model based on cubic Bézier curves. We provide a broad validation of the model parameters and show, that the shape deviation between fitted model and real droplet will be significantly decreased compared to state-of-the-art sphere section models. The proposed shape model can be easily integrated into existing raindrop recognition frameworks.

## 2 Related Work

The visual effects of rain are manifold and complex. Water droplets in the atmosphere lead to contrast attenuation in the far-field of the camera, whereas falling raindrops produce sharp intensity changes in image sequences. Adherent raindrops in front of the camera lens disturb the view from the camera and light reflections on the droplet surfaces additionally deteriorate computer vision algorithms.

Related work on dynamic weather effects like the appearance of falling raindrops in image sequences has been performed by [5,7]. They studied the influence of falling raindrops on the image acquisition process and introduced a photometric model for spherical raindrops in the atmosphere that is used for enhanced video processing like removing rain from image sequences [4] or rain streak rendering [6].

In the targeted context of outdoor navigation and surveillance, falling raindrops and rain streaks can be considered as atmospheric noise and are not the dominant disturbing effect. Stronger limitations are imposed by adherent water droplets on the glass surface covering the camera lens. Kurihata et al. [9] used a machine learning approach with raindrop templates to detect raindrops on windshields from inside a moving vehicle. Results within the sky area were promising, whereas the proposed method produced a large number of false positives within the non-sky regions of the image where raindrop appearance modeling becomes more challenging. In this work, in contrast, we aim to accurately exploit the physical relationship between droplets shape and their appearance. Zhang et al. [19] combined a wavelet transform for image blur detection with motion analysis using cumulative differences to recognize optical contaminations close to the camera. Their approach works well for rigid, opaque contaminations but fails in the presence of raindrops because their appearance strongly depends on the scene background. Yamashita et al. exploited hardware constraints like multiple cameras [13,18] or pan-tilt surveillance cameras [16,17] with known yaw rates in order to bypass the challenge of modeling the complex optical behavior of raindrops. Roser et al. [8,12] simulated spherical droplets on a glass surface using geometrical optics and out-of-focus blur for the task of raindrop detection. However, they lack a realistic shape parametrization of droplets that in practice are subject to gravity.

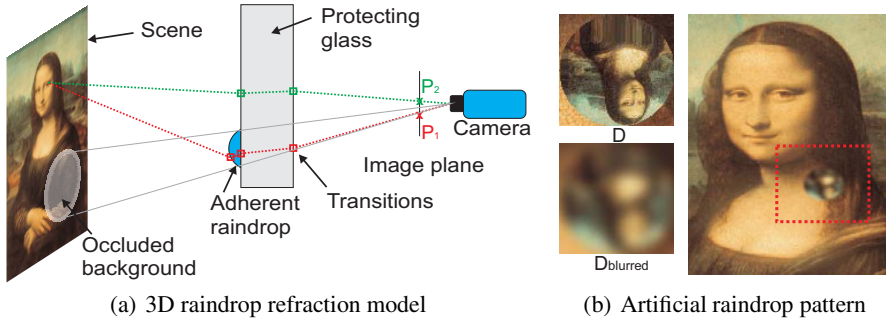
The remainder of this paper is structured as follows: Section 3 discusses the propagation of light rays through a droplet and shows how it can be used for raindrop detection. In Section 4 we propose a raindrop shape model based on a Bézier curve representation. Validation results on real data and a comparison to a sphere section model [8] are given in Section 5.

## 3 Raindrop Recognition

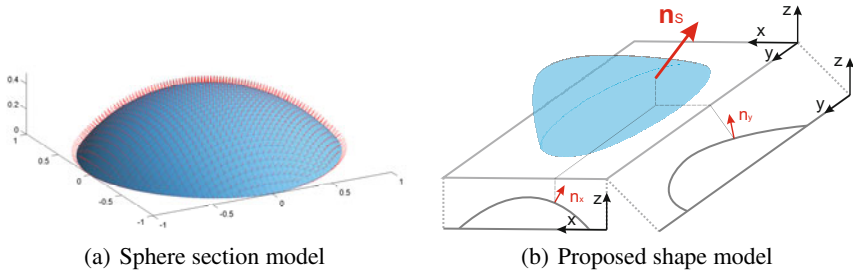
Given a set of  $n$  artificial raindrop hypotheses  $\mathbf{D}_1, \dots, \mathbf{D}_n$  for different image positions and presumed drop radii, the raindrop recognition task can be formulated as computing the MAP estimate of the conditional probability

$$p(\mathbf{d}|\mathbf{z}) \propto p(\mathbf{d})p(\mathbf{z}|\mathbf{d}) \quad (1)$$

with respect to the pattern  $\mathbf{d} \in \{\emptyset, \mathbf{D}_1, \dots, \mathbf{D}_n\}$ . Here  $\emptyset$  is a background pattern that models the case where the image region is not disturbed by a raindrop and  $\mathbf{z}(u, v, r)$



**Fig. 1. Droplet refraction model.** (a) depicts the image formation process in the presence of raindrops on a protecting glass in front of the camera, using geometrical optics. In (b) an artificial raindrop pattern  $D$  is rendered by tracing the light rays through the raindrop to the background and composing all found background pixels. For demonstration purposes, an additional out-of-focus blur is applied and the blurred pattern  $D_{\text{blurred}}$  is added to the original image.



**Fig. 2. Raindrop surface models.** (a) shows the droplet surface and its surface normals for a sphere section model. In (b) a 3D model is created by superposing two orthogonal Bézier curves.

denotes the observation at position  $u, v$  with radius  $r$  in form of local image statistics. It can be achieved densely as well as from preselected points of interest, like CenSurE [1] or SURF [2]. Fairly standard cost measures such as the Sum-of-Absolute Differences (SAD) or the Sum-of-Squared Differences (SSD) are applied for modeling the observation likelihood  $p(z|\mathbf{d})$ . The prior may model the occurrence probability for different raindrop sizes in various adverse weather conditions in an empirical Bayesian perspective according to [15].

Raindrop hypotheses for any circular region  $\mathbf{x} = (u, v, r)^T$  are achieved from observed points in the environment, using geometrical optics. As depicted in Fig. 1(a), a light ray emanating from a point in the environment will be refracted by the raindrop and the protecting glass surface multiple times and reaches the camera sensor at point  $P_1$ . Unless the raindrop does not occlude this environment point, it will be sensed a second time at point  $P_2$ . Note, that the droplet acts as a convex lens with a small focal length. Hence, for typical application in navigation and surveillance it is ensured that only a minor part of the environment points are occluded by the raindrop (see Fig. 1(a)). An accurate geometric relationship between  $P_1$  and  $P_2$  can be derived using Snell's

law of refraction as shown in Fig. 1(b). Note, that the refraction on the protecting glass occurs with respect to the (constant) plane normal  $n_p$  of the protecting glass, whereas a general drop surface  $S$  exhibits a normal field  $N_S$  that can be deduced from the chosen drop parametrization. Whereas [8,12] use simple sphere sections that are in general 3D surfaces of constant curvature as depicted in Fig. 2(a), this model approximates the raindrop shape only insufficiently and results in a high model deviation especially when dealing with tilted glass surfaces. For this reason, we derive a raindrop shape model regarding numerous tilt angles and drop sizes by using two orthogonal oriented Bézier curves as illustrated in Fig. 2(b).

## 4 Raindrop Shape Model

### 4.1 Bézier Representation

Droplets on a horizontally aligned surface are symmetrical and have equal contact angles. When neglecting any gravity, they can be characterized adequately, using the sphere section model as described in [8,12]. However, gravity leads to a flattened raindrop surface shape which results in an inaccurate droplet modelling when assuming sphere sections. On tilted surfaces the sphere section model assumption is violated even more, because the unsymmetrically applied gravity force will shift the droplet centroid towards the declining direction, which yields different contact angles and a distinctly bellied shape as illustrated in Fig. 5.

The shape of a raindrop can be described by parametric functions, like polynomials of arbitrary order, Taylor polynomials or Bézier curves [3]. Here we employ Bézier curves, since they describe real water droplets accurately and they provide an intuitive, low-dimensional parameter set with a credible physical interpretation. This makes the verification of the model and an approximation for different angles and drop volumes more transparent than interpreting the coefficients of a polynomial fit.

A Bézier curve of the degree  $n$  is characterized by a control polygon consisting of  $n + 1$  Bézier points  $(P_i)_{i=0}^n, P \in \mathbb{R}^2$ . It is defined in an interval  $t \in [0 \dots 1]$  as

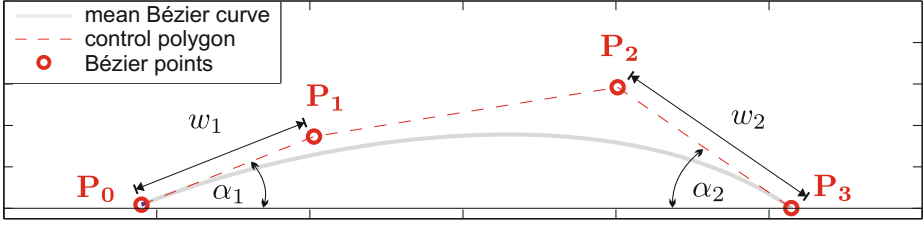
$$C(t) = \sum_{i=0}^n B_{i,n}(t)P_i, \tag{2}$$

whereas

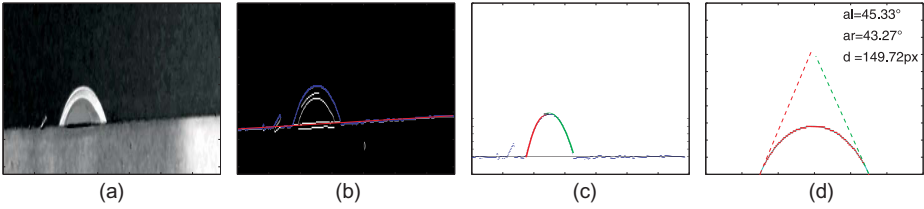
$$B_{i,n}(t) = \binom{n}{i} t^i(1-t)^{n-i} \tag{3}$$

indicates the Bernstein polynomial  $i$  of degree  $n$  [3].

A cubic Bézier curve ( $n = 3$ ) has sufficient degrees of freedom to describe the raindrop shape well. As depicted in Fig. 3, a capable interpretation of the Bézier points  $(P_i)_{i=0}^3$  can be achieved by transforming them to the contact angles  $\alpha_1, \alpha_2$  of the droplet that are originated from physics of boundaries and the weight factors  $w_1, w_2$  that are related to the centroid shift due to gravity.



**Fig. 3. Cubic Bézier curve representation.** The Bézier points are transformed physically interpretable:  $\alpha_1, \alpha_2$  represent the droplets contact angles and the weight factors  $w_1, w_2$  are related to the influence of gravity for inclined surfaces.



**Fig. 4. Image processing for drop shape extraction.** (a) shows the original image taken in the experimental setup. A distinction between surface plane and raindrop points is performed by RANSAC line fitting in the Canny image (b). In order to remove further outliers, two second order polynomials are fitted robustly to the left (red) and right (green) side of the raindrop (c). Finally, least squares Bézier curve fitting is performed on all inlier points (d).

$$\alpha_1 = \angle(\overline{\mathbf{P}_0\mathbf{P}_1}, \overline{\mathbf{P}_0\mathbf{P}_3}) \quad (4)$$

$$\alpha_2 = \angle(\overline{\mathbf{P}_2\mathbf{P}_3}, \overline{\mathbf{P}_0\mathbf{P}_3}) \quad (5)$$

$$w_1 = \overline{\mathbf{P}_0\mathbf{P}_1} \quad (6)$$

$$w_2 = \overline{\mathbf{P}_2\mathbf{P}_3} \quad (7)$$

Finally, the curvature normals of two orthogonal, cubic Bézier curves form a 3D droplet surface  $\mathbf{S}$  as illustrated in Fig. 2(b)

$$\mathbf{n}_S = \frac{1}{\| (n_x, 0, 1)^T + (0, n_y, 1)^T \|} \left( \begin{pmatrix} n_x \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ n_y \\ 1 \end{pmatrix} \right), \quad (8)$$

where one curvature represents the side view with the inclination angle of the lens-protecting glass and the other representing the top view with  $\theta = 0^\circ$ .

## 4.2 Bézier Curve Fitting

In order to characterize and describe the drop shape in terms of cubic Bézier curves, we performed an image pre-processing as described briefly in the following paragraph.



An overview of the image processing and curve fitting methods can be found in Fig. 4. The first step of extracting the drop shape is to take raw observations from the canny edge image (Fig. 4(b)). A robust RANSAC line fitting approach estimates the remaining glass surface direction and hence compensates small errors due to inaccuracies in the angular arrangement of glass plate and camera. In order to further remove outliers from the measurements, two parabolas were fitted through the remaining points, using RANSAC: one from the maximum to the left side (red line in Fig. 4(c)) and one to the right side (green line in Fig. 4(c)). Note, that we do not use the parameters of the parabola fits directly because the shape is neither described consistently nor interpretable in a physical way. Instead, a combination of all inliers gives a set of points that is used for the subsequent Bézier curve fitting as shown in Fig. 4(d). The Bézier curve fitting is performed in a least squares sense [14] by splitting (2) into two independent equations for the  $x$  and  $y$  coordinates

$$x = a_x t^3 + b_x t^2 + c_x t + d_x \quad (9)$$

$$y = a_y t^3 + b_y t^2 + c_y t + d_y, \quad (10)$$

and computing the Bézier points  $(\mathbf{P}_i)_{i=0}^3$  by comparing coefficients to (2). The factor  $t \in [0 \dots 1]$  corresponds to the normalized curvature length. For a curve described by  $N$  points  $t$  is approximated by

$$t(n) = \frac{\sum_{k=1}^n \sqrt{\Delta x(k)^2 + \Delta y(k)^2}}{\sum_{l=1}^N \sqrt{\Delta x(l)^2 + \Delta y(l)^2}}, \quad (11)$$

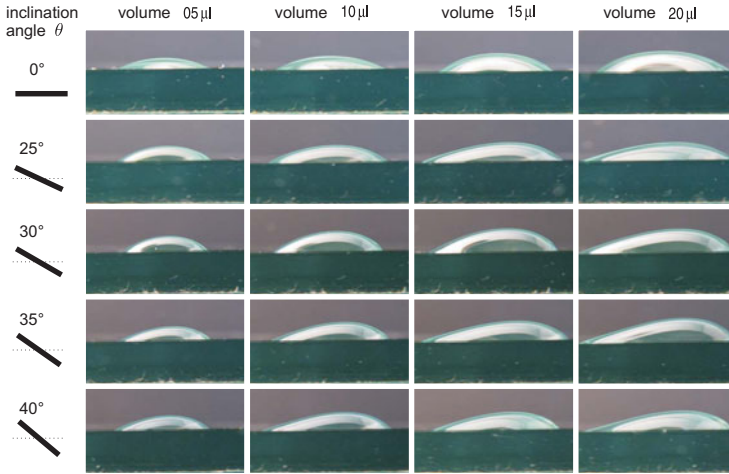
where  $\Delta x$  and  $\Delta y$  are the differences between two neighboring points.

Repeating the experiments  $M$  times for each drop volume and inclination angle configuration, we receive  $M$  different Bézier curve parameterizations. A mean Bézier curve is finally achieved by computing the mean of each Bézier point  $(\mathbf{P}_i)_{i=0}^3$ :

$$\mathbf{P}_i = \sum_{k=1}^M \frac{\mathbf{P}_i^k}{M}. \quad (12)$$

## 5 Results

For all experiments, a digital camera was mounted next to a tiltable glass plate to capture the shape of water droplets of different sizes under multiple inclination angles. We used an *Eppendorf Research Plus* pipette for all experiments in order to guarantee a precise but adjustable drop volume size. In the experimental setup all drops are illuminated by a lamp in front of a dark background to achieve a good contrast and ensure reliable shape extraction. As input for finding an empirical description of water droplets on a flat surface, multiple images with different drop sizes and surface inclination angles were taken. The chosen drop volumes for our test series were 5, 10, 15 and  $20\mu\text{l}$  and the inclination angles of the glass plate were  $0^\circ$ ,  $25^\circ$ ,  $30^\circ$ ,  $35^\circ$  and  $40^\circ$ . The chosen



**Fig. 5. Experiments.** Sample imagery for manifold drop volumes and surface inclination angles. For model estimation the mean fit of 20 images for each drop volume and inclination angle setting is used.

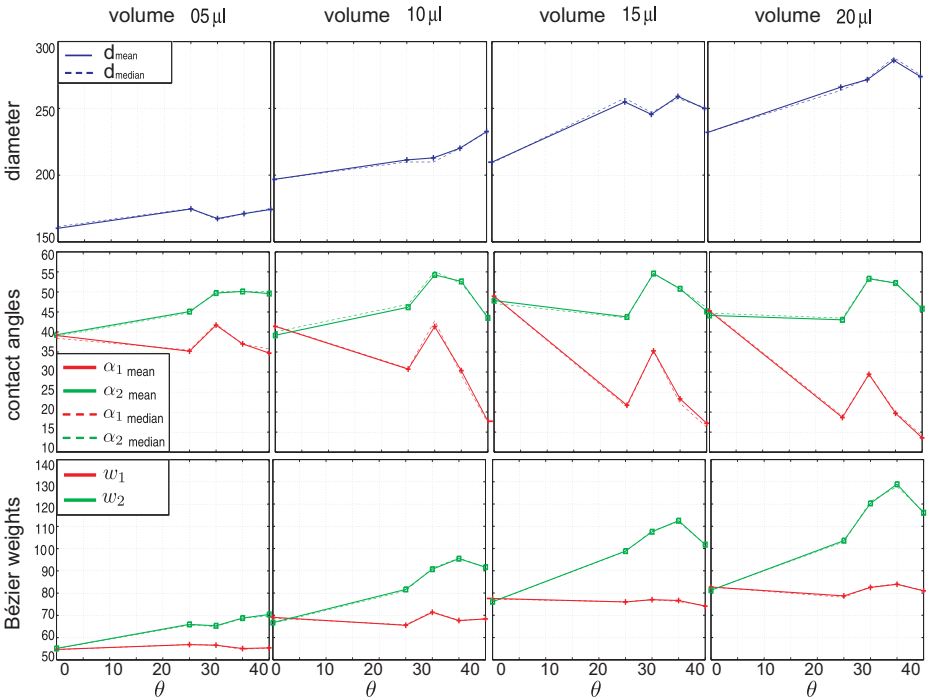
drop volumes in  $\mu\text{l}$  correspond to 1.06, 1.34, 1.53 and 1.7 mm drop radii of falling raindrops, which was motivated by [7] who proposes probable raindrop radii between 0.5 – 2.5 mm. The experiments were repeated 20 times for each drop volume and inclination angle configuration. Hence, 400 raindrop shape images were acquired in total. An overview of the different setup properties and their effects on the droplet shape is depicted in Fig. 5.

The results section is divided into two parts: First, we discuss the estimated raindrop parameters. Then we present a comparison of the proposed model with the sphere section model of [8].

## 5.1 Model Parameters

A model capable of generating realistic droplet surfaces demands for a low dimensional parametrization to avoid overfitting. In this section, we discuss the obtained dependencies of the Bézier curve based model with respect to the design parameters (drop volume and inclination angle).

Assuming the raindrop diameter  $d = |\overline{\mathbf{P}_0\mathbf{P}_3}|$ , the upper row in Fig. 6 shows the expected behavior that the drop radius increases with its volume. A tendency of increasing drop diameters for larger inclination angles exists, although it may not be the predominant effect. This phenomenon can be explained from the droplet area that loses its circular shape and develops a predominant direction with increasing inclination angles. Hence, even if the drop volume is not a-priori known like it is the case in image-based raindrop detection tasks, for a given surface inclination angle the volume can be estimated from the observed drop diameter. In principle, the scale of standard multi-scale interest point detectors like SURF [2] would provide sufficient information for that task.

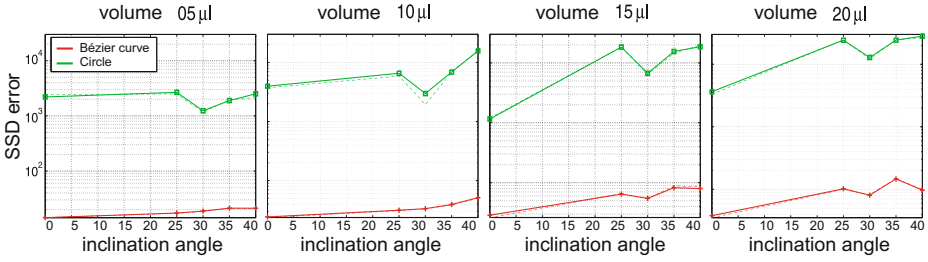


**Fig. 6. Model parameter.** The first row shows the averaged droplet diameter  $d$  as a function of the inclination angle  $\theta$  for different raindrop volumes (columns). The second and the third row depict the mean contact angles  $\alpha_1, \alpha_2$  and the mean Bézier weights  $w_1, w_2$ , respectively.

Tilting the glass surface leads to a deformation of the drop due to changed gravity influences. For this reason, we expect an increasing difference  $\Delta = |\alpha_2 - \alpha_1|$  between both contact angles. The middle row in Fig. 6 shows the expected behavior, although not all contact angles could be extracted accurately, throughout the experiments. However,  $\alpha_1$  tends to decrease with increasing inclination angle.  $\alpha_2$  shows a slight ascent but decreases for  $\theta = 40^\circ$ . This can be explained by having a deeper look at the performed experiments. We are only interested in stationary droplets. For  $\theta \approx 40^\circ$ , the drop begins rinsing down and hence we could not acquire representative imagery data.

The Bézier weight  $w_1$  remains constant for varying inclination angle and ascends with increasing drop volume. For the right side  $w_2$  increases with inclination angle and drop volume while for angles  $\theta \approx 40^\circ$  the drop begins to move, again. As discussed above, for these inclination angles, no reliable conclusion can be drawn.

In conclusion, a physically correct droplet shape can be derived, as soon as the inclination angle and the drop volume are given. The drop volume can be deduced from the observed raindrop diameter, whereas the surface inclination angle is given by the defined camera mounting. This makes the proposed droplet shape model applicable for an image-based raindrop detection approach.



**Fig. 7. Model accuracy.** (a)-(d) show the SSD error of a sphere section model and the proposed model, using cubic Bézier curves.

## 5.2 Comparison

For comparing the accuracy of the proposed method to state-of-the-art, a 2D cut of a sphere section was fitted to the extracted raindrop surface measurements using nonlinear Levenberg-Marquardt optimization [10]. The error measure is defined in terms of Sum-of-Squared Differences (SSD).

Fig. 7 shows the error generated using the sphere section model in comparison to the new Bézier curve based model. Even for flat surfaces ( $\theta \approx 0^\circ$ ) and small drop volumes, the proposed model has an SSD error which is three orders of magnitude smaller. This illustrates the importance to take into account the gravity force which flattens the drop surface. An increasing drop volume and inclination angle lead to unsymmetrical droplet deformation, which emphasizes the advantage of the proposed shape model with respect to the sphere section model.

## 6 Conclusion

In this paper we proposed a novel raindrop shape model based on cubic Bézier curves and showed its potential for its integration in image-based raindrop detection approaches. The model was deduced from numerous experiments on water drops of various volumes on a flat surface with different inclination angles. A physically correct droplet shape could be computed if just the inclination angle and the drop volume were given. The drop volume was deduced from the observed raindrop diameter. This makes the proposed droplet shape model applicable for an image-based raindrop detection approach. Finally, we showed that the shape deviation between the estimated Bézier curve based model and the real droplet was significantly decreased compared to state-of-the-art sphere section models.

## Acknowledgment

The authors would like to thank the Karlsruhe School of Optics and Photonics (KSOP).

## References

1. Agrawal, M., Konolige, K., Blas, M.R.: CenSurE: Center Surround Extremas for Realtime Feature Detection and Matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 102–115. Springer, Heidelberg (2008)
2. 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)
3. Farin, G.: Curves and surfaces for CAGD: a practical guide. Morgan Kaufmann Publishers Inc., San Francisco (2002)
4. Garg, K., Nayar, S.K.: Detection and removal of rain from videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004), vol. I, pp. 528–535 (June 2004)
5. Garg, K., Nayar, S.K.: When does a camera see rain? In: IEEE International Conference on Computer Vision (ICCV 2005), vol. 2, pp. 1067–1074 (October 2005)
6. Garg, K., Nayar, S.K.: Photorealistic rendering of rain streaks. ACM Transactions on Graphics (July 2006)
7. Garg, K., Nayar, S.K.: Vision and rain. *Internatl. Journal of Computer Vision* 75(1), 3–27 (2007)
8. Halimeh, J., Roser, M.: Raindrop detection on car windshields using geometric-photometric environment construction and intensity-based correlation. In: IEEE Intelligent Vehicle Symposium (IV 2009), Xi'an, China (2009)
9. Kurihata, H., Takahashi, T., Ide, I., Mekade, Y., Muraseand, H., Tamatsu, Y., Miyahara, T.: Rainy weather recognition from in-vehicle camera images for driver assistance. In: IEEE Intelligent Vehicles Symposium (IV 2005), pp. 205–210 (2005)
10. Marquardt, D.: An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics* 11, 431–441 (1963)
11. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. *International Journal of Computer Vision* 48(3), 233–254 (2002)
12. Roser, M., Geiger, A.: Video-based raindrop detection for improved image registration. In: IEEE Workshop on Video-Oriented Object and Event Classification (in conjunction with ICCV 2009) (2009)
13. Tanaka, Y., Yamashita, A., Kaneko, T., Miura, K.T.: Removal of adherent waterdrops from images acquired with a stereo camera system. *IEICE - Transactions on Information Systems* E89-D(7), 2021–2027 (2006)
14. Teunissen, P.J.G.: Adjustment theory: an introduction. Delft University Press, Postbus 98, 2600 MG Delft (2000)
15. Vasconcelos, N., Lippman, A.: Empirical bayesian em-based motion segmentation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Puerto Rico, pp. 527–532 (1997)
16. Yamashita, A., Fukuchi, I., Kaneko, T., Miura, K.T.: Removal of adherent noises from image sequences by spatio-temporal image processing. In: IEEE International Conference on Robotics and Automation (ICRA 2008), pp. 2386–2391 (May 2008)
17. Yamashita, A., Harada, T., Kaneko, T., Miura, K.T.: Removal of adherent noises from images of dynamic scenes by using a pan-tilt camera. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), September- October 2, vol. 1, pp. 1:437–1:442 (2004)
18. Yamashita, A., Kuramoto, M., Kaneko, T., Miura, K.T.: A virtual wiper - restoration of deteriorated images by using multiple cameras. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003), vol. 3, pp. 4:3126–4:3131 (October 2003)
19. Zhang, Y., Yang, J., Liu, K., Zhang, X.: Self-detection of optical contamination or occlusion in vehicle vision systems. *Journal of Optical Engineering* 47(6), 067006 (2008)

# Illumination Invariant Cost Functions in Semi-Global Matching

Simon Hermann\*, Sandino Morales, Tobi Vaudrey, and Reinhard Klette

.*enpeda*.. group, Dept. Computer Science, University of Auckland, New Zealand

**Abstract.** The paper evaluates three categories of similarity measures: ordering-based (census), gradient-based, and illumination-based cost functions. The performance of those functions is evaluated especially with respect to illumination changes using two different sets of data, also including real world driving sequences of hundreds of stereo frames with strong illumination differences. The overall result is that there are cost functions in all three categories that can perform well on a quantitative and qualitative level. This leads to the assumption that those cost functions are in fact closely related at a qualitative level, and we provide our explanation.

**Keywords:** cost functions, stereo matching, illumination invariance.

## 1 Introduction and Related Literature

Stereo algorithms typically solve the correspondence problem by using some cost function (usually called the *data cost*) to determine a good match between pixels, and a discontinuity condition (usually called the *smoothness cost*) to handle outliers and homogeneous areas of the data. The combined cost is then minimized using an optimisation strategy that yields either scan-line or global consistency. At the moment, state-of-the-art optimisation strategies can be split into four major groups: belief propagation [7], graph-cuts [3], dynamic programming [17] which has been extended to a semi-global-matching technique (SGM) [10], and variational techniques [4,23].

One major problem in stereo matching that affects, primarily, the data cost are illumination differences (between stereo images). This effect can have a major influence on the image data and therefore on the quality of the matching cost itself. This is especially prominent when it comes to real world image sequences [5]. One approach [11,20,21] to handle illumination changes is to decompose the input images into a structure and a texture component. The texture component tends to be robust against illumination changes.

Recent studies [11,12] evaluated the performance of cost functions under illumination changes and found the census [22] cost function to be very robust against lighting differences. However, a gradient-based measure was unfortunately not part of those evaluations. In [13] the gradient was employed as a

---

\* The author thanks the German Academic Exchange Service (DAAD) for financial support.

similarity measure that was additively incorporated into the sum of absolute difference (SAD) cost function (accumulated over a  $3 \times 3$  window). Another study used the same two cost functions (SAD and gradient) when creating a similarity measure, but used a multiplicative contribution along with the normalized cross correlation cost function [6]. The contribution of the gradient was shown to provide a more reliable cost function when analysed under different lighting conditions. However, none of those studies were using or analysing the gradient concept isolated from other cost functions to determine the performance contribution of the gradient.

In this paper, the performances of three cost functions are compared: SAD applied to residual images (RSAD), the census cost function (both were previously identified as being robust against illumination differences) and a gradient-based measure, each being a representative of different categories of cost functions. Census belongs to the non-parametric ordering-based cost functions. Distances of central differences as approximation of image derivatives define a gradient-based similarity measure. RSAD is used as an example of illumination-based cost functions. We also use the regular version, the SAD cost function, to evaluate a metric that purely relies on the assumption of intensity consistency. In the methodology presented below, the four cost functions are evaluated under differing illumination and exposure settings on data sets where ground truth is available. The performance comparison is done using SGM [10] as the optimization technique. This method has proven to be computationally efficient [8] and of high quality [14].

The main goal of the presented research is to improve the robustness of stereo algorithms when used in real-world applications. Therefore, a comparison of the performance of the selected cost functions using two real-world sequences is performed. The sequences were recorded using three synchronized (trinocular) cameras, so evaluation is possible using the prediction error technique as described in [15].

The following two sections introduce the matching costs, as well as the semi-global matching technique with implementation details used in the experiments. This is followed by the methodology, data sets, and testing measures used for evaluation. This leads onto a discussion of the experimental results, which is then finalised by conclusions.

## 2 Cost Functions

In a rectified stereo image pair we consider a *base* and a *match* image. The base image is assumed to be the left image  $L$ . The match image  $R$  is usually the right image. The images are of same size within the image domain  $\Omega$ . We only consider intensity images (ignoring colour) in this paper with values in the range  $[0, I_{\max}] \subset \mathbb{N}$ . Any cost function  $\Gamma$  defines a global mapping  $\Gamma(L, R) = C$  that takes rectified stereo images  $L$  and  $R$  as input, and outputs a 3D cost matrix  $C$  with elements  $C(i, j, d)$ . The cost matrix represents the cost when matching a pixel  $(i, j)$  in  $L$  with a pixel  $(i - d, j)$  in  $R$ , for any relevant disparity

$d$  in the range  $[1, d_{\max}] \subset \mathbb{N}$  (zero is used for an “invalid” disparity, such as an occlusion). The ranges for  $i$  and  $j$  are  $[0, n] \subset \mathbb{N}$  and  $[0, m] \subset \mathbb{N}$ , respectively. We simplify notation as we are working with rectified images (epipolar lines are aligned horizontally), and we consider a fixed image row  $j$  in both the base and match image. Let  $p_i$  denote a pixel location in  $L$  at column  $i$ . Let  $L_i$  be the value at this location in the base image;  $q_{i-d}$  denotes the pixel location  $(i-d, j)$  in the match image  $R$  with intensity  $R_{i-d}$ . The cost can be abbreviated to omit the row  $C(i, d)$ .

We identify three different categories of cost functions: ordering-based, gradient-based, and intensity-based. We now introduce one representative of each function category that we evaluate in this paper.

**Non parametric or ordering-based cost function.** The census [22] cost function was identified to be a very robust measure when it comes to illumination changes [12]. Its performance serves as a reference when compared to the other two cost functions. We use it based on the following definition:

$$C_{\text{census}}(i, d) = \sum_{(x,y) \in \mathcal{N} + \{p_i\}} \rho(x, y, d) \quad \text{with} \quad (1)$$

$$\rho(x, y, d) = \begin{cases} 0 & \text{if } L_{x,y} > L_i \text{ and } R_{x-d,y} > R_{i-d} \\ 0 & \text{if } L_{x,y} < L_i \text{ and } R_{x-d,y} < R_{i-d} \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

where  $\mathcal{N}$  denotes the set of all nine pixel locations of the used  $3 \times 3$  window when centred at reference point  $(0, 0)$ .

**Gradient-based cost function.** This cost function employs the spatial distance of the end points of the gradient vectors as the similarity measure. It is defined as:

$$C_{\text{GRAD}}(i, d) = |\nabla L_i - \nabla R_{i-d}|_1 \quad (3)$$

where  $\nabla$  is estimated using central differences [4] and  $|\cdot|_1$  is the  $L_1$ -norm. Using central differences also keeps the neighbourhood influence within a  $3 \times 3$  window.

**Intensity-based cost function.** The *absolute difference* (AD) of the base and match pixel is the simplest and cheapest (in terms of computational cost) intensity-based measure:

$$C_{\text{AD}}(i, d) = |L_i - R_{i-d}| \quad (4)$$

In order to make a comparison more fair to census and the gradient, which use information from a  $3 \times 3$  neighborhood, we choose to sum the absolute difference over a  $3 \times 3$  window. This extension is known to be the *sum of absolute differences* (SAD) cost function. We define this intensity-based representative as:

<sup>1</sup> Our experiments use central differences. However, other gradient operators may possibly provide even better results depending on given image data.



$$C_{\text{SAD}} = \frac{1}{|\mathcal{N}|} \sum_{(x,y) \in \mathcal{N} + \{p_i\}} |L_{x,y} - R_{x-d,y}| \quad (5)$$

with cardinality  $|\mathcal{N}| = 9$ . However, since SAD is known to perform bad when it comes to illumination differences, we also apply this cost function on the texture component of the input images. We calculate the residual image (texture component)  $T$  of an image  $I$  as

$$T(I) = I - S(I) \quad (6)$$

where  $S(I)$  denotes the smoothed image (in this case, a  $3 \times 3$  mean image) of  $I$ . We refer to this version as RSAD.

### 3 Semi-Global Matching

This paper uses the *semi-global matching* technique (SGM) [10]. The SGM algorithm approximates the minimum of a 2D energy function by minimizing multiple 1D energies, and employing a dynamic programming scheme. The energy function consists of a data term and a smoothness term. The smoothness term penalizes small disparity changes of neighbouring pixels with a rather low penalty  $c_1$  to allow slanted surfaces. A second penalty is applied for larger disparity changes with a higher penalty  $c_2$ . This second penalty is independent of the actual disparity change in order to preserve depth discontinuities. The previously mentioned 1D energies are defined as minimum cost paths  $L_{\mathbf{a}}$  that start at each border pixel of the image and are traversed in direction  $\mathbf{a}$ .

A direction is basically a digitized line, and all digital lines of identical slopes are considered to be equivalent. Usually eight directions (up, down, left, right, and the in-between angles) are sufficient in SGM to obtain high-quality results. For a digital line in direction  $\mathbf{a}$ , processed between image border and pixel  $p$ , we only consider the segment  $p_0, p_1, \dots, p_n$  of that digital line, with  $p_0$  on the image border, and  $p_n = p$ . The cost at pixel position  $p$  (for a disparity  $d$ ) on the path  $L_{\mathbf{a}}$  is recursively defined as follows (for  $i = 1, 2, \dots, n$ ):

$$L_{\mathbf{a}}(p_i, d) = C(p_i, d) + \min \left\{ \begin{array}{l} L_{\mathbf{a}}(p_{i-1}, d) \\ L_{\mathbf{a}}(p_{i-1}, d-1) + c_1 \\ L_{\mathbf{a}}(p_{i-1}, d+1) + c_1 \\ \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) + c_2 \end{array} \right\} - \min_{\Delta} L_{\mathbf{a}}(p_{i-1}, \Delta) \quad (7)$$

where  $C(p, d)$  corresponds to the data cost term and is the similarity cost of pixel  $p$  for disparity  $d$ . The costs of paths  $L_{\mathbf{a}}$ , for all (say, eight) directions  $\mathbf{a}$ , are accumulated at a pixel  $p$ , for all disparities  $d$  in the range  $[1, d_{\text{max}}] \subset \mathbb{N}$ , and the disparity  $d_{\text{opt}}$  with the lowest cost is finally selected. To adjust the second penalty, the magnitude of the forward difference is calculated at each pixel  $p_i$  in direction  $\mathbf{a}$ . The magnitude of the forward difference scales the penalty for each  $p_i$  with

$$c_2(p_i) = \frac{c_2}{|I(p_{i-1}) - I(p_i)|} \quad (8)$$

To enforce the uniqueness of a disparity map (for a given stereo pair), roles of base and match images are swapped, which allows the calculation of a second disparity image. In a final consistency check, a pixel is labelled valid only if the corresponding disparities are identical; otherwise the pixel is labelled invalid. This is often referred to as a left-right consistency check.

The implementation used in this paper follows the SGM description from the original paper, as outlined above. However, it deviates in the following three points. To achieve sub-pixel accuracy the original paper proposes the standard procedure to fit a quadratic curve through costs of disparities  $d_{opt} - 1$ ,  $d_{opt}$ , and  $d_{opt} + 1$ , and to take the disparity position of the minimum. Since a comparison of cost functions is the objective of this paper, generating disparities with sub-pixel accuracy is omitted, as results may differ depending on the nature of the cost function.

The second difference is omitting the use of median filtering to remove outliers, because this is considered a post processing technique to improve performance, and raw performance of the cost functions are of interest in this paper.

The third difference is that costs are not scaled to 11-bit. The intention of this scaling is to have similar settings of penalties when cost functions are exchanged. However, simple scaling may not be descriptive enough to have a fair parameter setting between cost functions. For example, consider the census function that produces discrete costs in the range of  $[0, 8] \subset \mathbb{N}$ . There are basically no outliers possible, because of the nature of this function, while one outlier in SAD may result in an unfortunate scaling. However, this is an interesting topic and with a deeper understanding of the characteristics of cost functions w.r.t. the data, it should be possible to derive parameter settings for the optimization techniques. This will be discussion for future work. – Our implementation uses eight accumulation paths with  $c_1 = 30$  and  $c_2 = 150$ .

## 4 Methodologies and Datasets

Illumination issues have been proven to cause major issues when it comes to stereo matching and may, in fact, be the worst type of noise for stereo matching [16]. The first methodology uses a data set where ground truth is available. It tests the presented cost functions under normal lighting conditions, as well as with different exposures and illuminations between the left and right camera. The calculated costs are then evaluated when applied to the SGM optimisation approach. The second methodology examines the behaviour of the analyzed cost functions in combination with SGM using real-world image sequences. To overcome the lack of ground truth correspondence, we evaluate the output of the stereo algorithm using a prediction error technique [15]; which is similar to the approach reported in [19] to evaluate optical flow techniques.

**Synthetic or engineered test data.** Such stereo images provide a way to obtain ground truth, but come with their specific [9]. Stereo images may be recorded under different lighting and exposure settings, to provide test data where illumination/exposure could cause issues. Figure 1 shows an example from



**Fig. 1.** Illumination and exposure differences for the *Art* [14] input pair. Left to right: left (base) reference image, and right (match) image with identical illumination/exposure, right image with illumination change, right image with exposure change.

the data set [14] used in this paper; the cost functions are tested against the following images from this dataset: *Art*, *Books*, *Dolls*, *Laundry*, *Moebius*, and *Reindeer*. For each image pair used, the base image is using the exposure setting of 1 and illumination setting of 2, as defined on [14]. The left image is kept at this setting, but both illumination and exposure are varied in the right hand image. For each measure (outlined below) three tests are performed using different right hand images:

1. *Reference*: Identical lighting conditions (exp. 1, illum. 2)
2. *Illumination*: Illumination difference (exp. 1, illum. 1)
3. *Exposure*: Exposure difference (exp. 0, illum. 2)

We calculate the *good pixel percentage* (GPP) for all datasets. The GPP is defined as follows. Let  $G$  be the ground truth image of the corresponding data set where  $G_i$  encodes the *true disparity* at pixel  $p_i$ . The *good pixel percentage* is defined below:

$$GPP = 100\% \times \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \begin{cases} 1, & \text{if } |d_{opt} - G_i| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where  $\Omega$  is the set of all pixels where  $G_i \neq 0$ , as 0 is used to identify occlusions.

In other words, if the optimal disparity  $d_{opt}$  is within one disparity distance of the ground truth, it is a good pixel. We take the mean GPP over all data sets for each illumination setting as quality measure for the cost functions. Results are shown in Figure 3 and discussed in Section 5.



**Fig. 2.** Sample stereo pairs from a real world data set on [5]. The first two images from the left are a stereo pair from the bird sequence. The last two images from the left are a sample stereo pair from the driving straight sequence.

**Real world test data.** We analysed two sequences (400 trinocular frames each) as available on [5], recorded within an urban scenario; both sequences, were recorded the same day with only a few minutes of difference. See Figure 2 for sample frames of both sequences. The first sequence, *bird*, was chosen due to the strong brightness difference between the stereo pairs and varies throughout the sequence. The second sequence, *driving straight*, was recording while driving on a straight road. It is a traffic sequence in which the brightness in both input images varies only slightly.

**Trinocular stereo evaluation.** The prediction error technique of [15] for stereo sequences requires at least three different images of the same scene (from different perspectives at the same time instance). The objective is to generate a *virtual image*  $V$  from the output of a stereo matching algorithm, and to compare this with an image recorded by an additional *control camera*, that was not used to generate the disparity map. We generate the virtual image by mapping (warping) each pixel of the reference image into the position in which it would be located in the *control image*  $N$  (image recorded with the control camera). Then,  $N$  and  $V$  are compared by calculating the *normalized cross correlation* (NCC) index between them as follows:

$$NCC(N, V) = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \frac{[N(i, j) - \mu_N][V(i, j) - \mu_V]}{\sigma_N \sigma_V} \quad (10)$$

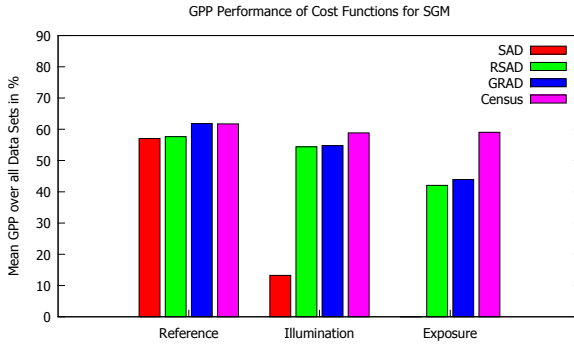
where  $\mu_N$  and  $\mu_V$  denote the means, and  $\sigma_N$  and  $\sigma_V$  the standard deviations of the control and virtual images, respectively. The domain  $\Omega$  is only for non-occluded pixels (i.e., pixels visible in the three images).

## 5 Results and Discussion

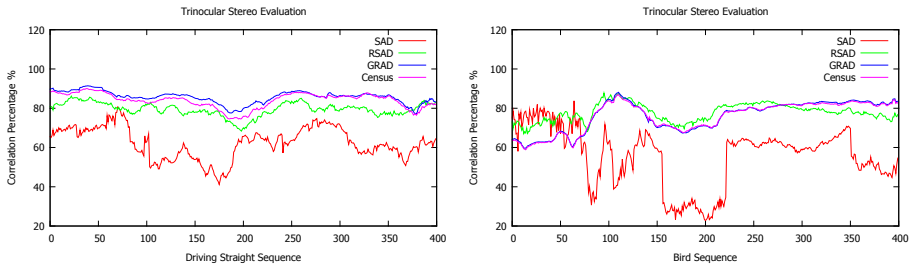
Figure 3 shows the mean GPP over the evaluated engineered test data for different illumination settings applied to SGM. From this evaluation, all cost functions seem to perform equally well, except for the pure SAD function, which is not surprising because the intensity consistency assumption is violated. It appears though that census is slightly more robust especially when looking at exposure changes, as the mean GPP does not change significantly when looking at different illumination settings. Conversely, RSAD and gradient both seem to be robust to illumination change, but not as much to exposure change.

Figure 4 shows the NCC percentage for all 400 frames of both real-world driving sequences. The overall performance on the driving straight (left) sequence is better than in the bird sequence (right). This may be explained because of the higher illumination variance between stereo frames in the bird sequence. We see that the overall quality of all cost functions is lower when illumination differences are strong; this is seen when we compare the driving straight (low changes) with the bird (high changes) sequence.

The gradient seems to outperform the census cost function for all but a few frames when looking at the driving straight scene (left). This may be due because illumination differences are not that strong. Otherwise performance appears to be almost identical.



**Fig. 3.** Results for ground truth evaluation on engineered test data. The GPP is a mean value over all six datasets evaluated.



**Fig. 4.** Trinocular prediction error NCC analysis plots for the real-world data set. Left: Bird sequence. Right: Driving straight sequence.

However, all curves roughly seem to follow the same pattern (this is discussed later in this section), except for the pure SAD cost function.

The major difference is the RSAD cost function. While performance is consistently lower than for the gradient and the census function in the driving straight scene (left), it seem to respond slightly differently to the data in the bird sequence (right).

However, all cost functions seem to perform equally well (again except for the standard SAD). This may not be surprising because all of them respond to relative intensity jumps in the underlying image data. The left  $3 \times 3$  window in Figure 5 shows a sample of a grey scale intensity image. The next window to the right shows the census transform when we choose 1 if the intensity increases from the centre pixel, and 0 otherwise. We gain from this transformation the signature vector  $(1, 0, 1, 1, 0, 1)$  when starting from the top left corner, and cycling clockwise. However, if we compute forward differences in all eight directions of the 8-neighbourhood of the central pixel (look at the window labelled gradient) and write down the results in a vector (starting top left and cycling clock-wise), we get  $(23, -41, 60, 47, 35, -10, 12, -31)$ . If we just look at the signs and represent a positive value as a 1 and a negative value as 0, the resulting signature vector is identical to the census signature vector. This makes a close relation between

Intensity		
177	113	214
123	154	201
166	144	189

Census		
1	0	1
0	X	1
1	0	1

Gradient		
23	-41	60
-31	X	47
12	-10	35

Residual		
12	-52	49
-42	X	36
1	-21	24

**Fig. 5.** From left to right: A  $3 \times 3$  window of a intensity image. Followed by the corresponding census transformation. This is followed by forward differences when computed in all directions of a 8-neighbourhood. Finally the zero-mean calculation.

derivative-based (or gradient-based) and the census-based data descriptors which are employed for cost functions.

We can also compute the mean of this window (which is 165) and subtract it from the intensity of each neighbour we perform the zero-mean transformation. This is closely related to the residual computation we applied for the SAD cost function used in this paper. The resulting vector is the vector from the gradient shifted by an offset of 11 and would be identical if the mean happened to be 154.

This analysis shows that all three cost functions are related. All fit into a first order data term category, as each of those functions represent a relative intensity change in the image. But this is nothing else than the derivative in a distinctive direction; and this is closely related to edge detection.

## 6 Conclusions

This paper shows that the performance of a gradient based cost function competes with the performance of cost functions already identified as being robust to illumination changes. A potential relation between the categories of those cost functions is established. One conclusion of this analysis could be that finding a good and robust cost function for real world applications reduces to the problem of finding a cost function that describes intensity changes appropriately w.r.t the underlying data. All of the illumination robust cost functions seem, in effect, to be related to the gradient. The census function describes the gradient in a rough sense, which makes it robust to noise. The gradient and RSAD function adds intensity information on a relative scale to the cost, which adds more descriptive information to the cost. However, this makes those functions more affected by noise than the census. This may explain the better performance of census on the engineered data as noise has a bigger influence when exposure is changed.

## References

1. Aujol, J.-F., Gilboa, G., Chan, T., Osher, S.: Structure-texture image decomposition – modeling, algorithms, and parameter selection. *Int. J. Computer Vision* 67, 111–136 (2006)
2. Barnard, S.T., Fischler, M.A.: Computational stereo. *ACM Computing Surveys* 14, 553–572 (1982)

3. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis Machine Intelligence* 23, 1222–1239 (2001)
4. Brox, T., Bruhn, A., Papenber, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Pajdla, T., Matas, J.(G.) (eds.) *ECCV 2004*. LNCS, vol. 3024, pp. 25–36. Springer, Heidelberg (2004)
5. .enpeda. image sequences analysis test site, <http://www.mi.auckland.ac.nz/EISATS>
6. El-Mahassani, E.D.: New robust matching cost functions for stereo vision. In: *Proc. DICTA*, pp. 144–150 (2007)
7. Felzenszwalb, P.F., Huttenlocher, D.: Efficient belief propagation for early vision. *Int. J. Computer Vision* 70, 41–54 (2006)
8. Gehrig, S.K., Eberli, F., Meyer, T.: A real-time low-power stereo vision engine using semi-global matching. In: *Proc. ICCV*, pp. 134–143 (2009)
9. Haeusler, R., Klette, R.: Benchmarking stereo data (Not the matching algorithms). In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) *Pattern Recognition*. LNCS, vol. 6376, pp. 383–392. Springer, Heidelberg (2010)
10. Hirschmüller, H.: Accurate and efficient stereo processing by semi-global matching and mutual information. In: *Proc. CVPR*, vol. 2, pp. 807–814 (2005)
11. Hirschmüller, H., Scharstein, D.: Evaluation of cost functions for stereo matching. In: *Proc. CVPR*, pp. 1–8 (2007)
12. Hirschmüller, H., Scharstein, D.: Evaluation of stereo matching costs on images with radiometric differences. *IEEE Trans. Pattern Analysis Machine Intelligence* 31, 1582–1599 (2009)
13. Klaus, A., Sormann, M., Karner, K.: Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In: *Proc. CVPR*, vol. 3, pp. 15–18 (2006)
14. Middlebury College, stereo vision page, <http://vision.middlebury.edu/stereo/>
15. Morales, S., Vaudrey, T., Klette, R.: A third eye for performance evaluation in stereo sequence analysis. In: Jiang, X., Petkov, N. (eds.) *CAIP 2009*. LNCS, vol. 5702, pp. 1078–1086. Springer, Heidelberg (2009)
16. Morales, S., Woo, Y.W., Klette, R., Vaudrey, T.: A study on stereo and motion data accuracy for a moving platform. In: Kim, J.-H., Ge, S.S., Vadakkepat, P., Jesse, N., Al Manum, A., Puthusserypady, S.K., Rückert, U., Sitte, J., Witkowski, U., Nakatsu, R., Braunl, T., Baltes, J., Anderson, J., Wong, C.-C., Verner, I., Ahlgren, D. (eds.) *Advances in Robotics*. LNCS, vol. 5744, pp. 292–300. Springer, Heidelberg (2009)
17. Ohta, Y., Kanade, T.: Stereo by two-level dynamic programming. In: *Proc. Int. Joint Conf. Artificial Intelligence*, vol. 2, pp. 1120–1126 (1985)
18. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: *Proc. ICCV*, pp. 7–42 (2002)
19. Szeliski, R.: Prediction error as a quality metric for motion and stereo. In: *Proc. ICCV*, pp. 781–788 (1999)
20. Vaudrey, T., Klette, R.: Residual images remove illumination artifacts! In: Denzler, J., Notni, G., Süße, H. (eds.) *Pattern Recognition*. LNCS, vol. 5748, pp. 472–481. Springer, Heidelberg (2009)
21. Vaudrey, T., Wedel, A., Klette, R.: A methodology for evaluating illumination artifact removal for corresponding images. In: Jiang, X., Petkov, N. (eds.) *CAIP 2009*. LNCS, vol. 5702, pp. 1113–1121. Springer, Heidelberg (2009)
22. Zabih, R., Woodfill, J.: Non-parametric local transform for computing visual correspondence. In: *Proc. ECCV*, vol. 2, pp. 151–158 (1994)
23. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) *DAGM 2007*. LNCS, vol. 4713, pp. 214–223. Springer, Heidelberg (2007)

# Relative Pose Estimation for Planetary Entry Descent Landing

Luca Zini<sup>1</sup>, Francesca Odone<sup>1</sup>, Alessandro Verri<sup>1</sup>,  
Piergiorgio Lanza<sup>2</sup>, and Alessandra Marcer<sup>2</sup>

<sup>1</sup> DISI - Università degli Studi di Genova, Italy

<sup>2</sup> Thales Alenia Space Italia S.p.A., Torino, Italy

**Abstract.** The paper is about the estimation of the relative position of a spacecraft, during the Entry Descent Landing (EDL) phase, by means of computer vision. A camera installed on board of the vehicle acquires images that are used for estimating the relative position of the camera between two consecutive images. A crucial point of the analysis, and the main objective of this work, is the estimation of the fundamental matrix  $F$ , considering the fact that in most cases we deal with a quasi-degenerate configuration. Indeed, the distance between the spacecraft (and the camera) and the planet surface, together with the morphology of the ground, make the problem difficult since most of the points will be extracted from a dominating plane. We discuss two different ways of addressing such degeneracy, while keeping the computational cost low, and present very promising results on synthetic as well as real image sequences.

## 1 Introduction

A common request for the future scientific missions is the exploitation of a precise landing approach. The Viking lander (1976), Mars PathFinder (1997), Mars Polar Lander(1999) and Mars Exploration Rovers(2003) landing ellipse was of the order of 100-300 km long. The main aim today is to improve this precision in order to achieve, within few years, ellipse landing of hundreds of meters. Because of long distance between the Lander and the Earth is not possible to teleoperate the Lander during the Entry Descent Landing (EDL) phase. Hence the Lander shall be in charge to choose autonomously the final point of landing avoiding any hazardous landing areas. Another further requirement is the acquisition of images during the mission landing phase and their transmission to the Earth Ground Station. On the basis of above consideration is quite clear that a vision system could constitute a new approach in charge to satisfy the above mentioned requirements. In particular the EDL vision based approaches are different from mission to mission because they are strictly depending on the environmental features of the mission. For instance the main environmental constraints are given by:

- Mars atmosphere: this problem is mainly due to the Martian wind and the dusty atmosphere in the low altitude to guide the Lander during the final approach.



- Moon illumination: the main problem is relevant to the different light conditions.
- Asteroid shape: the unknown terrain morphology constitutes the most critical problems during the landing

A historical milestone in the EDL vision based approach is constituted by the success of Mars Exploration Rovers (MER) landings based on the Descent Image Motion Estimation System (DIMES) by NASA [2]. DIMES constitutes the first use of computer vision approach to control a spacecraft during the planetary landing. DIMES was based on a space qualified camera, an IMU and a radar altimeter to evaluate the distance between Lander and terrain [2,3]. On the basis of this first positive feedback many studies have been carried out from NASA. In particular [9,4] define a better approach with respect to DIMES in terms of slope estimation and hazard detection. In Europe some studies have been carried out financed by ESA contracts: the Autonomous On Board Navigation for Interplanetary Missions (AutoNav) and then the Navigation for Planetary Approach and Landing (NPAL) [6]. In particular Thales Alenia Space (henceforth TAS-I) is involved as prime contractor in two studies: the Vision Aided Inertial Navigation (VISNAV) and in particular two recent studies have been commissioned by ESA: the Vision Aided Inertial Navigation (VISNAV) and Scalable EDL GNC & Avionics System Demonstrator (SAGE) study where a Scalable EDL is realized.

In the literature two different conceptual approaches on EDL Vision based are pursued: the first one is based on feature tracking integrated into the GNC Extended Kalman filtering; the second one is based on a separate Guide Navigation Control (GNC) where the camera instrumentation furnishes the inputs on the basis of image processing. The work presented here falls in the second approach, which is more appealing from industrial point of view because relies on well consolidate approach of GNC based on Inertial Mass Unit (IMU) and star sensors instrument. In this way the camera can be considered as further instrument in the GNC chain.

Following this approach at each time instant we process current video frame and compute sparse correspondences with a previous frame. Such correspondences are used to estimate the relative geometry between the two views. In this paper we discuss how to compute the camera 3D motion by estimating the fundamental matrix  $F$  of the epipolar system relating two consecutive frames acquired with the same camera. From the computer vision stand point the problem under analysis is rather standard. At the same time the specific applications setting poses many challenges, that will be addressed throughout the paper. As a first thing, considering the harsh environmental conditions data will be noisy and rich of outliers; also, given the relative distance between the camera and the observed scenario, the latter will always appear as a quasi-planar surface. This leads to a well-known degenerate configuration. Finally, in space applications, relying on hard radiation tolerant computers, the available computational power is very low with respect to today's PCs. Thus a reduced computational complexity is another crucial issue.

To deal with such problems the computer vision literature proposes many different approaches: for what concerns how to limit the effect of outliers most algorithms are variations of the popular RANSAC [7]. It is worth mentioning projection based M-estimators (PBM) [1] and MAPSAC [12]. A specific reference to degenerate configurations is done in QDegSac, a method designed to deal with degeneracy in a wide class of geometric problems [10]. All these methods are known to be accurate but are computationally expensive.

Our work is based on exploring simple techniques well known in the computer vision community [8] while exploiting at best all prior information available from the application under consideration. The first method (henceforth referred to as *translational model*) we analyse breaks the degeneracy of the quasi-planar case by adding a constraint on the camera motion, assuming that it is translational. This hypothesis is reasonable in our case, since the rotation component between consecutive views is usually very small. The second method, instead, is a *plane plus parallax model* and exploits the fact that most (but not all) observed points lie on a planar surface. From the numerical standpoint, both methods are very simple and thus computationally efficient. We present a detailed experimental analysis that shows how the two proposed methods outperform the MAPSAC approach. The experiments are based on two sets of data, a real set acquired by means of the EDL laboratory installed on TAS-I premises, and a synthetic set generated by the Pangu ESA software [11]. Different trajectory types have been taken into consideration, following the experiments described in [14]. The results show the translational model is very efficient, but the price we pay for the additional constraint is that we cannot estimate the spacecraft attitude. Instead the plane plus parallax model appears to be a very good compromise between effectiveness and efficiency.

## 2 Fundamental Matrix Computation for Quasi-Planar Surfaces

In this section we briefly review the degenerate configuration for the estimation of the epipolar geometry caused by a planar scene [8] and then describe the two models evaluated for the specific application setting.

### 2.1 The Degeneracy of Planar Surfaces

Let us first set the notation and consider a calibrated stereo system whose origin corresponds to the first camera. The projection matrices are

$$\begin{aligned} M &= K[I|\mathbf{0}] \\ M' &= K'[R|\mathbf{t}] \end{aligned}$$

where  $K$  and  $K'$  are the intrinsic parameters of the two cameras and  $R$  and  $\mathbf{t}$  are the rotation matrix and the translation vector relating the two views. A point  $\mathbf{P}$  of the 3-D world is projected into  $\mathbf{p} = M\mathbf{P}$  and  $\mathbf{p}' = M'\mathbf{P}$  respectively. The fundamental matrix  $F$  carrying information on the geometry of the two views satisfies the following equation

$$\mathbf{p}'^\top F \mathbf{p} = 0. \tag{1}$$

In this general case it can be seen that the fundamental matrix  $F$  may be written as

$$F = [\mathbf{e}']_{\times} K' R K^{-1} \tag{2}$$

where  $\mathbf{e}'$  is the epipole of the second camera.

If the 3-D scene is a plane we have a degenerate configuration of the epipolar geometry. Indeed, the two views are related by a homography, that is, for all pairs  $(\mathbf{p}, \mathbf{p}')$

$$\mathbf{p}' = H \mathbf{p} \tag{3}$$

Putting (3) and (2) together we obtain  $\mathbf{p}'^\top F H^{-1} \mathbf{p}' = 0$  that it is true for all skew-symmetric matrices  $F H^{-1}$  and does not depends on the points set. Thus the solution for  $F$  is any matrix  $F = S H$  where  $S$  is skew-symmetric. Since  $S$  has 2 degrees of freedom, considering the scaling factor of the projective transformation, the solution we obtain is a two parameters family of homogeneous matrices.

This degeneracy holds if all points are exactly on a planar surface and the camera is undergoing a general motion. In the remainder of the section we will see how to deal with such degeneracy and recover the geometry of two consecutive views in the application environment we are considering.

## 2.2 Pure Translation Model

A pure translational motion seems to be appropriate for the application under analysis since the rotation component is usually very small and could be modeled as noise of the system. In this section we discuss the fact that, if the camera motion is a pure translation, a planar scene does not cause any degeneracy.

Let us start by observing that, with a pure translation motion and assuming that intrinsic parameters do not change, Eq. (2) can be written as  $F = [\mathbf{e}']_{\times}$ . This shows immediately that in this case  $F$  is always skew-symmetric and a minimal solution to the problem can be obtained by 2 points correspondences. Considering that two 3-D points are always on a plane, it is clear that with this particular type of motion a planar surface does not represent any special case. Algorithm 1 reports a way to address this special case, based on setting explicitly the skew-symmetry of matrix  $F$ .

---

### Algorithm 1. Pure translation model

---

- 1: **input:**  $n$  point correspondences  $n \geq 2$
- 2: Construct the system  $\mathbf{p} F \mathbf{p}' = 0$  with  $F$  a skew-symmetric matrix. Let  $A$  be the  $n \times 3$  coefficients matrix, then each point correspondence  $i$  sets a row of  $A$ :

$$[(p_3)^i (p_2')^i - (p_2)^i (p_3')^i, (p_1)^i (p_3')^i - (p_3)^i (p_1')^i, (p_2)^i (p_1')^i - (p_1)^i (p_2')^i]$$

- 3: Use RANSAC or MSAC [13] to solve of a homogeneous system  $A \mathbf{f} = 0$
  - 4: **output:** Matrix  $F$  built from  $\mathbf{f}$  - rank 2 is imposed by construction.
-

Data normalization, useful to deal with numerical instabilities, should be performed with care in this case. Usually it is performed by centering and scaling points with respect to each image plane:  $\hat{\mathbf{p}} = T\mathbf{p}$  and  $\hat{\mathbf{p}}' = T'\mathbf{p}'$  respectively. The fundamental matrix  $\hat{F}$  is computed with respect to the normalized data and then denormalized  $F = T'^T \hat{F} T$ . This transformation breaks the skew-symmetric structure of the matrix that is responsible of the unicity of the solution. Thus we suggest to normalize w.r.t. a global transformation  $T$  computed over all points from both image planes. In this case both  $F$  and  $\hat{F} = T^{-T} \hat{F} T^{-1}$  are skew-symmetric.

### 2.3 Plane Plus Parallax Model

An alternative way to exploit the prior information available on the application environment is to base our estimation on the assumption that most (but not all) points in the scene lie on a planar surface. This assumption holds in our setting, considering the high distance between camera and planetary surface. This allows us to use the virtual parallax induced by the plane to estimate  $F$ . The method we rely on was originally proposed by [5] (see also [8]).

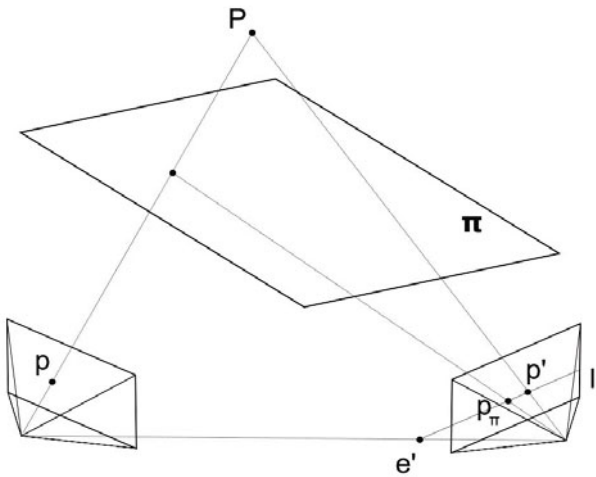


Fig. 1. The plane plus parallax geometry

The geometry we consider is depicted in Fig 1. Point  $\mathbf{P}$  (that does not lie on plane  $\Pi$ ) is projected on points  $\mathbf{p}$  and  $\mathbf{p}'$  respectively.  $\mathbf{p}_\pi = H\mathbf{p}$  is the mapping of point  $\mathbf{p}$  with respect to the homography on the second camera image plane.  $\mathbf{p}'$  and  $H\mathbf{p}$  lie on the same epipolar line,  $l'$ .

From a simple reasoning we may derive the relationship between  $F$  and the homography  $H$  induced by a plane. We consider the epipolar line of the second

camera and recall that it can be written as  $\mathbf{l}' = F\mathbf{p}$ . Also, since this line passes through the epipole  $\mathbf{e}'$ ,  $\mathbf{p}_\pi$  and  $\mathbf{p}'$  we may also write it as

$$\mathbf{l}' = \mathbf{e}' \times \mathbf{p}_\pi = [\mathbf{e}']_{\times} \mathbf{p}_\pi = [\mathbf{e}']_{\times} H\mathbf{p},$$

thus

$$F = [\mathbf{e}']_{\times} H. \quad (4)$$

Thus, since two 3-D points outside plane  $\Pi$  are enough to compute the position of the epipole  $\mathbf{e}'$  while with at least 4 coplanar points one may estimate  $H$ , we obtain a simple algorithm for the estimation of the fundamental matrix from 6 points only. Algorithm 2 summarizes the procedure.

---

**Algorithm 2.** Plane plus parallax model
 

---

- 1: **input:**  $n \geq 6$  points correspondences most on a plane – at least 4 on a plane and 2 outside the plane
  - 2: Construct the system from eq.  $\mathbf{p}'_i = H\mathbf{p}_i$ ,  $i = 1, \dots, n$ .
  - 3: Use RANSAC or MSAC to solve the homogeneous system  $A\mathbf{h} = \mathbf{0}$  - the entries of  $\mathbf{h}$  form  $H$
  - 4: Compute the set of inliers ( $\{\mathbf{p}\}_{in}, \{\mathbf{p}'\}_{in}$ ) and outliers ( $\{\mathbf{p}\}_{out}, \{\mathbf{p}'\}_{out}$ ) w.r.t.  $H$
  - 5: Construct the system  $\mathbf{p}'^\top [\mathbf{e}']_{\times} H\mathbf{p} = 0$  where  $\mathbf{p} \in \{\mathbf{p}\}_{out}$ . Let  $B$  be the  $n_{out} \times 3$  coefficients matrix
  - 6: Use RANSAC or MSAC so solve the system  $B\mathbf{e}' = \mathbf{0}$ .
  - 7: **output:** the matrix  $F$  computed as  $[\mathbf{e}']_{\times} H$
- 

We conclude by observing this algorithm can be seen as a particular case of QDegSAC [10] (or, otherwise, the latter is a generalization of the approach) but, since we are dealing directly with a specific case, the procedure we adopt is simpler and with a lower computational cost.

### 3 Comparative Analysis and Discussion

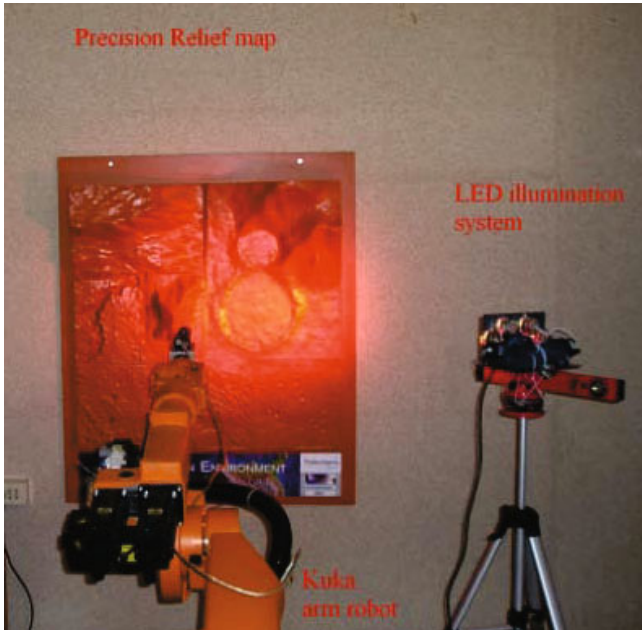
A preliminary comparative analysis among various methods from the literature (including 8 points with least squares estimation, with RANSAC [8], MSAC, Mlesac [13], MAPSAC [12], projection based M-estimators [1]) showed that MAPSAC (with a final non linear optimization) gave the best results, although its computational cost is not appropriate for the application under analysis. Thus this section aims at comparing its performance with the two methods proposed in Section 2.

#### 3.1 The Data

The experiments have been carried out on different image sequences. We have examined three different types of trajectories: vertical, polynomial of 4th order

---

<sup>1</sup> To the purpose of these experiments we use the implementation of MAPSAC available for download at CVonline [http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL\\_COPIES/TORR1/torrsam.zip](http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/TORR1/torrsam.zip)



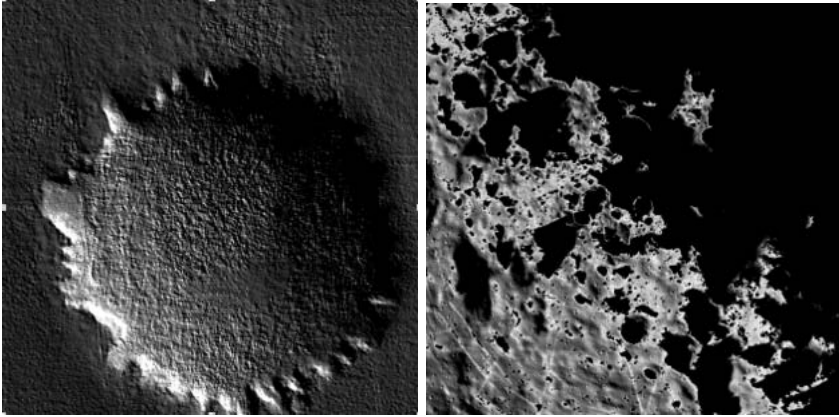
**Fig. 2.** TAS-I EDL laboratory. It is possible to see the Mars relief terrain, the Kuka robotic arm in the center and on the right the led illumination system.

with a limited curvature, polynomial of 4th order with enlarged curvature. These trajectories have been based on consideration and results presented in [14] and they have been thought and optimized on Mars landing approach. We consider two types of image sequences:

- Real image sequences acquired in EDL Laboratory, described below.
- Synthetic image sequences produced by means of Pangu ESA software [11].

The EDL laboratory on TAS-I premises (see Figure 2) is equipped with a Kuka robotic arm and a Mars relief terrain and led illumination equipment. The camera, a Marlin F-131 is mounted on the top of robotic arm. The arm moves with a precision lower than 1 mm along the three axes. The relief map has been built with a 3D printer and is made of nine panels of 300 x 300 mm each with a max height of 150 mm. Its precision is lower than 0.1 mm. The Pangu ESA software [11] processes Digital Elevation Map inputs, to visualize the terrain from different points of views and also to add boulders or craters realized by means of particular fractals functions. The reported experiments are based on two scenarios: a Crater Victoria DEM for Mars scenario and a South Pole DEM for the Moon scenario (sample frames are shown in Figure 3). Overall 20784 pairs of images have been tested for camera pose estimation which correspond to 7 real image sequences and 9 synthetic sequences by means of Pangu.

The validation process includes a feature extraction phase (for current experiments we adopt SURF features, a feature matching phase, and a postprocessing



**Fig. 3.** Sample images generated with Pangu ESA software. On the left Victoria Crater (Mars) on the right South Pole (Moon). Both scenarios are characterized by a sidelight illumination of 5 deg along the horizon.

that deletes spurious matches. Then for each image pair the fundamental matrix may be estimated. The latter constitutes the core of the experiments reported in the remainder of the section.

### 3.2 Results

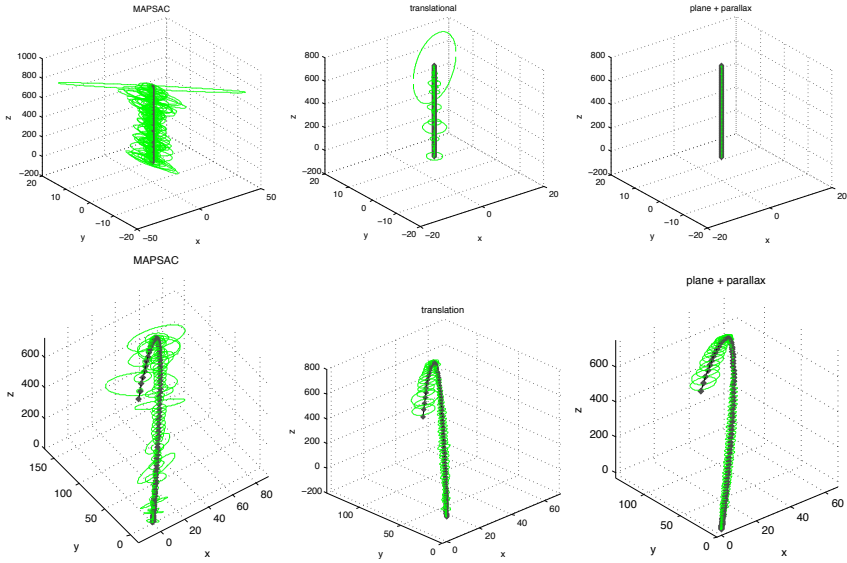
For each video sequence we consider consecutive frames and estimate their geometry. In Figure 4 we get a visual impression on how the three methods perform in the case of a vertical trajectory (top) and of a polynomial trajectory (bottom). The true trajectory is displayed in dark gray, and the computed steps are visualized as green circles overlaid to the trajectory. Each circle's radius is the computed step magnitude, while its normal vector is the computed camera's optical axis. By interpolating all the circles we obtain a "tube" inside which the estimated trajectory lies. The two proposed methods compare favorably with respect to MAPSAC.

A quantitative analysis is based on the relative error computed against the available ground truth:

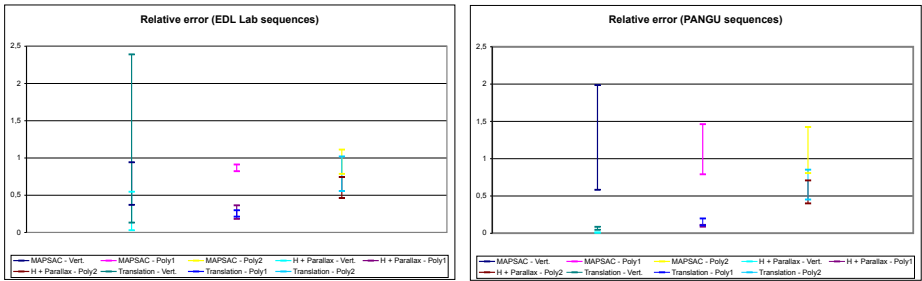
$$Error_i = \frac{\|TrueStep_i - ComputedStep_i\|_2}{\|TrueStep_i\|_2}.$$

Fig. 5 shows box plots of the three methods for each trajectory type. On the left are shown the results obtained with the EDL lab real video sequences; on the right the ones of the PANGU sequences. The plots clearly show that the plane plus parallax model consistently achieves a lower error and its solutions tend to be more stable. The translational model is also appropriate in almost all the cases.

We conclude by recalling that computer used for space missions must be hard radiation tolerant, therefore their computational performances are low with



**Fig. 4.** Estimated trajectories with respect to the ground truth (see text). Top: vertical trajectory; bottom: polynomial trajectory.



**Fig. 5.** Box plots showing the distribution of relative errors for the EDL lab real sequences (left) and the Pangu sequences (right). See text.

respect to commercial PCs. Thus it is very appreciated to control computational complexity avoiding any dynamic memory structure to increase the reliability of the software.

The computational advantage of employing one of the two proposed approaches is remarkable: an average estimate of the number of instructions executed by the three methods highlighted that, compared with the number of instructions executed by MAPSAC (without the final optimization step):

- the pure translation method executes *less than 1 %* instructions,
- the plane plus parallax *less than 6.2 %* instructions.



Thus, both methods are very appropriate for the task also from the computational standpoint. The plane plus parallax approach is preferred since it has a slightly superior performance and it also returns information on the spacecraft attitude, useful to the purpose of the mission.

## References

1. Chen, H., Meer, P.: Robust regression with projection based m-estimators. In: Proc. of IEEE ICCV, pp. 878–885 (2003)
2. Cheng, Y., Johnson, A., Matthies, L.: Mer-dimes: A planetary landing applications of computer vision. In: IEEE Proc. CVPR (2005)
3. Cheng, Y., Goguen, J., Johnson, A., Leget, C., Matthies, L., San Martin, M., Willson, R.: The mars exploration rovers descent image motion estimation system. IEEE Intelligent Systems 19(3), 13–21 (2004)
4. Cheng, Y., Johnson, A., Matthies, L., Wolf, A.: Passive imaging based hazard avoidance for spacecraft safe landing. In: Proc. iSAIRAS (2001)
5. Cross, G., Fitzgibbon, A.W., Zisserman, A.: Parallax geometry of smooth surfaces in multiple views. In: Proceedings of the 7th International Conference on Computer Vision, Kerkyra, Greece (1999)
6. Navigation for planetary approach a general approach and landing (May 2006)
7. 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(6), 381–395 (1981)
8. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press, Cambridge (2004), ISBN: 0521540518
9. Huertas, A., Cheng, Y., Madison, R.: Passive imaging based multicue hazard detection for safe spacecraft landing. In: Proc. IEEE/AIAA Aerosp. Conf., pp. 1–14 (2006)
10. Pollefeys, M., Frahm, J.-M.: Ransac for (quasi-)degenerate data (qdgsac). In: IEEE Proc. CVPR (2006)
11. Parkes, S., Martin, I., Dunstan, M.: Planet surface simulation with pangu. In: Eighth International Conference on Space Operations (2004)
12. Torr, P.H.S.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. IJCV 50(1), 35–61 (2002)
13. Torr, P.H.S., Zisserman, A.: MLESAC: A new robust estimator with application to estimating image geometry. Computer Vision and Image Understanding 78(1), 138–156 (2000)
14. Wong, E., Singh, G., Masciarelli, J.P.: Autonomous guidance and control design for hazard avoidance and safe landing on mars. In: AIAA Atmospheric Flight Mechanic's Conference (2002)

# AR Cultural Heritage Reconstruction Based on Feature Landmark Database Constructed by Using Omnidirectional Range Sensor

Takafumi Taketomi, Tomokazu Sato, and Naokazu Yokoya

Graduate School of Information Science, Nara Institute of Science and Technology  
8916-5 Takayama, Ikoma, Nara 630-0192, Japan  
{takafumi-t,tomoka-s,yokoya}@is.naist.jp

**Abstract.** This paper describes an application of augmented reality (AR) techniques to virtual cultural heritage reconstruction on the real sites of defunct constructs. To realize AR-based cultural heritage reconstruction, extrinsic camera parameter estimation is required for geometric registration of real and virtual worlds. To estimate extrinsic camera parameters, we use a pre-constructed feature landmark database of the target environment. Conventionally, a feature landmark database has been constructed in a large-scale environment using a structure-from-motion technique for omnidirectional image sequences. However, the accuracy of estimated camera parameters is insufficient for specific applications like AR-based cultural heritage reconstruction, which needs to overlay CG objects at the position close to the user's viewpoint. This is due to the difficulty in compensation of the appearance change of close landmarks only from the sparse 3-D information obtained by structure-from-motion. In this paper, visual patterns of landmarks are compensated for by considering local shapes obtained by omnidirectional range finder to find corresponding landmarks existing close to the user. By using these landmarks with local shapes, accurate geometric registration is achieved for AR sightseeing in historic sites.

## 1 Introduction

AR is a technique that enhances the real world by overlaying CG objects. In this study, AR techniques are used for virtual cultural heritage reconstruction on the real site of the defunct temple in ancient Japanese capital Asuka. By using our method, visitors can see virtually reconstructed buildings by CG at the original place as shown in Figure 1. To realize AR-based cultural heritage reconstruction, geometric registration of real and virtual worlds is required; that is, real and virtual world coordinates should be aligned. In the literatures, vision-based registration methods, which result in estimating extrinsic camera parameters, are extensively investigated [1,2,3,4,5,6] because they can achieve pixel-level geometric registration. These methods can be classified into the following two groups.

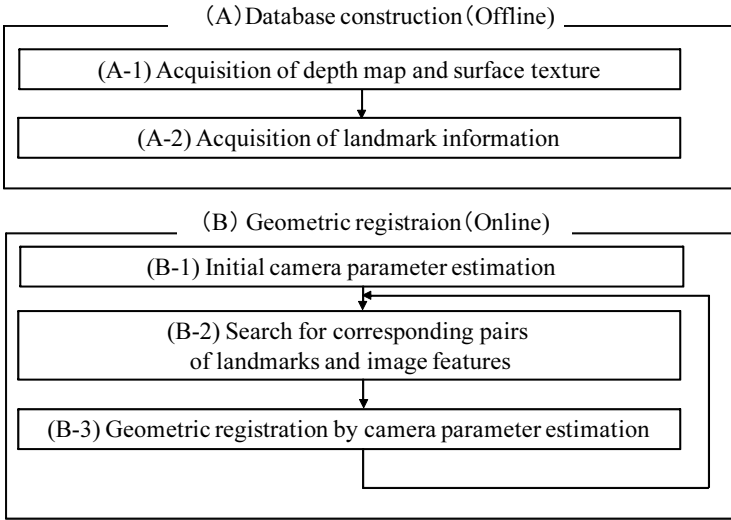


**Fig. 1.** AR sightseeing in historic site

One is a visual-SLAM based method [12] that estimates camera parameters without pre-knowledge of target environments. In this method, database construction and camera parameter estimation are carried out simultaneously. The problem of visual-SLAMs is that they only estimate relative camera motion. Thus, this approach cannot be used for position-dependent AR applications like navigation and landscape simulation.

The other uses some kinds of pre-knowledge of target environments such as 3-D models [3,4,5] and feature landmarks [6]. In this approach, camera parameters are estimated in the global coordinate system. However, a 3-D model based approach usually requires large human costs to construct 3-D models for large-scale environments. On the other hand, a feature landmark-based camera parameter estimation method [6] has been proposed. The method constructs a feature landmark database automatically by using structure-from-motion (SFM) in a large-scale environment. However, the accuracy of estimated camera parameters is insufficient for some kinds of AR applications like AR sightseeing where CG objects may be placed at the position close to the user's viewpoint as shown in Figure 1. This is due to the difficulty of matching feature landmarks that exist close to the user's position (close landmarks). In order to successfully compensate for the appearance change caused by the viewpoint change for close landmarks, sparse 3-D information by obtained the SFM process is not sufficient.

In this study, in order to improve the accuracy of vision-based geometric registration at the spot where CG objects of cultural heritage must be placed at the position close to the user, we newly compensate for visual patterns of landmarks using a dense depth map obtained by an omnidirectional laser range sensor. Figure 2 shows the flow diagram of the proposed vision-based registration method. The feature landmark-based geometric registration method is composed of two stages: the database construction stage in an offline process and the geometric registration stage in an online process. Although the framework of geometric registration method is basically the same as the method proposed



**Fig. 2.** Flow diagram of proposed vision-based registration

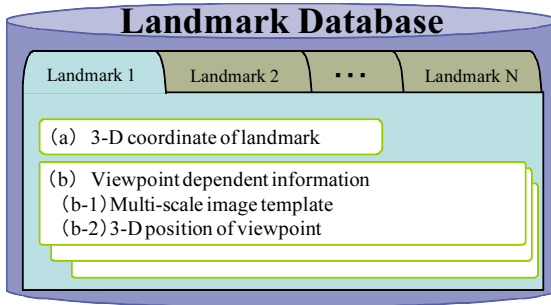
by Taketomi et al. [6], in our method, image templates of close landmarks are compensated by considering local 3-D structure around the landmark.

## 2 Landmark Database Construction Considering Local 3-D Structure

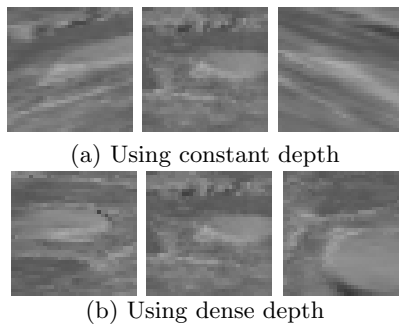
This section describes a feature landmark database construction process in the offline stage (A) in Figure 2. The feature landmark database must be constructed for the target environment before the online camera parameter estimation (B) is started for geometric registration. In this study, to compensate for image templates of landmarks, we use dense depth information obtained by the omnidirectional laser range sensor.

### 2.1 Acquisition of Depth Map and Surface Texture

Range data and texture are acquired using the omnidirectional laser range sensor and the omnidirectional camera in the target environment. In this scanning process, the geometrical relationship between these sensors is calibrated and fixed in advance. In the obtained depth map, some parts including the sky area cannot be measured by the laser range sensor. If we simply mask these unmeasurable areas in the pattern matching process, the aperture problem will easily be caused, especially for landmarks that exist at the boundary of the sky and landscape. It should be noted that such landmarks often become the key points for estimating the camera posture. To avoid the aperture problem, in the proposed method, infinite depth values are set to the sky area. Concretely, the largest region where depth values are not available in the omnidirectional image is determined as the sky area.



**Fig. 3.** Elements of landmark database



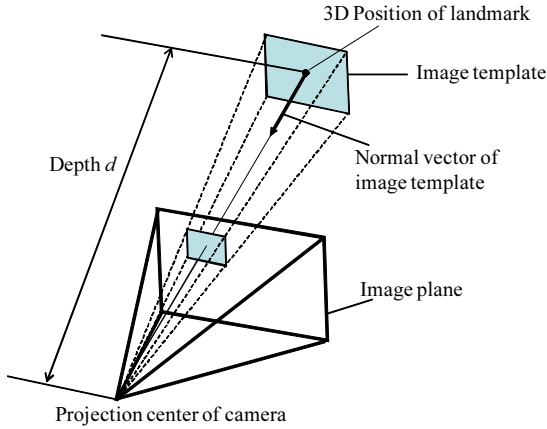
**Fig. 4.** Example of warped image patterns for some landmarks

## 2.2 Acquisition of Landmark Information

The feature landmark database consists of a number of landmarks as shown in Figure 3. Each landmark retains (a) a 3-D coordinate and (b) viewpoint dependent information.

**(a) 3-D Coordinate of Landmark:** 3-D positions of landmarks are used to estimate extrinsic camera parameters in the online stage (B). For all the feature points detected by using a Harris corner detector [7] from omnidirectional images, 3-D positions of feature points (a) are determined from the depth map obtained by the omnidirectional laser range sensor. These Harris corners are then registered as landmarks.

**(b) Viewpoint Dependent Information:** In this process, view dependent information is generated for every grid point that is placed on the ground plane around the sensor position. Figure 4(a) shows warped image patterns of close landmark that is stored in the database as image templates by using the SFM-based method [6]. In this figure, the second column shows the generated image pattern from the original viewpoint (position of the sensor). The first column and third column show warped image patterns where the viewpoints are set five meters to the right and forward from the original viewpoint, respectively. As can



**Fig. 5.** Generation of image template by conventional method

be seen in Figure 4(a), in the SFM-based database construction, warped images of landmarks that exist close to the user's position are largely distorted because image patterns are compensated with constant depth values  $d$  acquired for the landmark by SFM as shown in Figure 5.

In the proposed method, dense 3-D data obtained by the omnidirectional laser range sensor is used to correctly compensate for image templates of landmarks. Concretely, first, depth values  $d_i$  for each pixel  $i$  on the image template are obtained from range data. Next, pixel  $i$  values on the image template is determined by projecting an omnidirectional image using these depth values  $d_i$ . In this pattern generation, occluded areas in the image template are set as masked areas in order to ignore them in the pattern matching process. Figure 4(b) shows warped images obtained by using the proposed method. It can be observed that warped images are generated without large distortion by using dense 3-D information.

### 3 Geometric Registration: Extrinsic Camera Parameter Estimation Using Landmark Database

This section describes the camera parameter estimation stage in the online process (B) in Figure 2 for AR geometric registration. In this process, first, initial camera position and posture are estimated. Initial camera position and posture for the first frame of the input are assumed to be given by the landmark-based camera parameter estimation method for a still image input 8 (B-1). Next, search for corresponding pairs (B-2) and geometric registration (B-3) are repeated.

**Search for Corresponding Pairs:** In this process, corresponding pairs of landmarks and image features are searched for in the current frame. First, landmarks used to estimate camera parameters in the previous frame are selected

and tracked to the current frame. In the successive frames, visual aspects of landmarks hardly change. Thus, tracking of landmarks can be realized by a simple SSD (Sum of Squared Differences) based tracker. After landmark tracking, tentative camera parameters are determined using tracked landmarks. Image templates from the nearest viewpoint from the current camera position are then selected from the database. Finally, corresponding pairs between landmarks and image features are searched for using NCC (Normalized Cross-Correlation) with ignoring masked pixels.

**Geometric Registration by Camera Parameter Estimation:** After determining the corresponding pairs of landmarks and image features, extrinsic camera parameters are determined in the world coordinate system by solving the PnP problem [9] using these pairs. In order to remove outliers, the LMedS estimator [10] is employed in this process. After estimating extrinsic camera parameters, CG objects that are placed in the world coordinate system in advance are overlaid on the input image by using projection matrix computed by estimated camera parameters.

## 4 Experiments

To demonstrate the usefulness of the proposed method, first, the effectiveness of pattern compensation by considering local 3-D structure of the landmark is evaluated. Next, estimated camera parameters are compared with those by the SFM-based method [6].

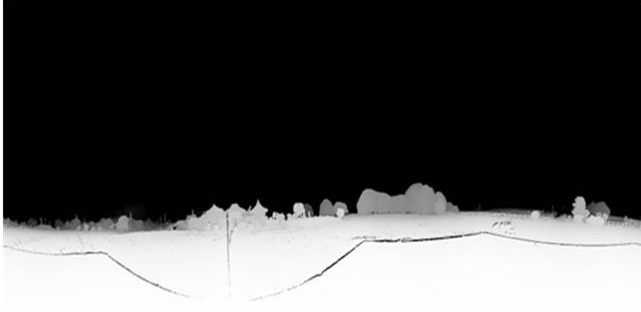
In this experiment, the landmark database is constructed for an outdoor environment using an omnidirectional multi-camera system (Point Grey Research Ladybug2) and an omnidirectional laser rangefinder (Riegl LMS-Z360). Figure 6 shows a panoramic image and corresponding depth map used for database construction. In this experiment, the ground plane of the target environment is divided into  $10 \times 10$  grid points at 1 meter intervals. To compare the accuracy of estimated camera parameters, the SFM-based feature landmark database is also constructed in the same place. For both methods, the same video image sequence ( $720 \times 480$  pixels, progressive scan, 15fps, 250 frames) captured in the target environment is used as the input video for the online camera parameter estimation. In this experiment, camera position and posture for the first frame are given manually.

### 4.1 Quantitative Evaluation of Pattern Compensation

Generated image templates of landmarks by the proposed and the previous methods are quantitatively evaluated by comparing to ground truth. To show the effectiveness of pattern compensation, compensated image templates of landmarks exemplified in Figure 4 are compared with image patterns of landmarks in input images. In this experiment, viewpoints for pattern compensation are given by estimating camera parameters with manually specified correspondences of landmarks in input images.



(a) Panoramic image taken by omnidirectional multi-camera system



(b) Depth map taken by omnidirectional laser range sensor

**Fig. 6.** Acquired omnidirectional data**Table 1.** Comparison of normalized cross-correlation value

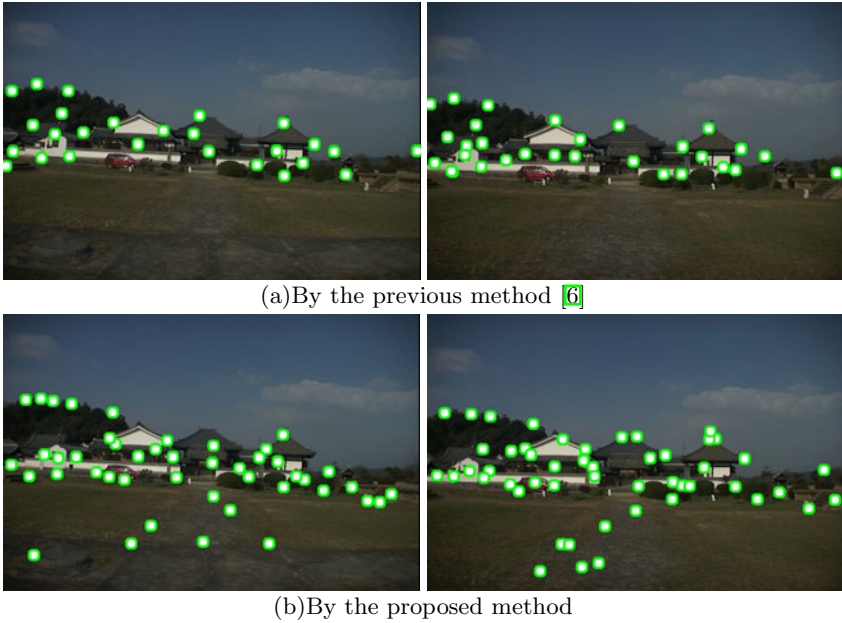
	Proposed method	Previous method [6]
Average	0.63	0.47
Standard deviation	0.039	0.052

Table 1 shows average and standard deviation of normalized cross-correlation values between compensated image templates and image patterns of landmarks in input images for 30 image templates of landmarks. In the proposed method which considers dense depth information, the average normalized cross-correlation value (0.63) is higher than that of the previous method (0.47) which does not consider the local 3-D structure around the landmark. From this result, we can confirm that compensated image templates are more similar to image patterns of landmarks in input images than that of the previous method.

## 4.2 Quantitative Evaluation of Estimated Camera Parameters

In the second experiment, the accuracy of estimated camera parameters is quantitatively evaluated and compared with the previous method [6]. Figure 7 shows

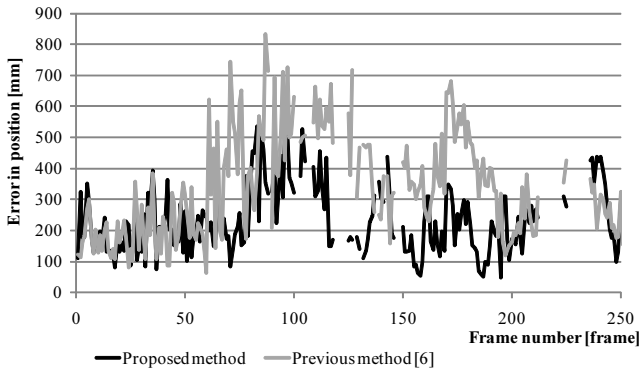


**Fig. 7.** Corresponded landmarks**Table 2.** Comparison of the accuracy

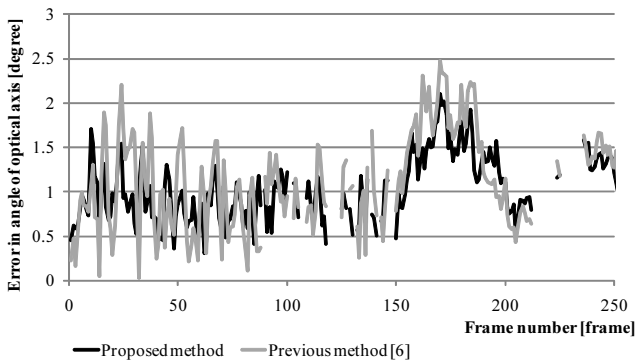
	Proposed method	Previous method [6]
Average of position error (mm)	231	342
Standard deviation of position error (mm)	107	164
Average of posture error (degree)	1.11	1.41
Standard deviation of posture error (degree)	0.52	0.46

landmarks in example frames that are used for camera parameter estimation. As shown in this figure, corresponding pairs of landmarks and feature points are successfully found for the ground part in the proposed method, while the previous method could not find any corresponding landmarks at the ground part of the images. This is regarded as the effect of appropriate pattern compensation using dense 3-D information.

Table 2 shows the accuracy of each method. To evaluate the accuracy of estimated camera parameters, we create the ground truth by estimating camera parameters with manually specified correspondences of landmarks. Note that we have removed several frames in which the reprojection error of the obtained ground truth is over 1.5 pixels. From this result, the accuracy of the proposed method has been proven to be improved than that of the previous method. Figures 8 and 9 illustrate errors in position and posture, respectively. In most frames, errors of the proposed method are the same or smaller than that of the previous method.



**Fig. 8.** Error in position for each frame



**Fig. 9.** Error in posture for each frame

Figure 10 shows examples of generated images using the proposed method for AR sightseeing in Asuka, Japan. Virtual objects are overlaid on the site of the old temple. We have confirmed that CG objects placed at the position close to the user's viewpoint are correctly registered.

## 5 Conclusion

In this paper, we have proposed a method to use dense depth information for landmark-based geometric registration for realizing AR sightseeing in the historic site. In this method, unlike other methods, the landmarks close to the user's viewpoint that effect the accuracy of geometric registration are aggressively used by compensating its visual patterns based on dense depth information acquired by using omni-directional range finder. Importance of close landmarks are validated quantitatively through the experiment. It should be noted that the proposed method is not for large-scale environments but for selected places where the accuracy of geometric registration largely depends on close landmarks. In

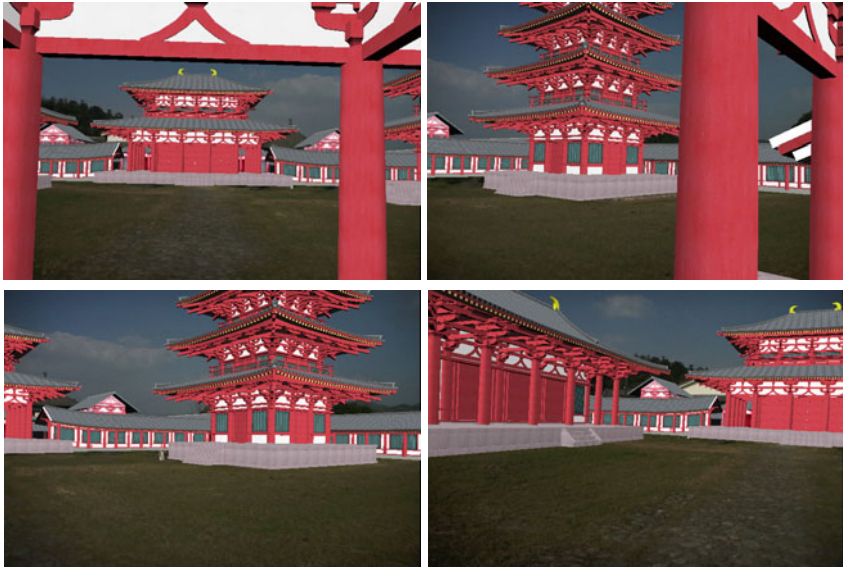


Fig. 10. User's views in AR sightseeing

future work, we will develop a method that uses both dense and sparse 3-D structures for efficiently constructing the database in large-scale outdoor environments.

## Acknowledgement

This research is supported in part by Core Research for Evolutional Science and Technology (CREST) of Japan Science and Technology Agency (JST), and the "Ambient Intelligence" project granted by the Ministry of Education, Culture, Sports, Science and Technology.

## References

1. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proc. Int. Symp. on Mixed and Augmented Reality, pp. 225–234 (2007)
2. Williams, B., Klein, G., Reid, I.: Real-time SLAM relocalisation. In: Proc. Int. Conf. on Computer Vision (2007)
3. Lepetit, V., Vacchetti, L., Thalmann, D., Fua, P.: Fully automated and stable registration for augmented reality applications. In: Proc. Int. Symp. on Mixed and Augmented Reality, pp. 93–102 (2003)
4. Vacchetti, L., Lepetit, V., Fua, P.: Combining edge and texture information for real-time accurate 3D camera tracking. In: Proc. Int. Symp. on Mixed and Augmented Reality, pp. 48–57 (2004)
5. Yang, G., Becker, J., Stewart, C.V.: Estimating the location of a camera with respect to a 3d model. In: Proc. Int. Conf. on 3-D Digital Imaging and Modeling, pp. 159–166 (2007)

6. Taketomi, T., Sato, T., Yokoya, N.: Real-time camera position and posture estimation using a feature landmark database with priorities. In: CD-ROM Proc. 19th IAPR Int. Conf. on Pattern Recognition (2008)
7. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. Alvey Vision Conf., pp. 147–151 (1988)
8. Susuki, M., Nakagawa, T., Sato, T., Yokoya, N.: Extrinsic camera parameter estimation from a still image based on feature landmark database. In: Proc. ACCV 2007 Satellite Workshop on Multi-dimensional and Multi-view Image Processing, pp. 124–129 (2007)
9. Klette, R., Schluns, K., Koschan, A. (eds.): Computer Vision: Three-dimensional Data from Image (1998)
10. Rousseeuw, P.J.: Least median of squares regression. *J. of the American Statistical Association* 79, 871–880 (1984)

# Augmented Reality-Based On-Site Tour Guide: A Study in Gyeongbokgung

Byung-Kuk Seo, Kangsoo Kim, and Jong-Il Park\*

Department of Electronics and Computer Engineering,  
Hanyang University, Seoul, Korea

{bkseo,vistavision}@mr.hanyang.ac.kr, jipark@hanyang.ac.kr

**Abstract.** This paper presents an on-site tour guide using augmented reality in which past life is virtually reproduced and visualized at cultural heritage sites. In the tour guide, animated 3-D virtual characters are superimposed on the cultural heritage sites by visually tracking simple geometric primitives of the sites such as rectangles and estimating camera poses (positions and orientations) that can be considered as a tourist's viewpoints. Contextual information, such as a tourist's locations and profiles, is used to support personalized tour guides. In particular, the tourist's locations are obtained by visually recognizing wooden tablets of the cultural heritage sites. The prototype of the augmented reality tour guide was tested at Gangnyeongjeon and Gytotaejeon in Gyeongbokgung, which is a symbolic cultural heritage site in Korea and its user evaluation is discussed.

## 1 Introduction

Various types of tour guides have been provided to tourists in cultural heritage sites. Booklets or tour maps are the most common and familiar type. Local tour guides who guide tour routes and orally explain historical information or give the background of cultural heritage sites are also popular. Multimedia tour guides have recently become attractive because they help tourists easily understand cultural heritage sites through audio or video contents.

Augmented reality (AR), which superimposes virtual information on real scenes, has provided good solutions for on-site tour guides. In contrast to the conventional types of tour guides, AR-based tour guides enable tourists to have intuitive and realistic experiences by overlaying virtual contents on cultural heritage sites. Many studies and research projects have been recently presented AR-based tour guides for cultural tourism [1]. For example, Papagiannakis *et al.* [2] developed an AR framework to revive life in ancient fresco paintings in ancient Pompeii and create narrative space. The revival was realized by superimposing 3-D virtual characters with body, speech, facial expression, and cloth simulation on the real environment. Augmented reality-based cultural heritage on-site guide (Archeoguide) [3], which is a system developed by its research project,

---

\* Corresponding author.

presented new ways to access information at cultural heritage sites. Archeoguide helps tourists navigate sites, visualizes AR reconstruction of ancient life, and offers user-friendly multimodal interaction. Moreover, Vlahakis *et al.* [4] showed various potentials for mobile AR tour guides by implementing it on different mobile units such as laptop, pen-tablet, and palmtop, and demonstrating it at Greece's Olympia archaeological site. Intelligent tourism and cultural information through ubiquitous services (iTacitus) [5] is another good example of AR-based tour guides. iTacitus overlays 3-D virtual models and multimedia contents, such as video and audio, on real scenes. Additionally, it offers context-awareness services based on a tourist's locations and interests, and supports an interactive itinerary planning tool to explore cultural heritage sites. The prototype of iTacitus was implemented on Ultra Mobile PCs and recently, smartphones. It was demonstrated at Reggia Venaria Reale in Italy and Winchester Castle's Great Hall in the UK [5,6]. The Ename 974 research project developed a spatially installed on-site AR system, called TimeScope 1 [7]. TimeScope 1 has operated at the archaeological park since 1997. It offers tourists a picture of ancient life at Ename by superimposing a 3-D model of the abbey church on its original site. Recently, Portalés *et al.* [8] applied an AR application to recreate former states of two features—a Baroque vault and a Renaissance reredos—above the high altar of Valencia Cathedral and reported its practical experiences with user tests.

In this paper, we present an on-site tour guide using AR in Gyeongbokgung, which is a representative cultural heritage site in Korea. Gyeongbokgung has already provided several tour guide services such as booklets, local guides, and portable audio devices. However, these services have mainly been used to offer and explain historical information of cultural heritage sites. In the proposed AR tour guide, intuitive and realistic experiences are provided to tourists by augmenting animated 3-D virtual characters, which reproduce past life on real sites. Historical information is also offered by narration synchronized with the 3-D virtual characters. Such on-site augmentation is performed by visually tracking simple geometric primitives such as rectangles, which are the bases of most man-made structures, without positional sensors or compasses. Contextual information such as a tourist's locations and profiles is used to support personalized tour guides. To obtain the tourist's locations, wooden tablets are visually recognized and it is accompanied by tourist's participation such as capturing the wooden tablets to get historical information. Finally, the prototype of the AR tour guide was tested and evaluated at Gangnyeongjeon and Gyotaejeon in Gyeongbokgung.

## 2 Methodology

### 2.1 Framework

The framework of the proposed AR tour guide consists of four parts: context-awareness, augmentation, and input/output agent. In the input agent, snapshot or live video images of target scenes are captured by a camera which is attached

on the AR tour guide. A tourist’s profiles are obtained by simply selecting graphic user interface (GUI) menus with a stylus pen or finger.

In the context-awareness part, a tourist’s locations are recognized by matching snapshot images of wooden tablets to their predefined reference images of a database. These are then sent to management agents: context management agent and map management agent. The context management agent organizes contextual information such as the tourist’s locations and profiles (age and language), and links to their corresponding information and content of each database. The map management agent defines location-related information, e.g. tourist’s locations, AR service zones, and tour paths, to display on the tour map of the AR tour guide.

Using live video images obtained by the input agent, in the augmentation part, natural scene information of cultural heritage sites is visually tracked and camera poses are estimated in real-time. Animated 3-D virtual characters are rendered on the real sites based on the estimated camera poses. The output agent displays information, contents, and a tour map on GUI windows of the AR tour guide. It also provides prerecorded narration synchronized to the rendered 3-D virtual characters through a speaker. The framework and dataflow of the AR tour guide are shown in Fig. 1. More details of the context-awareness and augmentation part are explained through the following subsections.

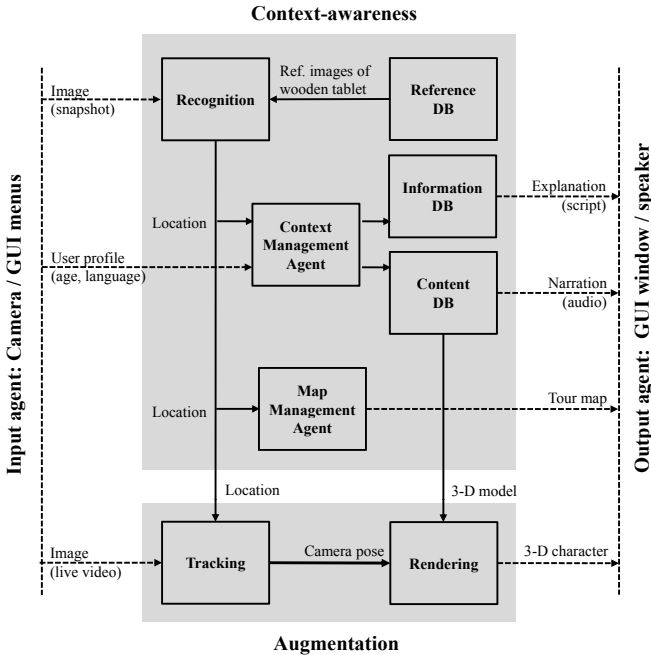


Fig. 1. Framework and dataflow of the AR tour guide

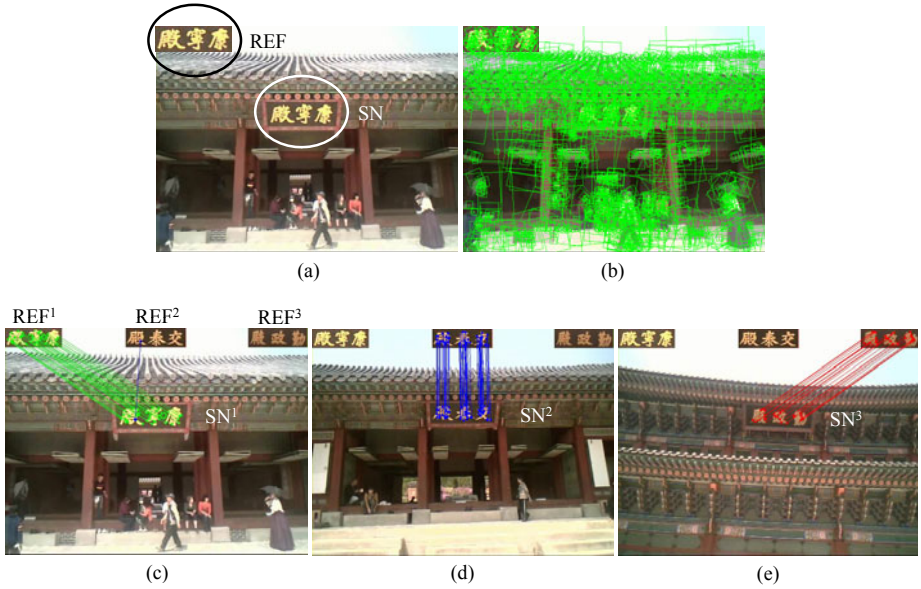
## 2.2 Context-Awareness

A variety of contextual information can be used for personalized tour guides on cultural heritage sites. The AR tour guide defines a tourist's location (where) and profile (who) as entities of the primary context. The tourist's location provides simple information, e.g., "Where am I?", and its relevant context, e.g., "What information or content is provided here?" or "Where are other AR service zones near here?" The tourist's profile significantly affects personal preference of the tour guide. In our approach, age and language are obtained for the tourist's profile and in particular, age plays an important role to decide the level or type of information and contents because most young people may require easily understood descriptions or contents to amuse them or stimulate their interests.

A tourist's profiles are obtained by simply selecting GUI menus in which short questions, e.g., "I am under 13" or "I prefer English language" are written. However, it is not easy to obtain the tourist's locations without sensor devices such as positional sensors and compasses. In the proposed AR tour guide, wooden tablets of cultural heritage sites are used as visual context cues. The tourist is guided to capture the wooden tablets using a camera. In Gyeongbokgung, all entrance doors and palaces or court buildings have wooden tablets on which their names are written. The wooden tablet's name identifies the heritage site and provides historical information, such as its meaning, history, or style of handwriting. For example, Gangnyeongjeon was a building used for the king's main sleeping and living quarters. Its wooden tablet was named "Gangnyeong", meaning health among five blessings: longevity, wealth, health, love of virtue, and peaceful death, and "Jeon", meaning a hall. Geunjeongjeon was the throne hall of Gyeongbokgung where the king granted audiences to his officials, and it was named "Geunjeong", meaning diligence helps governance. Therefore, the captured images of the wooden tablets are recognized by matching to their predefined reference images of a database, while the tourist participates in the tour by capturing the wooden tablets to obtain such historical information. After the tourist's locations are obtained by recognizing the wooden tablets, predefined AR service zones (target scene) are indicated on the GUI window of the AR tour guide and guided to the tourists with narration.

The recognition of wooden tablets is performed as follows. To match a snapshot image of a wooden tablet to its corresponding image, robust feature points are detected in the snapshot image. Then, their descriptors are computed using SIFT, which is a well-known descriptor [9] (see Fig. 2(b)). Finally, the descriptors are matched to descriptors of predefined reference images of a database using the k-nearest neighbor (KNN) algorithm. Here, the descriptors of the reference images are computed in the same way off-line in advance. Figure 2(c–e) show the matching results of feature points. Given three reference images (REF<sup>1</sup>, REF<sup>2</sup>, and REF<sup>3</sup> in the upper side of each figure), each snapshot image (SN<sup>1</sup>, SN<sup>2</sup>, and SN<sup>3</sup> in the bottom side of each figure) was correctly matched to its corresponding reference image. As some characters of the wooden tablets were the same, a few feature points of wrong reference images could be matched to the snapshot image (blue line in Fig. 2(c)) even though their shapes were slightly different.





**Fig. 2.** Recognition of wooden tablets: (a) snapshot image (SN) and reference image (REF), (b) descriptor computed from (a), (c–e) matching results against each snapshot image (REF<sup>1</sup>/SN<sup>1</sup>: Gangnyeongjeon, REF<sup>2</sup>/SN<sup>2</sup>: Gyotaejeon, REF<sup>3</sup>/SN<sup>3</sup>: Geunjeongjeon).

However, the number of correct matching was dominant (green line in Fig. 2(c)); thus, recognition was reliable.

Table 1 shows the number of matching when reference images were matched to different snapshot images. In the experiment, the snapshot images were captured by video sequences (resolution 640 by 480, 450 frames) where each reference image continuously appeared at different viewpoints, and matched to the three reference images. As shown in Table 1, the number of matching was much higher when each corresponding image was matched.

**Table 1.** Number of matching between reference images and snapshot images

Number of Matching (Std. Dev.)	Reference Image <sup>†</sup>		
	REF <sup>1</sup>	REF <sup>2</sup>	REF <sup>3</sup>
Snapshot Image <sup>‡</sup>	SN <sup>1</sup> 25.771(4.583)	0.329(0.553)	0.400(0.593)
	SN <sup>2</sup> 0.971(0.613)	35.567(6.127)	0.269(0.527)
	SN <sup>3</sup> 0.753(0.900)	0.638(0.812)	23.640(7.080)

<sup>†</sup>Each reference image is resolution 125 by 40.

<sup>‡</sup>Each snapshot image is resolution 640 by 480, 450 frames.

### 2.3 Augmentation

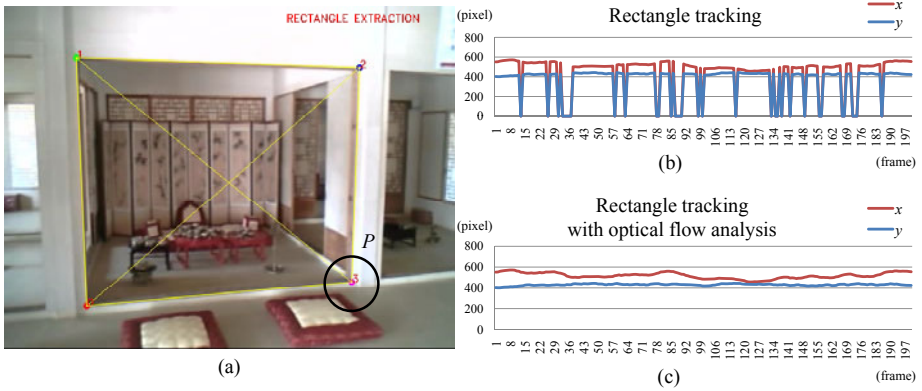
The augmentation part is divided into a tracking module and a rendering module. As a principal one, the tracking module localizes the camera relative to target scenes using the planar-based visual tracking method that is simple and robust. In this subsection, we explain how the tracking works in detail.

To achieve our goal, 3-D virtual models should be precisely superimposed and rendered on real scenes according to user's viewpoints. Generally, a camera pose (position and orientation) can be considered the user's viewpoints; it can be measured from positional sensors and compasses, or estimated by tracking visual information from real scenes. In our approach, the camera pose is estimated by visually tracking simple geometric primitives of target scenes without additional sensor devices. Gyeongbokgung has many palace and court buildings, and these buildings mainly consist of geometric primitives such as rectangles and line segments. Therefore, in the target sites—Gangnyeongjeon, which served as the king's living quarters and Gyotaejeon, which served as the queen's main residence, we track rectangles of their doorframes to estimate the camera pose in real-time. The details of the procedure are as follows:

1. Detect edges of a target scene using the Canny operator.
2. Find contours that can be candidates for a rectangle of a doorframe in the edges. Note that the edges are linked to minimize their discontinuity and find the contours reliably.
3. Extract the rectangle of the doorframe by searching the contours with some constraints: convexity, four corners, and the area of the rectangle. When this fails (the rectangle is not extracted), the four corners on the current image can be approximated from the previous ones if the camera's motion is not fast. Thus, we find the correspondences of the corners obtained from the previous image using optical flow analysis [10].
4. Estimate a camera pose using planar-based pose estimation [11] with the four corners of the extracted rectangle. Here, we assume that the four corners are coplanar.

With the estimated camera pose, 3-D virtual models are correctly augmented on the target scenes. Note that texture information inside the target scene is predefined in the reference database, and the doorframe of the target scene is distinguished from similar ones in adjacent scenes by the same process as the recognition of wooden tablets.

Figure 3(b,c) show  $x$  and  $y$  positions of the right-bottom corner ( $P$  point in Fig. 3(a)) of the extracted rectangle which were estimated during rectangle tracking (resolution 640 by 480, 200 frames). When the extraction of the rectangle failed, its corners were not found ( $x$  and  $y$  positions are zero in Fig. 3(b)), and it caused the augmented 3-D virtual characters to flicker. However, in



**Fig. 3.** Estimation of the camera pose (a) using the extracted rectangle of the doorframe. Comparison of (b) the rectangle tracking with (c) the rectangle tracking with optical flow analysis.

the tracking with optical flow analysis, both positions of the corners were approximated from previous ones and the tracking was good without discontinuity as shown in Fig. 3(c).

### 3 Demonstrations

The prototype of the AR tour guide was tested at Gangnyeongjeon and Gyotaejeon in Gyeongbokgung. As shown in Fig. 4, the prototype was implemented on a laptop (LG X-NOTE C1, Intel Core2 1.20 GHz) with a USB camera (MS LifeCam NX-6000, resolution 640 by 480, 15 fps). The GUI had four display windows: augmentation window, displaying cultural heritage sites and animated 3-D virtual characters that are superimposed on them; recognition window, displaying captured snapshot images of wooden tablets; tour map window, displaying a tourist’s locations and service zones; and information window, displaying explanation of cultural heritage sites. Additionally, there were several menus for executing its functions. They were easily activated by touching with a stylus or finger.

Figure 5 shows our demonstrations at Gangnyeongjeon and Gyotaejeon in Gyeongbokgung. As mentioned briefly above, Gangnyeongjeon and Gyotaejeon offer a glimpse into everyday life in the royal household. During summer, in particular, the palace buildings open for tourists so that they can enter the buildings and see exhibitions in the rooms. In both sites, our AR tour guide precisely tracked the doorframes of the rooms and successfully augmented the animated 3-D virtual characters (the king and the queen) on the real rooms. Historical information of the sites was also provided by narration synchronized to the characters.

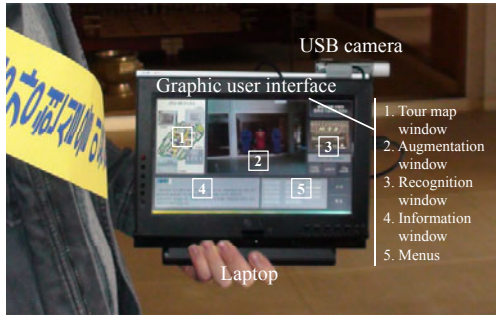


Fig. 4. Prototype of the AR tour guide



Fig. 5. Demonstrations of the proposed AR tour guide at (a) Gangnyeongjeon and (b) Gyotaejeon

## 4 Discussion

The proposed AR tour guide was tested and evaluated by seven tourist groups at Gangnyeongjeon. Each tourist group consisted of one or more tourists and in each group, randomly chosen tourists (8 males and 14 females, under 10 to over 50 years old, Korean and foreigner) answered five questions with a range of 1–5 points (higher score means good evaluation). The questionnaires and average scores are shown in Table 2. The results of the evaluations can be summarized in the following aspects of the proposed AR tour guide.

**Performance:** The performance of the AR tour guide mainly depends on the reliability of the tracking method. The participants gave above average ratings for question 1: “Did the AR tour guide work reliably?” (the average score was 3.28). Even though the tracking worked quite well, a few participants felt the augmented 3-D content was unstable when the device motion was fast or the lighting conditions of the target sites were changed (both sites were partially outdoor environments).



Fig. 6. Tests and evaluations by tourist groups

Table 2. Questionnaires and evaluations

Questionnaire	Ave. Score (Std. Dev.)
Q1. Did the AR tour guide work reliably? (1–5) <sup>†</sup>	3.28 (1.11)
Q2. Were the animated 3-D virtual characters realistic? (1–5)	2.71 (0.76)
Q3. Was the AR tour guide useful and helpful for your tour? (1–5)	4.57 (0.53)
Q4. Was the current prototype device convenient? (1–5)	2.71 (0.75)
Q5. Would it be better if the AR tour guide were served on mobile phones? (1–5)	4.14 (1.07)

<sup>†</sup>Score—1: very bad, 2: bad, 3: normal, 4: good, 5: very good.

**Usefulness:** The answers to question 3: “Was the AR tour guide useful and helpful for your tour?” were very positive (the average score was 4.57). Most participants said that it made their experiences interesting and helped them easily understand the cultural heritage sites. On the other hand, they said the animated 3-D virtual characters were insufficient to allow them to have a fully immersive experience (the average score for question 2 was 2.71). It means the quality and variety of the 3-D contents significantly affected users’ satisfaction as much as the performance of the AR tour guide did.

**Device compatibility:** Our prototype used a portable laptop with a wide screen so that users could easily look at the augmented 3-D contents. It also supported a pen-tablet-based GUI so that users could conveniently use the AR tour guide. However, the evaluations showed that several participants felt the prototype was heavy and uncomfortable, particularly for women and children (the average score for question 4 was 2.71). This aspect was also shown in question 5. The majority answered that they would like to experience the AR tour guide services on their mobile phones (the average score for question 5 was 4.14) because mobile phones tend to be light and compact.

Consequently, the tourists responded that the proposed AR tour guide was a well-suited framework for on-site tour guides, but they highly recommended 3-D contents to be more realistic and varying. They also commented that device types should be carefully considered based on service information or contents because

screen sizes of portable and light device types are relatively small to provide fully immersive experiences, even though the device types would be better for mobile tour guide services.

## 5 Conclusion

In this paper, we presented an AR tour guide that provides intuitive and realistic experiences at cultural heritage sites—Gangnyeongjeon and Gyotaejeon in Gyeongbokgung. The proposed AR tour guide correctly estimated camera poses by tracking simple geometric primitives of the sites (the rectangles of the door-frame), and successfully augmented the animated 3-D virtual characters on the real cultural heritage sites. Moreover, the AR tour guide supported personalized tour guides on the sites by utilizing contextual information such as a tourist's location and profile. Finally, the lessons learned from the tests and evaluations of tourist groups were briefly discussed.

Currently, we are improving the vision-based methods (tracking and recognition) and implementing the next version of our AR tour guide on smartphones.

**Acknowledgement.** This research was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2009 (2nd year).

## References

1. Noh, Z., Sunar, M.S., Pan, Z.: A review on augmented reality for virtual heritage system. In: Chang, M., Kuo, R., Kinshuk, Chen, G.-D., Hirose, M. (eds.) *Learning by Playing*. LNCS, vol. 5670, pp. 50–61. Springer, Heidelberg (2009)
2. Papagiannakis, G., Schertenleib, S., O'Kennedy, B., Arevalo-Poizat, M., Magnenat-Thalmann, N., Stoddart, A., Thalmann, D.: Mixing virtual and real scenes in the site of ancient Pompeii. *Computer Animation and Virtual Worlds* 16, 11–24 (2005)
3. Archeoguide, <http://archeoguide.intranet.gr/>
4. Vlahakis, V., Ioannidis, N., Karigiannis, J., Tstros, M., Gounaris, M., Stricker, D., Gleue, T., Daehne, P., Almeida, L.: Archeoguide: An augmented reality guide for archaeological sites. *IEEE Computer Graphics and Applications* 22, 52–60 (2002)
5. iTacitus, <http://www.itacitus.org/>
6. Zoellner, M., Keil, J., Drevensek, T., Wuest, H.: Cultural heritage layers: Integrating historic media in augmented reality. In: *International Conference on Virtual Systems and Multimedia*, pp. 193–196 (2009)
7. Pletinckx, D., Callebaut, D., Killebrew, A.E., Silberman, N.A.: Virtual-reality heritage presentation at Ename. *IEEE MultiMedia* 7, 45–48 (2000)
8. Portalés, C., Lerma, J.L., Pérez, C.: Photogrammetry and augmented reality for cultural heritage applications. *The Photogrammetric Record* 24, 316–331 (2009)
9. Lowe, D.: Distinctive image feature from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110 (2004)
10. Bouguet, J.Y.: Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm (2000)
11. Schweighofer, G., Pinz, A.: Robust pose estimation from a planar target. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 2024–2030 (2006)



# 3D Reconstruction of a Collapsed Historical Site from Sparse Set of Photographs and Photogrammetric Map

Natchapon Futragoon, Asanobu Kitamoto, Elham Andaroodi, Mohammad Reza Matini, and Kinji Ono

Insight Intellilearn, National Institute of Informatics,  
University of Tehran, University of Yazd

**Abstract.** This paper deals with the challenge of city-scale 3D reconstruction using computer vision techniques. Our method combines the photogrammetric map created from aerial photographs with photographs taken by the general public. The former gives the surface, while the latter gives the texture, and we make a 3D model step-by-step based on a semi-automatic process. We applied this method to the 3D reconstruction of the citadel of Bam, which is a collapsed historical site by the earthquake. Available photographs are limited because new images cannot be captured after the collapse, but we successfully produced a 3D model of the site with texture taken from the photograph. Our system is based on 3ds Max software with several MAXScript tools, such as automatic tools for generating mesh surface from wireframe by assuming walls, slopes and grounds, and assistance tools for a semi-automatic process of estimating camera parameters and transformation matrix.

## 1 Introduction

This paper deals with the challenge of city-scale 3D reconstruction using computer vision techniques. The uniqueness of this paper is that we deal with a city fully collapsed by the earthquake. The old city of Bam in Iran was famous for the largest mud brick structure in the world, but it was completely collapsed by the earthquake occurred in December 2003 (Fig. 1). After the earthquake, the city was declared as a UNESCO world heritage site in danger, but it was too late to make the complete documentation of the city. To reconstruct the city as it was before the earthquake, we need to take advantage of the resources left after the earthquake. This is the purpose of our project "Historical city of Bam" [1,2,3].

City-scale 3D reconstruction is a hot research topic in many fields such as e-heritage, but most research is based on the assumption that we can take a privilege of capturing massive amount of data "from now." This is not the case in Bam; the city was gone, and new data cannot be captured. This excludes the application of some state-of-the-art computer vision approaches such as laser scanning [4,5]. We need to take advantage of available resources such as a photogrammetric map created from aerial photographs, architectural document such



**Fig. 1.** Citadel of Bam : a view form the first city wall toward the bazaar, governor’s district and main tower (a) before the earthquake (b) after the earthquake

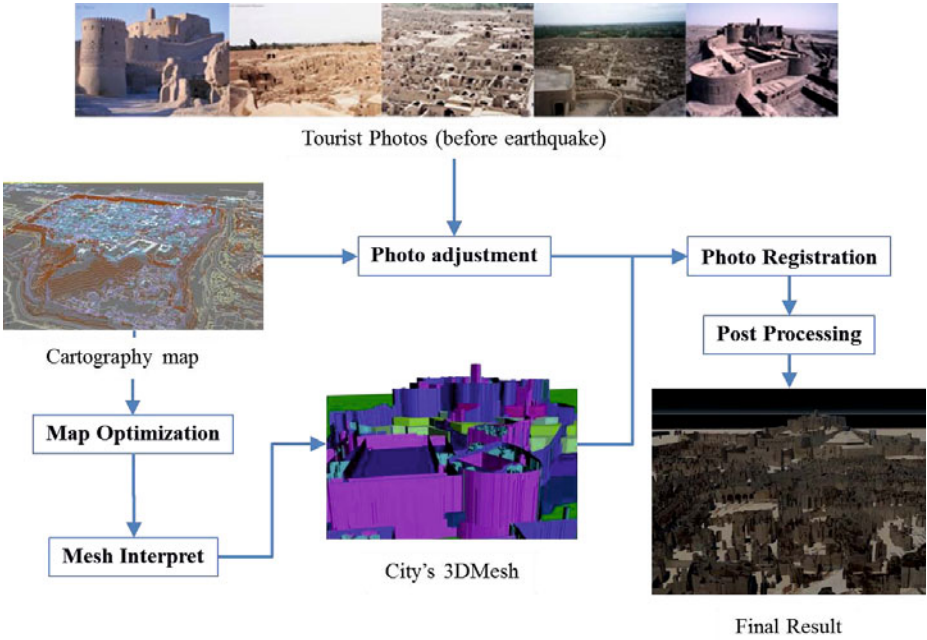
as plans, and tourist photographs we collected after the earthquake, and finally the memory of experts who worked in the city.

Our approach is similar to image-based modeling approaches [6,7,8], but is different due to the availability of the data. Many image-based modeling approaches aim at reconstructing 3D structures from multiple images, or namely structure from motion (SFM). For example, Agarwal et.al. [9] applied structure from motion algorithms for photographs collected from Flickr and realized a city-scale 3D reconstruction. This approach is especially effective for a place where many tourist photographs are uploaded to photo-sharing websites. In contrast, Bam city was a popular place for only a limited number of tourists, and it was collapsed in 2003, when digital camera and photo sharing was emerging. For this reason, digital resources available on the Internet are much less than Rome or other popular touristic sites.

Our approach can be compared with other image-based modeling approaches where limited number of photographs are effectively used to reconstruct the 3D model [10,11,12]. In fact we also tested these approaches and found out that our photograph collection has fatal problems as follows. Firstly, many photographs are taken by analog cameras with unknown parameters, and later digitized, so the situation is much harder than using photographs from digital cameras. Secondly, because photographs had been taken for a few decades, they are affected by the physical reconstruction of the site which had been going on until just before the earthquake. Even if multiple photographs capture the same frame, architecture inside the frame may take a different shape due to the renovation. This effect is significant when the number of photographs is limited, and we finally discarded this approach.

In our case, however, structure does not have to be estimated only from images. We have a photogrammetric map created from aerial photographs, and it can be used as a "2.5-D" model with contours of the surface. By taking advantage of this map, we can overlay the photographs on the photogrammetric map to create a city-scale 3D model with texture mapping. Hence we propose a semi-automatic approach to a city-scale 3D modeling of Bam, which is illustrated in Fig. 2. The main concept of the framework can be described as follows. Firstly





**Fig. 2.** Framework of the system. Photographs and the photogrammetric map is input to the system from *top left*, and the final result is shown at *bottom right*.

the system automatically interprets the photogrammetric map into the 3D mesh with solid value. Secondly, a user maps the texture using system tools that help the developer finding the camera position and registering photos automatically into the scene.

## 2 History of the Project

Citadel of Bam was almost completely collapsed by the earthquake occurred in December 2003. Just after the earthquake, we decided to start “Bam Project” which tries to keep the memory of Bam and collect accurate data for the physical reconstruction of the city in the future. Five days after the earthquake, the last day of 2003, we started a website “Bam, Heritage in Danger”<sup>1</sup> and asked for the general public in the world to send us photographs and videos taken before the earthquake. We received tens of responses from people who visited our website and agreed to donate their photographs taken at Bam. Our photograph collection has grown to more than 200 photographs and one video.

At the same time, we also started to collect remaining documentation that is useful for the reconstruction. One of the most important documentation is the photogrammetric map. The map is developed by Micro Station tool and

<sup>1</sup> <http://dsr.nii.ac.jp/bam/>.

imported as an AutoCAD file. Some parts were modified manually in "The Irano-French 3-D Cartographic Agreement on Bam (IFCA) between CNRS (Centre National de la Recherche Scientifique) and the Iranian National Cartographic Centre (NCC)". Modification are mostly done in important parts e.g. Citadel and main gate. The map is a wireframe which does not have surface nor solid volumes. Hence we need to add surface to the wireframe to use it as a 3D model. This process is explained in Section 3.

Because of the complexity and scale of the city, we divided the city into regions with three levels of architectural importance, and applied different approaches of modeling. For the most important regions, we used a completely manual modeling approach [1][2][3]. This is because, for these important regions, the accuracy of the 3D model is our most important focus, while the automation of the process is out of concern. This in fact involves large amount of manual process, such as interview and discussion with experts about the parts without accurate data. In contrast, for the least important regions, the automation of the process is important because the cost of manual modeling is prohibitive. Hence computer vision techniques are applied to those regions to create a model that may be less accurate but with less cost. In the future, we are planning to merge those three model types to create a unified model of the city. We believe that this is a good combination of accuracy and cost.

### 3 Creating 3D Models from Photographs and the Photogrammetric Map

Our method is based a semi-automatic process due to the limited availability of data. We start with the automatic process of making 3D models from the photogrammetric map by building mesh from the contour map. Then we move on to manual steps of matching photographs with 3D models with a few assistance tools developed by us. Then we refine the 3D models with a few steps. Although manual steps, our assistance tool helps improve the efficiency and accuracy of the task which may otherwise be a tedious and laborious task. Hence we call it a semi-automatic process of creating 3D models.

Assistance tools are developed using Autodesk 3ds Max software<sup>2</sup>. Most of the steps is performed on 3ds Max framework.

#### 3.1 Optimizing Splines

The 3D photogrammetric map is composed of splines in 3D space. Splines are generated using the photogrammetric techniques from aerial photographs, but they are not optimized for making 3D models from them. We hence reduce the complexity of splines while maintaining the shape of the map by a procedure for rearranging data into a closed shape structure. The procedure works as follows.

<sup>2</sup> Autodesk 3ds Max, formerly 3D Studio MAX, is a modeling, animation and rendering package developed by Autodesk Media and Entertainment.

<http://usa.autodesk.com/>

First, all the vertexes staying in the middle of a straight line are removed. Second, if distance between the starting and the ending vertex of a spline is less than a threshold, the spline is marked as a close spline. Lastly, the starting and the ending vertex are connected together with a short line.

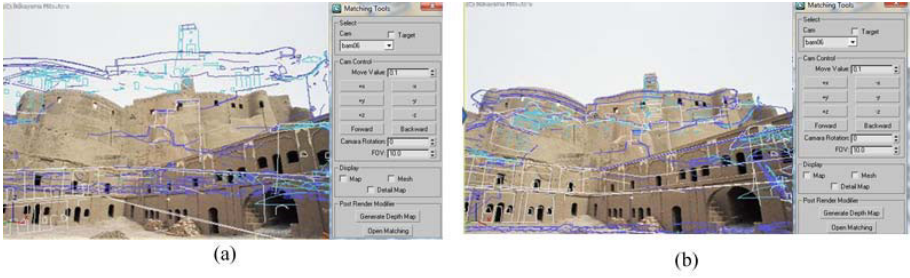
### 3.2 Building Mesh from Splines

Splines should be converted into a mesh with surfaces for mapping texture. Since each spline represents the highest point of the geometry, reasonable surface can be obtained by sweeping splines downward with the following steps.

1. Apply extrude modifier : Extrude modifier is a modifier provided by 3ds Max. This modifier sweeps splines downward. Distance of sweeping is set to a high value in order to make sure all the splines are lower than the ground plane.
2. Slice mesh at ground plane : Extruded mesh is sliced by slicer modifier. Slicer plane is set at the ground plane. This cuts the mesh that exceeds the ground plane. Note that this process does not reduce the number of faces or vertexes.
3. Apply normal modifier : Mesh produced by extrude modifier contains normal vector with arbitrary direction. Many of them are incorrect. This error is obvious when light is applied to the scene, and the color of adjacent faces is not continuous. Therefore, unify normal and smoothing normal vector filter is applied to solve this problem, but some errors still remain.
4. Face optimization : Number of faces can be further reduced by applying mesh optimization. Mesh optimization modifier merge nearby faces into a single face while maintaining the shape of the mesh.
5. Making the ground and hill : Due to the sparseness of the map, ground surface cannot be generated automatically, hence all the hills and grounds are generated manually. A hill is made by using the ground photogrammetric map as a guide line. Some part of the map is missing, however, and missing parts are filled by observing the surrounding line. The ground of the city can be made from a single plane, but this made the ground plane unrealistically flat, so further modifications is required in the post-processing stage.

### 3.3 Estimating Camera Parameters

Camera parameters of each photograph are estimated manually in our approach. Automatic registration of the photograph may be possible in theory using approaches proposed in the literature such as [13], but this is difficult in our case because the 3D model created from the photogrammetric map has lower resolution than photographs. Matching feature points in 3D models with feature points of photographs and identify the viewpoint should deal with matching features across different scales. Instead, we estimate the camera transformation in 3D visual world, by moving a virtual camera around the virtual city and try to find the viewpoint where the virtual scene and the scene of the photograph



**Fig. 3.** Camera parameters estimated by the matching tool. (a) A snap shot before adjusting camera parameters (b) A snap shot after matching camera parameters.

makes a good matching. From the caption of photographs, we can move directly to a neighbor of the true viewpoint, and we then search the best viewpoint in the neighborhood. This process is a crucial step because it affects directly to the overall result. This step is not easy, however, because of the large degree of freedom in choosing camera parameters. We therefore developed a matching tool for this task.

This matching tool helps to determine four main parameters of the camera. That is, camera position, camera target position, lens size of the field of view, and camera rotation. The tool provides a graphical interface to render the result of camera motion in real-time. Then a user can match the edge of the model with the edge of the photograph. Fig. 3 shows a snap shot of the matching tool. The edge of the model is shown as wireframe with different colors. The matching tool was developed using MAXScript<sup>3</sup>.

### 3.4 Estimating Transformation Matrix

In order to deal with the distortion of camera, we developed a tool for adjusting transformation of photographs. Transformation is computed using control points on the photograph and 3D scene from the virtual camera. The photograph is used as the base coordinate and scene from the virtual camera is used as the target coordinate. Control points are manually assigned on both images so that they correspond across both images. At least six control points are required. Warping coordinate from the base image coordinate  $(x, y)$  into the target coordinate  $(u, v)$  can be represented by the second order polynomial equation.

$$[u \ v] = [1 \ x \ y \ xy \ x^2 \ y^2] \times T_{inv} \quad (1)$$

where  $T_{inv}$  is a 6-by-2 unknown coefficient matrix. All six control points is used to derive  $T_{inv}$ . Using this transformation matrix, the photograph is matched with the scene on a virtual camera, so it can be projected in the scene to server as a texture map.

<sup>3</sup> MAXScript is a built-in scripting language for 3ds Max.

### 3.5 Photograph Registration

We finally obtained both camera parameters and a transformation matrix for each photograph. The next step is the registration of a photograph, but it is straightforward using camera parameters and the transformation matrix. For this task, we used two of 3ds Max features, namely composite mapping and camera map per pixel.

1. Composite mapping : This mapping allows a user to map multiple texture into a single material. We need to map multiple photographs on a single material, so composite mapping is required.
2. Camera map per pixel : This map is used for projecting a map from the direction of a particular camera. This map requires a camera object, a depth map and a bitmap texture. Here the bitmap texture is the photograph transformed and the depth map can be obtained by rendering the scene of the virtual camera. Note that texture tiling must not be used in this case.

### 3.6 Post Processing

In the post processing stage, we apply additional modifiers to improve the visibility and reality of rendering,

1. Filling occluded texture : Some parts of the city are not visible in the available photographs, so mud-like texture is filled for those parts. This texture is based on the noise map filter. Base color is the average color of the structure in photographs. Noise color is also the average color of the darker places in photographs. The application of this modifier makes the surface of occluded parts look more natural without increasing memory usage.
2. Modifier ground surface : As mentioned before, ground surface cannot be generated automatically from the photogrammetric map. We need to modify some parts of the ground in order to match the photogrammetric map with photographs. We define that the base ground is at plane  $z = 0$ , and we manually move some parts of the ground by increasing  $z$  until the ground look natural. Finally noise modifier is applied.
3. Lighting system : A fundamental problem of image-based modeling is the problem of illumination. Estimating illumination condition from a photograph and removing the effect of illumination is a difficult problem. This is especially important in our work, because we try to combine many photographs taken in different illumination conditions, namely different time of the day, different camera aperture setting, and so on. Ideally, illumination conditions should be calibrated so that all photographs are combined with the same illumination condition, but we take a simpler approach of reducing the effect of different illumination conditions. The technique is to remove photograph's light and add a new lighting system into the scene. In 3ds Max, this can be performed by setting self-illumination to a small value (approximately the value of illumination coming from scattering light). After that, we add a lighting system object called "Sky light," which is a standard 3ds Max object. This includes both scattering and direct lights.

## 4 Result and Discussion

We applied the proposed algorithm to the 3D reconstruction of Bam. The photographs were selected from the collection of tourist photographs gathered from the world. We have more than 200 photographs, but we found out that useful photographs are limited. Many tourist photographs were taken from similar viewpoints with similar frames, which is typical for tourist photographs. For efficient mapping of photographs, we selected photographs so that it covers the wide area of the city with fewer photographs. Close-up photographs were also discarded because they are sometimes difficult to determine camera parameters. As a result, we can use only 22 photographs for our experiment. [4](#) shows some of the photographs used.

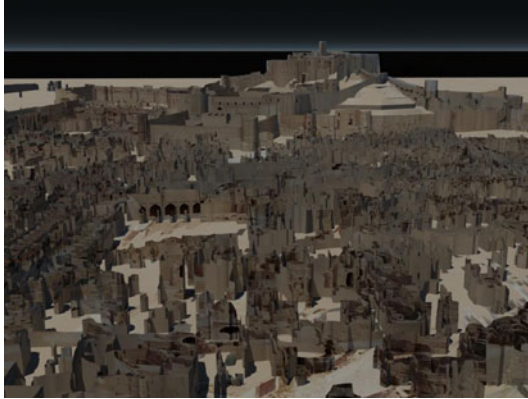
[Fig. 5](#) shows the final result of city-scale 3D modeling and rendering for the citadel of Bam. We generated a realistic scene with texture maps extracted from real architecture in photographs. Our approach provides a solution for the reconstruction of a large scene with limited resources such as tourist photographs and the photogrammetric map.

This method gives a simple solution for a quick 3D modeling and rendering of a large scene. Firstly, we make a photogrammetric map from aerial photographs. The photogrammetric map has information on the height of place, and it may contain texture information of the top surface. However, it does not have information of horizontal view of the building (such as facade). We then overlay photographs on the photogrammetric map as texture mapping to give more realistic view of buildings. This method is applicable to a collapsed historical site such as the citadel of Bam, where only a limited amount of data is available. This is in contrast to massive-scale image-based modeling that requires intensive image capturing activity.

Future work includes the improvement of surface and texture. First, surface can be more intelligently generated. Due to the structure of buildings in Bam, many buildings do not have top surface or roof, and this makes a view from above look poor. We need an algorithm to search for closed surface and interpret as a roof, thus generating mesh on the surface. Second, texture should be improved for occluded parts or parts without reference photographs. In our current implementation, as addressed in [Section 3.6](#), those parts are filled with mud-like texture, but obviously this is not a sophisticated solution, and may be improved



**Fig. 4.** Selected photographs. (a) Main gate toward the citadel (b) Citadel taken from a helicopter (c) Main gate toward the east side (d) Citadel from under.



**Fig. 5.** Final result of the framework

by a context-aware texture generation or instance-based texture generation referring to the database of texture. Another challenge in terms of texture is to remove the effect of illumination from each photograph and merge them without artifacts due to illumination differences.

Our future goal of Bam project is to integrate 3D models created by many types of methods, such as manual, semi-automatic and automatic methods. The choice of methods is related to required accuracy, availability of data, and the size of the model. The combination of various methods may lead to cost-effective 3D model generation with importance-based accuracy. 3DCG08

## Acknowledgments

The supporting research project, 3D CG reconstruction of the Citadel of Bam is a collaborative project between Digital Silk Road Project of NII and Iranian Cultural Heritage, Handicraft and Tourism Organization (ICHHTO). The 3D photogrammetric material is provided to NII by Professor Chahryar ADLE from CNRS and ICHHTO.

## References

1. Ono, K., Andaroodi, E., Einifar, A., Abe, N., Matini, M.R., Bouet, O., Chopin, F., Kawai, T., Kitamoto, A., Ito, A., Mokhtari, E., Eomofar, S., Beheshti, S.M., Adle, C.: 3DCG reconstitution and virtual reality of UNESCO world heritage in danger. *Journal of Progress in Informatics* 5 (2008)
2. Matini, M.R., Andaroodi, E., Kitamoto, A., Ono, K.: Development of CAD-based 3D drawing as a basic resource for digital reconstruction of Bam's Citadel (UNESCO world heritage in danger). In: *Conference on Virtual Systems and Multimedia (VSMM 2008) Volume Full Papers*, pp. 51–58 (2008)

3. Matini, M.R., Andaroodi, E., Kitamoto, A., Ono, K.: Digital 3D reconstruction based on analytic interpretation of relics; case study: Bam Citadel. In: 22nd International Symposium on Digital Documentation, Interpretation and Presentation of Cultural Heritage, CIPA 2009 (2009)
4. Gruen, A.F., Zhang, L.: Image-based reconstruction and modeling of the great buddha statue in Bamiyan, Afghanistan. *Remote Sensing and Spattial Information Sciences (XXXIV-5/W10)*
5. Ikeuchi, K., Nakazawa, A., Hasegawa, K., Ohishi, T.: The Great Buddha Project: Modeling cultural heritage for VR systems through observation. In: ISMAR 2003: Proceedings of the 2nd IEEE/ACM International Symposium on Mixed and Augmented Reality, vol. 7. IEEE Computer Society, Washington, DC, USA (2003)
6. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. In: SIGGRAPH 2005: ACM SIGGRAPH 2005 Papers, pp. 577–584. ACM, New York (2005)
7. Saxena, A., Chung, S.H., Ng, A.Y.: 3-D depth reconstruction from a single still image. *Int. J. Comput. Vision* 76, 53–69 (2008)
8. Xiao, J., Fang, T., Zhao, P., Lhuillier, M., Quan, L.: Image-based street-side city modeling. In: SIGGRAPH Asia 2009: ACM SIGGRAPH Asia 2009 Papers, pp. 1–12. ACM, New York (2009)
9. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building rome in a day. In: 12th International Conference on Computer vision, pp. 72–79 (2009)
10. Thormählen, T., Seidel, H.P.: 3D-modeling by ortho-image generation from image sequences. In: SIGGRAPH 2008: ACM SIGGRAPH 2008 Papers, pp. 1–5. ACM, New York (2008)
11. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In: SIGGRAPH 1996: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 11–20. ACM, New York (1996)
12. van den Hengel, A., Dick, A., Thormählen, T., Ward, B., Torr, P.H.S.: Videotrace: rapid interactive scene modelling from video. In: SIGGRAPH 2007: ACM SIGGRAPH 2007 Papers, vol. 86. ACM, New York (2007)
13. Stamos, I., Liu, L., Chen, C., Wolberg, G., Yu, G., Zokai, S.: Integrating automated range registration with multiview geometry for the photorealistic modeling of large-scale scenes. *Int. J. Comput. Vision* 78, 237–260 (2008)



# Recognition and Analysis of Objects in Medieval Images

Pradeep Yarlagadda, Antonio Monroy, Bernd Carque, and Björn Ommer

Interdisciplinary Center for Scientific Computing, University of Heidelberg, Germany  
{pyarlagadda, amonroy, bcarque, bommer}@iwr.uni-heidelberg.de

**Abstract.** Rapid and cost effective digitization techniques have led to the creation of large volumes of visual data in recent times. For providing convenient access to such databases, it is crucial to develop approaches and systems which search the database based on the representational content of images rather than the textual annotations associated with the images. The success of such systems depends on one key component: category level object detection in images.

In this contribution, we study the problem of object detection in the application context of digitized versions of ancient manuscripts. To this end, we present a benchmark image dataset of medieval images with groundtruth information for objects such as ‘crowns’ in the image dataset. Such a benchmark dataset allows for a quantitative comparison of object detection algorithms in the domain of cultural heritage, as illustrated by our experiments. We describe a detection system that accurately localizes objects in the database. We utilize shape information of the objects to analyze the type-variability of the category and to manually identify various sub-categories. Finally, we report a quantitative evaluation of the automatic classification of object into various sub-categories.

## 1 Introduction

Large scale digitization efforts in the field of cultural heritage have led to the accumulation of vast amounts of visual data in recent times. For a systematic access to such collections, it is necessary to develop algorithms that search the database based on the representational content of the images. For this, it is necessary to go beyond a mere analysis of individual image pixels onto a stage where the semantics of images can be modeled and analyzed. In contrast to this semantics based indexing, the current retrieval systems depend almost exclusively on queries which are directed at the textual metadata. Textual annotations provide only limited search options because of the infeasibility of comprehensive manual indexing. To make image databases accessible in a quicker, more reliable and detailed way, semantics based indexing is indeed necessary. The key for such algorithms is category level object detection.

In this contribution, we explore the question of category level object detection in the context of a benchmark dataset for cultural heritage studies. This

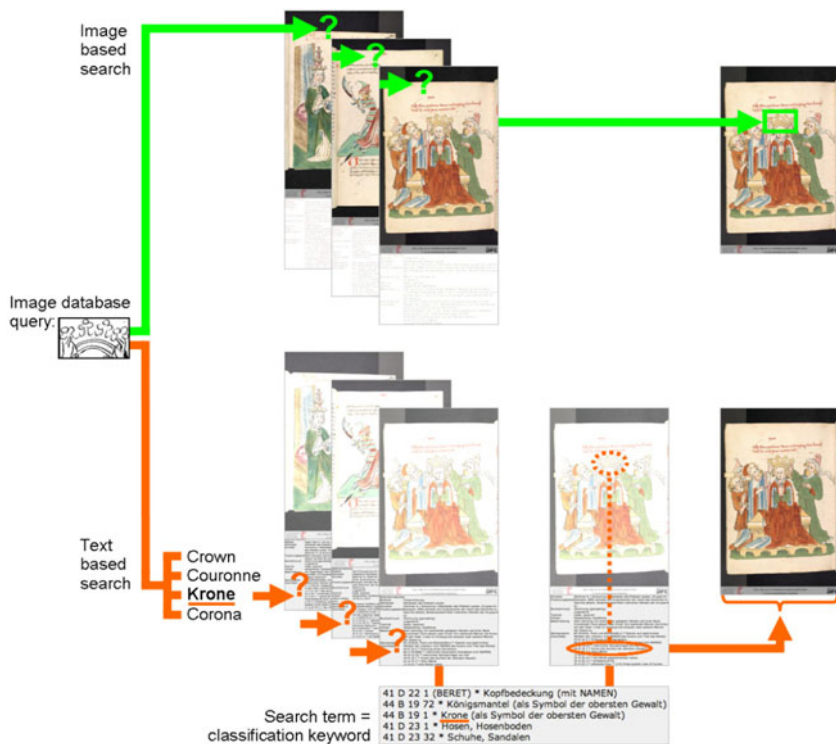


Fig. 1. Text based vs image based retrieval

dataset is highly significant because of its completeness of late medieval workshop production and also it is the first of its kind to enable benchmarking of object detection and retrieval in pre-modern tinted drawings. We also present a statistical analysis of the variability and relations within object categories i.e medieval crowns. The analysis yields a single 2-d visualization of the diversity found among large numbers of instances of a category. Such a visualization can be augmented to the search results of a semantics based query system and the amount of insight it provides into the database cannot be matched by a text based query system.

## 2 Related Work

Image databases in the field of cultural heritage are normally made accessible via textual annotations referring to the representational content of the images [1]. Therefore, content-based image retrieval depends on either the controlled vocabularies of the used classification systems or the textual content of free descriptions. In both cases only that can be found what has been considered in the process of manual indexing; and it can only be found in the specific form in which it has been verbalized. The inevitability of textual descriptions generates numerous problems, for example concerning the scope and detailedness of

the taxonomies, their compatibility beyond linguistic [2], professional or cultural boundaries, their focus on specific aspects of the content according to specific scientific interests or not least the qualification and training of the cataloguer. One of the most sophisticated classification systems is ICONCLASS [3]. Yet, despite its high level of differentiation it has severe limits in a global perspective because it was developed only to cover Western art and iconography. Therefore its ability to index for instance transcultural image resources such as the database of the Cluster of Excellence Asia and Europe in a Global Context at the University of Heidelberg [4] is limited. Furthermore, object definition schemes are featuring a very limited differentiation. In our showcase ‘crown’ the hierarchy of objects ends with this general notion and does not offer varying types of crowns. To focus the object retrieval on subtypes is, in contrast, possible in the case of REALonline, the most important image database in the field of medieval and early modern material culture [5]. Here, the controlled vocabulary contains a few compounds like ‘Buegelkrone’ or ‘Kronhut’. But whereas the main division ‘Kleidung–Amtstracht’ is searchable in German and in English, these subdivisions are available only in German, thus raising difficulties of translation. Problems such as the lack of detail and connectivity are even greater in the case of heterogeneous databases, which are –like HeidICON [6], Prometheus [7] or ARTstore [8] –generated by the input from different institutional and academic contexts. In such cases, the cataloguing of the image content is almost arbitrary due to the uncontrolled textual descriptions. Finally, a basic problem of all these databases is the fact that –due to the serious efforts of manual indexing in terms of cost and time –the fast-growing number of images that are available in a digital format can hardly be itemized in detail and thus cannot be used efficiently in the long term. To overcome these restrictions, we present a system that directly searches the visual data thereby circumventing the need for detailed textual annotations.

Compared to standard benchmark datasets used in computer vision (e.g. [9,10]), we present a database with a high degree of background clutter, scale variation, and within-class variability. Being close to the needs in the field of cultural heritage, this image collection is highly challenging for categorization algorithms, e.g. [9], voting methods for detection such as [11,10,12], and sliding window based classifiers [13].

### 3 Benchmarking, Analysis, and Recognition

#### 3.1 Database and Benchmarking

We have assembled a novel, annotated benchmark image dataset for cultural heritage from a corpus of 27 late medieval paper manuscripts, held by Heidelberg University Library [14]. Produced between 1417 and 1477 in three important Upper German workshops, this corpus is rare in its magnitude and, in addition, offers an exceptional homogeneity concerning its date of origin, its provenance and its technical execution. More than 2,000 half- or full-page tinted drawings illustrate religious and devotional texts, chronicles and courtly epics. Their content has been itemized by means of ICONCLASS, so that we are able to evaluate



Fig. 2. Sample images from the late medieval manuscripts

the capability of the classification system and to detect its desiderata. For this purpose we built a unique dataset of annotations, which covers object categories in a more detailed way than any existing taxonomy, e.g. more than 15 different subtypes of crowns. Thus, the demands on our object retrieval system can be defined precisely. Although our approach is quite generic which can be applied to different object categories, we start from the category which has a high semantic validity since it belongs to the realm of medieval symbols of power [15]. This ensures that our analysis has the highest possible connectivity to research in the humanities, e.g. to art history and history with a focus on ritual practices [16] or on material culture.

*Breakthroughs entailed by a novel benchmark dataset:* Our motivation for introducing a novel benchmark dataset is spurred by the influence the Berkeley Segmentation Dataset (BSDS) [17] has had on the development and evaluation of segmentation algorithms. Before BSDS, measuring segmentation performance was mostly subjective and algorithms were difficult to compare. The new BSDS dataset with its groundtruth annotation has, for the first time, provided an objective performance measure for segmentation. This has stimulated algorithm development which led to previously unexpected breakthroughs in segmentation performance. The F-measure, which is a suitable metric for comparing the performance of segmentation algorithms, has only seen a slight increase in the years before BSDS. Early segmentation algorithms such as Roberts (1965) [18] and Canny (1986) [19] achieved F-measures of 0.47 and 0.53, respectively. In the short time since the introduction of BSDS in 2001, contributions such as [20] have increased the performance to 0.7 while human performance stands at 0.79.

*Annotating the data:* In order to generate groundtruth localizations for objects in the images, we developed an interactive annotation system. Using the expertise of an art historian we have gathered groundtruth annotations. Cubic splines are used to fit a bounding region to the principal curvature of an object. This helps excluding more background from the bounding boxes compared to rectangular bounding boxes.

### 3.2 Object Analysis

The most basic component for object analysis and object recognition is choosing an appropriate mathematical representation for objects which lays the foundation for recognition and further analysis. We utilize a shape based representation of objects since shape is an important cue in these medieval manuscripts.

*Extracting artistic drawings to represent shape:* We have discovered from experiments that the images when represented in HSV color space, particularly the saturation component, provide a good starting point for object boundary extraction. Object boundaries are essentially ridges in an image with few pixels thickness. To detect such ridges, we apply a filter which smoothes the image along the direction orthogonal to the ridge and sharpens the image along the direction of the ridge, called the ridge detection filter [21]. It is defined by the following formula.

$$G(x, y, \sigma_x, \sigma_y) = \frac{1}{\pi * \sigma_x^2} * \left(1 - \frac{x^2}{2 * \sigma_x^2}\right) * \exp\left(-\frac{x^2}{\sigma_x^2} - \frac{y^2}{\sigma_y^2}\right) \quad (1)$$

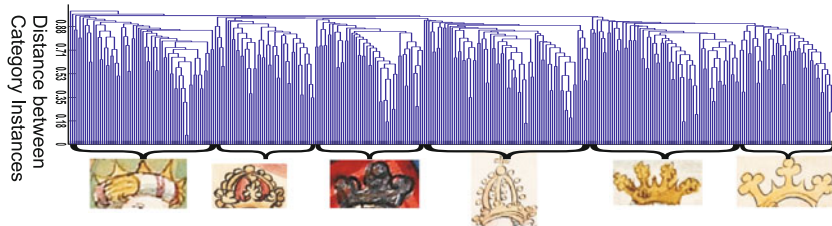
Coordinates  $x, y$  denote image location,  $\sigma_x, \sigma_y$  determine the support of the ridge filter along the x and y directions. Equation 1 defines the ridge filter assuming that the ridge is oriented along the x-axis. This formula is easily extended for detecting ridges at an orientation  $\theta$ .

At each point in the image, optimization over the parameters  $\sigma_x, \sigma_y$  and  $\theta$  yields the maximal filter response. Images marked 1 and 2 in fig. 1 shows an input image and the result of applying the ridge filter to the input.

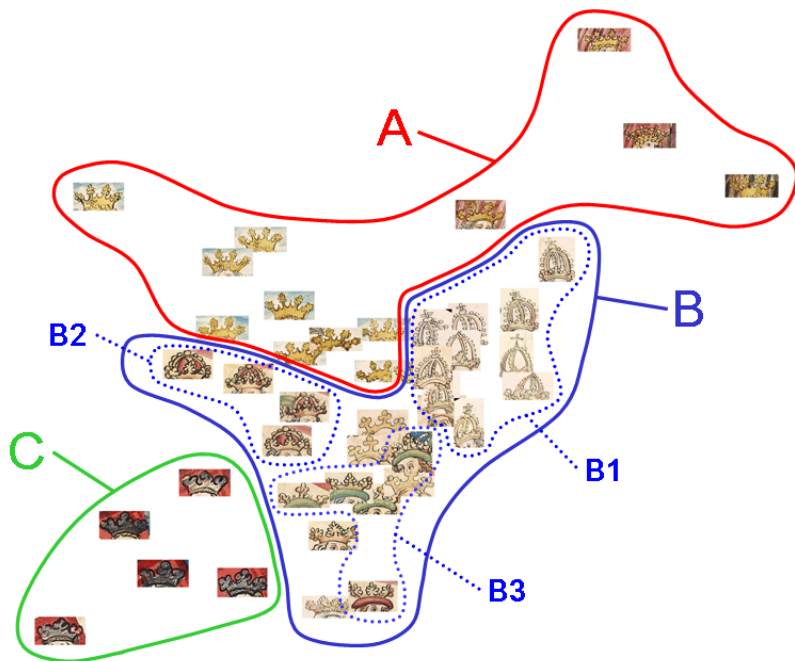
*Shape representation:* Ridges are represented using orientation histograms. We compute these Histograms of Oriented Gradients (HoG) [22] on a dense grid of uniformly spaced cells in the image. We combine histograms from 4 different scales and 9 orientations into a 765 dimensional feature vector.

*Automatic discovery of intra-category structure:* We capture the relationship between various object instances in the database in a single plot by embedding high dimensional HoG feature vectors into a low dimensional space. Such a plot makes it convenient for researchers from cultural heritage to discover relationships without having to study thousands of images. In a first step pairwise clustering based on HoG descriptors is employed to discover the hierarchical substructure of crowns. Then we compute the pairwise distances for samples in the vicinity of the cluster prototypes. Thereafter, a distance preserving low-dimensional embedding is computed to project the 765 dimensional feature vectors onto a 2-d subspace that is visualized in fig. 4. This procedure has extracted relationships, variations and substructure of an object category out of hundreds of images and makes these directly apparent.

The plot displays two central findings of our recognition system and thus reveal the potential of the approach: i) the high type-variability within a category and ii) the different principles of artistic design. In particular, our clusters for the category ‘crown’ show that to the simple crown circlet (A) varied elements like arches (B1), lined arches (B2), torus-shaped brims (B3), hats, or helmets are added. Thus, objects provide advanced semantic information concerning e.g.



**Fig. 3.** Hierarchy of substructure in object category ‘crown’



**Fig. 4.** Visualization of Intra-category variability and substructure of crowns. Group A shows the Swabian workshop of Ludwig Henfflin. Group B shows the Hagenau workshop of Diebold Lauber with the subgroups of crowns with arches (B1), crowns with lined arches (B2) and crowns with torus-shaped brims (B3). Group C shows the Alsatian workshop of 1418

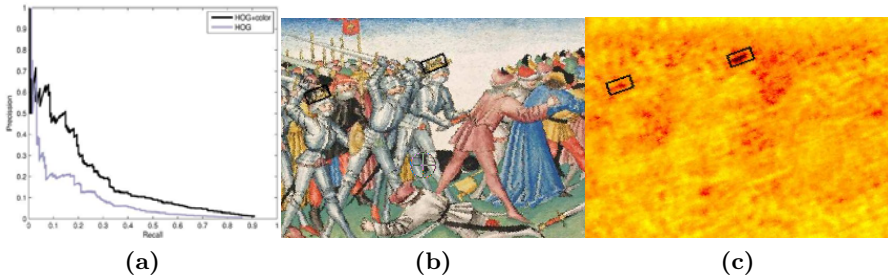
social hierarchies, which is not displayed by the common taxonomies. Since an automated image-based search does not suffer from the desiderata of annotation taxonomies, it becomes a crucial instrument to assist with the detailed differentiation of such subtypes, combining data from large numbers of images and organizing the compositional complexity of objects into a hierarchy of formal variants. Moreover, the clustering and visualization in a MDS-plot (fig. 4) features different principles of artistic design, which are characteristic for different workshops engaged with the illustrations. Group (B) indicates the concise

and accurate style, mainly based on definite contours, of the Hagenau workshop of Diebold Lauber [23], group (A) the more delicate and sketchy style of the Swabian workshop of Ludwig Henfflin, and group (C) the particular summary style of the so-called ‘Alsatian Workshop of 1418’. This detection of specific drawing styles is a highly relevant starting point to differentiate large-scale datasets by workshops, single teams within a workshop, or even by individual draftsmen.

### 3.3 Object Recognition

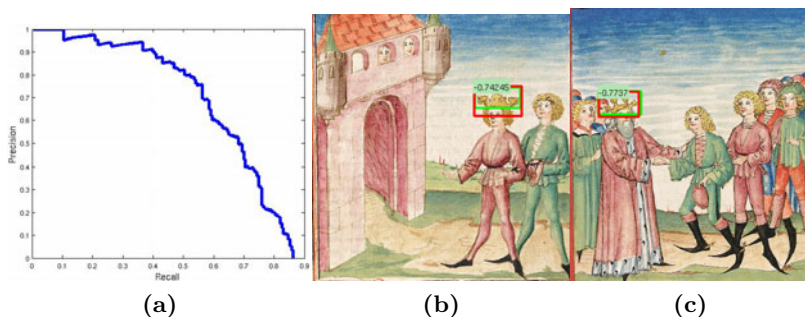
Objects are detected by classifying image regions as object or background using a support vector machine with intersection kernel [24]. This detection algorithm scans the image on multiple scales and orientations. Image regions are represented using the shape representation from subsection 3.2 and a color histogram. The necessary codebook of representative colors is obtained by first quantizing training image using minimum variance quantization into a set of 100 prototypical clusters per image. The bias towards large, homogenous regions is resolved by clustering all these prototypes into an overall set of 30 prototypical colors. We count an object hypothesis as correct if  $\frac{A_h \cap A_g}{A_h \cup A_g} \geq 0.4$  where  $A_h$  and  $A_g$  is the area of the predicted and the groundtruth bounding box, respectively. The precision-recall curve in part a) of fig. 5 shows the detection performance achieved by the presented approach.

The precision recall curves in fig. 5 show scope for improvement as the curves are far from reaching the saturation stage. A closer look at the detection results revealed a lot of false positives in the images which were not sufficiently represented during the training stage of the SVM. To deal with this issue, we have incorporated a bootstrap training procedure to focus on difficult negative samples as is motivated by [25,26]. Training starts as before by learning an SVM model on all positive training samples and an equally sized, random set of negative samples, i.e. bounding boxes drawn from the background. In the next round, negative samples which are either incorrectly classified by the model or fall inside the margin (defined by the SVM classifier) are added to the training set. Also, positive samples which are classified correctly and fall outside the margin are



**Fig. 5.** a) Precision recall curve for crowns obtained from HoG and HoG plus color features. b) Crowns detected in a test image. c) Response of our object detector at each image location.





**Fig. 6.** a) Precision recall curve for crowns obtained by using a bootstrapping training procedure. b) and c) Crowns detected in test images along with the SVM scores.

**Table 1.** Classification results on the crowns from workshops corresponding to groups A, B and C in fig. 4. Columns are the predicted workshop labels and rows are the correct labels. A: Swabian workshop of Ludwig Henfflin, B: Hagenau workshop of Diebold Lauber and C: Alsatian workshop of 1418. The average classification accuracy is  $97.67 \pm 1.7$  %.

Workshops pred.: correct:	A	B	C
A	0.9836	0.0163	0
B	0.0365	0.9634	0
C	0.0083	0.0083	0.9833

removed from the training set. This process is repeated iteratively until there are no new hard negative samples that can be added to the training set. This iterative training procedure resulted in a significant improvement in the detection performance and the resulting PR curves are presented in fig. 6 along with two examples of detections in test images.

Accurate localization of objects within the images as shown in fig. 5 makes complex representations like battle scenes or coronations with several symbols of power more easily readable. Textual annotations do not provide localization information so that object detection and reasoning about the spatial relationship between objects or about their performative context [27] remains impossible.

In section 3.2, we have presented an unsupervised approach to identify category substructure which has then lead to a visualization (fig. 4) of the different artistic workshops that have contributed to the Upper German manuscripts. Based on this visualization, art historians have provided us with groundtruth information so that we can conduct a quantitative evaluation: they have labeled all crowns in the dataset with the workshop that they come from based on formal criteria [23]. There are 137 crowns in our dataset that belong to group A (the workshop of Ludwig Henfflin), 106 crowns belong to group B (the workshop of Diebold Lauber) and 23 crowns belong to group C (the Alsatian workshop). We then incorporate a discriminative approach for predicting the workshop that a crown belongs to. This multi-class classification problem is tackled using the



features from before and incorporating SVM in a one-versus-all manner. For evaluation, we apply 10-fold cross-validation: In each round, 50 % of the crowns from each group have been used for training and the remaining 50 % of the crowns are used for testing by holding back their labels. The classification results of the crowns according to the workshops are presented in table 11 in the form of a confusion matrix.

## 4 Discussion and Conclusions

The present case study on the Upper German manuscripts of Heidelberg University Library shows the detection results that can be obtained by state-of-the-art category level object recognition techniques in the context of cultural heritage. It is now possible to automatically discover the substructure of object categories which is, for instance, caused by different subtypes or principles of artistic design. In order to refine our method, we will apply it in a second step to the entire corpus of the Upper German manuscripts and, in a third step, to the remaining c. 5,000 images of the *Codices Palatini germanici* (28), which have, for the most part, not previously been labeled.

## References

1. Baca, M., Harpring, P., Lanzi, E., McRae, L., Whiteside, A.: *Cataloging Cultural Objects. A Guide to Describing Cultural Works and Their Images* (2006)
2. Kerscher, G.: *Thesaurus-Verwendung und internationalisierung in Bilddatenbanken*. *Kunstchronik* 57, 606–608 (2008)
3. van Straten, R.: *Iconography, Indexing, ICONCLASS. A Handbook* (1994)
4. <http://www.asia-europe.uniheidelberg.de/research/heidelberg-research-architecture/hra-databases-1/transcultural-image-database/the-image-database>
5. <http://www.imareal.oeaw.ac.at/realonline/>
6. <http://heidicon.ub.uni-heidelberg.de>
7. <http://www.prometheus-bildarchiv.de>
8. <http://www.artstor.org>
9. Fergus, R., Perona, P., Zisserman, A.: *Object class recognition by unsupervised scale-invariant learning*. In: *Intl. Conf. on Comp. Vision and Pat. Rec.* (2003)
10. Ferrari, V., Jurie, F., Schmid, C.: *From images to shape models for object detection*. In: *Intl. Journal of Comp. Vision* pp. 40–82 (2009)
11. Leibe, B., Leonardis, A., Schiele, B.: *Combined object categorization and segmentation with an implicit shape model*. In: *Europ. Conf. on Comp. Vision* (2004)
12. Ommer, B., Malik, J.: *Multi-scale object detection by clustering lines*. In: *Intl. Conf. on Comp. Vision and Pat. Rec.* (2009)
13. Lampert, C., Blaschko, M., Hofmann, T.: *Beyond sliding windows: Object localization by efficient subwindow search*. In: *Intl. Conf. on Comp. Vision and Pat. Rec.* (2008)
14. Pietzsch, E., Effinger, M., Spyra, U.: *Digitalisierung und Erschließung spätmittelalterlicher Bilderhandschriften aus der Bibliotheca Palatina*. In: Thaller, H. (ed.) *Digitale Bausteine FÜR Die Geisteswissenschaftliche Forschung*

15. Schramm, P.E.: *Herrschaftszeichen und Staatssymbolik*, vol.3 (1954)
16. Schwedler, G., Meyer, C., Zimmermann, K. (eds.): *Rituale und die Ordnung der Welt* (2008)
17. Fowlkes, C., Tal, D., Martin, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: *Intl Conf. on Comp. Vision* (2001)
18. Roberts, L.: Machine perception of three-dimensional solids. *Optical and Electro-Optical Information Processing*, 159–197 (1965)
19. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pat. Analysis and Machine Intelligence*, 679–714 (1986)
20. Arbelaez, P., Fowlkes, C., Maire, M., Malik, J.: Using contours to detect and localize junctions in natural images. In: *Intl. Conf. on Comp. Vision and Pat. Rec.* (2008)
21. Kovesi, P.D.: MATLAB and Octave functions for computer vision and image processing, <http://www.csse.uwa.edu.au/~pk/research/matlabfns/>
22. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *Intl. Conf. on Comp. Vision and Pat. Rec.* (2005)
23. Saurma-Jeltsch, L.E.: *Spätformen mittelalterlicher Buchherstellung. Bilderhandschriften aus der Werkstatt Diebold Laubers in Hagenau*, vol.2 (2001)
24. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: *Intl. Conf. on Comp. Vision and Pat. Rec.* (2008)
25. Davison, A., Hinkley, D.: *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge(1997)
26. Felzenszwalb, P.F., Girshick, R.B., Mcallester, D., Ramanan, D.: Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2010)
27. Petersohn, J.: Über monarchische Insignien und ihre Funktion im mittelalterlichen Reich. *Historische Zeitschrift* 266, 47–96 (1998)
28. <http://www.ub.uni-heidelberg.de/helios/digi/codpalgerm.html>

# 3D Shape Restoration via Matrix Recovery

Min Lu<sup>1</sup>, Bo Zheng<sup>1</sup>, Jun Takamatsu<sup>2</sup>, Ko Nishino<sup>3</sup>, and Katsushi Ikeuchi<sup>1</sup>

<sup>1</sup> Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

{lumin,zheng,katsu}@cvl.iis.u-tokyo.ac.jp

<sup>2</sup> Nara Institute of Science and Technology, Nara, Japan

j-taka@is.naist.jp

<sup>3</sup> Department of Computer Science, Drexel University, Philadelphia, USA

kon@drexel.edu

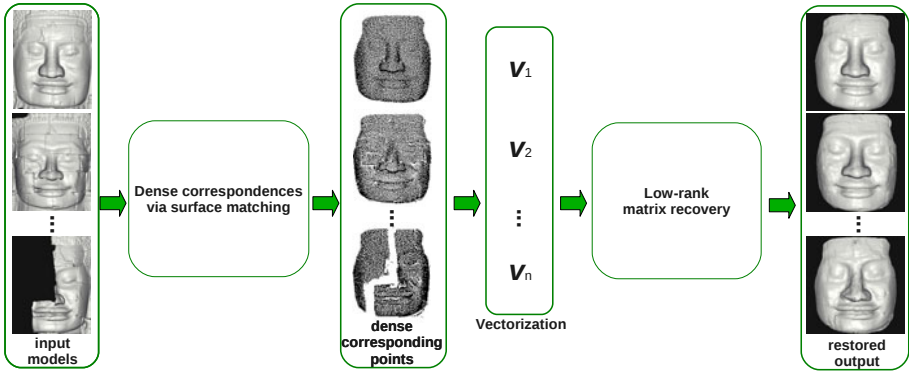
**Abstract.** Cultural relics are often damaged and incomplete due to various reasons. For the purpose of helping archaeological studies, we present a novel method for simultaneously restoring the original shapes of a group of similar objects. Based on the assumption that similar shapes are approximately linearly correlated, we use a matrix recovery technique to achieve the restoration. In order to represent input shapes in a matrix form, vectorization of each aligned sample is carried out by stacking coordinates of dense corresponding points that are generated by a surface matching scheme using non-rigid deformation. An experiment using 3D scans of facial sculptures from *Bayon* is conducted, and the result verifies the feasibility and effectiveness of our method.

## 1 Introduction

Three-dimensional digital replicas play an increasingly important role in cultural heritage preservation. With current 3D data acquisition technology, such as laser rangefinders, the geometric information of real-world objects can be accurately and reliably digitized. These 3D digital models can then be used for various archaeological studies. For example, a 3D shape comparison technique was used to help archaeologists understand the meaning of four-faced towers in the temple *Bayon* at *Angkor* [1].

Due to natural and human factors, e.g., weathering and vandalism, historic cultural relics are often partially damaged (as an example see Figure 2b). Even for complete objects, sometimes it is difficult to acquire all the shape information, because of self-occlusion or some special physical properties of the surface. Therefore, 3D shape completion or restoration becomes a problem of practical significance.

Several approaches have been proposed for 3D shape restoration. For instance, one may focus on the smoothness of the underlying surface, using localized geometric constraints to achieve a smooth continuation [2,3,4]. This kind of completion methods is suitable for filling holes, but when the missing part contains a lot of details and structural information, it will become ineffective. An alternative approach is to use a *copy and paste* scheme. Patches with similar surface



**Fig. 1.** An overview of our shape restoration pipeline. For the input shapes, we first generate dense correspondences among them; then, by stacking coordinates of these corresponding points, input samples are represented as fixed-length vectors; in the end, the restoration process is accomplished by a matrix recovery procedure.

characteristics could be selected from either the incomplete object itself [5,6], or analogous candidate models [7,8].

In this paper, we focus on a specific instance of the shape recovery problem: given a group of similar objects, where many of them are partially damaged and incomplete, we aim to restore all these objects simultaneously, using the common shape structures. Notice that this specific problem setting is not so unusual in heritage conservation. For example, there are usually many similar god statues excavated from the same place, keeping a unified style.

We present a new shape restoration method based on a matrix recovery method [9]. We formulate the shape restoration task as a low-rank matrix recovery problem, that we solve using convex optimization. A simple but effective dense correspondence scheme for shape vectorization is also proposed, where a deformation-based surface matching method is used. Figure 1 depicts an overview of our proposed method. Given a group of similar shapes, we first generate dense correspondences among all samples. Then each sample is represented as a fixed-length vector, using coordinates of corresponding points. Finally, input samples are restored to their original shapes using matrix recovery.

The remainder of this paper is organized as follows: Section 2 first gives a brief introduction of matrix recovery theory, and then formulates the task of restoring a group of similar shapes as a low-rank matrix recovery problem; Section 3 introduces the procedure of acquiring dense correspondences among all input samples, which is a crucial step for shape vectorization; Section 4 presents results of an experiment using real world relics, demonstrating the effectiveness of the proposed method; and Section 5 concludes with a discussion of the limitations and promising directions of our method.

## 2 Low-Rank Matrix Recovery

*Matrix recovery*, also known as *robust principal component analysis* (Robust PCA), was first introduced in [9]. The essential idea of this theory is to recover corrupted entries of a matrix using structural information of the matrix itself. Compared to ordinary principal component analysis, this method is more robust to outlying and corrupted observations, and it can handle such complex problems as background modeling [10] and batch image alignment [11]. In this section, we first give a brief introduction of matrix recovery theory, and then we explain how this method is used to solve our shape restoration problem.

### 2.1 Problem Statement

Given the observed data matrix  $D \in \mathbb{R}^{m \times n}$ , generated by corrupting some of the entries of an unknown low-rank matrix  $A \in \mathbb{R}^{m \times n}$ , let an error matrix  $E \in \mathbb{R}^{m \times n}$  represent the corruption.  $E$  is also unknown but supposed to be sparse. The goal is to recover  $A$ .

Robust principal component analysis [9] solves this problem by seeking the lowest rank  $A$  that could have generated the observation  $D$ , while subjecting the error matrix  $E$  to a sparseness constraint:  $\|E\|_0 \leq k$ . Here the  $L^0$  norm is employed to measure the matrix sparseness. Thus the initial problem becomes an optimization:

$$\min_{A,E} \text{rank}(A) + \gamma \|E\|_0, \quad \text{s.t.} \quad A + E = D, \quad (1)$$

where  $\gamma$  is the weighting parameter that trades off the rank of the solution and the sparseness of the error.

As detailed in [9], the optimization problem (1) is highly non-convex, and currently with no efficient solution. A tractable optimization, however, can be obtained by relaxing the original problem. By replacing the  $L^0$  norm with the  $L^1$  norm, and by measuring the rank with the nuclear norm  $\|A\|_*$ , problem (1) can be converted to a tractable convex optimization:

$$\min_{A,E} \|A\|_* + \lambda \|E\|_1. \quad \text{s.t.} \quad A + E = D. \quad (2)$$

Here the nuclear norm of a matrix is defined as the sum of its singular values:  $\|A\|_* \doteq \sum_i \sigma_i(A)$ . And the weighting parameter  $\lambda$  is in the form  $c/\sqrt{m}$ , where  $c$  is a constant, and typically set to be around 1. Notice that the new objective function in problem (2) is continuous and convex, so it can be solved efficiently [9,12].

### 2.2 Applying to the Shape Restoration Problem

Now let us describe how we formulate our shape restoration problem as matrix recovery. Given an object category  $\mathbf{C}$ , in which many samples are partially damaged and incomplete, let  $\{\mathbf{s}_i\}_{i=1}^n$  denote a group of observed 3D shape and

$\{\mathbf{s}_i^0\}_{i=1}^n$  denote the corresponding original shapes without corruption. As we assume that all these samples are of similar shapes and structures, i.e. they are drawn from the same category, we may assume that they belong to a same linear subspace  $\mathbb{S}$ . In other words, as long as  $n$  is sufficiently large, an arbitrary sample  $\mathbf{s}^0$  from the same category  $\mathbf{C}$  will approximately lie in the linear span of the samples  $\{\mathbf{s}_i^0\}_{i=1}^n$ :

$$\mathbf{s}^0 \approx \sum_{i=1}^n \alpha_i \mathbf{s}_i^0, \quad (3)$$

where  $\{\alpha_i\}_{i=1}^n \in \mathbb{R}$  are coefficients. In our method, this assumption of linear correlation is the only prior knowledge we rely on to restore the corrupted samples.

As in [11], we define an operator  $vec : \mathbf{C} \rightarrow \mathbb{R}^m$  that extracts an  $m$ -dimensional feature vector from a 3D model  $\mathbf{s}_i$ . In our shape restoration case, this operation can be accomplished by simply stacking the  $(x, y, z)$  coordinates of the points of interest. We will discuss how to achieve this via non-rigid registration in Section 3. This results in a matrix  $A$  that represents all the observed samples:

$$A \doteq [vec(\mathbf{s}_1^0) | \cdots | vec(\mathbf{s}_n^0)] \in \mathbb{R}^{m \times n}. \quad (4)$$

According to the linear correlation assumption (Eq. (3)), matrix  $A$  should be approximately low-rank.

For an observation sample  $\mathbf{s}_i$ , let  $\mathbf{e}_i$  denote the corrupted or missing component from the original shape  $\mathbf{s}_i^0$ , so  $\mathbf{s}_i = \mathbf{s}_i^0 + \mathbf{e}_i$ . Using the operator  $vec$  we defined above, the corrupted observation can then be written as

$$D \doteq [vec(\mathbf{s}_1) | \cdots | vec(\mathbf{s}_n)] = A + E, \quad (5)$$

where matrix  $A$  is a low-rank matrix defined in Eq. (4), revealing the common shape information of this category, and  $E \doteq [vec(\mathbf{e}_1) | \cdots | vec(\mathbf{e}_n)] \in \mathbb{R}^{m \times n}$  is the error matrix, representing the shape corruption. As we assume that the corruption is partial and localized, the error matrix  $E$  should be sparse, which means most of its entries are zero. Thus, the task of restoring the shape of similar objects in the same category becomes a matrix recovery problem as defined in Section 2.1.

### 3 Shape Correspondence and Vectorization

In the whole process of shape restoration via matrix recovery, a crucial step is to properly represent the shape of each sample using a fixed-length vector, so that accurate correspondences are established among all input objects. Recall that in Section 2.2, we introduced an operator  $vec$  to extract an  $m$ -dimensional feature vector from a 3D shape. In this section, we describe this procedure in detail.

#### 3.1 Sparse Correspondences

First, let us consider establishing a group of sparse corresponding points. Obviously, a trivial solution is to manually specify the corresponding points. Although

there are several automatic methods [13,14,15,16], using human assistance is still the most reliable approach for finding correspondences, especially when data corruption and high scanning noise exist, which is common for historical objects in the outdoors. If the output correspondence is acceptable, automatic shape registration methods could be chosen as well. Notice that the methods in [15,16] can also be used to generate dense correspondences.

In our work, we chose to leverage manual intervention. Among all input shapes, a relatively complete sample is chosen as a template (Figure 2a). We predetermine a group of feature points (Figure 2c) and manually select these points on each sample (Figure 2d). If certain points are missing due to shape corruption, we simply mark them as null points.

Then we adopt a rigid registration process for all samples (Figure 2f). The posture of the template is fixed, and all other samples are aligned to the template using *Iterative closest point* (ICP) [17] algorithm. Note that the initial posture estimation could be calculated from the sparse corresponding points.

### 3.2 Dense Correspondences and Vectorization

Based on the sparse correspondences, a sampling strategy while keeping the correct correspondence could be carried out to obtain dense correspondences among all input samples. The uniform remeshing method in [18] is a workable choice, but here we use a surface matching scheme based on shape deformation:

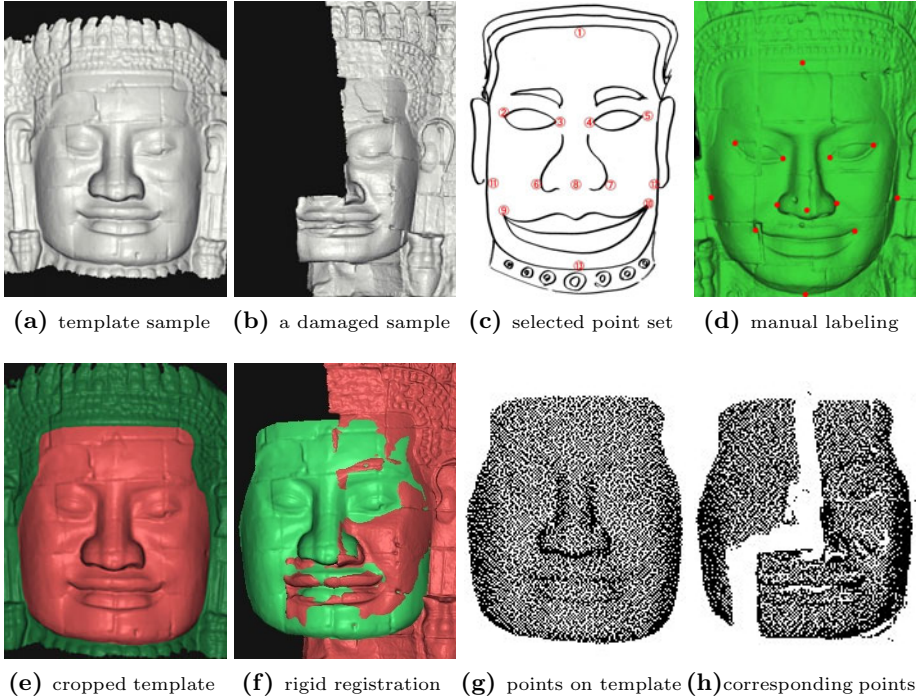
1. Adopt a uniform sampling on the template sample to create a *final template* with an adequate number of points (Figure 2g).
2. Deform the final template to fit each sample using the sparse correspondences we manually selected before as control handles for the non-rigid surface deformation process.
3. Search the closest point on the destination sample for each point of the final template, and label the result as the approximate corresponding point. Here we set a distance threshold: if there is no point within this threshold, correspondence is marked as a null point.

For the shape deformation phase, a *moving least squares* (MLS) deformation similar to [19] is employed:

Given a set of  $N$  control points (in our case the corresponding points), let  $\{\mathbf{p}_i\}_{i=1}^N \in \mathbb{R}^3$  be the original positions on source model  $S_0$ , and  $\{\mathbf{q}_i\}_{i=1}^N \in \mathbb{R}^3$  be the corresponding deformed positions on destination shape  $S_d$ . Consider an arbitrary point  $\mathbf{x} \in \mathbb{R}^3$  on the source model  $S_0$ , let  $F_{\mathbf{x}}$  denote the transformation that gives the corresponding position of point  $\mathbf{x}$  on  $S_d$  after the deformation. According to the MLS theory,  $F_{\mathbf{x}}$  could be determined by solving an optimization:

$$\min_{F_{\mathbf{x}}} \sum_{i=1}^N \frac{1}{d(\mathbf{x}, \mathbf{p}_i)^{2\alpha}} \|F_{\mathbf{x}}(\mathbf{p}_i) - \mathbf{q}_i\|_2, \quad (6)$$

where  $d(\mathbf{x}, \mathbf{p}_i)$  is the distance between  $\mathbf{x}$  and  $\mathbf{p}_i$ ,  $\alpha$  is a system parameter.



**Fig. 2.** Establishing dense correspondences. (a) and (b) are two illustrations of input shapes, where (a) is relatively complete and selected as the template, while (b) is a heavily damaged sample; (c) shows the point set chosen for sparse correspondences and (d) is an example of manually labeled points; (e) is a cropped template sample that keeps the region of interest only; (f) shows the rigid registration procedure between the template and other samples before shape vectorization; (g) shows selected points on the template via uniform sampling, and (h) is the corresponding points on example (b).

In order to get better deformation results, geodesic distances are used in the weight function. Given one 3D shape represented by a triangle mesh, the geodesic distance between two points on its surface can be approximated with the length of the shortest path from one to the other, which can be calculated by Dijkstra's algorithm. Moreover, the mapping  $F_{\mathbf{x}}$  is assumed to be an affine transformation, consisting of a linear transformation  $M$  followed by a translation  $T$ :  $F_{\mathbf{x}}(\mathbf{x}) = M\mathbf{x} + T$ .

So far, we have obtained a set of dense sampling points with correct correspondences for all input samples. As for the vectorization of each sample, the  $(x, y, z)$  coordinates of all selected points are stacked to form a vector that represents the 3D shape. Obviously, all these vectors are of the same length as the number of sampling points are fixed. Notice that points corresponding to damaged parts may be marked as null points in our scheme. These null points could be substituted with nearby points on the object's convex hull or bounding box





**Fig. 3.** Some example 3D shapes from the 3D database of facial sculptures in *Bayon*

for actual calculation. Figure 2h shows an example where points corresponding to the missing right side of face are chosen from the bounding box instead.

## 4 Experimental Results

We conducted experiments to restore the 3D shapes of real-world cultural relics to validate the proposed method. A group of 151 scanned models of facial sculptures in the temple *Bayon* (Figure 3) were used. Due to weathering, vandalism, and some other reasons, many sculptures are incomplete, and some of them are damaged so heavily that only a small part is preserved (e.g. Figure 2b).

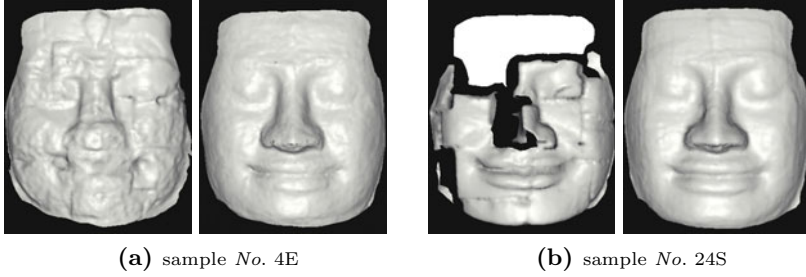
Each sample contains around 500,000 points and 1,000,000 triangles in average. A relatively complete sample, *No. 15N* (Figure 2a), is chosen as the template and 13 feature points (apex of nose, corners of eyes and mouth, etc.) for sparse correspondences were chosen (Figure 2c). These feature points were manually localized on each sample.

Compared to the outer part of a facial sculpture, such as the ears and the headwear, the inner part (the face) contains more information we are interested in. Taking this into consideration, before generating dense correspondences, the outer part of the template is masked out (Figure 2e).

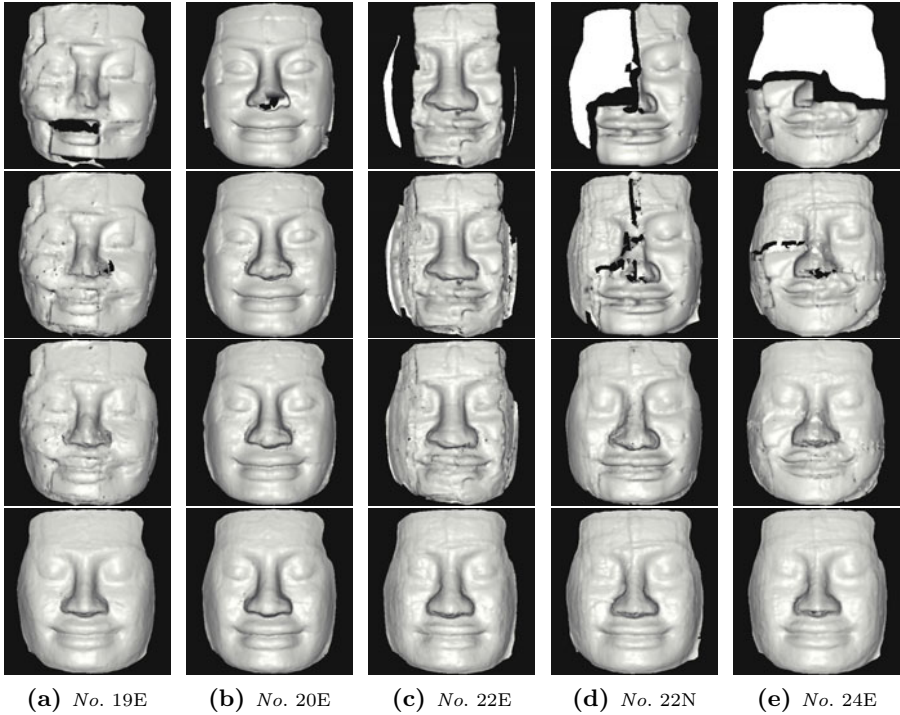
In the dense correspondence phase, all samples were downsampled to 10,000 points, which makes the observation matrix  $D$  30,000 rows and 151 columns. The *Augmented Lagrange Multiplier* (ALM) method [12] is employed to solve the convex optimization problem (2). On a common PC platform, the processing time for solving Eq. (2) was within a few minutes.

Figure 4 shows two restoration examples. Sample *No. 4E* is so severely damaged that it is difficult to identify facial features, while the situation of sample *No. 24S* is even worse: several parts, including the whole forehead and half the nose, are missing. In spite of that, our restoration method still gives satisfactory restoration results.

In the convex optimization process (Eq. (2)), there is a weighting parameter  $\lambda$  that trades off the rank of the solution versus the sparseness of the error. As



**Fig. 4.** Two restoration results with parameter  $c$  set to 1. In each group, the picture on the left side shows the observed geometry, and the other one shows the restored output. Parameter  $c$  is a scaled version of parameter  $\lambda$  in Eq. (2):  $\lambda = c/\sqrt{m}$ , where  $m$  is the length of the input vectors.



**Fig. 5.** Five different restoration results, with three different values of parameter  $c$ . Each column belongs to the same sample, and the first row shows the original inputs. The remaining rows demonstrate the outputs under different values of parameter  $c$ , 2, 1.6 and 1, respectively, from top to bottom.

we mentioned, parameter  $\lambda$  is in the form  $c/\sqrt{m}$ , where  $c$  is a constant, typically set to 1.  $m$  is the length of the input vectors, fixed to three times the number of corresponding points in our experiment. Therefore the constant  $c$  could be used as a scaled version of the parameter  $\lambda$ . Notice that for our shape restoration problem, this parameter  $c$  trades off the similarities of all input models versus the characteristics of each sample: the larger  $c$  is, the more individual characteristics, as well as the error caused by shape incompleteness, will be kept and vice versa. Figure 5 illustrates the effect of changing the value of parameter  $c$ . The result shows that the typical value 1 seems to be a good trade-off for parameter  $c$  in our shape restoration task.

## 5 Conclusion and Discussion

We have proposed a novel method for 3D shape restoration. We focused on a group of similar shapes, aiming to restore them simultaneously. The key idea is to make use of shape similarities, which is handled by a matrix recovery procedure. Experimental results on facial sculptures from *Bayon* verify the effectiveness of our method. Although it is difficult to evaluate the accuracy of our restoration output, as there is no ground truth available, we believe the method is of significant importance for meaningful and feasible archaeological studies, especially when the shapes of a group of similar relics are needed to be restored.

The method, as it currently stands, has a few limitations. First, the scheme for acquiring dense shape correspondences is inefficient. Currently we use the closest point after shape deformation as an approximation of the correspondence, which is not guaranteed to be accurate and reliable, especially when significant non-rigid deformation exists. The choice of the template sample may also affect the result. Second, some shape details are lost after restoration. This is caused by the downsampling process and the parameter selection in matrix recovery. As an immediate future work, we plan to investigate methods to distinguish corrupted and missing data so that different strategies can be employed to restore each.

## References

1. Kamakura, M., Oishi, T., Takamatsu, J., Ikeuchi, K.: Classification of Bayon faces using 3D models. In: The 11th International Conference on Virtual Systems and Multimedia, VSMM 2005 (2005)
2. Chalmovianský, P., Jüttler, B.: Filling holes in point clouds. In: Wilson, M.J., Martin, R.R. (eds.) *Mathematics of Surfaces*. LNCS, vol. 2768, pp. 196–212. Springer, Heidelberg (2003)
3. Ju, T.: Robust repair of polygonal models. *ACM Trans. Graph.* 23, 888–895 (2004)
4. Nooruddin, F.S., Turk, G.: Simplification and repair of polygonal models using volumetric techniques. *IEEE Transactions on Visualization and Computer Graphics* 9, 191–205 (2003)
5. Sharf, A., Alexa, M., Cohen-Or, D.: Context-based surface completion. In: *ACM SIGGRAPH 2004* (2004)

6. Breckon, T.P., Fisher, R.B.: Three-dimensional surface relief completion via non-parametric techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 2249–2255 (2008)
7. Pauly, M., Mitra, N.J., Giesen, J., Gross, M., Guibas, L.J.: Example-based 3D scan completion. In: *SGP 2005: Proceedings of the Third Eurographics Symposium on Geometry Processing* (2005)
8. Kraevoy, V., Sheffer, A.: Template-based mesh completion. In: *SGP 2005: Proceedings of the Third Eurographics Symposium on Geometry Processing* (2005)
9. Wright, J., Ganesh, A., Rao, S., Peng, Y., Ma, Y.: Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In: *Advances in Neural Information Processing Systems 22*. MIT Press, Cambridge (2009)
10. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? Submitted to *Journal of the ACM* (2009)
11. Peng, Y., Ganesh, A., Wright, J., Xu, W., Ma, Y.: RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR* (2010)
12. Lin, Z., Chen, M., Wu, L., Ma, Y.: The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Submitted to *Mathematical Programming* (2009)
13. Zhang, H., Sheffer, A., Cohen-Or, D., Zhou, Q., van Kaick, O., Tagliasacchi, A.: Deformation-driven shape correspondence. In: *Computer Graphics Forum (Special Issue of Symposium on Geometry Processing 2008)*, vol. 27, pp. 1431–1439 (2008)
14. Lipman, Y., Funkhouser, T.: Möbius voting for surface correspondence. In: *ACM SIGGRAPH 2009* (2009)
15. Zeng, W., Zeng, Y., Wang, Y., Yin, X., Gu, X., Samaras, D.: 3D non-rigid surface matching and registration based on holomorphic differentials. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part III*. LNCS, vol. 5304, pp. 1–14. Springer, Heidelberg (2008)
16. Zeng, Y., Gu, X., Samaras, D., Wang, C., Wang, Y., Paragios, N.: Dense non-rigid surface registration using high-order graph matching. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR* (2010)
17. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 239–256 (1992)
18. Li, X., Jia, T., Zhang, H.: Expression-insensitive 3D face recognition using sparse representation. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition, CVPR* (2009)
19. Alvaro, C., Claudio, E., Antonio, O., Paulo, R.C.: 3D as-rigid-as-possible deformations using MLS. In: *The 25th Computer Graphics International Conference, CGI 2007* (2007)

# A Development of a 3D Haptic Rendering System with the String-Based Haptic Interface Device and Vibration Speakers

Kazuyoshi Nomura<sup>1</sup>, Wataru Wakita<sup>2</sup>, and Hiromi T. Tanaka<sup>2</sup>

<sup>1</sup> Graduate School of Science and Engineering, Ritsumeikan University, Kusatsu, Japan

<sup>2</sup> Department of Human and Computer Intelligence, College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan  
{nomura,wakita,hiromi}@cv.ci.ritsumei.ac.jp

**Abstract.** We propose a haptic rendering system for a 3D noh-cloth model based on the measurement with the string-based haptic interface device and vibration speakers. In the field of digital archives, high-definition measurement, modeling and rendering of the cultural heritages are very important elements. However, it is not allowed to touch valued cultural heritages in general and it is impossible to measure parameter of these cultural heritages by contiguous measure method. Therefore, we propose a novel system to display both of tactile and kinematic sense for 3D model based on parameters of actual model by noncontact measurement method. This paper describes a measurement and modeling of the noh-cloth with OGM(Optical Gyro Measuring Machine) and a rendering system for a 3D noh-cloth model based on the measurement with the string-based haptic interface device and vibration speakers.

## 1 Introduction

In recent years, it is getting possible to present visually high-definition digital archives of tangible cultural heritages by Computer Vision(CV) and Computer Graphics(CG) technologies [1]. Moreover, the concept of "Digital Museum" has been generated. In this concept, it is expected that visitors can see information about the exhibited object using Augmented Reality(AR) and Virtual Reality(VR) techniques, and can experience interactively touch and feel the exhibited objects by not only vision sense but also haptic sense and audio sense [2].

In the field of digital archive, high-definition measurement, modeling and rendering of the cultural heritages are very important elements. To preserve and reproduce the high-definition model of cultural heritages, it is necessary to display not only based on vision but also haptic sense and it is necessary to create more real and interactive exhibition system. However, it is not allowed to touch valued cultural heritages in general and it is impossible to measure parameter of these cultural heritages by contact-type measure method.

Also in the haptic field, the standard device such as audio speaker and graphical display in the field of acoustic sense and visual sense is required for display

of the tactile and kinematics sense. In the previous work, we proposed the technique for displaying the roughness of textures by using vibration generated by normal of the texture [3]. However, our system was applied the technique to only 2D texture and user can only feel tactile feeling. Therefore, we developed the novel system to display both of tactile and kinematic sense based on noncontact parameters of actual model by noncontact method.

We firstly capture information of surface structure with OGM(Optical Gyro Measuring Machine) and generate the normal map [4] which has asperity information. Secondly, we model tactile sense by vibration signals based on normal map. For display tactile sense of roughness vibration for users, we use vibration speaker which is reasonable and small. Finally, we propose a haptic rendering system for a 3D noh-cloth model based on the measurement with the string-based haptic interface device and vibration speakers.

## 2 Noncontact Method of Measuring and Modeling of Asperity Information

The technique to estimate the normal vector of the surface by analyzing multi-illuminated images has been established in computer vision research [5]. Moreover, it is possible to estimate meso-structures of fabric from high-definition images captured using lens of high power [4].

### 2.1 Environment of Noncontact Measuring

We used OGM (see Fig. 1) and capture the asperity information. The OGM can capture images in any position of incidence and view using totally 4 axis degree of rotational freedom (a two-axis light, a one-axis camera and one-axis stage). Images were captured in a darkened room and metal halide lamp(LS-M180FB) which is close to natural light was used as a light source.

We used a digital camera which has 3888x2592 resolution (Cannon EOS Kiss Digital X), EF100mm F2.8 Macro USM lens and Kenko digital tereplus Teleplus PRO300 for Cannon. Finally getting the camera close up to the object by minimum focal length of macro lens by 0.31m, we captured high-definition image(1 pixel = 0.005mm).

### 2.2 Estimation of Asperity Information

It is known that the half vector between eye vector and light incidence vector is identical to the surface normal vector of the surface. The normal vector  $\mathbf{n}$  is determined by the light position vector  $\mathbf{l}$ (maximum value of reflectance ratio in obtained reflection data) and the camera position vector  $\mathbf{v}$ .

$$\mathbf{n} = \frac{(\mathbf{l} + \mathbf{v})}{|\mathbf{l} + \mathbf{v}|} \quad (1)$$

The obtained normal vectors of the each pixel  $\mathbf{n} = [n_x, n_y, n_z]$  are normalized to  $[0, 255]$  and outputted to the two dimension image called normal map.

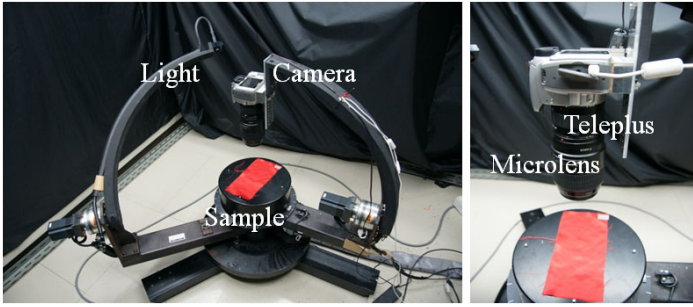


Fig. 1. Optical Gyro Measuring Machine(OGM)

### 3 Displaying Tactile and Haptic Sense Based on Measuring

In this study, we used texture based method[6][7][8]. The algorithm of display haptic sense was shown in Fig. 2

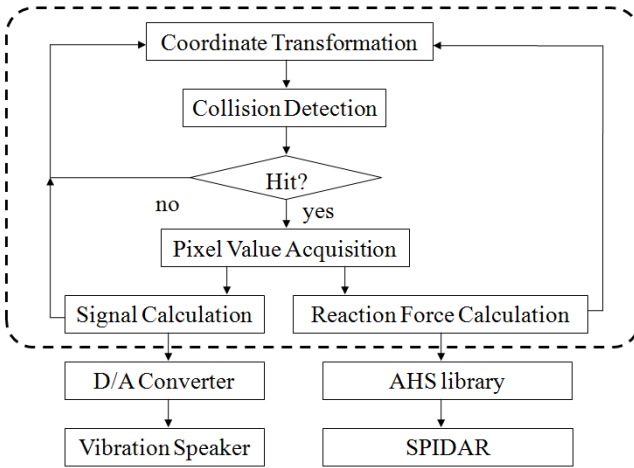


Fig. 2. Haptic Process

#### 3.1 Preprocessing

To touch virtual objects naturally with a haptic device, the tip position of the grip of the haptic device in the device space is required to match the camera position and direction in virtual space. Therefore, the tip position of the grip in device space is transferred to suit the camera space in the virtual space.

### 3.2 Somatosensory Rendering

To display the shape of 3-D virtual object, the collision detection between the tip position of the grip  $\mathbf{p}$  and the object surface is required. We use Möller et al.'s method for the intersection detection [9]. Detecting if the segment from tip position of the grip  $\mathbf{p}$  to the face normal vector of each polygon of virtual object, the polygon(active polygon) which is nearest position from tip position of the grip are exited. If the tip of the grip touches a polygon, the pixel value is taken to use for tactile rendering according to normal map. In our system, the collision detection is done from around the active polygon and calculate the reaction force based on Constraint-based God-object Method [10]. The reaction force for expression of shapes is calculated by the Equ. 2.

$$\mathbf{V}_j = \mathbf{n}_j S d_j + \mathbf{n}_j D \frac{d_j - d_{j-1}}{\Delta t} \tag{2}$$

Where,  $j$  is the update counter of haptic process,  $\mathbf{n}$  is the pixel value of the normal map,  $S$  is given hardness,  $d$  is the penetration depth,  $D$  is dumper invariable,  $\Delta t$  is update counter of haptic process.

### 3.3 Tactile Rendering

It is said that there are four types of mechanoreceptors in the human skin. These receptors are stimulated by mechanical stimulation which is caused by the deformation of the skin and human feel a variety of tactile feeling according to the frequency of the stimulus [11]. This means that human's tactile receptors respond to the changing and human recognize roughness by rather the tiny vibration caused by tracing surface of objects than the 3D shapes itself. It is also said that coetaneous sensory function properly works when it is actively touched to object and human need to trace the object's surface to recognize the texture properly [12]. When the finger traces the object's surface, the normal force is changed according to the surface structure (see Fig. 3).

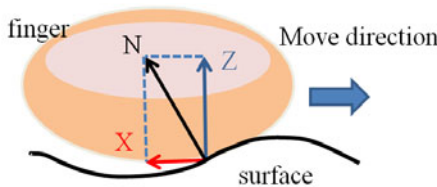


Fig. 3. The Finger Tracing Direction and Normal Force

Therefore, in this study, we generate the vibration signal from the surface normal vector in the direction of finger tracing and display it to the human finger and excite the mechanoreceptors in the skin to display the roughness texture feeling.



In our system, we display the vibration signal transform to voltage and input to vibration speaker. The signal  $s$  is calculated by the inner product finger tracing direction  $\mathbf{m}$  and the normal vector in the contact point of normal map  $\mathbf{n}$  in Equ. 3.

$$s = \mathbf{n}_j \cdot \mathbf{m}_j \tag{3}$$

Where,  $j$  is the update counter of haptic process.

### 4 Multisensory Display System

The overall view of our system is shown in Fig. 4. Our system is composed of a haptic display, a graphical display, an application and a DA converter. Haptic display is composed of two vibration speaker and two kinematic displays SPIDAR-4 [13]. We used TDA-770PCI (Mirco Science Corporation) for Digital Analog conversion. The resolution of output voltage is 12bit, the range of output is -10 to +10V and the maximum update rate is 450kHz.

Users attach the vibration speakers on their index finger and thumb, also attach the cap in the tip of SPIDAR on their fingers.

Users can feel kinematic sense and tactile sense moving the 3D cursor and tracing the 3D virtual objects surface in the graphical display. In the application, user's finger positions and move directions are calculated. The output signal is calculated with the finger's move direction vector and the normal vector of collision detected point in normal map. The calculated signal is converted to

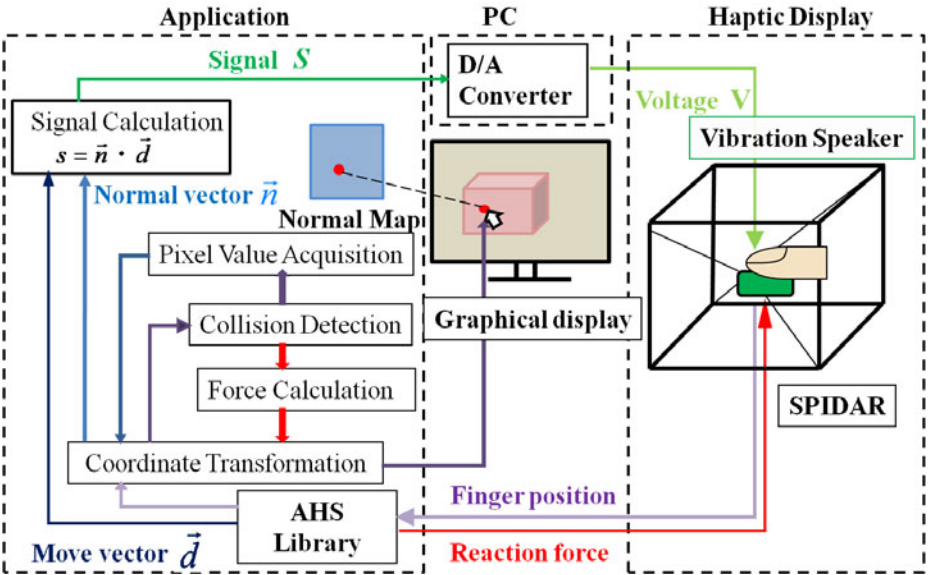


Fig. 4. Multisensory Display System

voltage by DA converter and output from vibration speakers to human skin of fingers as coetaneous sense. Moreover, the reaction force is output to user's finger as somesthetic sense calculating the value of normal and amount of infiltration at the collision point.

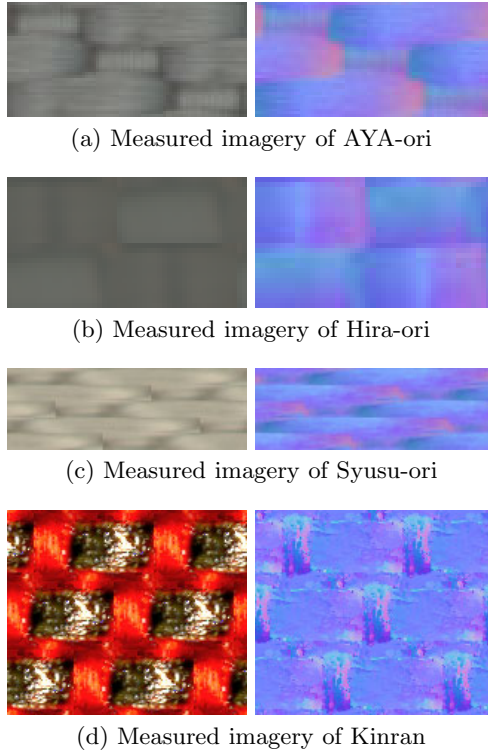
## 5 Multisensory Display for Digital Archive Model

We applied our system to a 3D digital archive of tangible cultural heritage.

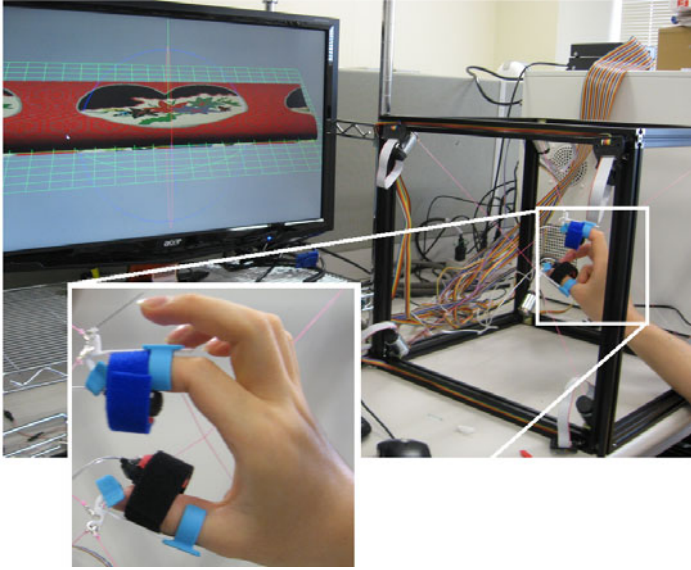
### 5.1 Modeling of Noh-cloth Based on Measuring

We obtained the surface normal map from a piece of actual noh-cloth by analyzing images captured by OGM. The diffuse map and normal map of several fabric structures are shown in Fig. 5.

Fig. 6 shows a 3D noh-cloth model based on the measurement data and our system overview. Users can feel 3D shapes by thumb and index finger with SPIDAR and feel difference of texture by thumb and index finger with vibration speaker.



**Fig. 5.** Measured Imagery of Noh-cloth



**Fig. 6.** Multi-sense Display for Digital Archive Model

This system displays the vibration signal generated from only one point where the cursor contacts.

When users slide their fingers on the texture, vibration is generated according to the normal vector of contact point. However, when they stop their fingers, they feel noting. That is because only one-point information of the texture is used for generating signal. In the real world we contact the surface of objects by the face of finger.

Also in this system, user can feel same feeling at the begging and end of sliding. But in the real world, it is known that the frictional force become altered by velocity of the objects.

Therefore, using the normal information of neighborhood to generate vibration signal and display the change of friction force according to user's finger velocity are future works.

## 6 Conclusion

We proposed the novel method to measure parameters of actual model by non-contact method and we developed a system to display both of tactile and kinematic sense based on measurement. Specifically, we firstly captured information of surface structure with OGM and generate the normal map which has asperity information. Secondly, we modeled tactile sense by vibration signals based on normal map. Moreover, we used the vibration speaker which is reasonable and small to display roughness to the finger of human. Finally, we developed a haptic rendering system for a 3D noh-cloth model based on the measurement

with the string-based haptic interface device and vibration speakers. Our system enabled users to feel 3D shapes by thumb and index finger with SPIDAR and feel difference of texture by thumb and index finger with vibration speaker.

As future works, we consider how to render the sense of face-contact, how to capture and display the friction and elasticity by noncontact method and apply this system for various textures.

## References

1. Takeda, Y., Tanaka, H.T.: Multi-resolution anisotropic btf modeling of gold brocade fabrics based on multi-illuminated hdr image analysis. *The Transactions of the Institute of Electronics, Information and Communication Engineers* 91, 2729–2738 (2008)
2. Hirose, M.: Digital museum project. *The Journal of the Institute of Image Information and Television Engineers* 64, 783–788 (2010)
3. Nomura, K., Yin, X., Sakaguchi, Y., Tanaka, H.T.: Modeling system of tactile feeling based on photometric images analysis. *The Transactions of the Institute of Electronics, Information and Communication Engineers D* 93 (2010) (in press)
4. Ozaki, R., Nishiwaki, Y., Takeda, Y.: Weave pattern modeling of silk-like fabrics from multi-illuminated hdr image analysis. *Transactions of the Virtual Reality Society of Japan* 14, 315–324 (2009)
5. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. *Optical Engineering* 19, 139–144 (1980)
6. Wakita, W., Ido, S.: A haptic rendering for high polygon model using distance map and normal map. *Transactions of Information Processing Society of Japan* 49, 2509–2517 (2008)
7. Wakita, W., Ido, S.: A material system under haptic rendering for pseudo-roughness. *The Transactions of the Institute of Electronics, Information and Communication Engineers J91-D*, 2061–2070 (2008)
8. Wakita, W., Murakami, K., Ido, S.: Development of a texture-based haptic modeling system. *The Transactions of the Institute of Electronics, Information and Communication Engineers J91-D*, 2773–2780 (2008)
9. Möller, T., Trumbore, B.: Fast, minimum storage ray-triangle intersection. *Journal of Graphics Tools* 2, 21–28 (1997)
10. Zilles, C., Salisbury, K.: A constraint-based god-object method for haptic display. In: *Proc. IEEE/RSJ International Conference Intelligent Robots and Systems*, vol. 3, pp. 146–151 (1995)
11. Iwamura, Y.: Touch. Igaku-Shoin Ltd. (2001)
12. Iwamura, Y.: Neural mechanisms and modeling of active touch. *Journal of the Society of Instrument and Control Engineers* 41, 728–732 (2002)
13. Sato, M., Hirata, Y., Kawaharada, H.: Space interface device for artificial reality—spidar—. *The Transactions of the Institute of Electronics, Information and Communication Engineers J74-D-II*, 887–894 (1991)

# A Texture-Based Direct-Touch Interaction System for 3D Woven Cultural Property Exhibition

Wataru Wakita<sup>1</sup>, Katsuhito Akahane<sup>2</sup>,  
Masaharu Isshiki<sup>3</sup>, and Hiromi T. Tanaka<sup>1</sup>

<sup>1</sup>Department of Human and Computer Intelligence, College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan

{wakita,hiromi}@cv.ci.ritsumei.ac.jp

<sup>2</sup>Precision and Intelligence Lab, Tokyo Institute of Technology, Yokohama, Japan

kakahane@hi.pi.titech.ac.jp

<sup>3</sup>Department of Electrical and Electronic Engineering and Computer Science, Ehime University, Matsuyama, Japan

isshiki@cs.ehime-u.ac.jp

**Abstract.** We propose a texture-based direct-touch interaction system for 3D woven cultural property exhibition of the “Tenmizuhiki” tapestries “Hirashaji Houou Monyou Shishu” of “Fune-hoko” of “Gion Festival in Kyoto”. In the field of digital archive, it is important to archive and represent the cultural property at the high-definition. To archive the shape, color and texture of the cultural property, it is important to archive and represent not only visual effect but haptic impression. Recently, in the field of haptics, various haptic rendering devices have been developed, and various haptic rendering techniques to touch the virtual object have been proposed. In haptic rendering for the high-definition virtual object, it is difficult to render the haptic impression smoothly and calculate the reaction force at realtime. Therefore, it is require the realtime and high-definition haptic rendering techniques. In our previous work, we proposed a texture-based haptic modeling and rendering techniques at realtime and with high-definition. However, our techniques are not based on the measurement. Moreover, in the field of digital archive, it is necessary to represent the digital archived cultural property intuitively and interactively. Therefore, we applied our texture-based haptic modeling and rendering techniques for the digital archive, and we developed a realtime and direct-touch interaction system for 3D cultural property exhibition.

## 1 Introduction

In the field of digital archive, it is important to archive and represent the cultural property at the high-definition. To archive the shape, color and texture of the cultural property, it is important to archive and represent not only visual effect but haptic impression. Recently, various haptic rendering devices have been

developed, and various haptic rendering techniques to touch the virtual object have been proposed.

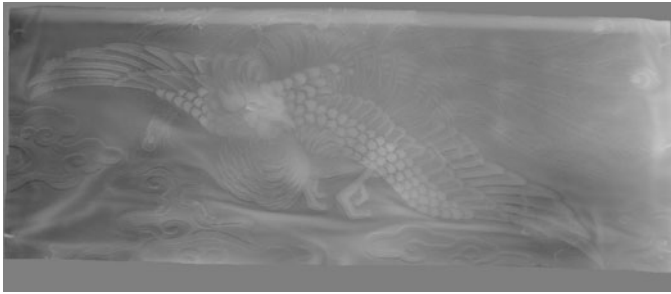
Penalty-based haptic rendering technique [1][2] is a basic approaches to represent the polygon wall, has several problems such as passing through, discontinuous force and vibration. To solve these problems, Zilles *et al.* proposed a constraints-based God-object method [3]. However, their method has the same problems such as passing through, discontinuous force and vibration in haptic rendering for the high-definition virtual object. In haptic rendering for the high definition virtual object, it is difficult to render the haptic impression smoothly and calculate the reaction force at realtime. On the other hand, several texture-based haptic rendering techniques have proposed to represent the asperity of the interior of the polygon according to the 2D image. Stanney proposed a force mapping technique [4] which enables representation of the gradient of the object surface according to the force map. In his approach, the direction of the reaction force is dynamically perturbed according to the pixel value of the interior of the polygon which mapped the force map. Theoktisto *et al.* proposed a height field mapping technique [5] which enables representation of the height of the object surface according to the height field map. In their approach, the surface height is dynamically changed according to the pixel value of the interior of the polygon which mapped the height field map. We proposed a texture-based haptic rendering technique for the pseudo-roughness on the surface of the low-polygon virtual object using height map and normal map [6], and we developed a material system under haptic rendering for pseudo-roughness on the low-polygon object surface [7]. In this system, difference of the haptic impression is represented by changing magnitude and/or direction of the reaction force dynamically according to the pixel value of the object surface which mapped the special texture images which converted asperity, stiffness and friction into the 2D image. Moreover, we have proposed a realtime haptic rendering technique for representation of the shape of the high-definition virtual object using the low-definition virtual object, distance map and normal map [8]. In this approach, the reaction force is calculated according to the pixel value of the low polygon object surface which mapped the special texture image which converted the geometric difference of the high polygon model and the low polygon model into the 2D image.

However, these techniques are not based on the measurement. To represent the high-definition virtual object, it is necessary to model the virtual object based on the measurement. The same can be said for digital archive. Moreover, in the field of digital archive, it is necessary to represent the digital archived cultural property intuitively and interactively.

Therefore, we applied our texture-based haptic modeling and rendering techniques for the digital archive, and we developed a realtime and direct-touch interaction system for 3D cultural property exhibition. Specifically, firstly we measured and modeled the woven cultural property “Hirashaji Houou Monyou Shishu” of “Fune-hoko” of “Gion Festival in Kyoto” [9]. Secondly, we developed an exhibition system with the stereoscopic projector and string-based haptic interface device “SPIDAR” based on our texture-based techniques.

## 2 Digital Archiving and Modeling

We used the laser range scanner “VIVID” [10] for the measurement of the shape, and we used a high-resolution multiband imaging camera for measurement of the color and spectral reflectance. The measured range data have  $1193 \times 512$  vertices, and measured color image data have  $13650 \times 5370$  pixels. We converted the measurement range data to a height map. Fig. 1 shows measurement data of the woven cultural property “Hirashaji Houou Monyou Shishu”. Fig. 1(a) shows a height image data (height map) that was generated from measured range data by the laser range scanner, and Fig. 1(b) shows a color image data (color map) by the multiband camera.



(a) measured height image data



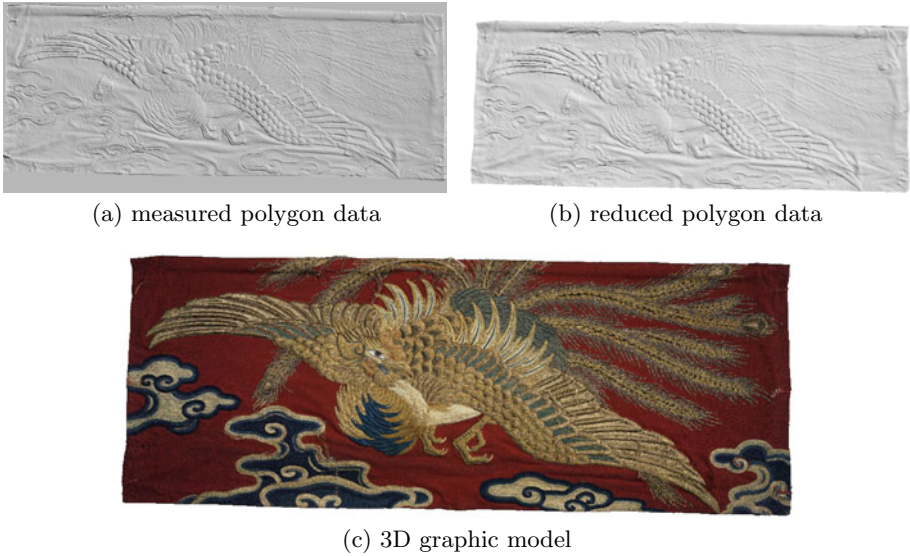
(b) measured color image data

**Fig. 1.** Measured 2D Image Data of the “Hirashaji Houou Monyou Shishu”

### 2.1 3D Graphic Modeling

We created a graphic model based on measurement data. Firstly, we converted a 2D height map to a 3D polygon model which have 612,522 vertices and 1,221,632 triangles (see Fig. 2(a)). Secondly, to reduce the graphic rendering cost, we reduced a 3D polygon model to 132,554 vertices and 263,484 triangles (see Fig. 2(b)). Finally, we mapped a measured 2D color map to a reduced 3D polygon model. Fig. 2(c) shows a 3D graphic model of the “Hirashaji Houou Monyou Shishu”.



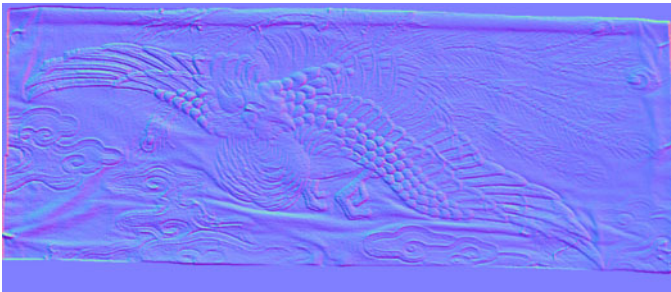


**Fig. 2.** 3D Graphic Model of the “Hirashaji Houou Monyou Shishu”

## 2.2 3D Haptic Modeling

To reduce the haptic rendering cost, we used our texture-based haptic modeling and rendering technique [11].

Firstly, we created a surface gradient image data (normal map) (see Fig. 3) from height map (see Fig. 1(a)).



**Fig. 3.** Normal map

This normal map is used to represent the surface gradient, where the RGB values correspond to the XYZ coordinates of the normal vector. The height map is used to represent the surface height, where the surface height is changed according to the grayscale value (white is high and black is low elevation).

Secondly, we created a friction map (see Fig. 4) with a polarization plate, camera, and multiple light source.



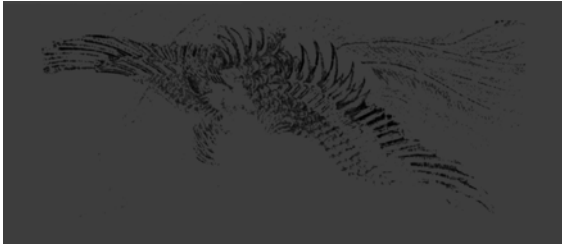


Fig. 4. Friction map

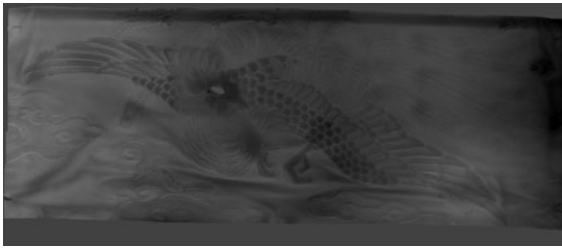
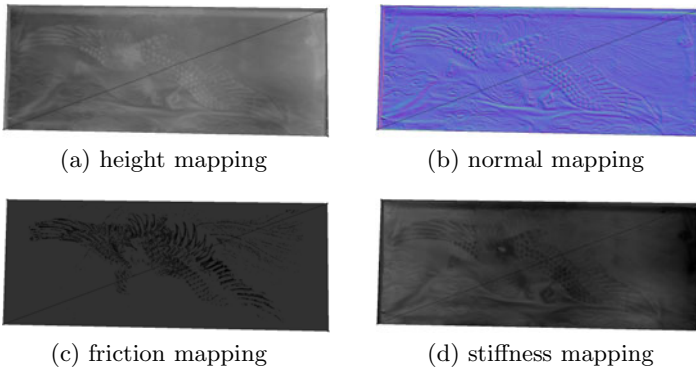


Fig. 5. Stiffness map



(a) height mapping

(b) normal mapping

(c) friction mapping

(d) stiffness mapping

Fig. 6. 3D Haptic Model of the “Hirashaji Houou Monyou Shishu”

This friction map is used to represent the surface friction, where the surface friction is changed according to the grayscale value (white is rough and black is smooth). Generally, the tangible haptic sensor is used for the friction measurement. However, it is difficult to touch the valuable cultural properties. Therefore, noncontact measurement method is necessary. Currently, we estimate the friction parameter based on the reflection component of the cultural property by the noncontact measurement. For example, easily reflectable (high specular reflection) area is flat and smoothly touchable, and low specular reflection area is rough.

Thirdly, we created a stiffness map (see Fig. 5) based on the height map (see Fig. 4(a)).

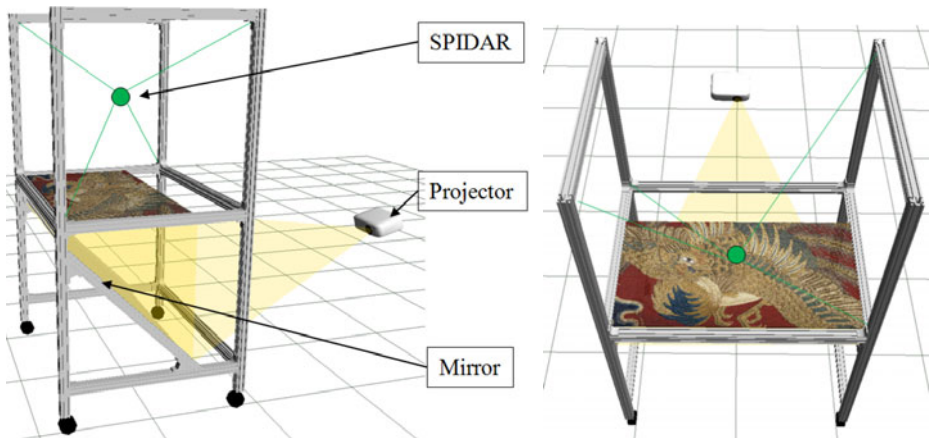
This stiffness map is used to represent the surface stiffness, where the surface stiffness is changed according to the grayscale value (white is hard and black is soft). Currently, we have not measured stiffness parameter by the noncontact method. “Hirashaji Houou Monyou Shishu” is wadded with cotton and eye part is made of glass. Therefore, we assume that the woven cultural property is placed on the floor, and we currently estimate that higher area (cotton) is soft and lower area (floor) and eye part are hard.

We mapped these maps to 2 polygons square model (see Fig. 6). This square model is used for the haptic rendering, and is not used for the graphic rendering.

### 3 Realtime and Direct-Touch Interaction System for 3D Woven Cultural Property Exhibition

#### 3.1 System Architecture

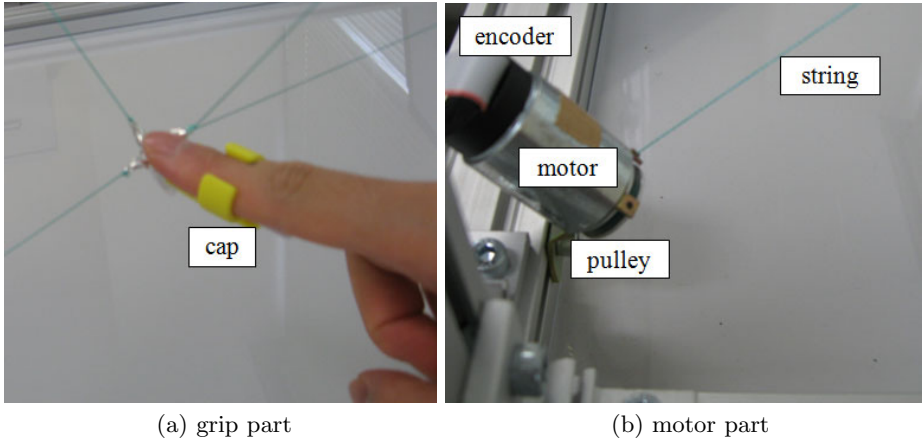
We developed a direct-touch interaction system for the digital archived 3D woven cultural property “Hirashaji Houou Monyou Shishu” exhibition (see Fig. 7).



**Fig. 7.** Direct-touch Interaction System for 3D Woven Cultural Property “Hirashaji Houou Monyou Shishu” exhibition

Our system is composed of a display system and an 3D application. A display system consists of a graphic part and haptic part. In graphic part, we used a rear projector screen (1000mm×750mm) and the stereoscopic projector “DepthQ HD”. The stereoscopic vision is projected to the bottom projector screen with a mirror.

A haptic part is on top of a projector screen, and at one with a graphic part. In haptic part, we used the string-based interface device “SPIDAR-4” [12]. The



**Fig. 8.** Haptic Part with the String-based Haptic Interface Device “SPIDAR-4”

SPIDAR-4 has ability to control the 3DOF position and to present the 3DOF forces. We used a finger cap to grip part. A tip of the cap is attached to 4 strings from 4 motors with an encoder (see Fig. 8). The strings length got from each encoder’s data is used to measure the grip’s position. The strings tension from each motor is displayed the feedback forces.

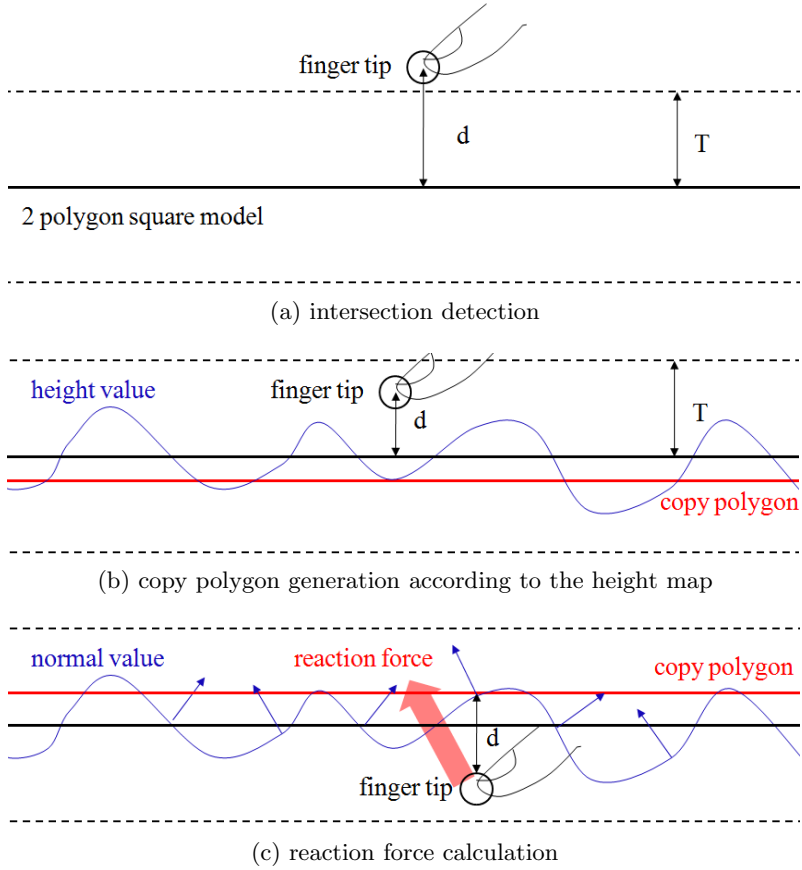
In 3D application, we used the OpenGL and OpenCV library for the graphic API, and we used the AHS library for the haptic API. SPIDAR is controlled via the AHS library.

### 3.2 Direct-Touch and Texture-Based Haptic Rendering

To direct-touch the virtual objects naturally with the SPIDAR, the finger tip in the device space is required to match the camera space (camera position and direction) in virtual space. Therefore, the finger tip position in device space is converted to the camera space in the virtual space.

In haptic rendering, the reaction force is calculated by using a 2 polygons square model which mapped haptic textures. Our haptic rendering technique is based on a constraint-based God-object method [3].

Firstly, the intersection is detected between the finger tip and an invisible 2 polygons square model on the screen (see Fig. 9(a)). We used Möller *et al.*’s method [13] for the intersection detection. Secondly, if they are crossed in the intersection detection and have the possibility of contact ( $d < T$ ), the polygon height is changed according to the pixel value of the height map in relation to intersection point and the polygon is replicated (see Fig. 9(b)). Finally, the intersection is detected again between the finger tip and a copy polygon, and the reaction force is calculated according to the penetration depth and the pixel value of the friction map and stiffness map. The direction of the reaction force is perturbed according to the pixel value of the normal map (see Fig. 9(c)).



**Fig. 9.** Direct-touch and Texture-based Haptic Rendering Technique

If  $(r_j \leq p_j)$ , the reaction force  $F$  is calculated as follows:

$$N_j = n_j s_j d_j + n_j C \frac{d_j - d_{j-1}}{\Delta t} \quad (1)$$

$$S_j = m_r r_j + m_j C \frac{r_j - r_{j-1}}{\Delta t} \quad (2)$$

$$F_j = N_j - S_j \quad (3)$$

else,

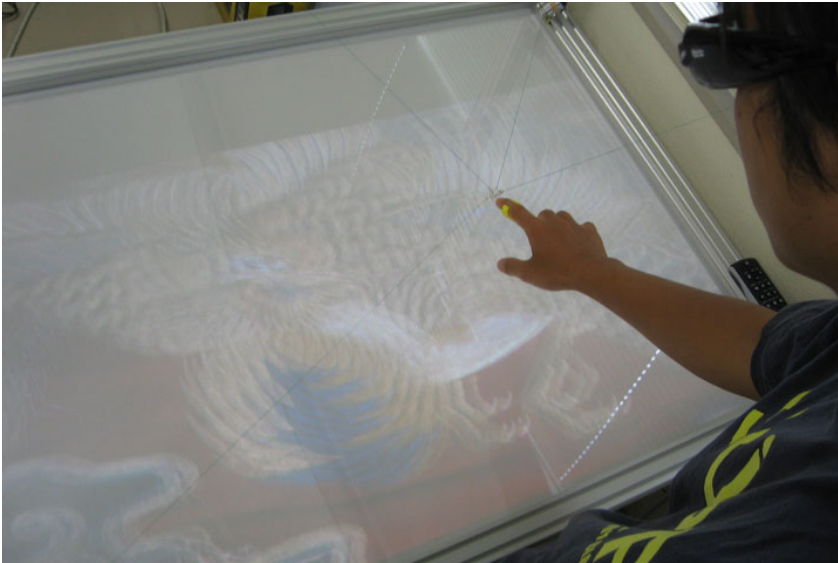
$$F_j = N_j \quad (4)$$

where,  $N_j$  is the normal force for the shape rendering,  $S_j$  is the friction force,  $j$  is the update counter in the haptic process,  $r$  is the horizontal penetration depth in static point,  $p$  is the pixel value of the friction map,  $n$  is the pixel value of the normal map,  $s$  is the pixel value of the stiffness map,  $d$  is the penetration depth,

$C$  is the constant of damper,  $\Delta t$  is the update rate of haptic process, and  $m$  is the slip direction.

## 4 Results

Fig. 10 shows our realtime and direct-touch interaction system for the 3D woven property “Hirashaji Houou Monyou Shishu” exhibition based on our texture-based haptic modeling and rendering technique. In our system, we used two 2.33 GHz Intel(R) Xeon(R) CPU E5410, NVIDIA Quadro FX 580 graphics card with 512MB video memory, 16GB RAM, Windows VISTA 64bit, and NVIDIA 3D Vision. The graphic process is 120Hz update rate, and haptic process is 1kHz update rate. Our system enabled a realtime and direct-touch for the stereoscopic vision comes to the surface on the screen with SPIDAR.



**Fig. 10.** Realtime and Direct-touch Interaction System for 3D Woven Cultural Property “Hirashaji Houou Monyou Shishu”.

## 5 Conclusion and Future Work

We proposed a texture-based direct-touch interaction system for the 3D woven cultural property exhibition. Specifically, firstly we archived the cultural property “Tenmizuhiki” tapestries “Hirashaji Houou Monyou Shishu” of “Fune-hoko” of “Gion Festival in Kyoto”. Secondly, we developed a exhibition system with the stereoscopic projector and string-based haptic interface device “SPIDAR” based on our texture-based technique. However, in our system, the stiffness properties are not based on the measurement data. Therefore, we plan to measure various materials such as cloth, plastic, soil, and etc. of the stiffness properties.

## References

1. Mark, W.R., Randolph, S.C., Finch, M., Verth, J.M.V., Taylor II, R.M.: Adding force feedback to graphics systems: Issues and solutions. In: Proceedings of SIGGRAPH 1996, pp. 447–452 (1996)
2. Massie, T.H., Salisbury, J.K.: The phantom haptic interface: A device for probing virtual objects. In: Proceedings of ASME Winter Annual Meeting, Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, vol. 55-1, pp. 295–301 (1994)
3. Zilles, C., Salisbury, K.: A constraint-based god-object method for haptic display. In: Proceedings of IEEE/RSJ International Conference Intelligent Robots and Systems, pp. 146–151 (1995)
4. Stanney, K.: Handbook of Virtual Environments, pp. 117–134. Lawrence Erlbaum Associates, Inc., Mahwah (2001)
5. Theoktisto, V., Fairén, M., Navazo, I., Monclús, E.: Rendering detailed haptic textures. In: Proceedings of 2nd Workshop in Virtual Reality Interactions and Physical Simulations (VRIPHYS 2005), pp. 16–23 (2005)
6. Wakita, W., Mitani, H., Ido, S.: Haptic rendering for pseudo-roughness using height map and normal map. *Informaton Technologies Letters* 6, 347–350 (2007)
7. Wakita, W., Ido, S.: A material system under haptic rendering for pseudo-roughness. *The Transactions of the Institute of Electronics, Information and Communication Engineers J91-D*, 2061–2070 (2008)
8. Wakita, W., Ido, S.: A haptic rendering for high polygon model using distance map and normal map. *IPSJ Journal* 49, 2509–2517 (2008)
9. Tanaka, H.T., Yano, K., Hachimura, K., Nishiura, T., Choi, W., Fukumori, T., Furukawa, K., Wakita, W., Tsuchida, M., Saiwaki, N.: “gion festival in kyoto”: Reproduction of “fune-boko” float of the gion festival parade in “virtual kyoto”. In: Proceedings of ASIAGRAPH (2010)
10. VIVID, <http://www.ksdl.co.jp/product/scanner/vivid.html>
11. Wakita, W., Murakami, K., Ido, S.: Development of a texture-based haptic modeling system. *The Transactions of the Institute of Electronics, Information and Communication Engineers J91-D*, 2773–2780 (2008)
12. Sato, M., Hirata, Y., Kawaharada, H.: Space interface device for artificial reality—spidar—. *The Transactions of the Institute of Electronics, Information and Communication Engineers J74-D-II*, 887–894 (1991)
13. Möller, T., Trumbore, B.: Fast, minimum storage ray-triangle intersection. *Journal of Graphics Tools* 2, 21–28 (1997)

# High Dimensional Correspondences from Low Dimensional Manifolds – An Empirical Comparison of Graph-Based Dimensionality Reduction Algorithms

Ribana Roscher, Falko Schindler, and Wolfgang Förstner

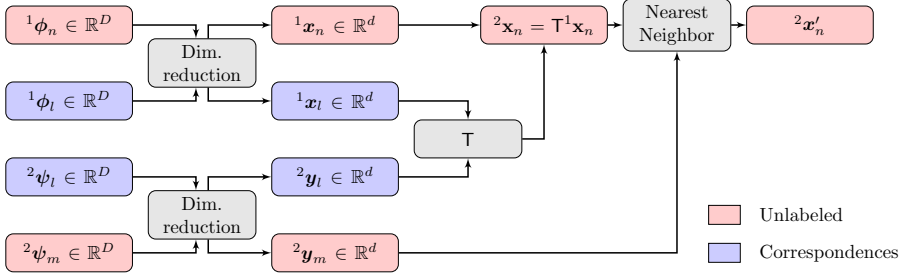
Department of Photogrammetry, Institute of Geodesy and Geoinformation,  
University of Bonn

**Abstract.** We discuss the utility of dimensionality reduction algorithms to put data points in high dimensional spaces into correspondence by learning a transformation between assigned data points on a lower dimensional structure. We assume that similar high dimensional feature spaces are characterized by a similar underlying low dimensional structure. To enable the determination of an affine transformation between two data sets we make use of well-known dimensional reduction algorithms. We demonstrate this procedure for applications like classification and assignments between two given data sets and evaluate six well-known algorithms during several experiments with different objectives. We show that with these algorithms and our transformation approach high dimensional data sets can be related to each other. We also show that linear methods turn out to be more suitable for assignment tasks, whereas graph-based methods appear to be superior for classification tasks.

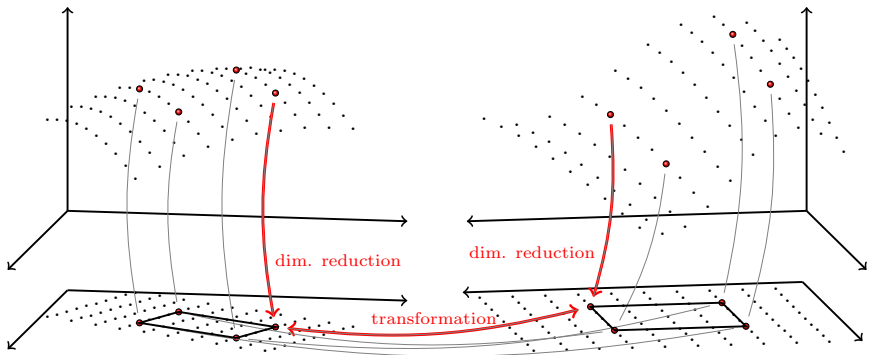
## 1 Introduction

Applications of methods for dimensionality reduction are widely spread in the field of computer vision, namely concerning problems of detection, tracking, recognition, segmentation and reconstruction [1,2,3,4]. Establishing the correspondence between two sets of data points in high dimensional spaces can efficiently be achieved by dimensionality reduction of both sets and establishing the correspondence in the subspaces using a transformation between these subspaces. A suitable transformation is a hyper-plane-preserving mapping between the subspaces, e. g. an affine mapping. The goal of this paper is to compare the quality of six well known algorithms for dimensionality reduction based on an affine transformation between the corresponding subspaces. Only a small subset of the two sets is assumed to be known, however the dimensionality reduction schemes make use of all data points in a subset.

Our problem shown schematically in Fig. 1 can be stated as follows: We have given a set  $\mathcal{S}_1$  containing  $N$  unlabeled data points  $\{\phi_n\}$ ,  $n = 1, \dots, N$  and  $L$



**Fig. 1.** Schematic overview for putting data points in high dimensional spaces into correspondence by learning an affine transformation between correspondences on a lower dimensional structure



**Fig. 2.** We apply dimensionality reduction on two similar data sets living in two different high dimensional spaces, exploiting the structure of all available data. We estimate a unique, affine transformation between both reduced spaces using only few correspondences. The transformation is illustrated by the affine-distorted polygon. This way a unlabeled test point from one data set can be assigned to its nearest neighbor within the other set without labeling all data.

labeled correspondences  $\{\phi_l^1\}$ ,  $l = 1, \dots, L$  in  $\mathbb{R}^D$  and a set  $\mathcal{S}_2$  containing  $M$  unlabeled data points  $\{\psi_m^2\}$ ,  $m = 1, \dots, M$  and  $L$  labeled correspondences  $\{\psi_l^2\}$ ,  $l = 1, \dots, L$  in  $\mathbb{R}^D$ . We want to find lower dimensional data points  $\{x_n^1, x_l^1\}$  from  $\mathcal{S}_1$  and  $\{y_m^2, y_l^2\}$  from  $\mathcal{S}_2$  in  $\mathbb{R}^d$  with  $d \ll D$  so that  $\{x_n^1, x_l^1\}$  is an appropriate representation of  $\{\phi_n^1, \phi_l^1\}$  and  $\{y_m^2, y_l^2\}$  of  $\{\psi_m^2, \psi_l^2\}$ .

Having points  $\{x_l^1\}$ ,  $\{y_l^2\}$  from two similar data sets, a unique hyper-plane preserving transformation  $T$  from one space into another is to be determined. Therefore, a small amount of labeled correspondences  $\{y_l^2, x_l^1\}$  can be used. Fig. 2 shows how both data sets can be related via the low dimensional space. The unlabeled, transformed points  $\{x_n^2\}$  from set  $\mathcal{S}^1$  can be assigned to its nearest neighbor within  $\mathcal{S}_2$  yielding  $\{x_n^2'\}$ .

We investigate and evaluate six popular, linear and graph-based methods for dimensionality reduction in an extensive test framework. Assuming to have only



few correspondences with geometric information, yet a large number of unlabeled data, we consider only unsupervised dimensionality reduction algorithms.

We relate both reduced data sets to each other with a linear, affine transformation matrix. This matrix can be determined in case at least  $d(d+1)$  common data points are available. More complex transformations can also be considered, but this is not scope of this work.

Using the dimensional reduction followed by an affine transformation we can relate every new test point to its nearest neighbor within the other data set using Euclidean distances, without the need of a complete set of labels and a full  $D \times D$  transformation matrix. We investigate the quality of these assignments depending on the dimensionality reduction algorithm and several parameters.

Our contribution is to show that using such algorithms and the presented transformation approach we can relate high dimensional data sets to each other with a minimal amount of correspondences. This can be used for the assignment of data points from different spaces as well as for the classification of images. The approach is independent of the used dimensionality reduction algorithm and can be used with affine as well as more complex transformations. Our further contribution is to show empirical investigations on the standard dimensionality reduction algorithms.

In Section 2 we give an overview of the related work with special focus on spectral methods. Section 3 briefly illustrates representative linear and graph-based state-of-the-art spectral methods, which we analyze in our experiments. In Section 4 we explain how to relate two data sets using an affine transformation. Section 5 compares six spectral methods in order to find preferably robust and low dimensional structures in a semi-supervised manner. We evaluate the algorithms on handwritten and computer digits and on cartoon images with and without glasses. In the last section we discuss our results and provide an outlook to future work.

## 2 Related Work

The field of computer vision offers numerous feature selection algorithms and extraction methods. Spectral methods constitute a group specialized in reducing the dimensionality of data. We distinguish between two major types of algorithms:

1. Linear methods including Principal Component Analysis (PCA) [5] and Multidimensional Scaling (MDS) [6] and
2. Nonlinear methods including graph-based methods (e. g. [7,8,9,10,11]), and kernel methods (e. g. [12,13]). In contrast to linear methods the nonlinear methods perform better on complex nonlinear data structures as they occur in real world data sets.

In the last years further algorithms closely related to the latter have been developed focusing on acceleration (e. g. [14,15]) and qualitative improvement (e. g. [16,17,18]).

Typical data sets used in the aforementioned methods are pictures of an object subject to changing illumination, angle, translation or other varying characteristics. The data points defined by the vectorized pixel intensities of each image vary smoothly so that they define a manifold in a high dimensional space.

For example, Roweis and Saul [19] demonstrate the Isomap algorithm employing it on face images with varying illumination conditions and angles, on hand images with natural hand movements and also on handwritten digits.

A preliminary work on learning high dimensional correspondences from low dimensional manifolds has been done by Ham et al. [20], who extends Locally Linear Embedding to handle constraints introduced by correspondences. They show on several datasets that the constrained Locally Linear Embedding gives better reconstruction errors than supervised algorithms, factor analysis and a model similar to the bilinear model proposed by Tenenbaum and Freeman to separate style from content [21]. Also De la Torre and Black [22] proposed a method to find a common manifold for learning asymmetrically coupled linear models. We follow the approach of Wang and Mahadevan [23], which uses Laplacian Eigenmaps to determine the low dimensional points for each data set and Procrustes analysis to align these to each other. We extend this approach using an affine transformation, which preserves hyper-planes and compare well-known dimensionality reduction algorithms regarding accuracy in two experiments.

### 3 Spectral Methods for Dimensionality Reduction

Spectral methods are a class of techniques used for dimensionality reduction. The reduction is done by detecting a low dimensional structure in a higher-dimensional space by decomposing a specially constructed matrix, which is mostly a weighted graph of the initial data. Spectral methods are convex and therefore optimize an objective function globally.

In contrast to manifold learning, where some representation for the underlying manifold  $f : f(\phi) = 0$  is estimated, dimensionality reduction only considers the estimation of lower-dimensional data points  $\{\mathbf{x}_n\}$  from the input data points  $\{\phi_n\}$ . Consequently a transformation back into high dimensional space is non-trivial but not necessary in our context. We stay with the output points and assign nearest neighbors after transforming one set of points into the other lower dimensional space.

#### 3.1 Linear Methods

Generally, linear methods retrieve a structure of the lower dimensional data points  $\{\mathbf{x}_n\}$  lying close to a linear affine subspace of the high dimensional space. The methods yield data points  $\mathbf{x}_n = \mathbf{r}_1^* \phi_{n,1} + \mathbf{r}_2^* \phi_{n,2} + \dots + \mathbf{r}_M^* \phi_{n,M} = \mathbf{R}^* \phi_n$ , which are  $d$ -dimensional linear combinations of the original  $D$ -dimensional data points  $\phi_n$  with  $\mathbf{R}^*$  being the  $(d \times D)$ -dimensional matrix for the linear transformation. The star indicates the reduced dimensionality in contrast to  $\mathbf{R}$  being a square matrix. For the combined point matrices we obtain  $\mathbf{X} = \mathbf{R}^* \Phi$ .

We consider two state-of-the-art subspace methods: Principal Component Analysis (PCA) [5], and Metric Multidimensional Scaling (MDS) [6]. Since many years they are used widely in the field of pattern recognition.

*PCA* reduces the dimensionality while preserving the global covariance structure of all data points. We can compute the lower dimensional data points  $\mathbf{x}_n$  by mapping them onto the  $M$  basis vectors  $\mathbf{r}$  with the largest eigenvalues  $s$ :  $\mathbf{R}^* = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_M]^\top$ . The latter are derived from the eigen decomposition of the covariance matrix  $\Sigma_{\phi, \phi} = \mathbf{R} \mathbf{S}^2 \mathbf{R}^\top$ .

*MDS* reduces the dimensionality while preserving the inner products between the data points by decomposing the Gram matrix:  $K_{nm} = \phi_n \cdot \phi_m$ , having the same eigenvalues as the covariance matrix of the PCA up to a constant in the classical setup. Therefore, the output of classical MDS is identical to that of the PCA. Modern MDS algorithms use iterative methods, so that the points are better arranged.

The main drawback of both methods is that they retain large distances, which may not reflect the correct metric or even maybe outliers, and do not consider the local distribution of the neighborhood around data points. Therefore important structures can be lost like in the Swiss roll data set [7].

### 3.2 Graph-Based Methods

If the structure underlying the data is not affine, linear methods can fail. Graph-based methods can find this structure even if the data is lying within or close to a low dimensional manifold. The key aspect of these algorithms is to preserve local topological and geometrical properties.

These methods can be divided into three parts:

1. Construct a graph  $\mathcal{G}$  with nodes representing the data points  $\Phi$  and edges defining relations between them. Each node is connected to all data points within a local  $\epsilon$ -neighborhood or to its  $k$ -nearest neighbors.
2. A matrix  $W$  is derived from the graph  $\mathcal{G}$  by choosing weights, e. g.  $w_{nm} = 0$  if there is no connection between points  $n$  and  $m$  and  $w_{nm} = 1$  or some distance measure  $w_{nm} = d(\phi_n, \phi_m)$  if there is one.
3. In the last step a matrix including the weights  $W$  is decomposed. The way of how to use  $W$  mainly makes up the difference between the algorithms.

We consider four representative state-of-the-art graph-based methods: Isometric Mapping (Isomap) [7], Locally Linear Embedding (LLE) [8,19], Laplacian Eigenmaps [9] and Local Tangent Space Alignment (LTSA) [17].

*Isomap* preserves pairwise distances between data points  $\{\phi_n\}$  along an estimated manifold. In principle the Isomap algorithm equals MDS, whereby the Euclidean distances are replaced by geodesic distances. Isomap may suffer from holes within the data structure and so called short-circuiting, e. g. misleading connections to topologically separated points.

*LLE* preserves local linear structure of nearby data points. After decomposing the matrix  $M = (I - W)^T(I - W)$  with  $I$  being the identity matrix the largest eigenvector is discarded and the remaining ones yield the lower-dimensional data points. The neighborhood of every point is assumed to be planar. Despite of its good performance in a wide variety of applications LLE tends to cluster dense regions of the data and can hardly handle holes.

*Laplacian Eigenmaps* preserves so called proximity relations: Nearby input data points  $\{\phi_n\}$  are projected to nearby output data points  $\{x_n\}$ . They minimize the gradient norm in a least squares sense by decomposing the matrix  $L = I - D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ , which is also called Graph Laplacian. Using the  $d + 1$  largest eigenvectors of the matrix yield the  $M$ -dimensional data points, whereby the largest eigenvector is discarded. The diagonal matrix  $D$  has elements  $D_{nm} = \sum_m W_{nm}$ . The algorithm suffers from similar drawbacks like LLE.

*LTSA* preserves the geometry within the tangent space at each data point. The method approximates the local tangent space of each neighborhood of a point. The local tangent space is aligned and embedded in a global coordinate system. As we can conclude from our experiments described in Section 5 LTSA as well as LLE cause high computational costs due to their complexity.

## 4 Transformations between Different Subspaces

Given two sets of data points written with homogeneous coordinate vectors  ${}^1x_l^T = [{}^1x_l^T, 1]$ ,  ${}^2y_l^T = [{}^2y_l^T, 1]$ ,  $l = 1, \dots, L$  an linear, affine transformation  $T : {}^2y_l = T^1x_l$  is to be determined. In terms of the combined coordinate matrices  ${}^1X$  and  ${}^2Y$  the homogeneous representation is  ${}^2Y = T^1X$ .

Since the transformation is affine, we need at least corresponding  $d(d + 1)$  points. We can multiply with the pseudo inverse of  ${}^2X$  to obtain  $T = {}^1Y^2X^+$ .

## 5 Experiments

In our experiments we compare six well-known subspace methods in our test framework described in Fig. 2: PCA, MDS, Isomap, LLE, Laplacian Eigenmaps and LTSA. We analyze the influence of changing parameters like the number of labeled points, the number of used neighbors  $k$  and the target dimension  $d$ . The implementations of all graph-based algorithms are kindly provided by the authors. For the PCA we use a fast implementation from Mark Tygert [1] and the MDS implementation is an iterative version written by Michael Lee [24].

As depicted in Fig. 3 we use two pairwise similar data sets in our experiments.

*Glasses* concerns the problem of occlusions containing 1626 semi-automatically created Simpsons avatars [2], each pair containing one face with and without

<sup>1</sup> <http://www.mathworks.de/matlabcentral/fileexchange/21524-principal-component-analysis>

<sup>2</sup> <http://www.simpsonsmovie.com/>



**Fig. 3.** Example images for both data sets used in this paper: faces with/without glasses (left) and handwritten/digital digits (right)

glasses. By searching for the nearest neighbor in the space of non-glass images we remove glasses from these cartoon faces.

*Digits* are 1900 handwritten digits [25] and 2940 digital ones. We use dimensionality reduction methods and the estimated transformation to classify handwritten images by finding corresponding digital versions.

We reduce the dimensionality of both data sets to the same target dimension and compute the transformation between both subsets.

After determining the transformation between both low dimensional spaces the unlabeled test images from one subspace are transformed into the other subspace. There they are assigned to their nearest neighbors to yield an approximate relative position to other images within the high dimensional space. The labels, i. e. the unique ID, of each paired test images are compared to compute an accuracy measure between 0 and 100 %. In case of a classification task like the classification of handwritten digits only the classes of a test image pair are compared, since the specific image ID is not relevant.

### 5.1 Assignment Accuracy of Cartoon Faces with and without Glasses

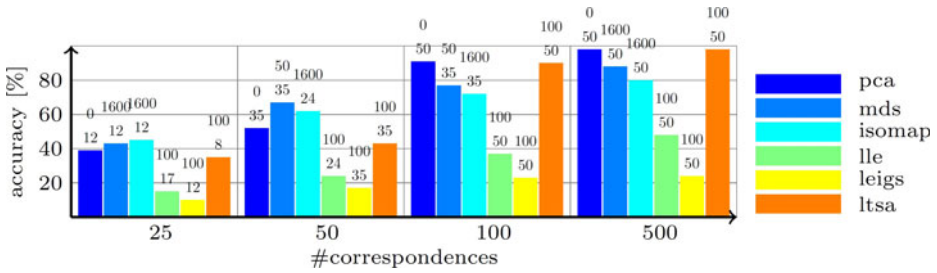
In our first experiment we choose two data sets with the same image content and feature dimension concerning the problem of occlusions. We create 1626 pairs of randomly assembled Simpsons avatars with and without glasses. The occlusion caused by the glasses is about 5 to 10 % of the image. We run all experiments with an image size of  $60 \times 40$  pixels, different target dimensions  $d = \{2, \dots, 80\}$ , different neighborhood sizes  $k = \{5, \dots, 1600\}$  and a varying number of labeled points  $L$ .

Fig. 4 shows that the accuracy for all methods increases with the number of labeled data points. With the highest number of given correspondences  $L = 100$ , LTSA and PCA outperform all other methods. In a further experiment we observed that for all methods the correct correspondence was within the five nearest neighbors.

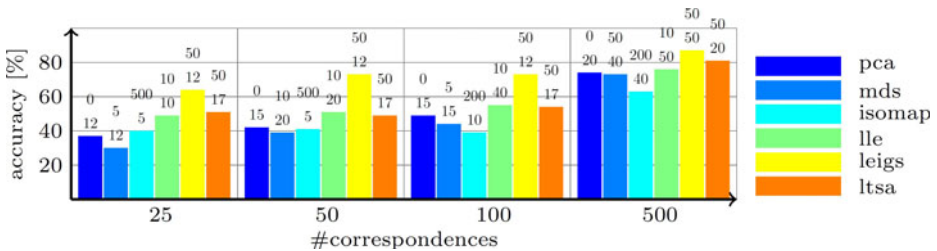
We observed that Isomap yields better results at a large number of neighbors  $k$  and the accuracy increases with increasing number of neighbors.

### 5.2 Classification Accuracy of Handwritten and Digital Digits

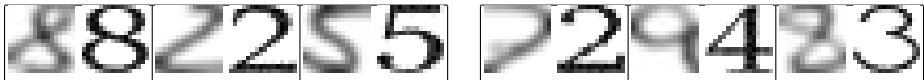
In this experiment we use 1900 handwritten digits [25] and 2940 digitally generated digits for a classification task. Both data sets include gray-valued images of size  $16 \times 16$  pixels. Some of the handwritten and all of the digital digits are



**Fig. 4.** Number of labeled points  $L$  versus assignment accuracy of cartoon faces with and without glasses. For each algorithm the lowest error rate out of different neighborhood sizes  $k$  and dimensionalities  $d$  is plotted. Above each bar there is the number of neighbors  $k$  (top) and the dimensionality  $d$  (below).



**Fig. 5.** Results for the handwritten and digital digits in analogy to Fig. 4



**Fig. 6.** Example results for the assignment of handwritten and digital digits: correct assignments (left) and false assignments (right)

labeled with the number shown in the image. Given unlabeled test images of handwritten digits can be assigned to labeled digital digits and classified. The advantage of the classification procedure over others is that the data sets can be easily extended to capture more variability of the classes.

Fig. 5 shows that the Laplacian Eigenmaps outperform all other algorithms in all cases. For comparison a discriminative linear classifier achieves an average accuracy of {37%, 40%, 43%, 71%} for {25, 50, 100, 500} correspondences. Again, Isomap needs much more neighbors than other graph-based algorithms.

It can be seen that at a certain number of labeled points  $L$  the classification accuracy does not improve significantly anymore.

In this experiment, LTSA and LLE show long running times. We quit their calculations with more than 100 neighbors after several hours.

As Fig. 6 illustrates most erroneous classifications arise from similarities between specific numbers, e. g. 3, 8, 0.

## 6 Discussion and Outlook

We propose a method to learn high dimensional correspondences from low dimensional manifolds by determining an affine transformation between a few labeled correspondences. We tested well-known dimensionality reduction algorithms regarding the accuracy in both an assignment and a classification task.

We showed that in an assignment task concerning occlusions the linear methods are more robust and have a higher accuracy. The Isomap algorithm only performs well using a number of neighbors in order of the size of used data points, which is intractable for large data sets.

In the classification task nearly all nonlinear methods perform better if the number of correspondences is low. If the number of correspondences is high, the linear methods perform comparably to the most nonlinear methods. But the Laplacian Eigenmaps, which perform worst for the assignment task, outperform all other methods for all given number of correspondences.

The nonlinear methods suffer from the aspect that unknown parameters like the number of neighbors have to be chosen carefully to achieve good results. We show that in our application especially Isomap tends to achieve better results with a very high number of neighbors. For the digit data set we observed LTSA and LLE not being practicable regarding the computational time if the size of the dataset is high.

The proposed framework can be applied to further fields of machine learning dealing with high dimensional data. The part of the dimensionality reduction and the transformation can be replaced by other methods depending on the given task. In future work we will address the reduction to one common manifold with different reduction algorithms, which is an alternative way to learn high dimensional correspondences from low dimensional manifolds.

## References

1. Bach, F.R., Jordan, M.I.: Spectral Clustering for Speech Separation. Wiley, Chichester (2009)
2. Mittal, A., Monnet, A., Paragios, N.: Scene Modeling and Change Detection in Dynamic Scenes: A Subspace Approach. In: CVUI, vol. 113 (2009)
3. Rao, S., Tron, R., Vidal, R., Ma, Y.: Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In: CVPR, vol. 37, p. 18 (2008)
4. Murase, H.: Moving Object Recognition in Eigenspace Representation: Gait Analysis and Lip Reading. *Pattern Recognition Letters* 17, 155–162 (1996)
5. Jolliffe, I.T.: Principal Component Analysis. Springer, Heidelberg (2002)
6. Cox, T.F., Cox, M.A.: Multidimensional Scaling, vol. 30. Chapman & Hall, Sydney (1994)
7. Tenenbaum, J.B., Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 2319 (2000)
8. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* (2000)

9. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* 15, 1373–1396 (2003)
10. Nadler, B., Lafon, S., Coifman, R.R.: Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems. *Applied and Computational Harmonic Analysis* 21, 113–127 (2006)
11. Weinberger, K.Q., Saul, L.K.: Unsupervised Learning of Image Manifolds by Semidefinite Programming. *IJCV* 70, 77–90 (2006)
12. Ham, J., Lee, D.D., Mika, S., Schölkopf, B.: A Kernel View of the Dimensionality Reduction of Manifolds. In: *ICML*, vol. 47 (2004)
13. Schölkopf, B., Smola, A., Müller, K.: *Kernel Principal Component Analysis*. MIT Press, Cambridge (1999)
14. De Silva, V., Tenenbaum, J.B.: Global versus Local Methods in Nonlinear Dimensionality Reduction. In: *NIPS* (2003)
15. Weinberger, K.Q., Packer, B.D., Saul, L.K.: Nonlinear Dimensionality Reduction by Semidefinite Programming and Kernel Matrix Factorization. In: *International Workshop on Artificial Intelligence and Statistics*, pp. 381–388 (2005)
16. Chang, H., Yeung, D.Y.: Robust Locally Linear Embedding. *Pattern Recognition* 39, 1053–1065 (2006)
17. Zhang, Z., Zha, H.: Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment. *SIAM Journal of Scientific Computing* (2004)
18. Donoho, D.L., Grimes, C.: Hessian Eigenmaps: Locally Linear Embedding Techniques for High-Dimensional Data. *National Academy of Sciences* 100 (2003)
19. Saul, L.K., Roweis, S.T.: Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds. *JMLR* 4, 119–155 (2003)
20. Ham, J., Lee, D., Saul, L.: Learning High Dimensional Correspondences from Low Dimensional Manifolds. In: *ICML* (2003)
21. Tenenbaum, J., Freeman, W.: Separating Style and Content with Bilinear Models. *Neural Computation* 12 (2000)
22. De la Torre, F., Black, M.: Dynamic coupled component analysis. In: *CVPR* (2005)
23. Wang, C., Mahadevan, S.: Manifold Alignment Using Procrustes Analysis. In: *ICML* (2008)
24. Lee, M.: *Algorithms for Representing Similarity Data* (1999)
25. Seewald, A.K.: Digits–A dataset for Handwritten Digit Recognition. TR (2005)



# Multi-label Classification for Image Annotation via Sparse Similarity Voting

Tomoya Sakai<sup>1</sup>, Hayato Itoh<sup>2</sup>, and Atsushi Imiya<sup>3</sup>

<sup>1</sup> Faculty of Engineering, Nagasaki University, Japan  
tsakai@ieee.org

<sup>2</sup> Graduate School of Science and Technology, Chiba University, Japan  
hayato-itoh@graduate.chiba-u.jp

<sup>3</sup> Institute of Media and Information Technology, Chiba University, Japan  
imiya@faculty.chiba-u.jp

**Abstract.** We present a supervised multi-label classification method for automatic image annotation. Our method estimates the annotation labels for a test image by accumulating similarities between the test image and labeled training images. The similarities are measured on the basis of sparse representation of the test image by the training images, which avoids similarity votes for irrelevant classes. Besides, our sparse representation-based multi-label classification can estimate a suitable combination of labels even if the combination is unlearned. Experimental results using the PASCAL dataset suggest effectiveness for image annotation compared to the existing SVM-based multi-labeling methods. Nonlinear mapping of the image representation using the kernel trick is also shown to enhance the annotation performance.

## 1 Introduction

This paper addresses multi-label classification for annotating images of multiple objects. Multi-labeling is a fundamental functionality of a multi-class classifier for the automatic image annotation. The classifier is required to assign multiple labels of objects to an image of those objects.

*Prior Work on Multi-class Classification and Multi-Labeling.* A popular approach to the image-based object recognition and annotation is to employ a discriminative model using bag-of-features image representation [1] in learning and labeling phases. One-vs-rest SVM [2,3] and one-vs-one SVM [4] consist of two-class SVM classifiers, each of which learns a margin between object classes. A test image to be annotated, however, has mixture of features of multiple objects in it. The two-class classifiers have to be able to discriminate the individual objects by the mixture. Multi-label ranking (MLR) [5] fixes this problem by simultaneously learning from multi-label data so as to minimize the classification error for all classes in total. MLR is shown to outperform the state-of-the-art multi-labeling SVM algorithms in the bag-of-features image classification task, but its performance for test images with unlearned combinations of labels is not guaranteed.

The image annotation based on multi-label classification is essentially a problem of finding a combination of learned objects whose features can synthesize the mixture of features of a test image. An important fact is that among the learned classes a few of them are relevant to a test image. Sparse representation-based classification (SRC) [6] takes advantage of this fact by representing a test image as a sparse linear combination of training images. The SRC achieves robust single-labeling for face recognition. For the image annotation task, Wang *et al.* [7] proposed multi-label sparse coding (MSC) in the same manner as the SRC together with linear embedding into a discriminative space learned from the training images and their sparse labels. Hsu *et al.* [8] have exploited the sparsity of the classifier output by the compressed sensing technique [9,10,11,12,13] for reducing computational expense of multi-label classification with linear regression.

*Our Method.* In this paper, we propose a substantial method of multi-labeling on the basis of the sparse representation and accumulation of similarities. Our method consists of the following steps:

**Sparse representation:** explain concisely the test image by the training images, i.e., find sparse coefficients  $\hat{\alpha}_j$  such that

$$\phi(\text{test image}) \approx \sum_j \hat{\alpha}_j \phi(j\text{-th training image})$$

where  $\phi$  indicates a high dimensional representation of the input image, e.g., a histogram of visual words.

**Similarity measurement:** compute similarities

$$w_j \sim \hat{\alpha}_j \kappa(j\text{-th training image, test image})$$

where  $\kappa$  calculates an inner product.

**Voting:** classes indicated by the labels of the  $j$ -th training image receive the votes of  $w_j$ .

Preliminary details of the sparse representation are provided in Section 2. Differing from the existing multi-label methods exploiting sparsity, our method does not use the labels of training images for the computation of the sparse coefficients  $\hat{\alpha}_j$ . While the use of the labels in the training phase would refine the classification performance for a test image to give a learned combination of labels, it could degrade the generalization capabilities of the sparse representation for most of the label combinations unlearned in practice. After the sparse representation, our method measures the similarities because we must not assemble the output labels by directly using the coefficients  $\hat{\alpha}_j$  as done in the MSC. We also introduce the kernel trick to improve the classification performance. Our algorithms and the kernelization are described in Section 3. We experimentally show the ability to find unlearned label combinations as well as the outperformance of our method in Section 4.

## 2 Sparse Representation for Multi-labeling

### 2.1 Multi-class Classification and Multi-labeling

Multi-label classification is a task of assigning a suitable number of class labels to unlabeled test data. A training dataset  $S \subset \mathbb{R}^d$  with a collection of labels  $Y \subset \{0, 1\}^l$  is available for the classification. The labels of a training data  $\mathbf{s}_j \in \mathcal{S}$  are represented as a binary vector  $\mathbf{y}_j = [y_1, \dots, y_l]^\top$  where  $y_i \in \{0, 1\}$ .

The binary classification is the case of  $l = 1$ , and the case of  $l > 1$  is known as the multi-class classification. In the prediction of a label  $\hat{\mathbf{y}} \in \{0, 1\}^l$  for a given test data  $\mathbf{x} \in \mathbb{R}^d$ , the multi-class classification under the constraint  $\|\hat{\mathbf{y}}\|_0 \leq 1$  is called the single-labeling. Here,  $\|\cdot\|_0$  denotes the  $l^0$  norm, which counts the nonzero components. The multi-class classification without the constraint is the multi-labeling. There are possibly  $2^l$  combinations of labels.

### 2.2 Sparse Representation of Test Data

Let  $\mathbf{S} \in \mathcal{R}^{d \times n}$  be a matrix of  $d$ -dimensional  $n$  column vectors of training data  $\mathbf{s}_j$ , and let  $\mathbf{Y} \in \{0, 1\}^{l \times n}$  be the matrix with corresponding label vectors  $\mathbf{y}_j$  in its columns. Supposing the linear vector space model and given an enough number of training data, one can represent a test data  $\mathbf{x} \in \mathbb{R}^d$  as a linear combination of the vectors of training data.

$$\mathbf{x} = \sum_{j=1}^n \alpha_j \mathbf{s}_j = \mathbf{S}\boldsymbol{\alpha} \quad (1)$$

Here,  $\boldsymbol{\alpha} \in \mathbb{R}^n$  is the vector of  $n$  combination coefficients  $\alpha_j$  to be estimated.

The solution  $\boldsymbol{\alpha}$  to Equ. (1) exists if the test data  $\mathbf{x}$  lies in  $\text{span}\mathbf{S}$ , i.e., the subspace spanned by the training data. We would like to assign labels to the test data according to the solution to Equ. (1). If no solution exists, one should not assign any label, i.e.,  $\hat{\mathbf{y}} = \mathbf{0}$ . This is the case where the training dataset is insufficient for representing the test data. If a sufficient number of training data are given, Equation (1) has non-unique solutions. We require regularization to select a unique solution. From the viewpoint of classification, a test data should be concisely explained by relevant training data. A sparse solution whose nonzero components indicate a few relevant classes to the test data would be preferable.

Finding a sparse solution is formulated as a  $l^0$ -minimization problem:

$$\min \|\boldsymbol{\alpha}\|_0 \quad \text{subject to} \quad \mathbf{x} = \mathbf{S}\boldsymbol{\alpha}. \quad (2)$$

The  $l^0$ -minimization is a NP-hard problem, which is often relaxed to a convex problem:

$$\min \|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad \mathbf{x} = \mathbf{S}\boldsymbol{\alpha}. \quad (3)$$

One can find literature on the uniqueness of the sparse solution and on the equivalence between the  $l^0$ - and  $l^1$ -minimization problems [12, 14, 15]. The uniqueness

of the solution, for example, is guaranteed under the condition called the restricted isometry property (RIP). The RIP condition with parameters  $(m, \delta)$  for a matrix  $\Theta$  is described as

$$(1 - \delta)\|\beta\|_2 \leq \|\Theta\beta\|_2 \leq (1 + \delta)\|\beta\|_2 \quad \forall \beta \in \{\mathbf{b} \mid \|\mathbf{b}\|_0 \leq m\}.$$

A vector  $\mathbf{b}$  is called  $m$ -sparse if  $\|\mathbf{b}\|_0 \leq m$ . It is known that the  $l^0$ -minimization problem (2) has a unique  $m$ -sparse solution if the matrix  $\mathbf{S}$  satisfies the RIP condition with  $(2m, \delta < 1)$ . The  $m$ -sparse solution is equivalent to the  $l^1$ -minimizer for (3) if  $\mathbf{S}$  satisfies the RIP condition with  $(2m, \delta < \sqrt{2} - 1)$  [12].

### 2.3 Dimensionality Reduction

One can reduce the computational cost of dealing with high-dimensional training and test data by linear projection. The compressed sensing methodology shows that a small number of projections of a high-dimensional vector can contain salient information about its sparse representation enough to recover it with regularization that promotes sparsity [9, 11, 16]. Random projection is known to be a universal way of dimensionality reduction.

Let  $\mathbf{R}$  be a  $d_c \times d$  random matrix. A training dataset  $\mathbf{S}$  and a test data  $\mathbf{x}$  are compressed by random projection as  $\mathbf{x}_c = \mathbf{R}\mathbf{x} \in \mathbb{R}^{d_c}$  and  $\mathbf{S}_c = \mathbf{R}\mathbf{S} \in \mathbb{R}^{d_c \times n}$ . Equation (1) is rewritten as  $\mathbf{x}_c = \mathbf{S}_c\boldsymbol{\alpha}$ . It is known that the  $m$ -sparse vector  $\boldsymbol{\alpha}$  can be reconstructed from  $\mathbf{x}_c$  with probability  $1 - e^{-\mathcal{O}(d_c)}$  by the sparse regularization if  $d_c \geq d_0 = \mathcal{O}(m \log(d/m))$  [17, 18].

### 2.4 Multi-label Estimation by Similarity Voting

We describe how to assign labels to a test data via sparse representation. Let  $\hat{\mathbf{x}}$  be a reconstructed test data using the training data matrix  $\mathbf{S}$  and a sparse solution  $\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}$ .

$$\hat{\mathbf{x}} = \mathbf{S}\hat{\boldsymbol{\alpha}}$$

We measure the similarity between the test data  $\mathbf{x}$  and its reconstruction  $\hat{\mathbf{x}}$  as

$$\cos \theta = \frac{\mathbf{x}^\top \hat{\mathbf{x}}}{\|\mathbf{x}\|_2 \|\hat{\mathbf{x}}\|_2} = \frac{\mathbf{x}^\top \mathbf{S}\hat{\boldsymbol{\alpha}}}{\|\mathbf{x}\|_2 \|\hat{\mathbf{x}}\|_2} = \sum_{j=1}^n w_j.$$

Here,

$$w_j = \frac{\hat{\alpha}_j \mathbf{s}_j^\top \mathbf{x}}{\|\mathbf{x}\|_2 \|\hat{\mathbf{x}}\|_2} \tag{4}$$

is the similarity between the test data and the  $j$ -th component of the reconstructed test data on the basis of training data. Note that  $\mathbf{w} = [w_1, \dots, w_n]^\top$  is as sparse as  $\hat{\boldsymbol{\alpha}}$ . Regarding  $w_j$  as the partial membership value for a combination of classes labeled as  $\mathbf{y}_j$ , we estimate the multi-label  $\hat{\mathbf{y}}$  for the test data by accumulating the labels as

$$\hat{\mathbf{y}} = \sum_{j=1}^n w_j \mathbf{y}_j = \mathbf{Y}\mathbf{w}.$$

This accumulation is interpreted as label voting with the weight  $w_j$ . One can determine the labels for the test data by thresholding or ranking the magnitudes of the vector components of  $\hat{\mathbf{y}}$ .

### 3 Algorithms

#### 3.1 Multi-label Classification

Our multi-labeling algorithm is summarized in Algorithm 1.

---

**Algorithm 1.** Multi-label classification (main algorithm in linear case)

---

**Input:**  $\mathbf{x} \in \mathbb{R}^d$ : test data,  $\mathbf{S} \in \mathbb{R}^{d \times n}$ : matrix of training data,  $\mathbf{Y} \in \{0, 1\}^{l \times n}$ : matrix of labels;

- 1 normalize the columns of  $\mathbf{S}$  to have unit  $l^2$  norm;
- 2 perform dimensionality reduction of  $\mathbf{S}$  and  $\mathbf{x}$  if the dimensionality  $d$  is intractably high;
- 3 decompose  $\mathbf{x}$  with respect to  $\mathbf{S}$  under sparse regularization to obtain the sparse solution  $\hat{\boldsymbol{\alpha}}$ ;
- 4 compute the similarities  $\mathbf{w} = [w_1, \dots, w_n]^\top$ ;

**Output:**  $\hat{\mathbf{y}} \leftarrow \mathbf{Y}\mathbf{w}$ : label estimates.

---

The classification does not involve any expensive computation for training. We do not have to solve a quadratic programming problem like support vector machines or an eigenvalue problem for subspace methods. Algorithm 1 can start testing soon after loading the training data. It is therefore easy to append and remove the data before testing if necessary. We would also remark that Algorithm 1 can answer unlearned combinations of labels when the relevant training data can sparsely represent the test data.

#### 3.2 Sparse Decomposition

There are basically two types of algorithms for solving the minimization problem (2). One is called the basis pursuit (BP) [19], which relaxes the  $l^0$  to  $l^1$  minimization problem. Linear programming can solve the  $l^1$  minimization problem in (3). One can find some algorithms [20,21,22,23] for the related convex problems

$$\min \|\mathbf{x} - \mathbf{S}\boldsymbol{\alpha}\|_2 \quad \text{subject to} \quad \|\mathbf{x}\|_1 \leq \tau \quad (5)$$

$$\min \|\boldsymbol{\alpha}\|_1 \quad \text{subject to} \quad \|\mathbf{x} - \mathbf{S}\boldsymbol{\alpha}\|_2 \leq \varepsilon \quad (6)$$

to obtain robust solution against noise.

The other type is the greedy algorithms [24,25,26,27], which greedily seek for the nonzero components. Matching pursuit (MP) [28] selects a column vector  $\mathbf{s}_j$  in  $\mathbf{S}$  which is most coherent to the residual of  $\mathbf{x}$ , and removes from the residual the component in the direction of  $\mathbf{s}_j$ , iteratively. Orthogonal matching pursuit (OMP) [24] instead removes the component in the subspace spanned

by previously selected column vectors. Regularized orthogonal matching pursuit (ROMP) [26] is guaranteed to recover any  $m$ -sparse solution for a matrix satisfying the RIP condition with  $(2m, 0.03/\sqrt{\log m})$ . The greedy algorithms are very simple to implement and faster than BP. In this paper, we employ ROMP.

### 3.3 Kernelization

The above formulation assumes the linear relationship as in Equ. (1). Although Algorithm 1 can benefit from the sparsity of the linear representation, we would like to translate our framework into a nonlinear version hoping to improve the classification performance. We map the data in the nonlinear input space  $\mathbb{R}^d$  to an Affine space using a nonlinear function  $\phi$ , assuming the linear relationship between training data and test data as

$$\phi(\mathbf{x}) = \sum_{j=1}^n \alpha_j \phi(\mathbf{s}_j). \quad (7)$$

We apply the kernel trick using a kernel function  $\kappa(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^\top \phi(\mathbf{y})$  and kernel matrix  $\mathbf{K}(\mathbf{X}_1, \mathbf{X}_2) \in \mathbb{R}^{n_1 \times n_2}$  whose  $ij$ -th entry is the inner product of the  $i$ -th and  $j$ -th column vectors of the matrices  $\mathbf{X}_1 \in \mathbb{R}^{d \times n_1}$  and  $\mathbf{X}_2 \in \mathbb{R}^{d \times n_2}$ , respectively.

---

#### Algorithm 2. Kernelized ROMP

---

**Input:**  $\mathbf{x} \in \mathbb{R}^d$ : test data,  $\mathbf{S} \in \mathbb{R}^{d \times n}$ : matrix of training data,  $m_0$ : sparsity level,  $\varepsilon_0$ : tolerance;

1 initialize  $\mathcal{I} \leftarrow \emptyset$  and  $\hat{\boldsymbol{\alpha}} \leftarrow \mathbf{0}$ ;

2 **repeat**

3  $\mathbf{u} \leftarrow \mathbf{K}(\mathbf{S}, \mathbf{x}) - \mathbf{K}(\mathbf{S}, \mathbf{S}_{\mathcal{I}}) \hat{\boldsymbol{\alpha}}_{\mathcal{I}}$ ;

4  $\boldsymbol{\gamma} \leftarrow [|u_1|, \dots, |u_n|]^\top$ ;

5 let  $\mathcal{J}$  be a set of indices of the  $m_0$  biggest components of  $\boldsymbol{\gamma}$ , or all of its nonzero components, whichever set is smaller;

6 sort  $\mathcal{J}$  in descending order of the components  $\gamma$ ;

7 among all subsets  $\mathcal{J}_0 \subset \mathcal{J}$  such that  $\gamma_i \leq 2\gamma_j$  for all  $i < j \in \mathcal{J}_0$ , choose  $\mathcal{J}_0$  with the maximal energy  $\|\boldsymbol{\gamma}_{\mathcal{J}_0}\|_2^2 = \sum_{k \in \mathcal{J}_0} \gamma_k^2$ ;

8  $\mathcal{I} \leftarrow \mathcal{I} \cup \mathcal{J}_0$ ;

9  $\hat{\boldsymbol{\alpha}}_{\mathcal{I}} \leftarrow \arg \min_{\boldsymbol{\alpha}_{\mathcal{I}}} \|\mathbf{r}(\boldsymbol{\alpha}_{\mathcal{I}})\|_2^2$ ;

10 **until**  $\|\mathbf{r}(\hat{\boldsymbol{\alpha}}_{\mathcal{I}})\|_2^2 / \|\mathbf{x}\|_2 \leq \varepsilon_0$  or  $\text{card} \mathcal{I} \geq 2m_0$ ;

**Output:**  $\hat{\boldsymbol{\alpha}}$ : sparse solution.

---

We present a kernelized version of ROMP for nonlinear structure of the input space. The kernelized ROMP is described as Algorithm 2. The vector  $\boldsymbol{\alpha}_{\mathcal{I}}$  indicates a vector with the components of  $\boldsymbol{\alpha}$  specified by  $\mathcal{I}$ . At Step 9, one can easily obtain  $\hat{\boldsymbol{\alpha}}_{\mathcal{I}}$  by solving least squares problem without explicitly computing the residual vector  $\mathbf{r}$ , since the squared norm is a quadratic form

$$\|\mathbf{r}(\boldsymbol{\alpha}_{\mathcal{I}})\|_2^2 = \kappa(\mathbf{x}, \mathbf{x}) - 2\boldsymbol{\alpha}_{\mathcal{I}}^\top \mathbf{K}(\mathbf{S}_{\mathcal{I}}, \mathbf{x}) + \boldsymbol{\alpha}_{\mathcal{I}}^\top \mathbf{K}(\mathbf{S}_{\mathcal{I}}, \mathbf{S}_{\mathcal{I}}) \boldsymbol{\alpha}_{\mathcal{I}}. \quad (8)$$

As the ROMP works in linear time with respect to  $n$  and  $d$  [26], our kernelized ROMP also works in linear time.

After running the kernelized ROMP, the similarities are measured as

$$w_j = \frac{\hat{\alpha}_j \kappa(\mathbf{s}_j, \mathbf{x})}{\sqrt{\kappa(\mathbf{x}, \mathbf{x}) \hat{\alpha}_{\mathcal{I}}^\top \mathbf{K}(\mathbf{S}_{\mathcal{I}}, \mathbf{S}_{\mathcal{I}}) \hat{\alpha}_{\mathcal{I}}}}. \quad (9)$$

Equation (9) coincides with Equ. (4) if one utilizes the linear kernel  $\kappa(\mathbf{x}, \mathbf{x}) = \mathbf{x}^\top \mathbf{x}$  and  $\mathbf{K}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{X}_1^\top \mathbf{X}_2$ . Algorithm 2 with the linear kernel is also equivalent to the original ROMP.

## 4 Experiment

*Data* We apply our multi-label method to image annotation. We used PASCAL VOC 2009 dataset [29]. The VOC 2009 dataset has 3,473 training images and 3,581 validation images of twenty object classes. Each image is annotated by one or more object class labels. We chose the 2,236 training images with single labels as the training data in order to assess the ability to find suitable combinations of labels without using multi-label training data. We randomly selected half of the validation images for tuning the classifier parameters and the other half for testing. A standard bag-of-features model [1] was used to represent the images in this experiment. We extracted SIFT descriptors [30] from every training image in grayscale, and clustered these features into 1,000 clusters by the  $k$ -means clustering. Each image was represented as a tf-idf vector.

*Evaluation and Procedure.* We characterize the performance of multi-label classification as receiver operating characteristic (ROC) curve and the area under the curve (AUC). Our ROC evaluates the ranking performance: how high the correct labels are ranked. We calculate the true positive ratio (TPR) and false positive ratio (FPR) by changing the number of top labels indicated by the label estimates  $\hat{\mathbf{y}}$ . The same evaluation metric is used for MLR [5]. We did not invoke the dimensionality reduction in Algorithm 1. The input parameters of Algorithm 2 were tuned and set as  $m_0 = 35$  and  $\varepsilon_0 = 10^{-2}$ .

*Results.* Table 1 shows the AUC of rank ROC. Our method provides a comparative AUC to MLR with the linear kernel. The AUC is improved by the kernelization in both methods. Our method with a Gaussian kernel achieves slightly better performance than MLR. MLR has been shown to outperform the existing multi-label SVMs [5]. We deduce from these results that our method is highly effective for the image annotation tasks.

Figure 1 shows some examples of multiply annotated images and annotations by our method with the Gaussian kernel. Note that we used only single-label images for training. We could observe that the relevant object labels are ranked high. Algorithm 1 with MATLAB implementation took about 0.1 (linear) and 0.5 (kernelized) seconds per test image using a CPU single core.

**Table 1.** AUC of rank ROC for PASCAL VOC 2009

Kernel	Proposed	MLR
Linear	74.0%	74.1%
Nonlinear	78.1%	76.3%

				
aeroplane, car	bird, boat	dog, person, sofa	bus, car, person	chair, person, sofa, tvmonitor
aeroplane, car	<b>bird, boat</b>	<b>person, cat, sofa</b>	<b>bus, car,</b> train, <b>person</b>	<b>person,</b> tvmonitor, chair

**Fig. 1.** Multi-labeling results. First row: test images, second row: ground-truth labels, third row: labels by our method. The true positive labels are in bold.

## 5 Concluding Remarks

Assigning multiple labels of objects to an unlabeled test image is a problem of finding a combination of learned objects which can synthesize the mixture of features of objects in the test image. We casted this problem as a sparse decomposition of image representation. Our method decomposes the bag-of-features representation of a test image into those of labeled training images as concisely as possible via sparse regularization. This enables us to detect the relevant training images even if all the combinations of objects are not learned from the training images. As suggested in Section 3.1, our method does not have any intensive computation in training. Of course the sparse decomposition for testing requires considerable time, but we have many advantages: easy update of training data, capability to answer unlearned label combinations, and robustness against noise or clutter. We should investigate the performance of our method on large-scale dataset. The performance would be further improved by incorporating co-occurrence statistics of objects and features.

**Acknowledgments.** The first author was partially supported by the Grant-in-Aid for Young Scientists, from the Ministry of Education, Culture, Sports, Science and Technology of Japan under MEXT KAKEN 22700163.

## References

1. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV (2004)



2. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) Proceedings of ECML-1998, 10th European Conference on Machine Learning, pp. 137–142. Springer, Heidelberg (1998)
3. Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 13, 415–425 (2002)
4. Kressel, U.H.G.: Pairwise classification and support vector machines. MIT Press, Cambridge (1999)
5. Bucak, S.S., Mallapragada, P.K., Jin, R., Jain, A.K.: Efficient multi-label ranking for multi-class learning: approach to object recognition. In: International Conference on Computer Vision (2009)
6. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 210–227 (2009)
7. Wang, C., Yan, S., Zhang, L., Zhang, H.J.: Multi-label sparse coding for automatic image annotation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 0, pp. 1643–1650 (2009)
8. Hsu, D., Kakade, S., Langford, J., Zhang, T.: Multi-label prediction via compressed sensing. In: 23rd Annual Conference on Neural Information Processing Systems (2009)
9. Donoho, D.: Compressed sensing. *IEEE Trans. Information Theory* 52, 1289–1306 (2006)
10. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory* 52, 489–509 (2006)
11. Candès, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. *Comm. on Pure and Applied Math.* 59, 1207–1223 (2006)
12. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématique* 346, 589–592 (2008)
13. Candès, E.J., Wakin, M.B.: An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 21–30 (March 2008)
14. Gribonval, R., Nielsen, M.: Sparse representations in unions of bases. *IEEE Transactions on Information Theory* 49, 3320–3325 (2003)
15. Donoho, D., Elad, M.: Optimally sparse representation in general (non-orthogonal) dictionaries via  $l^1$  minimization. *Proc. the National Academy of Sciences of the United States of America*, 2197–2202 (2003)
16. Candès, E.J., Tao, T.: Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory* 52, 5406–5425 (2006)
17. Candès, E.J., Tao, T.: Decoding by linear programming. *IEEE Transactions on Information Theory* 51, 4203–4215 (2005)
18. Rudelson, M., Vershynin, R., Rudelson, M., Vershynin, R.: Geometric approach to error correcting codes and reconstruction of signals. *Int. Math. Res. Not.* 64, 4019–4041 (2005)
19. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20, 33–61 (1998)
20. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58, 267–288 (1996)
21. Kim, S.J., Koh, K., Lustig, M., Boyd, S., Gorinevsky, D.: An interior-point method for large-scale  $l_1$ -regularized least squares. *IEEE Journal on Selected Topics in Signal Processing* 1, 606–617 (2007)

22. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing* 1, 586–597 (2007)
23. Tomioka, R., Sugiyama, M.: Dual augmented lagrangian method for efficient sparse reconstruction. Technical report, arXiv:0904.0584, (preprint, 2009)
24. Pati, Y.C., Rezaeiifar, R., Rezaeiifar, Y.C.P.R., Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, pp. 40–44 (1993)
25. Tropp, J.A., Anna, G.C.: Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Information Theory* 53, 4655–4666 (2007)
26. Needell, D., Vershynin, R.: Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics* 9, 317–334 (2009)
27. Needell, D., Tropp, J.A.: CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis* 26, 301–321 (2009)
28. Mallat, S., Zhang, Z.: Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing* 41, 3397–3415 (1993)
29. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 303–338 (2010)
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110 (2004)

# Centered Subset Kernel PCA for Denoising

Yoshikazu Washizawa<sup>1</sup> and Masayuki Tanaka<sup>2</sup>

<sup>1</sup> Brain Science Institute, Riken

<sup>2</sup> Tokyo Institute of Technology

**Abstract.** Kernel PCA has been applied to image processing, even though, it is known to have high computational complexity. We introduce centered Subset KPCA for image denoising problems. Subset KPCA has been proposed for reduction of computational complexity of KPCA, however, it does not consider a pre-centering that is often important for image processing. Indeed, pre-centering of Subset KPCA is not straightforward because Subset KPCA utilizes two sets of samples. We propose an efficient algorithm for pre-centering, and provide an algorithm for pre-image. Experimental results show that our method is comparable with a state-of-the-art image denoising method.

## 1 Introduction

Principal component analysis (PCA) and related statistical approaches have been widely used for image analysis (e.g., image compression, restoration, analysis, understanding and denoising). PCA is a very fundamental and simple linear approach, however, its performance is limited due to its linearity. Kernel trick is one of the methods that extend from a linear approach to a non-linear approach, and it has been used widely in machine learning area such as support vector machines (SVM). Kernel trick implicitly utilizes a non-linear pre-mapping  $\Phi$  that maps from  $d$ -dimensional input space  $\mathbb{R}^d$  to higher dimensional Hilbert space  $\mathcal{F}$  that is called feature space. Actual computation is done by using kernel function  $k(\cdot, \cdot)$  that satisfies  $k(\mathbf{x}_1, \mathbf{x}_2) = \langle \Phi(\mathbf{x}_1) | \Phi(\mathbf{x}_2) \rangle$  for all  $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ , where  $\langle \cdot | \cdot \rangle$  is the inner product [1].

PCA has also been extended by the kernel trick, it is called Kernel PCA (KPCA) [2]. KPCA has also been applied to image analysis, and showed higher performance [3,4]. However, KPCA is known to have high computational complexity, that is the eigenvalue decomposition of which size equals to the number of samples  $N$ . In image analysis, we often divide one image to many small blocks that are called patches. In case that we use 5x5 pixel patches from 512x512 pixel image, the number of patches is  $N = (512 - 5 + 1)^2 \simeq 2.5 \times 10^5$ . If we use four-bytes floating-point system, KPCA requires at least  $4 \times N(N + 1)/2 \simeq 1.3 \times 10^2$  Giga bytes memory for the kernel Gram matrix, and we have to obtain eigenvalues/eigenvectors of the matrix. This is infeasible to obtain in a current computational environment.

Recently, Subset KPCA (SubKPCA) that reduces computational complexity of KPCA, has been proposed [5]. SubKPCA utilizes subset of samples for basis,

and all samples for estimation. By using all samples for estimation, SubKPCA always shows higher approximation error than KPCA that only uses subset of samples. In this paper, we apply SubKPCA to image denoising.

When we apply PCA or related statistical approach to image analysis, we often remove DC components (averages of training vectors) before applying PCA. This pre-centering enhances its performance in many cases. However, in kernel PCA this pre-processing is sometimes skipped (e.g., [3]). In SubKPCA, since a set of basis is given as a subset of training samples, the definition of mean vector is not straightforward. In this paper, we introduce an efficient definition of mean vector by using linear combination of the samples in the subset.

Experimental results show denoising by centered SubKPCA is comparable with state-of-art image de-nosing technique, especially, SubKPCA show better performance in texture area.

## 2 KPCA and Subset KPCA

We here review KPCA and SubKPCA briefly. Let  $\mathbf{x}_1, \dots, \mathbf{x}_N$  be samples. KPCA obtains eigenvalues and eigenvectors of the correlation matrix (operator),  $R_\Phi$  in the feature space  $\mathcal{F}$ ,

$$R_\Phi = \sum_{i=1}^N \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_i)^\top, \quad (1)$$

If the dimension of the feature space is infinite, the outer product should be expressed by ket-bra  $|\Phi(\mathbf{x}_i)\rangle\langle\Phi(\mathbf{x}_i)|$ , or Neumann-Shatten product  $\Phi(\mathbf{x}_i) \otimes \overline{\Phi(\mathbf{x}_i)}$ , however, in the interest of brevity, we use notations of the finite dimensional space. Since it is difficult to obtain the eigenvalue decomposition (EVD) of  $R_\Phi$  directly, we calculate EVD of the kernel Gram matrix  $K \in \mathbb{R}^{N \times N}$ ,  $(K)_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ . We let a matrix  $S = [\Phi(\mathbf{x}_1) \dots \Phi(\mathbf{x}_N)]$ ,  $K = S^\top S$ . Suppose that the  $i$ th eigenvalue and eigenvector of  $K$  are  $\mathbf{v}_i$  and  $\lambda_i$ , then the  $i$ th eigenvalue of  $R_\Phi$  is  $\lambda_i$  and the  $i$ th eigenvector of  $R_\Phi$  is given by  $\mathbf{u}_i = \frac{1}{\sqrt{\lambda_i}} S \mathbf{v}_i$ . Let  $V = [\mathbf{v}_1 \dots \mathbf{v}_r]$ ,  $U = [\mathbf{u}_1 \dots \mathbf{u}_r]$ ,  $\Lambda = \text{diag}[\lambda_1, \dots, \lambda_r]$ , the projection onto the subspace is given by

$$P_{\text{KPCA}} = UU^\top = SVA^{-1}V^\top S^\top. \quad (2)$$

One of the problems of KPCA is computational complexity because EVD has high computation cost that increases with  $N^3$ . Moreover when we obtain projection of an input vector  $\mathbf{x}$ , we have to evaluate values of the kernel function of the input vector and all training samples  $\mathbf{x}_1, \dots, \mathbf{x}_N$ , and therefore we have to store all training samples. SubKPCA approximates KPCA and reduces these computational complexities.

PCA and KPCA are characterized by minimization of mean squared error between samples and transformed samples under the rank constraint,

$$\begin{aligned} \min_X \quad & \frac{1}{N} \sum_{i=1}^N \|\Phi(\mathbf{x}_i) - X\Phi(\mathbf{x}_i)\|^2 \\ \text{Subject to} \quad & \text{rank}(X) \leq r, \mathcal{N}(X) \supset \mathcal{R}(S)^\perp, \end{aligned} \quad (3)$$

where  $\mathcal{R}(\cdot)$  and  $\mathcal{N}(\cdot)$  denote the range and the null space respectively. The cost function is minimized by a projection that is obtained by KPCA [5]. The problem is in the higher dimensional Hilbert space  $\mathcal{F}$ , however, the dimension of the space spanned by samples is at most  $N$ . Therefore the problem can be transformed to the dual problem in  $\mathbb{R}^N$ , and the problem is reduced to EVD of an  $N \times N$  matrix. SubKPCA minimizes the same cost function in smaller dimensional subspace. Suppose that  $M$  is the dimension of the smaller subspace ( $M < N$ ). Let  $\mathbf{y}_1, \dots, \mathbf{y}_M$  be a subset of samples that spans the smaller dimensional subspace, i.e.,  $\Phi(\mathbf{y}_1), \dots, \Phi(\mathbf{y}_M)$  are basis of the space. The subset is selected by a clustering or a forward search. Let  $T = [\Phi(\mathbf{y}_1), \dots, \Phi(\mathbf{y}_M)]$ , then the problem of SubKPCA is

$$\begin{aligned} \min_X \quad & \frac{1}{N} \sum_{i=1}^N \|\Phi(\mathbf{x}_i) - X\Phi(\mathbf{x}_i)\|^2 \\ \text{Subject to} \quad & \text{rank}(X) \leq r, \\ & \mathcal{N}(X) \supset \mathcal{R}(T)^\perp, \mathcal{R}(X) \subset \mathcal{R}(T). \end{aligned} \tag{4}$$

Let matrices  $K_y = T^\top T \in \mathbb{R}^{N \times N}$  and  $K_{xy} = S^\top T \in \mathbb{R}^{N \times M}$  be  $(K_y)_{ij} = k(\mathbf{y}_i, \mathbf{y}_j)$ ,  $(K_{xy})_{ij} = k(\mathbf{x}_i, \mathbf{y}_j)$ , and let  $\mathbf{z}_i$  be the  $i$ th eigenvector of the generalized eigenvalue problem,  $K_{xy}^\top K_{xy} \mathbf{z} = \lambda K_y \mathbf{z}$ . Suppose that the norm of  $\mathbf{z}_i$  is normalized by  $\mathbf{z}_i \leftarrow \mathbf{z}_i / \sqrt{\langle \mathbf{z}_i | K_y \mathbf{z}_i \rangle}$  that satisfies  $\langle \mathbf{z}_i | K_y \mathbf{z}_i \rangle = 1$ . Let  $Z = [\mathbf{z}_1 \dots \mathbf{z}_r] \in \mathbb{R}^{M \times r}$ , then the cost function of (4) is minimized by  $P_{\text{SubKPCA}} = T(\sum_{i=1}^r \mathbf{z}_i \mathbf{z}_i^\top) T^\top = T Z Z^\top T^\top$ .

Let  $P_{\text{SubKPCA}} = U U^\top$ ,  $U = T Z$ , then the transform of an input vector  $\mathbf{x}$  is given by

$$U^\top \Phi(\mathbf{x}) = [\langle \mathbf{z}_1 | \mathbf{h}_\mathbf{x} \rangle \langle \mathbf{z}_2 | \mathbf{h}_\mathbf{x} \rangle \dots \langle \mathbf{z}_r | \mathbf{h}_\mathbf{x} \rangle]^\top \in \mathbb{R}^r, \tag{5}$$

where  $\mathbf{h}_\mathbf{x} = T^* \Phi(\mathbf{x}) = [k(\mathbf{y}_1, \mathbf{x}), \dots, k(\mathbf{y}_M, \mathbf{x})]^\top$ .

For the sample selection, [5] proposed i) Clustering, such as K-means, ii) Random sample consensus (RANSAC) approach, iii) forward (incremental) search. In this paper, we employ i) K-means clustering approach.

In SubKPCA, the size of the generalized eigenvalue problem is  $M \times M$ , and when we evaluate an input vector, the kernel function is evaluated  $M$  times. The computational complexity that depends on the number of basis,  $M$ , and the accuracy of SubKPCA are trade-off. If  $M = N$ , SubKPCA is equivalent with KPCA. If we use smaller  $M$ , the value of the cost function (4) becomes larger.

### 3 Centered Subset KPCA

We often remove mean component of samples before we apply PCA. In other words, we use  $\mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_N - \bar{\mathbf{x}}$ , ( $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$ ) instead of  $\mathbf{x}_1, \dots, \mathbf{x}_N$ .

Suppose that  $\mathbf{f}$  is an original image,  $\mathbf{n}$  is a noise vector, and  $\mathbf{g} = \mathbf{f} + \mathbf{n}$  is an observed image. Let  $E[\cdot]$  be the ensemble mean. If the noise has zero mean,

$E[\mathbf{n}] = 0$ , we have  $E[\mathbf{g}] = E[\mathbf{f}]$ . Let us consider two restorations  $\hat{\mathbf{f}}_1 = A_1\mathbf{g}$  and  $\hat{\mathbf{f}}_2 = A_2(\mathbf{g} - E[\mathbf{g}]) + E[\mathbf{g}]$ . In PCA, we impose the rank constraint on  $A_1$  and  $A_2$ . Therefore,  $\hat{\mathbf{f}}_1$  is in  $\mathcal{R}(A_1)$  that is limited subspace. On the other hand, since  $E[\mathbf{g}]$  is noise free ( $E[\mathbf{g}] = E[\mathbf{f}]$ ),  $E[\mathbf{g}]$  is not needed to be removed noise components.  $A_2$  only removes noise in  $\mathbf{g} - E[\mathbf{g}]$ , but not in  $E[\mathbf{g}]$ . Thus  $\hat{\mathbf{f}}_2$  is expected to be better performance than  $\hat{\mathbf{f}}_1$ .

In Kernel PCA, this preprocessing is sometimes ignored (e.g., [3] assumes the mean vector is zero). Although the centroid of mapped samples cannot be calculated explicitly on RAM of PC, this preprocessing can be built into its algorithm. However, in SubKPCA, the definition of mean vector is not straightforward. Let  $\mathbf{1}_N$  be a  $N$ -dimensional vector that has an element ‘1’ in each dimension. We here consider three kinds of centroids vector,

1. mean of all training samples  $\bar{\Phi}_1 = \frac{1}{N} \sum_{i=1}^N \Phi(\mathbf{x}_i) = \frac{1}{N} S\mathbf{1}_N$ .
2. mean of the subset of samples  $\bar{\Phi}_2 = \mathbf{y}_1, \dots, \mathbf{y}_M = \frac{1}{M} T\mathbf{1}_M$ .
3.  $\bar{\Phi}_3 = \min_{\mathbf{v} \in \mathcal{R}(T)} \|\mathbf{v} - \frac{1}{N} \sum_{i=1}^N \Phi(\mathbf{x}_i)\|$ .

The first one is the simple mean of the all training samples. This is a natural definition for the centroid, however, if  $N$  is very huge, computational complexity is very large.  $\bar{\Phi}_2$  is the simple mean of the subset.  $\bar{\Phi}_3$  is the best approximation of  $\bar{\Phi}_1$  in the space spanned by the mapped subset  $\Phi(\mathbf{y}_1), \dots, \Phi(\mathbf{y}_M)$  in the sense of Euclidean distance. Since SubKPCA is constrained to the space spanned by the mapped subset  $\Phi(\mathbf{y}_1), \dots, \Phi(\mathbf{y}_M)$ ,  $\bar{\Phi}_1$  and  $\bar{\Phi}_3$  are equivalent in SubKPCA. However, for example, in the pre-imaging stage, as we described in the next section, we use the other cost function e.g., (15). In this case, the problem is not limited in the space, and  $\bar{\Phi}_1$  and  $\bar{\Phi}_3$  are not equivalent.

Let us consider the case of  $\bar{\Phi}_3$ . The linear combination of  $\Phi(\mathbf{y}_1), \dots, \Phi(\mathbf{y}_M)$  is

$$\sum_{i=1}^M \alpha_i \Phi(\mathbf{y}_i) = T\boldsymbol{\alpha}, \tag{6}$$

where  $\alpha_1, \dots, \alpha_M$  and  $\boldsymbol{\alpha}$  are coefficients and coefficient vector respectively. We obtain the optimum  $\boldsymbol{\alpha}$  that minimizes the distance between  $T\boldsymbol{\alpha}$  and the centroid of all training samples,  $\frac{1}{N} \sum_{i=1}^N \Phi(\mathbf{x}_i)$  in the feature space,

$$\min_{\boldsymbol{\alpha}} \|T\boldsymbol{\alpha} - \frac{1}{N} \sum_{i=1}^N \Phi(\mathbf{x}_i)\|^2. \tag{7}$$

This is a simple least square problem, and the solution can be calculated easily. If  $K_y$  is not singular, the cost function is minimized by

$$\boldsymbol{\alpha}^* = \frac{1}{N} K_y^{-1} K_{xy}^\top \mathbf{1}_N. \tag{8}$$

Consequently, we have the approximated mean vector,

$$\bar{\Phi} = T\boldsymbol{\alpha}^* = \frac{1}{N} T K_y^{-1} K_{xy}^\top \mathbf{1}_N. \tag{9}$$

Suppose that centered samples are  $\bar{\Phi}(\mathbf{x}_i) = \Phi(\mathbf{x}_i) - \bar{\Phi}$ , and let  $\bar{S} = [\bar{\Phi}(\mathbf{x}_1) \dots \bar{\Phi}(\mathbf{x}_N)]$ . Then we have only to replace  $\Phi(\mathbf{x}_i)$  by  $\bar{\Phi}(\mathbf{x}_i)$  in the cost function of (4), i.e., we have only to replace  $S$  by  $\bar{S}$ .

$$\begin{aligned} \bar{K}_{xy} &= \bar{S}^\top T, \\ \bar{K}_{xy}^\top \bar{K}_{xy} &= K_{xy}^\top K_{xy} - \frac{1}{N} (K_{xy}^\top \mathbf{1}_N)(\mathbf{1}_N^\top K_{xy}). \end{aligned}$$

Therefore, the solution of centered SubKPCA is given by eigenvectors of the generalized eigenvalue problem,

$$\left( K_{xy}^\top K_{xy} - \frac{1}{N} (K_{xy}^\top \mathbf{1}_N)(\mathbf{1}_N^\top K_{xy}) \right) \mathbf{z} = \lambda K_{xy} \mathbf{z}. \tag{10}$$

Let  $\tilde{\mathbf{z}}_1, \dots, \tilde{\mathbf{z}}_r$  be the sorted eigenvectors, and  $Z_2 = [\tilde{\mathbf{z}}_1 \dots \tilde{\mathbf{z}}_r] \in \mathbb{R}^{M \times r}$ , then the transform and the projection of centered SubKPCA are

$$U_2 = TZ_2 \tag{11}$$

$$P_2 = U_2 U_2^\top = T \sum_{i=1}^r \tilde{\mathbf{z}}_i \tilde{\mathbf{z}}_i^\top T^\top = TZ_2 Z_2^\top T^\top. \tag{12}$$

### 4 Pre-image of Centered SubKPCA

A projection of an input pattern in the feature space is also in the feature space. Indeed, we can investigate its properties such as norm, inner product with the other samples, and so forth. However, in many applications such as denoising, we have to pull back the projection to the  $d$ -dimensional input space,  $\mathbb{R}^d$ . In [3], pre-image of the projection of an input vector  $\mathbf{x}$ , is obtained by following criterion,

$$\min_{\mathbf{z} \in \mathbb{R}^d} \rho(\mathbf{z}) = \|P_{\text{KPCA}} \Phi(\mathbf{x}) - \Phi(\mathbf{z})\|^2. \tag{13}$$

The criterion seeks a vector  $\mathbf{z}$  such that  $\Phi(\mathbf{z})$  is the closest to  $P_{\text{KPCA}} \Phi(\mathbf{x})$ . If we use the Gaussian kernel function,  $k(\mathbf{x}_1, \mathbf{x}_2) = \exp(-c\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$ , we have  $k(\mathbf{z}, \mathbf{z}) = 1$ . In such a case, if we let  $\boldsymbol{\gamma} = V\Lambda^{-1}V^\top S^\top \Phi(\mathbf{x})$ , we have  $\rho(\mathbf{z}) = -2 \sum_{i=1}^N \gamma_i k(\mathbf{x}_i, \mathbf{z}) + \text{cost.}$ , and an iterative procedure

$$\mathbf{z}_{t+1} = \frac{\sum_{i=1}^N \gamma_i \exp(-c\|\mathbf{z}_t - \mathbf{x}_i\|^2) \mathbf{x}_i}{\sum_{j=1}^N \gamma_j \exp(-c\|\mathbf{z}_t - \mathbf{x}_j\|^2)}. \tag{14}$$

This iteration minimizes  $\rho(\mathbf{z})$  [3].

On the other hand, when we use the centered KPCA or the centered SubKPCA, the criterion should be [4],

$$\min_{\mathbf{z} \in \mathbb{R}^d} \rho_2(\mathbf{z}) = \|P(\Phi(\mathbf{x}) - \bar{\Phi}) - (\Phi(\mathbf{z}) - \bar{\Phi})\|^2. \tag{15}$$

We here provide its solution for the centered SubKPCA, and  $\bar{\Phi}$  is given by eq. (9). From  $P_2 = TZ_2Z_2^T T^T$ , and  $\bar{\Phi} = \frac{1}{N}TK_y^{-1}K_{xy}^T \mathbf{1}_N$ ,  $\rho_2(\mathbf{z})$  yields

$$\begin{aligned} \rho_2(\mathbf{z}) = & \langle (-2Z_2Z_2^T T^T \Phi(\mathbf{x}) + \frac{2}{N}Z_2Z_2^T K_{xy}^T \mathbf{1}_N \\ & - \frac{2}{N}K_y^{-1}K_{xy}^T \mathbf{1}_N) | T^T \Phi(\mathbf{z}) \rangle + \text{const.} \end{aligned} \quad (16)$$

If we let  $\tilde{\gamma} = -2Z_2Z_2^T T^T \Phi(\mathbf{x}) + \frac{2}{N}Z_2Z_2^T K_{xy}^T \mathbf{1}_N - \frac{2}{N}K_y^{-1}K_{xy}^T \mathbf{1}_N$ , we have  $\rho_2(\mathbf{z}) = \sum_{i=1}^M \tilde{\gamma}_i k(\mathbf{y}_i, \mathbf{z}) + \text{const.}$ , and iterative procedure by replacing  $\gamma_i$  by  $\tilde{\gamma}_i$  in eq. (14).

## 5 Experiment

### 5.1 Preliminary Experiment I – Centered vs. Non-Centered

Before we demonstrate the proposed method, we here compare denoising results of centered KPCA and non-centered KPCA.

We used the standard images “Lena” and “Barbara.” The procedure of the experiment is as follows;

1. Add Gaussian distributed noise.
2. Make 5x5 pixel patches from the noisy images.
3. Obtain representative  $K$  samples by K-means clustering.
4. Obtain projector of standard KPCA using the representative samples.
5. Obtain pre-image of all noisy patches.
6. Reconstruct images from obtained patches.

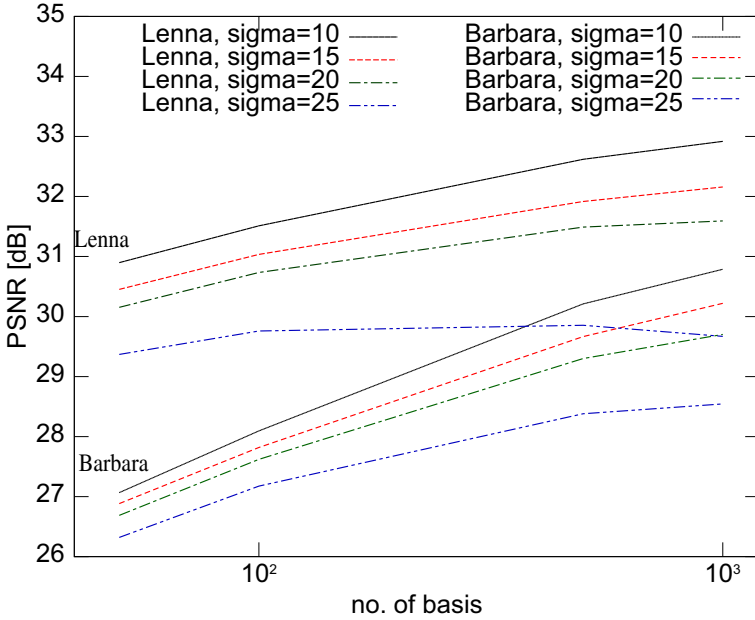
The parameters we used were as follows; 1) the number of representative samples,  $K = 500$ ; 2) the number of principal components,  $r = 50$ ; 3) the parameter of the Gaussian kernel function,  $c = 100$ .

Table 1 shows the results of the experiment. We tried four different standard deviations (SD), the values in the table are PSNR (peak signal to noise ratio) in dB. “Noisy” is the PSNR of each noisy image. From Table 1, centered KPCA always shows higher PSNR than non-centered KPCA. The pre-processing that removes mean vector enhances denoising performance.

**Table 1.** PSNR [dB] of preliminary experiment I, centered vs. non-centered. (L) stands for “Lena,” and (B) stands for Barbara.

SD of Noise	10	15	20	25
Noisy (L)	28.13	24.60	22.10	20.17
Non-centered (L)	28.87	29.00	30.32	29.54
Centered (L)	32.61	31.93	31.49	29.87
Noisy (B)	28.13	24.60	22.10	20.17
Non-centered (B)	28.51	27.12	27.62	27.40
Centered (B)	30.18	29.65	29.29	28.37





**Fig. 1.** Result of preliminary experiment II, relation between the number of samples and performance

## 5.2 Preliminary Experiment II – Relation between the Number of Samples and Performance

Next, we investigate the relation between the number of representative samples and the denoising performance. Settings of the experiment are the same as the previous experiment. Figure 1 shows the result of the experiment. It can be seen that the more samples gives the better performance expect the case Lena, SD=25. It is expected that if we use more samples, the denoising performance will be increased. However, as we described, KPCA has high computational complexity that depends on the number of samples  $N$ , and it increases with  $N^3$  order.

The exception, Lena SD=25, is due to too large noise. Since K-means extracts the representative patches that are centroids of several samples. Therefore, the representative samples have less noise than original patches, if the number of centroids,  $K$  is sufficiently small. K-means itself works for denoising in this case, however, this efficiency is limited, and in many cases, the performance is better if we use larger number of the representative samples.

## 5.3 Centered Subset KPCA

From previous two preliminary experiments, we found that 1) pre-centering enhances the performance of denoising with KPCA; 2) the more number of samples shows that better denoising performance in many cases. However, KPCA has large computational complexity. The proposed method, centered SubKPCA solves this problem.

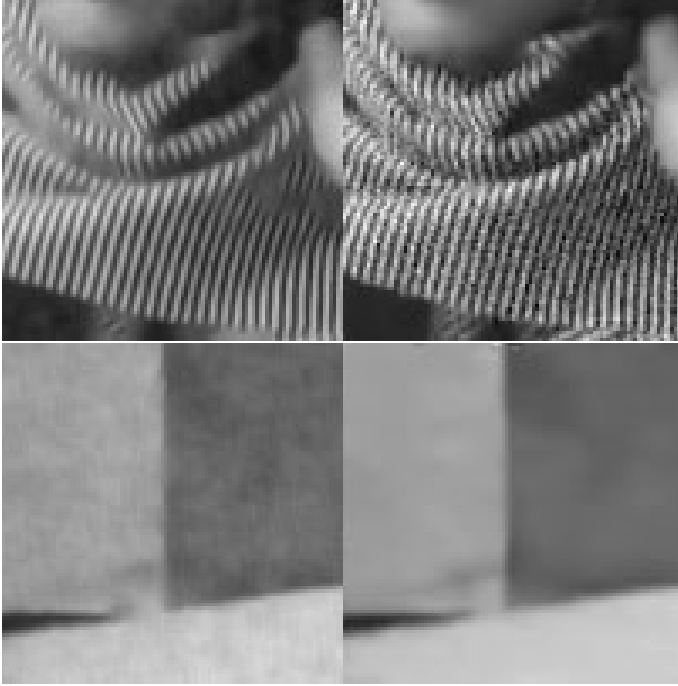
**Table 2.** Result of Denoising performance, PSNR in dB, Lena, “nc” stands for non-centered

SD of Noise	10	15	20	25
SubKPCA	32.84	32.00	30.57	30.06
SubKPCA(nc)	29.96	28.89	29.76	28.65
FoE <a href="#">[6]</a>	35.04	33.27	31.92	30.82

**Table 3.** Result of Denoising performance, PSNR in dB, Barbara, “nc” stands for non-centered

SD of noise	10	15	20	25
SubKPCA	31.31	30.53	28.52	27.86
SubKPCA(nc)	24.67	25.30	26.56	26.29
FoE <a href="#">[6]</a>	32.83	30.22	28.32	27.04

**Fig. 2.** Denoising results: Barbara SD of noise=20, Top-left: Noisy image, Top-right: Proposed, Bottom-left: FoE ([http://www.gris.informatik.tu-darmstadt.de/~sroth/research/foe/denoising\\_results.html](http://www.gris.informatik.tu-darmstadt.de/~sroth/research/foe/denoising_results.html)), Bottom-right: Wiener filter (PSNR: 27.12 [dB])



**Fig. 3.** Enlarges of denoising images, Barbara, SD of noise=20: left: proposed, right: FoE

It should be noted that the number of patches is  $N = 2.5 \times 10^5$  since the images are  $512 \times 512$  pixels, and if we use all patches, the kernel Gram matrix requires about  $N(N + 1)/2 \times 4 \simeq 130$  Giga bytes in four-bytes floating point system. It is almost impossible to store and obtain EVD of such matrix. Even though when  $N = 2.5 \times 10^5$ , we can obtain the matrix  $K_{xy}^\top K_{xy} \in \mathbb{R}^{M \times M}$  and the vector  $K_{xy}^\top \mathbf{1}_N$  by the segmentation technique. The dominant calculation times of our method are 1) K-means:  $1.0 \times 10^4$  seconds; 2)  $K_y$ :  $1.2 \times 10^0$  seconds; 3)  $K_{xy}^\top \mathbf{1}_N$ :  $1.7 \times 10^2$  seconds; 4)  $K_{xy}^\top K_{xy}$ :  $3.7 \times 10^2$  seconds; 5) EVD:  $1.6 \times 10^1$  seconds; 6) Pre-image:  $2.9 \times 10^2$  seconds, on quad-core 2.66GHz Intel CPU. K-means has the highest computational complexity, however, this may be reduced by speeding up techniques such as early stopping. Other KPCA based approaches [4,3] use noise free database. On the other hand, our algorithm makes and select patches only from the noisy image.

Tables 2 and 3 show denoising results in PSNR [dB] of Lena and Barbara. Although we just apply simple Subset KPCA, the proposed method is comparable with the Fields of Experts (FoE) [6] for denoising of Barbara. From the tables, it can be seen that the pre-centering enhances the denoising performance, and proposed approximation of centroids works well.

Figure 2 compares denoising images of Barbara and Lenna, respectively when (SD of noise)=20. From the figure, comparing to FoE, the proposed method

restores texture areas well, but not for flat area. We show enlarged images in Figure 3. The stripe pattern of Barbara is restored by our method well, while we can see remaining noise in the flat wall part. Since Lena has less texture part and the larger flat part, our method does not show good result. However, this problem may be solved by using sub-band decomposition techniques e.g., wavelet decomposition. Then it is expected that the performance of the proposed centered SubKPCA will be improved.

## 6 Conclusion

We proposed the centered Subset KPCA in this paper. In order to apply pre-centering to the Subset KPCA, we have to obtain centroid of the patterns. When the number of samples  $N$  is also very large, its computational complexity will be very large when we use a centroid of all samples. We introduced efficient approximation technique for this problem.

Experimental results showed that even simple applications of Subset KPCA for denoising is comparable with state-of-art denoising method, Field of Experts (FoE). It is expected that the performance of the proposed centered SubKPCA will be enhanced by sub-band decomposition techniques such as wavelet decomposition.

## References

1. Schölkopf, B., Smola, A.J.: Learning with Kernels; Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
2. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10, 1299–1319 (1998)
3. Mika, S., Schölkopf, B., Smola, A.J., Müller, K.R., Scholz, M., Rätsch, G.: Kernel PCA and de-noising in feature spaces. In: Kearns, M.S., Solla, S.A., Cohn, D.A. (eds.) *Advances in Neural Information Processing Systems 11*. MIT Press, Cambridge (1999)
4. Kim, K.I., Franz, M.O., Schölkopf, B.: Iterative kernel principal component analysis for image modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence* 27, 1351–1366 (2005)
5. Washizawa, Y.: Subset kernel principal component analysis. In: *Proc. of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP 2009)*, p. 40 (2009)
6. Roth, S., Black, M.J.: Fields of experts. *International Journal of Computer Vision* 82, 205–229 (2009)

# On the Behavior of Kernel Mutual Subspace Method

Hitoshi Sakano<sup>1</sup>, Osamu Yamaguchi<sup>2</sup>, Tomokazu Kawahara<sup>3</sup>, and Seiji Hotta<sup>4</sup>

<sup>1</sup> NTT Communication Science Lab.

2-4, Hikaridai, Seika-cho, gKeihanna Science Cityh Kyoto 619-0237 Japan

sakano.hitoshi@lab.ntt.co.jp

<sup>2</sup> Power and Industrial System R&D Center, Toshiba Corporation Power Systems Company

1, Toshiba-cho, Fuchu-Shi, Tokyo, 183-8511, Japan

osamu1.yamaguchi@toshiba.co.jp

<sup>3</sup> Corporate R&D Center, Toshiba Corporation

1 Komukai Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan

tomokazu.kawahara@toshiba.co.jp

<sup>4</sup> The Graduate School of Engineering, Tokyo University of Agriculture and Technology,

2-24-16 Naka-cho, Koganei-shi, Tokyo, 184-8588 Japan

s-hotta@cc.tuat.ac.jp

**Abstract.** Optimizing the parameters of kernel methods is an unsolved problem. We report an experimental evaluation and a consideration of the parameter dependences of kernel mutual subspace method (KMS). The following KMS parameters are considered: Gaussian kernel parameters, the dimensionalities of dictionary and input subspaces, and the number of canonical angles. We evaluate the recognition accuracies of KMS through experiments performed using the ETH-80 animal database. By searching exhaustively for optimal parameters, we obtain 100% recognition accuracy, and some experimental results suggest relationships between the dimensionality of subspaces and the degrees of freedom for the motion of objects. Such results imply that KMS achieves a high recognition rate for object recognition with optimized parameters.

## 1 Introduction

In recent decades, various types of mutual subspace method (MSM) [1,2] have been proposed for object recognition. These methods classify a set of test samples using the angles between subspaces spanned by test and training samples. Such approaches may improve recognition accuracy but their accuracies deteriorate when we apply them to samples that are difficult to classify linearly. To overcome this difficulty, one of the present authors has proposed a simple extension of MSM called kernel mutual subspace method (KMS) [3,4,5]. Owing to the high accuracy of KMS method, it is widely applied to real-life problems such as object recognition with large pose variation, lip movement recognition [6], the surveillance of vehicles and walking humans [7], speaker recognition [8], and space craft anomaly detection [9].

Because in many cases KMS can achieve a higher accuracy than the original MSM, some types of KMS [10] have been studied experimentally. However, one question arises, namely have the KMS parameters really been optimized in previous research? Because KMS has many parameters, parameter exploration a combinatorial explosion,

i.e., we should search for the optimum combination of parameters in a direct product space spanned by a kernel parameter, the dimensionalities of test and training subspaces, and the number of canonical angles. Moreover, nobody has yet verified the relationship between the kernel parameters and the other parameters.

To confirm the relationship between KMS parameters, we evaluate the recognition accuracies of KMS through experiments employing the ETH-80 [11] animal database. By searching exhaustively for the optimum parameters, we obtain 100% recognition accuracy, and certain experimental results indicate the relationships between the dimensionality of subspaces and the degrees of freedom for the motions of objects. Such results suggest that KMS achieves a high recognition rate for object recognition with optimized parameters. This paper is organized as follows. First, KMS is introduced in section 2. Next, we report and discuss experimental results obtained with the ETH-80 dataset in section 3. Finally we conclude and summarize this paper.

## 2 Kernel Mutual Subspace Method

Before introducing KMS, we summarize MSM to provide some background. Let  $c$  be the numbers of classes denoted as  $\{\omega_1, \dots, \omega_c\}$ . Suppose that the  $i$ th training object has a single class label  $y_i \in \{\omega_1, \dots, \omega_c\}$ . In our study, several images (samples) are observed of a single object. We represent one of them as a  $d$ -dimensional vector  $x = (x_1, \dots, x_d)^\top$ . As a result, our aim is to classify a test (unknown) object appropriately class using sets of samples.

### 2.1 Mutual Subspace Method (MSM)

In MSM, we first represent sets of test and training samples as linear subspaces. For this, principal component analysis (PCA) is applied to training samples belonging to the same class (training phase), and it is also adopted for test samples obtained from an unknown object. Note that the mean vector of the training samples is zero. Let  $\mathbf{U} = \{u_1|u_2|\dots|u_r\} \in \mathbb{R}^{d \times r}$  and  $\mathbf{V}_j = \{v_1|v_2|\dots|v_m\} \in \mathbb{R}^{d \times m}$  be the transform matrices obtained by applying PCA to test samples and training samples belonging to class  $j$ , respectively. The  $i$ th components of  $\mathbf{U}$  and  $\mathbf{V}_j$  (i.e.,  $u_i$  and  $v_i$ ) are the  $d$ -dimensional eigenvectors corresponding to the  $i$ th largest eigenvalues of their covariance matrices. Note that we assume  $r \leq m$  in all cases for simplicity, where  $r$  and  $m$  are the dimensionalities of the test and training subspaces, respectively.

MSM classifies an unknown object based on a similarity defined as the angles between  $\mathbf{U}$  and  $\mathbf{V}$ . To measure this similarity, we construct the following  $r \times r$  matrix:

$$\mathbf{Z}_j = \mathbf{U}^\top \mathbf{V}_j \mathbf{V}_j^\top \mathbf{U}, \quad Z_{ij} = \sum_{l=1}^m (u_i^\top v_l) \cdot (v_l^\top u_j). \tag{1}$$

In linear algebra, the eigenvalues of  $\mathbf{Z}_j$  indicate  $\cos^2 \theta_j^2$ s between  $\mathbf{U}$  and  $\mathbf{V}_j$ , i.e., the largest eigenvalue is equal to the maximum  $\cos^2 \theta_j^2(1)$ , and the second largest eigenvalue is the second largest  $\cos^2 \theta_j^2(2)$  etc. These  $\cos^2 \theta^2(\cdot)$ s are called canonical angles [12].

In the original MSM, the following classification rule is introduced: The class of the unknown object (denoted by  $\omega$ ) is determined as

$$\max_{j=1,\dots,c} \{\cos \theta_j^2(1)\} = \cos \theta_{j^*}^2(1) \Rightarrow \omega = \omega_{j^*}. \tag{2}$$

As previously reported [12], the number of canonical angles and how they are used have an effect on recognition accuracy. For example, we can improve the accuracy in some cases by using the mean values of the canonical angles ( $\sum_{i=1}^r \cos \theta_j^2(i)/r$ ) instead of using the largest one only. On the other hand, Kim has proposed a canonical angle fusion method using Adaboost [13] to improve accuracy. Maeda tried to clarify the role of canonical angles for object recognition, i.e., the second canonical angle related to the direction of motion. The hypothesis may not yet be definitely confirmed [14] but some experimental results suggest that the second or later canonical angles are important.

**2.2 Kernel Mutual Subspace Method (KMS)**

Before deriving KMS, a summary of kernel principal component analysis (KPCA) [15] may help us to understand KMS. KPCA is performed by carrying out singular value decomposition in a functional space  $\mathcal{F}$  for a given set of samples  $x_i, i = 1, \dots, m$  in a  $d$ -dimensional feature space  $\mathbb{R}^d$ . We can define a functional space  $\mathcal{F}$ , which is related to the feature space, possibly by non-linear mapping:

$$\Psi : \mathbb{R}^d \rightarrow \mathcal{F}, \quad x \rightarrow X. \tag{3}$$

Note that there is a possibility that the functional space  $\mathcal{F}$  will have infinite dimensionality. In the functional space  $\mathcal{F}$ , a covariance matrix can be written as follows:

$$\mathbf{C} = \frac{1}{m} \sum_{i=1}^m (\Psi(x_i)\Psi(x_i)^\top). \tag{4}$$

Its eigenvectors may be given by diagonalization but the matrix is too large (sometimes infinitely) to solve it with practical computation cost. To overcome this difficulty, we use an  $m \times m$  kernel matrix defined as follows:

$$K_{ij} = \Psi(x_i)^\top \Psi(x_j), \tag{5}$$

For computing desired eigenvectors, we first solve the following eigenvalue problem:

$$m\lambda \alpha = \alpha K, \tag{6}$$

where  $\alpha = (\alpha_1, \dots, \alpha_m)^\top$  is a column vector whose components are coefficients for corresponding samples  $x_1, \dots, x_m$ .

To extract the principal components, we have to compute the orthogonal projection onto eigenvectors  $\mathbf{V}_j$  in  $\mathcal{F}$ . Let  $\Psi(x)$  be a sample in  $\mathcal{F}$ . The orthogonal projection of  $\Psi(x)$  onto  $\mathbf{V}_j$  can be calculated by

$$\mathbf{V}_j^\top \Psi(x) = \frac{1}{\lambda} \sum_{i=1}^m \alpha_i \Psi(x_i)^\top \Psi(x). \tag{7}$$

This vector is called a nonlinear principal component corresponding to  $\Psi$ . As mentioned above, the computation cost extremely large (or infinite), so Schölkopf has introduced the Mercer kernel that satisfies

$$k(x, y) = \Psi(x)^\top \Psi(y). \tag{8}$$

By using this trick, the computation of a dot product  $\Psi(x)^\top \Psi(y)$  can be replaced with  $k(x, y)$ , i.e.,

$$\mathbf{V}_j^\top \Psi(x) = \frac{1}{\lambda} \sum_{i=1}^m \alpha_i k(x_i, x). \tag{9}$$

This result shows that we can calculate a projection for the nonlinear principal components in finite time without an explicit form of  $\mathbf{V} \in \mathcal{F}$ .

Now we can derive KMS as combination of MSM and KPCA, i.e., we can define a similarity measure for KMS in a functional space  $\mathcal{F}$ . Practical applications demand lower computational costs, so we have to prove that KMS takes a finite time to compute angles in a functional space  $\mathcal{F}$ .

Let  $\mathbf{U}$  and  $\mathbf{V}_j$  be matrices formed by eigenvectors obtained from  $r$  test samples  $\Psi(x_1), \dots, \Psi(x_r)$  and  $m$  training samples belonging to class  $j$ , i.e.,  $\{\Psi(x_1), \dots, \Psi(x_m)\} \in \omega_j$ . They can be represented as

$$\mathbf{U} = \sum_{l=1}^r \alpha_l \Psi(x_l), \quad \mathbf{V}_j = \sum_{i=1}^m \alpha_i \Psi(x_i). \tag{10}$$

The similarity between them can be computed using the dot product  $\mathbf{U}^\top \mathbf{V}_j$ :

$$\begin{aligned} \mathbf{U}^\top \mathbf{V}_j &= \left( \sum_{l=1}^r \alpha_l \Psi(x_l) \right)^\top \left( \sum_{i=1}^m \alpha_i \Psi(x_i) \right) \\ &= \sum_{l=1}^r \sum_{i=1}^m \alpha_l \alpha_i \Psi(x_l)^\top \Psi(x_i) = \sum_{l=1}^r \sum_{i=1}^m \alpha_l \alpha_i k(x_l, x_i). \end{aligned} \tag{11}$$

Since the numbers of  $r$  and  $m$  are limited, this dot product of two subspaces takes a finite time to compute. To obtain the angles between two subspaces, substitute (11) into (1).

### 3 Experiment on ETH80 Animal Database

#### 3.1 Experimental Setup

We used the open database ETH-80 (11), and selected 30 classes from it namely dogs, cows, and horses consisting of 10 classes each. They all have very similar forms (Fig. 1). Each class consists of images of three-dimensional models from 41 view-points (Fig. 2). The viewpoints are the same for all the classes. We separated 41 images of each class into 21 training images with odd numbers, and 20 validation images with even numbers (Fig. 2). Therefore, the viewpoints of the training images were different from those of the validation images. On the other hand, the validation data consisted of 10 images



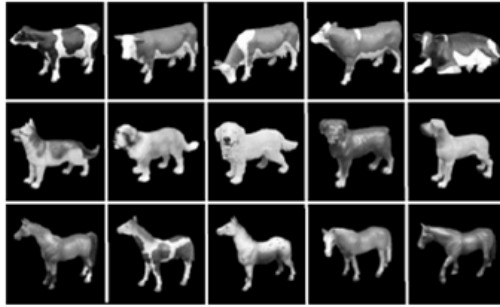


Fig. 1. Examples of animal images from ETH-80 [11]

whose frame numbers were from  $i$  to  $i + 9$ . We prepared creating validation data 10 times by varying  $i$  from 1 to 10. Consequently, the number of validations was 300 ( $= 10 \times 30$ ). This setup was the same as that reported in [16].

We used MATLAB7.6 and an image processing toolbox corresponding to this version on a standard PC that had a 3.2GHz CPU and 12Gb RAM.

### 3.2 KMS Parameters

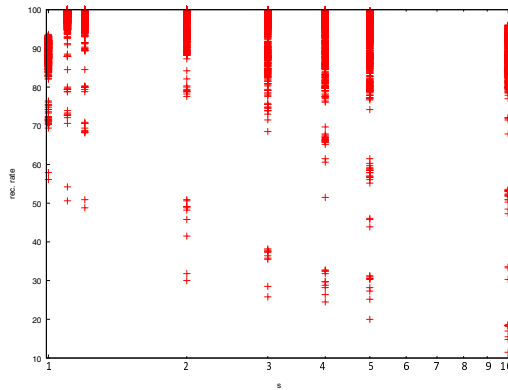
KMS has four parameters: the dimensionalities of subspaces  $r$  and  $m$ , the number of canonical angles (denoted as  $n_{ca}$ ), and a kernel parameter. As the kernel function, we used the RBF kernel:  $k(x, y) = \exp(-s \times a \times \|x - y\|^2)$ , where  $s$  is scale parameter (described later). The kernel parameter of KMS is related to the complexity of the distribution of objects in a feature space, hence we should adjust parameter  $a$  empirically according to the sample distribution complexity. In the following experiments, we determined  $a$  using a heuristic search based on the mean of the average Euclidean distances between all the training samples in individual classes [18] with a scale factor  $s$ . In subspace methods, the dimensionality of a subspace spanned by samples dominates the representation capacity of variations of an object and the approximation error of an object in a feature space. Therefore, high dimensional subspaces may represent large object variations but their recognition accuracies will deteriorate because the intersection of subspaces of different classes will be considerable. However, this problem hardly ever occurs on a kernel-based subspace method. In contrast, the physical meaning of the number of canonical angles is not clear. There have been few experimental results implying that the canonical angles have physical meanings, and the optimum combination of canonical angles for recognition remains an open issue. In the following experiments, we investigated all combinations of  $r$ ,  $m$ , and  $n_{ca}$  for every scale parameters  $s = 1.0 \sim 10.0$ . Consequently, the number of combinations were 924.

### 3.3 Experimental Results

First, we analyzed the dependence of the scale parameter  $s$  on KMS. Figure 3 is a plot in which each point indicates the recognition rate with respect to  $s$ , i.e., the horizontal axis indicates the value of the scale parameter  $s$ . As shown in the figure, the recognition



**Fig. 2.** All samples of dog1. Training samples are enclosed by dotted lines.



**Fig. 3.** Kernel parameter dependence on recognition rates

rates around  $s = 1.0$  to  $2.0$  were higher than the others. To obtain more detail, we plotted the frequency of parameters whose recognition rates were high in Fig. 4. The red, green and blue lines show the number of parameters that achieved recognition rates of over 98%, 99%, and 100%, respectively. These results were obtained by varying  $s$  from 1.1 to 2.0. As shown in this figure, the recognition rates of half of the parameters in this area at least than 98%. These results were much better than those previously reported in [16,17]. On the basis of these results, we investigated other parameters limited to only these areas in further experiments.

Figure 5 shows the properties of dependence with respect to  $r$  and  $m$  on recognition rates with  $s = 1.1$ . This indicates that KMS has asymmetric dimensionalities: There were some cases where the recognition rates  $r = 1$  were higher than others, however,

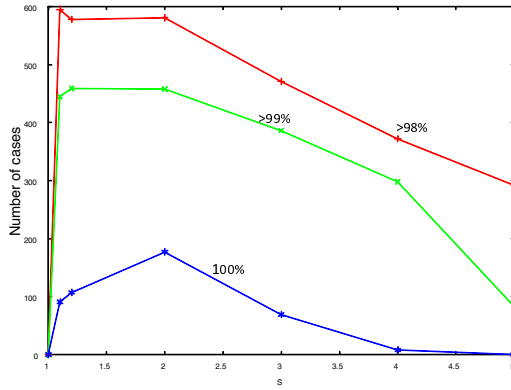


Fig. 4. Kernel parameter dependence on recognition rates (frequency)

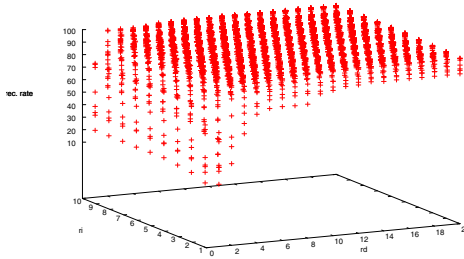
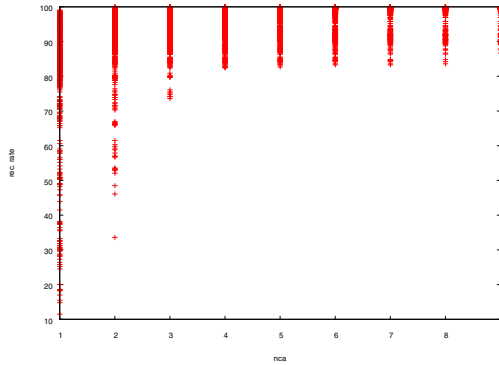


Fig. 5. Dimensionality dependence on recognition rates

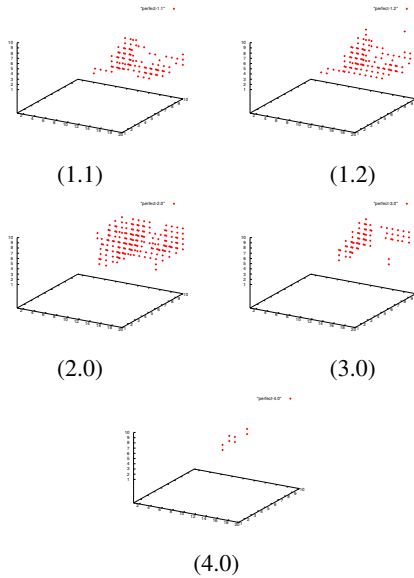
recognition rates  $m = 1$  were never higher than others. This result supports the hypothesis described in [3] namely that the optimum dimensionality corresponds to the degrees of freedom of objects. In other words, the above result implies that the degree of freedom of motion on an unknown object that consist of few motions can be approximated by a 1-dimensional subspace. In contrast, those of training objects that consist of various motions cannot be approximated solely by a 1-dimensional subspace.

Figure 6 shows the relation between the number of canonical angles and the recognition rates. As shown in this figure, larger numbers of canonical angles achieved better recognition rates. This fact cannot be confirmed solely from only result because the dimensionalities of the test and training subspaces are large when the number of canonical angles is large. Note that there were no cases that achieved 100% recognition accuracy with  $n_{ca} = 1$ .

Finally, we analyzed those cases that achieved 100% recognition rates. Figure 7 shows the three parameters  $r$ ,  $m$ , and  $n_{ca}$  that achieved 100% recognition rates with scale parameters 1.1, 1.2, 2.0, 3.0, and 4.0. As shown in this figure, the largest number of parameters that achieved 100% accuracy were 177 ( $s = 2.0$ ). In addition, the number of such parameters decreased as the scale factor increased.



**Fig. 6.** Recognition rates with respect to number of canonical angles



**Fig. 7.** Parameters that achieved 100% recognition rates

## 4 Conclusions

This paper reported an experimental evaluation of the parameters variation of KMS, i.e., Gaussian kernel parameters, dimensionalities of dictionary and input subspaces, and the number of canonical angles. After an exhaustive search for the optimum parameters, we obtained 100% recognition accuracy with the ETH-80 animal database. Some of our experimental results suggested a relationship between the dimensionalities of subspaces and the degrees of freedom of object motions. Such results implied that KMS

will achieve high recognition rates for object recognition with optimized parameters. We will now attempt to evaluate the effectiveness of feature extraction for parameters and verify the recognition accuracy using other datasets.

## References

1. Maeda, K.-i., Watanabe, S.: Pattern matching method with local structure. Trans. on IEICE (D) 68-D(3), 345–352 (1985) (Japanese Edition); Cipolla, R., et al. (ed.) recent English version is, Computer Vision: Detection, Recognition and Reconstruction (Studies in Computational Intelligence) Part 5. From the Subspace Methods to the Mutual Subspace Method. SCI. Springer Heidelberg (2010)
2. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: Proc. of IEEE 4th Int'l. Conf. on Face and Gesture Recognition, pp. 318–323 (1998)
3. Sakano, H., Mukawa, N.: Kernel mutual subspace method for robust facial image recognition. In: Proc. of 4th Int'l. Conf. on Knowledge based Engineering System, Brighton, vol. 1, pp. 245–248 (2000)
4. Sakano, H., Mukawa, N., Nakamura, T.: Kernel mutual subspace method and its application for object recognition. Electronics and Communications in Japan E88(6), 45–53 (2005)
5. Sakano, H., Suenaga, T.: Classifiers under continuous observations. In: Caelli, T.M., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 798–805. Springer, Heidelberg (2002)
6. Ichino, M., Sakano, H., Komatsu, N.: Speaker recognition using kernel mutual subspace method. In: Proc. of ICARCV (2004)
7. Zhang, B., Park, J., Ko, H.: Combination of self-organization map and kernel mutual subspace method for video surveillance. Advanced Video and Signal Based Surveillance, 123–128 (2007)
8. Ichino, M., Sakano, H., Komatsu, N.: Text-indicated speaker recognition using kernel mutual subspace method. In: Proc. of ICARCV, Singapore, p. 11027 (2008)
9. Fujimaki, R., Yairi, T., Machida, K.: An approach to spacecraft anomaly detection problem using kernel feature space. In: 11th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2005), pp. 401–410 (2005)
10. Fukui, K., Yamaguchi, O.: A Theoretical Extension of the Subspace Method and its Application for 3D Object Recognition. IPSJ Transactions on Computer Vision and Image Media 46(SIG 15(CVIM 12)), 21–34 (2005) (in Japanese)
11. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: CVPR, pp. 409–415 (2003)
12. Chatelin, F.: Veleurs propres de matrices. Masson, Paris (1988) (in French)
13. Kim, T.-K., Arandjelovic, O., Cipolla, R.: Boosted manifold principal angles for image set-based recognition. Pattern Recognition 40(9), 2475–2484 (2007)
14. Maeda, K., Yamaguchi, O., Fukui, K.: Towards 3-dimensional pattern recognition. In: SSPR/SPR 2004, pp.1061–1068 (2004)
15. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10, 1299–1319 (1998)
16. Fukui, K., Stenger, B., Yamaguchi, O.: A framework for 3D object recognition using the kernel constrained mutual subspace method. In: Proceedings of Asian Conference on Computer Vision, pp. 315–324 (2006)

17. Fukui, K., Yamaguchi, O.: The kernel orthogonal mutual subspace method and its application to 3D object recognition. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) ACCV 2007, Part II. LNCS, vol. 4844, pp. 467–476. Springer, Heidelberg (2007)
18. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. *Int'l J. of Computer Vision* 73(2), 213–238 (2007)

# Compound Mutual Subspace Method for 3D Object Recognition: A Theoretical Extension of Mutual Subspace Method

Naoki Akihiro and Kazuhiro Fukui

Graduate School of Systems and Information Engineering,  
University of Tsukuba, Japan

**Abstract.** In this paper, we propose the Compound Mutual Subspace Method (CPMSM) as a theoretical extension of the Mutual Subspace Method, which can efficiently handle multiple sets of patterns by representing them as subspaces. The proposed method is based on the observation that there are two types of subspace perturbations. One type is due to variations within a class and is therefore defined as “within-class subspace”. The other type, named “between-class subspace”, is characterized by differences between two classes. Our key idea for CPMSM is to suppress within-class subspace perturbations while emphasizing between-class subspace perturbations in measuring the similarity between two subspaces. The validity of CPMSM is demonstrated through an evaluation experiment using face images from the public database VidTIMIT.

## 1 Introduction

In this paper, we propose the Compound Mutual Subspace Method (CPMSM), which has the ability to classify similar sets of patterns accurately. Then we apply it in a face recognition experiment based on multiple images.

Subspace-based methods have recently attracted attention from many researchers who are interested in recognition of 3D objects, such as faces. The mutual subspace method (MSM) [1] is one of the most effective and efficient methods for object recognition, as it can efficiently handle multiple images [2] [3] [4]. In subspace-based methods, including MSM, a pattern composed of  $n \times n$  pixels is usually regarded as a vector  $\mathbf{x}$  in  $n^2$ -dimensional space. MSM represents a set of patterns  $\{\mathbf{x}\}$  from each class through a low-dimensional linear subspace generated from the set by using the Karhunen-Loève (KL) expansion, which is also known as principal component analysis (PCA). Finally, the similarity between two sets of patterns can be readily measured by using canonical angles  $\theta_i$  between two subspaces, as shown in Fig. 1.

Even though MSM is capable of absorbing differences in appearance caused by changes in view point or illumination, compared with conventional methods using a single input image, such as the subspace method [5], the classification performance of MSM is still not sufficiently high. One reason for this is that a

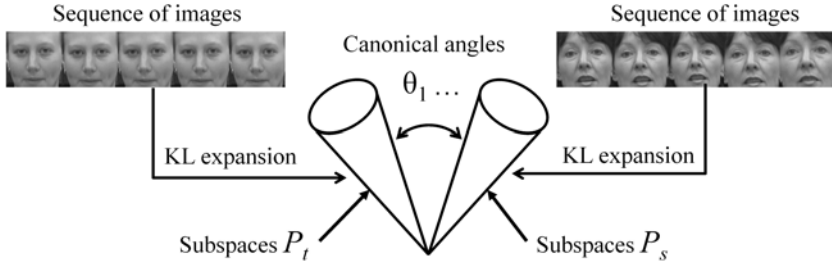


Fig. 1. Concept of Mutual Subspace Method

subspace which provides a satisfactory representation of the distribution of the training patterns in terms of a least-mean-square approximation is not always optimal in terms of classification performance. Many extended methods have been proposed [6] [7] [8] for improving the classification performance of MSM, including the nonlinear extensions [9] [10] [11] [12] using a kernel trick. In this paper, we focus on the Constrained MSM (CMSM) and the Orthogonal MSM (OMSM) [13] since they have been used in the development of the recognition engine of the state-of-the-art face recognition system “FacePass” and have achieved extremely high scores in the Face Recognition Vendor Test (FRVT) 2006 [14].

The essence of these methods is to apply MSM to sets of class subspaces which have been orthogonalized with respect to each other in advance. The implementation of orthogonalization is different in the two methods. In OMSM, all the class subspaces are orthogonalized by using the Fukunaga-Koontz framework [15]. The kernel OMSM executes this operation in extremely high-dimensional feature space in order to ensure complete orthogonalization. CMSM achieves approximate orthogonalization of all the class subspaces by projecting them onto the generalized difference subspace  $\mathcal{D}$ . The kernel CMSM executes the projection in a high-dimensional feature space.

In this paper, we also aim to improve the performance of MSM by introducing the concept of “difference subspace” between two subspaces. This approach is notably different from the orthogonalization operation used in CMSM and OMSM. Our approach is based on the observation that there are two types of subspace perturbations. One type occurs due to differences within a class, while the other is due to differences between separate classes. In this paper, we refer to the former as “within-class subspace  $\mathcal{D}_W$ ” and the latter as “between-class subspace  $\mathcal{D}_B$ ”.

It should be noted that MSM does not distinguish within-class subspace perturbations from between-class subspace perturbations. Thus, MSM cannot distinguish an input subspace between a subspace of a rival class and a subspace of the same class when they have the same canonical angles as a similarity to the input subspace.

This leads us to develop a proper strategy for suppressing within-class subspace perturbations while emphasizing between-class subspace perturbations.



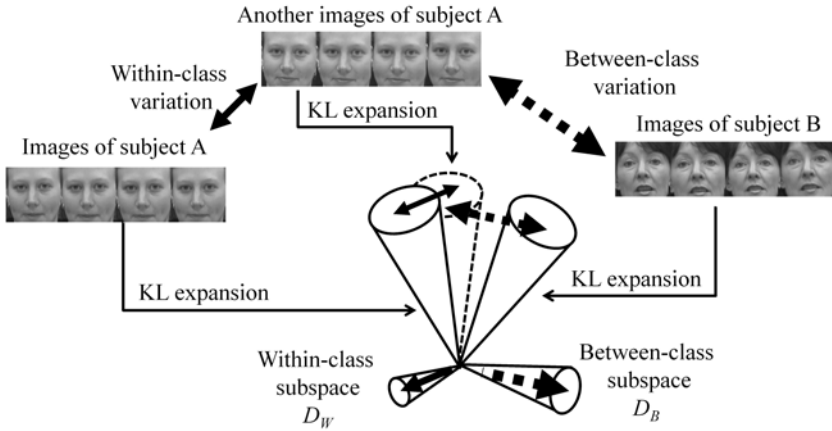


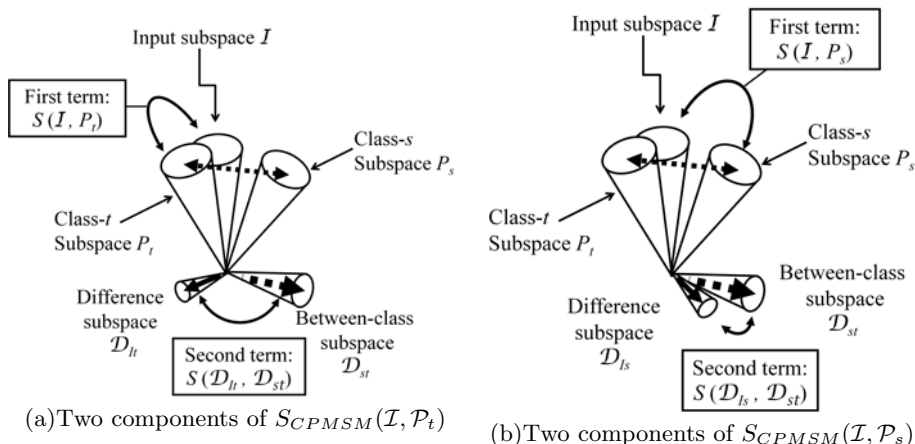
Fig. 2. Two types of difference subspaces

To realize such a strategy, we introduce the concept of “difference subspace” between two subspaces. The concept of difference subspace is a natural extension of the difference vector between two vectors. We can obtain a within-class subspace  $\mathcal{D}_W$  as the difference subspace between two subspaces of the same class, as shown in Fig. 2. On the other hand, we can obtain a between-class subspace  $\mathcal{D}_B$  as the difference subspace between subspaces belonging to different classes.

The essence of the proposed method is to classify a difference subspace  $\mathcal{D}_I$  between an unknown input subspace  $\mathcal{I}$  and a class- $t$  subspace  $\mathcal{P}_t$  into one of two types of subspaces  $\mathcal{D}_W$  and  $\mathcal{D}_B$  by using canonical angles. The similarity obtained through this classification is used to correct the similarity obtained with MSM. We refer to the MSM which takes into account  $\mathcal{D}_W$  and  $\mathcal{D}_B$  utilizing difference subspaces as the “Compound Mutual Subspace Method” (CPMSM).

The advantage of the proposed method is that it can be applied only to limited pairs of class subspaces which are too close and can be easily misclassified. This restriction can reduce the computation time as compared to both CMSM, which projects all class subspaces onto the constraint subspace, and OMSM, which performs orthogonalization of all class subspaces. In addition, the proposed method can be used as post-processing for existing methods, such as MSM, CMSM, and OMSM. Here, we evaluate CPMSM by applying it to a face recognition experiment using a public database containing face images (VidTIMIT audio-video database) [17].

This paper is organized as follows. In Section 2, we explain the concept behind the proposed method and describe the algorithm of CPMSM. In Section 3, the effectiveness of our method is demonstrated through evaluation experiments using a public database containing face images. Finally, Section 4 concludes the paper.



**Fig. 3.** Similarity of the input subspace  $\mathcal{I}$  to each class subspace. This figure shows the case that the input subspace belongs to class  $t$ , (a) the terms of the similarity to  $\mathcal{P}_t$ , (b) the terms of the similarity to  $\mathcal{P}_s$ .

## 2 Compound Mutual Subspace Method (CPMSM)

In this section, we first explain the basic principle of CPMSM. Then, we define a new similarity for CPMSM based on the concept of difference subspace.

### 2.1 The Basic Principle of CPMSM

The basic principle of CPMSM can be explained as follows. When the difference subspace  $\mathcal{D}_I$  between  $\mathcal{I}$  and the class- $t$  subspace  $\mathcal{P}_t$  is similar to the between-class subspace  $\mathcal{D}_B$  and dissimilar to the within-class subspace  $\mathcal{D}_W$ , the input subspace  $\mathcal{I}$  should be classified into class  $t$ . On the other hand, when  $\mathcal{D}_I$  is similar to the within-class subspace  $\mathcal{D}_W$  and dissimilar to  $\mathcal{D}_B$ , the input subspace  $\mathcal{I}$  can be considered to belong to some similar rival class rather than to the class- $t$  subspace  $\mathcal{P}_t$ . The similarity between difference subspaces can be measured by using canonical angles since a difference subspace is a linear subspace, as will be mentioned later.

In practical calculation of the similarity, it is only necessary to measure the similarity between the subspaces  $\mathcal{D}_I$  and  $\mathcal{D}_B$  since  $\mathcal{D}_I$  is projected onto an orthogonal complement of  $\mathcal{P}_t$  in such a way that the projected  $\mathcal{D}_I$  has no components belonging to the within-class subspace  $\mathcal{D}_W$ .

### 2.2 Calculation of Similarity in CPMSM

The similarity  $S_{CPMSM}$  consists of two terms, as follows:

$$S_{CPMSM}(\mathcal{I}, \mathcal{P}_t) = (1 - \mu)S(\mathcal{I}, \mathcal{P}_t) - \mu S(\mathcal{D}_{It}, \mathcal{D}_{st}), \tag{1}$$

where  $\mu$  is a weighting parameter which should be determined experimentally.

In the above equation, the first term  $S(\mathcal{I}, \mathcal{P}_t)$  indicates the similarity between the input subspace  $\mathcal{I}$  and the class- $t$  subspace  $\mathcal{P}_t$ . This similarity is obtained by using MSM. The second term  $S(\mathcal{D}_{It}, \mathcal{D}_{st})$  is the regulation term, which can be obtained as the similarity between two difference subspaces  $\mathcal{D}_{It}$  and  $\mathcal{D}_{st}$ , where  $\mathcal{D}_{st}$  is the difference subspace between the subspace of class  $t$  and that of its similar rival class  $s$ .

In the following paragraphs, we will explain how to apply the above similarity to the task of classifying an input subspace into one of two similar classes, subspace  $\mathcal{P}_t$  and  $\mathcal{P}_s$ , by using Fig. 3. In this case, we can obtain the following two similarities for the input subspace.

$$S_{CPMSM}(\mathcal{I}, \mathcal{P}_t) = (1 - \mu)S(\mathcal{I}, \mathcal{P}_t) - \mu S(\mathcal{D}_{It}, \mathcal{D}_{st}) , \tag{2}$$

$$S_{CPMSM}(\mathcal{I}, \mathcal{P}_s) = (1 - \mu)S(\mathcal{I}, \mathcal{P}_s) - \mu S(\mathcal{D}_{Is}, \mathcal{D}_{st}) , \tag{3}$$

where the former is the similarity for class  $t$  and the latter is that for class  $s$ . The input subspace is classified into the class with higher similarity.

The proposed idea of similarity shares common features with the method used in Bayesian face recognition [18] in that it is based on the analysis of image differences, that is, a difference vector between two image pattern vectors. However, that method can not handle complex situations, such as the relation between two sets of image pattern vectors. In addition, a single image is used as an input in the Bayesian method.

### 2.3 Measure of Similarity between Two Subspaces

The measure of similarity between two subspaces is defined through canonical angles. Assume that we have an  $N$ -dimensional subspace  $\mathcal{P}_t$  and an  $M$ -dimensional subspace  $\mathcal{P}_s$  (assume  $N \leq M$  for convenience). In this case, we can obtain  $N$  canonical angles  $\theta_i$ , ( $i = 1 \sim N$ ) between  $\mathcal{P}_t$  and  $\mathcal{P}_s$  by solving the eigenvalue equation of the following matrix  $\mathbf{S}$  [4]:

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a} . \tag{4}$$

$$S_{ij} = \sum_{l=1}^N (\Phi_i \cdot \Psi_l)(\Psi_l \cdot \Phi_j) , \tag{5}$$

where  $\Phi_i$  and  $\Psi_i$  are the  $i$ -th orthonormal basis vectors that span subspace  $\mathcal{P}_t$  and  $\mathcal{P}_s$ , respectively. The value of  $\cos^2\theta_i$  for the  $i$ -th smallest canonical angle  $\theta_i$  is obtained as the  $i$ -th largest eigenvalue of the matrix  $\mathbf{S}$ . Finally, the measure of similarity between two subspaces is defined with  $n$  canonical angles as the following equation (this measure of similarity is used for MSM):

$$S[n] = \frac{1}{n} \sum_{i=1}^n \cos^2\theta_i . \tag{6}$$

## 2.4 Definition of Difference Subspace

The difference subspace is considered a natural generalization of the difference vector between two vectors [16]. A difference subspace is spanned by a set of difference vectors  $\mathbf{d}_i$  between canonical vectors,  $\mathbf{u}_i$  and  $\mathbf{v}_i$ , which form the  $i$ -th canonical angle. The canonical vectors are calculated from the following equations:

$$\mathbf{u}_i = \sum_{l=1}^N a_{kl} \Psi_l \quad . \quad (7)$$

$$\mathbf{v}_i = \sum_{l=1}^N a'_{kl} \Phi_l \quad . \quad (8)$$

In the above equations, the coefficient  $a_{kl}$  is the  $l$ -th element of the  $k$ -th eigenvector  $\mathbf{a}_k$ , corresponding to the  $k$ -th smallest eigenvalue of matrix  $\mathbf{S}$  in Eq. (4). Furthermore, the coefficient  $a'_{kl}$  is the  $l$ -th element of the  $k$ -th eigenvector  $\mathbf{a}'_k$  of matrix  $\mathbf{S}'$ , where  $S'_{ij} = \sum_{l=1}^M (\Psi_i \cdot \Phi_l)(\Phi_l \cdot \Psi_j)$ .

## 2.5 Flow of the Classification Process Using Similarity in CPMSM

The process of classifying an input image set by using CPMSM is given as follows.

- Learning
  - Apply KL expansion on classes  $s$  and  $t$  of training image sets to obtain the reference subspaces  $\mathcal{P}_t$  and  $\mathcal{P}_s$ .
  - Obtain the difference subspace  $\mathcal{D}_{st}$  by using Eqs. (7) and (8).
- Testing
  - step 1**
    - Apply KL expansion on input image set to obtain the input subspace  $\mathcal{I}$ .
    - Calculate the similarities  $S(\mathcal{I}, \mathcal{P}_t)$  and  $S(\mathcal{I}, \mathcal{P}_s)$  by using Eq. (6).
  - step 2**
    - Obtain the difference subspaces  $\mathcal{D}_{I_s}$  and  $\mathcal{D}_{I_t}$  by using Eqs. (7) and (8).
  - step 3**
    - Calculate the similarities  $S(\mathcal{D}_{I_s}, \mathcal{D}_{st})$  and  $S(\mathcal{D}_{I_t}, \mathcal{D}_{st})$  by using Eq. (6).
  - step 4**
    - Combine  $S(\mathcal{I}, \mathcal{P}_t)$  with  $S(\mathcal{D}_{I_t}, \mathcal{D}_{st})$  to obtain  $S_{CPMSM}(\mathcal{I}, \mathcal{P}_t)$  in Eq. (2).
    - Combine  $S(\mathcal{I}, \mathcal{P}_s)$  with  $S(\mathcal{D}_{I_s}, \mathcal{D}_{st})$  to obtain  $S_{CPMSM}(\mathcal{I}, \mathcal{P}_s)$  in Eq. (3).
- Identification
  - Compare the obtained similarity  $S_{CPMSM}(\mathcal{I}, \mathcal{P}_t)$  with  $S_{CPMSM}(\mathcal{I}, \mathcal{P}_s)$ . The input subspace is classified into the class which has higher similarity.

### 3 Validation of the Proposed Method by Using a Database Containing Face Images

The proposed method was designed to distinguish classes that are difficult to distinguish with MSM. To demonstrate the validity of the proposed method, it is necessary to find such pairs in the data set in advance. For this purpose, we carried out a face recognition experiment using MSM and selected pairs that were frequently misclassified, after which we applied the proposed method to those pairs.

#### 3.1 Setup of Experiment for Face Recognition

We used face images from the VidTIMIT audio-video database [17]. This database contains face data for 43 subjects. Three sequences of images are available for each subject. In order to conduct a face recognition experiment, the face region was extracted from each of these images by using the face detection function distributed with OpenCV ver. 1.0. We carefully removed false positives and obtained 140 images for each sequence of images. These cropped face images were converted into  $15 \times 15$  pixels grayscale images, and 225 dimensional vectors were obtained.

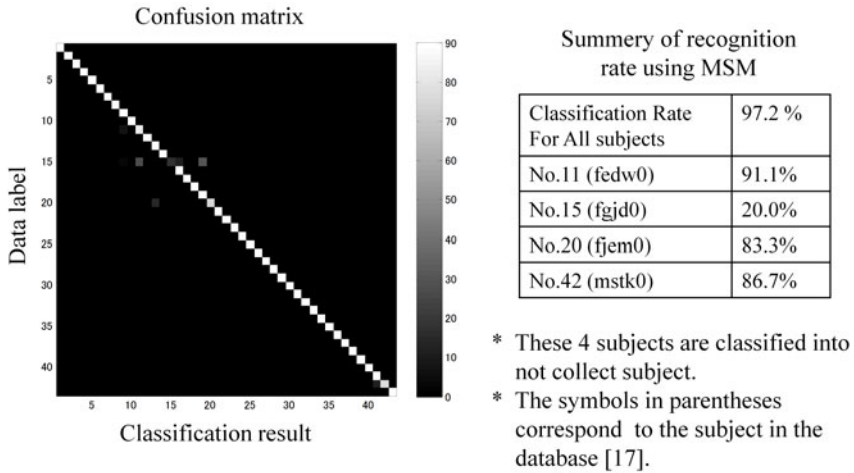
For the classification experiment in this paper, one sequence of images was used to prepare test data sets, and the others were used to prepare training data sets. Every third frame of the image sequence was used as a starting image of the test data set.

The parameters for the experiments of one class were set as follows. Number of training images used to generate reference subspace is 280. Number of testing images used to generate testing subspace is 30. Dimension of the reference subspace is 20. Dimension of the testing subspace is 7. Dimension of the between-class subspace is 20. Dimension of the difference subspace between the input subspace and either reference subspace is 7. Number of trials is 90.

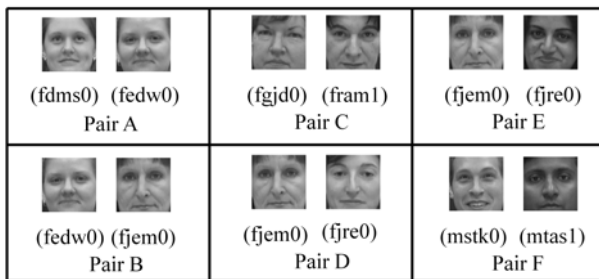
#### 3.2 Extraction of Frequently Misclassified Pairs

We conducted a classification experiment for all subjects contained in the database. To examine which input data is classified into which class, we constructed a confusion matrix. The confusion matrix is a table with a horizontal axis representing the results from the classifier and a vertical axis representing the labeled class. The classification frequency was plotted on this table.

The results from this experiment are plotted in Fig. 5. The color codes for the frequency are given in the legend on the right. From this confusion matrix, we can see that misclassification occurs only in certain specific similar pairs, namely, the six pairs that involve subjects No.11, No.15, No.20 and No.42, as shown in Fig. 5. The total recognition rate for all 43 subjects was 97.2%, as shown in Fig. 4. By contrast, the recognition rate of all subjects except the mentioned four subjects was 100%.



**Fig. 4.** Summary of classification with Mutual Subspace Method for 43 subjects



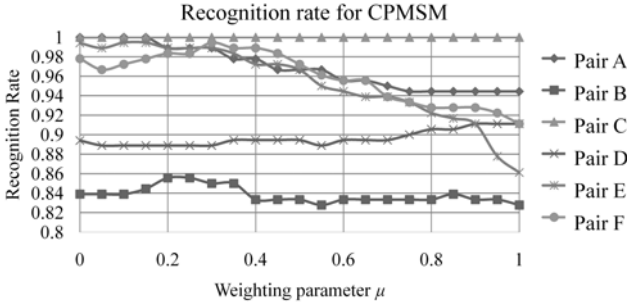
**Fig. 5.** Frequently misclassified pairs

### 3.3 Classification Results for Frequently Misclassified Pairs

To evaluate the validity of CPMSM, we compared the performance of CPMSM with that of MSM and CMSM. These methods were applied in distinguishing between pairs as obtained in the previous section. To compare the performance of these methods, we used recognition rate and EER. EER is the error rate at the threshold value where the false accept rate (FAR) is equal to the false reject rate (FRR).

The performance of CPMSM depends on the weighting parameter  $\mu$  in Eq. (3). We select the optimal value experimentally for each pair, as shown in Fig. 6. Note that when the weighting parameter  $\mu$  is 0, CPMSM is equivalent to MSM.

From Tables 1 and 2, it can be seen that the recognition rate and EER in CPMSM have been improved in comparison to those in MSM for all pairs. The average recognition rate for all pairs increased from 0.950 to 0.959, and the average EER decreased from 0.218 to 0.096. From these results, we can confirm the validity of CPMSM and its ability to improve the performance of MSM.



**Fig. 6.** Relation between recognition rate and  $\mu$ . In the case of  $\mu = 0$ , CPMSM is equivalent to MSM.

**Table 1.** Recognition rate

Confused Pairs	CPMSM	CMSM	MSM
Pair A	1.0	0.961	1.0
Pair B	0.856	0.850	0.839
Pair C	1.0	0.961	1.0
Pair D	0.911	0.911	0.894
Pair E	0.994	0.989	0.994
Pair F	0.994	0.961	0.978
Average	0.959	0.939	0.950

**Table 2.** Equal Error Rate

Confused Pairs	CPMSM	CMSM	MSM
Pair A	0.078	0.103	0.217
Pair B	0.150	0.217	0.286
Pair C	0.006	0.067	0.156
Pair D	0.094	0.139	0.222
Pair E	0.106	0.072	0.211
Pair F	0.144	0.139	0.217
Average	0.096	0.123	0.218

## 4 Conclusions

In this paper, we have proposed the Compound Mutual Subspace Method (CPMSM) for face recognition. The advantage of CPMSM is its strong ability to distinguish between specific highly similar pairs among a large number of combinations of subjects. This characteristics can reduce the computation time and can improve the overall recognition rate by improving the performance with respect to a small number of pairs. The strong ability to distinguish between similar pairs was achieved by introducing a regulation term into the measure of similarity in MSM. The validity of the proposed method has been demonstrated through evaluation experiments with face images taken from the VidTIMIT public database.

## References

1. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: Proc. 3rd. Int. Conf. on Face & Gesture Recognition, pp. 318–323 (1998)
2. Sakano, H., Suenaga, T.: Classifiers under continuous observation. Structural, Syntactic, and Statistical Pattern Recognition 2396, 631–663 (2009)

3. Faggian, N., Paplinski, A., Chin, T.-J.: Face Recognition From Video using Active Appearance Model Segmentation. In: 18th International Conference on Pattern Recognition, vol. 1, pp. 287–290 (2006)
4. Beveridge, J.R., Draper, B.A., Chang, J.-M., Kirby, M., Kley, H., Peterson, C.: Principal Angles Separate Subject Illumination Spaces in YDB and CMU-PIE. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31, 351–363 (2009)
5. Watanabe, S., Pakvasa, N.: Subspace method of pattern recognition. In: Proc. 1st. Int. J. Conf. on Pattern Recognition (1973)
6. Li, X., Fukui, K., Zheng, N.: Boosting constrained mutual subspace method for robust image-set based object recognition. In: Proc. 21st. Int. Joint Conference on Artificial Intelligence, pp. 1132–1137 (2009)
7. Kim, T.-K., Kittler, J., Cipolla, R.: Learning discriminative canonical correlations for object recognition with image sets. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006. LNCS*, vol. 3953, pp. 251–262. Springer, Heidelberg (2006)
8. Chin, T.-J., Suter, D.: A new distance criterion for face recognition using image sets. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) *ACCV 2006. LNCS*, vol. 3851, pp. 549–558. Springer, Heidelberg (2006)
9. Sakano, H., Mukawa, N.: Kernel Mutual Subspace Method for Robust Facial Image Recognition. In: Proc. 4th. Int. Conf. Knowledge-Based Intelligent Engineering Systems and Allied Technologies, vol. 1, pp. 245–248 (2000)
10. Wolf, L., Shashua, A.: Learning over Sets using Kernel Principal Angles. *Journal of Machine Learning Research* 4, 913–931 (2003)
11. Fukui, K., Yamaguchi, O.: The Kernel Orthogonal Mutual Subspace Method and Its Application to 3D Object Recognition. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) *ACCV 2007, Part II. LNCS*, vol. 4844, pp. 467–476. Springer, Heidelberg (2007)
12. Fukui, K., Stenger, B., Yamaguchi, O.: A Framework for 3D Object Recognition Using the Kernel Constrained Mutual Subspace Method. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) *ACCV 2006. LNCS*, vol. 3852, pp. 315–324. Springer, Heidelberg (2006)
13. Kawahara, T., Nishiyama, M., Kozakaya, T., Yamaguchi, O.: Face Recognition based on Whitening Transformation of Distribution of Subspaces. In: Workshop on ACCV2007, Subspac 2007, pp. 97–103 (2007)
14. Phillips, P., Scruggs, W., OfTools, A., Flynn, P., Bowyer, K., Schott, C., Sharpe, M.: *FRVT 2006 and ICE 2006 Large-Scale Results*. Technical Report NISTIR 7408, NIST (2007)
15. Fukunaga, K., Koontz, W.L.G.: Applications of the Karhunen-Loève Expansion to Feature Selection and Ordering. *IEEE Trans. Computers* C-19, 311–318 (1970)
16. Fukui, K., Yamaguchi, O.: Face Recognition Using Multi-viewpoint Patterns for Robot Vision. *Robotics Research* 15, 192–201 (2005)
17. Sanderson, C.: *Biometric Person Recognition: Face, Speech and Fusion*. VDM-Verlag (2008)
18. Moghaddam, B., Jebara, T., Pentland, A.: Bayesian Face Recognition. *J. of Pattern Recognition* 33, 1771–1782 (2000)
19. Chatelin, F.: *Eigenvalues of matrices*. John Wiley and Sons, Chichester (1993)
20. Basri, R., Jacobs, D.: Lambertian Reflectance and Linear Subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 218–233 (2003)



# Dynamic Subspace Update with Incremental Nyström Approximation

Hongyu Li and Lin Zhang

School of Software Engineering, Tongji University, Shanghai, China

**Abstract.** Low rank approximation methods, e.g. the Nyström method, are often used to speed up eigen-decomposition of kernel matrices. However, it cannot effectively update the extracted subspaces when datasets dynamically increase with time. In this paper, we propose an incremental Nyström method for dynamic learning. Experimental results demonstrate the feasibility and effectiveness of the proposed method.

## 1 Introduction

Kernel methods are attracting in the fields of machine learning and pattern recognition due to their advantages in modeling the highly complex, non-linear structures of objects. Such methods generally require the eigen-decomposition of a kernel matrix during computation. The eigen-decomposition, however, is actually the bottleneck of computation in practical applications as its time complexity of  $O(n^3)$  is very high. In the cases where only several bottom (top) eigenvectors are needed and the spectra of kernel matrices rapidly decay, low-rank approximation methods, like sampling-based methods, are often used to speed up the eigen-decomposition.

The Nyström method [1-4] is one of the popular sampling-based methods for handling batch data. When datasets increase dynamically, the original Nyström method must compute and decompose the kernel matrix of all data once again while discarding the previous result of eigen-decomposition, which is thus called the batch Nyström (B-Nyström). In such long observation applications as object tracking, the direct application of B-Nyström is obviously in low efficiency. Therefore, the Nyström method needs to be modified for the dynamic cases. To do this, this paper proposes an incremental version of the Nyström method, called the incremental Nyström. In the proposed approach, the B-Nyström is first used to approximate the eigen-decomposition of the kernel matrix constructed with initial data. For newly coming data, the eigenvectors are updated through keeping the old part and merely approximating the new part, which avoids repetitive computation when pooling all data. In addition, to improve the approximation accuracy, the orthogonal iteration algorithm [5] is adopted after the initial approximation. To maintain non-increasing memory usage and update duration, the reduced set (RS) expansions [6] is brought in I-Nyström.

In sum, the contribution of this study mainly includes three aspects:

1. The incremental Nyström method is proposed for dynamic subspace update, which has the high efficiency of Nyström in approximating eigen-decomposition and the good flexibility in practical applications.
2. The drift caused by the incremental data is successfully avoided with the Orthogonal-Iteration algorithm.
3. The memory usage and update duration maintain stable and non-increasing in the incremental Nyström method.

## 2 Nyström Approximation

The Nyström method is originated from the numerical treatment of the following integral equation [7],

$$\int p(y)k(x, y)\phi_i(y)dy = \lambda_i\phi_i(x), \tag{1}$$

where  $p$  is the probability density function,  $k$  the positive semi-definite kernel function,  $\lambda$  and  $\phi$  the eigenvalue and eigenfunction respectively. Given a set of samples  $\{x_1, x_2, \dots, x_n\}$  generated from function  $p$ , to approximately estimate  $\lambda$  and  $\phi$ , Eq. (1) is changed with the empirical average:

$$\frac{1}{n} \sum_{k=1}^n k(x, x_k)\phi_i(x_k) \simeq \lambda_i\phi_i(x). \tag{2}$$

This actually is a standard eigen-decomposition problem  $KU \simeq U\Lambda$  if replacing  $x$  with samples  $\{x_1, x_2, \dots, x_n\}$ . Here  $K$  is a positive semi-definite kernel matrix,  $U$  is with orthogonal columns, and  $\Lambda$  is a diagonal matrix.

The basic idea of the Nyström method is to approximate the eigenvectors of a kernel matrix with few samples. The following explains the implementation procedure of this method in detail. To decompose a kernel matrix  $K \in R^{n \times n}$  constructed with  $n$  data points, we first divide the matrix in four parts,

$$K = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}, \tag{3}$$

where  $A \in R^{m \times m}$ ,  $B \in R^{m \times k}$  and  $C \in R^{k \times k}$ . The numbers  $m, n, k$  satisfy the condition,  $m + k = n$ , and to construct matrix  $A$ ,  $m$  samples is first chosen from the initial  $n$  points. Since  $m$  is generally quite small, the eigen-decomposition of  $A$  is efficient and fast,

$$A = U\Lambda U^T.$$

Based on the eigen-decomposition of  $A$ , matrix  $K$  can be approximately decomposed as follows,

$$K \simeq \tilde{U}\tilde{\Lambda}\tilde{U}^T, \tag{4}$$

where

$$\tilde{U} = \begin{pmatrix} U \\ B^T U \Lambda^{-1} \end{pmatrix}, \tilde{\Lambda} = \Lambda. \tag{5}$$

Obviously, the approximation  $\tilde{K}$  of matrix  $K$  takes the following form,

$$\tilde{K} = \tilde{U}\tilde{\Lambda}\tilde{U}^T = \begin{pmatrix} A & B \\ B^T & A^{-1}B \end{pmatrix}. \tag{6}$$

It is easy to find from Eqs. (3) and (6) that  $C$  is approximately equal to  $B^T A^{-1}B$ . As a result, the approximation error of decomposing the kernel matrix  $K$  can be quantified as Schur complement,

$$e = \|C - B^T A^{-1}B\|_F,$$

where  $\|\cdot\|_F$  denotes the Frobenius norm.

The eigenvectors in expression (5) may not be orthonormal, therefore, we should orthogonalize them. One way to solve this problem [3] is stated below: Let  $Z = \tilde{U}\Lambda^{-\frac{1}{2}}$ ,  $F \sum F^T$  denotes the diagonalization of  $Z^T Z$  and  $V = ZF \sum^{\frac{1}{2}}$ . Consequently, the matrix  $V$  are the leading orthogonal eigenvectors of  $\tilde{K}$  and satisfies the condition  $VV^T = I$ . This orthogonalization method will suffer a time cost of  $O(n \times m^2 + m^3)$ , where  $m \ll n$  always holds for low-rank approximation. Therefore, the orthogonalization method is efficient for small  $m$ .

### 3 Incremental Subspace Update

Although the Nyström method can work well in the batch mode for eigen-decomposition, it cannot dynamically update the learned subspace with the change of datasets. This section extends the batch Nyström method and proposes an incremental Nyström method to handle dynamic data.

Given an initial data set  $S = \{x_1, x_2, \dots, x_n\}$ , we can construct the kernel matrix  $K$  and approximate the eigenvectors and eigenvalues of  $K$  with batch Nyström. When new data  $D = \{d_1, d_2, \dots, d_r\}$  come, the kernel matrix  $K'$  will be updated as follows,

$$K' = \begin{pmatrix} K & P \\ P^T & Q \end{pmatrix}, \tag{7}$$

where  $P$  is a  $n \times r$  matrix constructed with  $S$  and  $D$ ,  $Q$  a  $r \times r$  matrix constructed with  $D$ . Since the eigen-decomposition of matrix  $K$  has been already approximately computed at this moments, the simple and straight way of decomposing  $K'$  is to repeat the computing procedure described in Section 2. That is, since it is known that  $K \simeq \tilde{U}\tilde{\Lambda}\tilde{U}^T$ , we can easily get

$$K' \simeq U' \Lambda' U'^T, \tag{8}$$

where

$$U' = \begin{pmatrix} \tilde{U} \\ P^T \tilde{U} \Lambda^{-1} \end{pmatrix}, \Lambda' = \Lambda. \tag{9}$$

The orthogonalization method of  $U'$  is as above. As a consequence, the subspace spanned with leading eigenvectors is easily updated with  $U'$ . Since the

approximation error is accumulated at each incremental step, the extracted subspace will become more and more inaccurate along with the increase of new data. The total approximation error can be evaluated with the following expression,

$$E = \sum_{i=1}^t e_i, \tag{10}$$

where  $e_i$  denotes as the approximation error at the  $i$ -th step,

$$e_i = \|Q - P^T \tilde{K}^{-1} P\|_F. \tag{11}$$

## 4 Refinement Strategies

To reduce the approximation error, this study proposes to adopt the orthogonal iteration algorithm to refine the eigenvectors at each step. In addition, to compress storage and maintain constant speed, we bring the idea of reduced set in this work.

### 4.1 Error Reduction

As discussed above, the potential problem of the incremental Nyström method is that the accumulate error gets larger at each incremental step, causing large drift in approximation. This problem has a strong impact on applications that require long observation and update. To control the accumulate error and improve the quality of the proposed method, this section proposes a strategy based on the orthogonal iteration algorithm [5].

---

#### Algorithm 1. Orthogonal Iteration

---

```

 $U_0 = U'$  //the obtained eigenvectors at some incremental step
for  $t = 1$  to  $\dots$  do
     $\hat{Q}_t R_t = U_{t-1}$  //compute reduced QR factorization
     $U_t = K \hat{Q}_t$ 
end for

```

---

For the  $p$ -rank decomposition of a  $n \times n$  kernel matrix  $K$ , the procedure of orthogonal iteration is outlined in Algorithm 1, where  $\hat{Q}_t R_t$  is the reduced QR factorization of  $U_{t-1}$  and  $t$  denotes the iteration step.  $\hat{Q}_t$  is an  $n \times p$  matrix having orthonormal columns and  $R_t$  is a  $p \times p$  upper triangular matrix. After several iterations, the matrix  $U_t$  converges to an  $n \times p$  matrix  $\hat{U}$  whose columns correspond to the  $p$  largest eigenvectors of  $K$  and form the basis of an invariant subspace .

Although the orthogonal iteration algorithm is effective in approximating the eigen-decomposition, it generally converges slowly. More specifically, the convergence speed depends on the initial guess  $U_0$ . That is, the better the initial guess

$U_0$ , the faster this algorithm converges. As incremental Nyström approximates the eigen-decomposition of kernel matrices, the initial guess  $U_0$  in the orthogonal iteration algorithm can be assigned the obtained eigenvectors at some incremental step  $U'$ . After several iterations, the spectra of kernel matrices can be more accurate.

The time complexity in orthogonal iteration is  $O(t \times (p^3 + n^2p))$ . With the proper choice of the number  $t$  of iterations, the computation time will be completely determined by the size  $n$  of datasets. If  $n$  is very large, it remains expensive for orthogonal iteration. However, if we can find a proper measure to compress the dataset, which can reduce  $n$ , the performance will definitely get better. The compression strategy is from the idea of reduced set, which is discussed in the next subsection.

## 4.2 Compression Strategy

With the increase of new data, the kernel matrix will get larger and the data storage cost will become higher, which cause difficulties in real applications. For example, old and new data are respectively denoted as  $S$  and  $D$ , the eigen-decomposition of  $S$  has been completed, and we want to incrementally update the eigen-decomposition of the kernel matrix  $[S \ D]$  according to Eqs. 7 and 9. It is clearly unavoidable to save all the old and new data for update, which will ultimately influence the update speed.

To solve this problem, we employ an effective compression strategy by constructing the reduced set (RS) expansions. Due to the space limitation, please refer to 6 for more details about the construction of reduced set.

## 5 Experimental Results

This section evaluates the performance of the proposed method: the incremental Nyström method (I-Nyström) with orthogonal iteration (I-Nyström-Iter) and reduced set (I-Nyström-Iter-RS). The parameters involved in the experiments are listed below:

- $p$ : the number of principal components of kernel matrices.
- $\tau$ : the number of pre-images in compression for each feature vector.
- $N_{upd}$ : the number of data updates in the incremental procedure.
- $N_{iter}$ : the number of iterations during orthogonal iteration.

### 5.1 Accuracy and Efficiency

This part examines the performance of I-Nyström, I-Nyström-Iter through comparing them with I-KSVD, B-Nyström-Iter. Considering the subspace generated by batch KSVD (B-KSVD) as the baseline, we compute the distance of two subspaces extracted from each method to B-KSVD. The distance measure is based on kernel principle angles 8 between two subspaces:

$$d(\text{span}(\zeta_1), \text{span}(\zeta_2)) = \sqrt{\sum_{i=1}^p \theta_i^2} \quad (12)$$

**Table 1.** Datasets used in our experiments

Dataset	n	d
cpusmall	3000	12
letter	3000	16
EYFDB	2000	10304

where  $\text{span}(\zeta_1)$  means a  $p$ -dimensional subspace spanned by  $\zeta_1$  and  $\theta_i$  denotes as the  $i$ -th principal angle between two subspaces.

In this experiment, the Gaussian kernel with  $\sigma = 1$  is used and parameter  $p = 30$  for all tested methods. The parameter  $N_{upd}$  is set 40 for all the datasets, which means that each dataset is divided into 40 parts. The other parameter  $N_{iter}$  involved in I-Nyström-Iter is set 4.

The used datasets are listed in Table 1, where  $d$  represents the number of used features. Among the datasets, cpusmall and letter are from the benchmark datasets of LibSVM<sup>1</sup>; EYFDB means the dataset of Extended Yale Face Database B [9,10]. Only a part of data in each dataset are used here and the size  $n$  of each dataset also lists in Table 1.

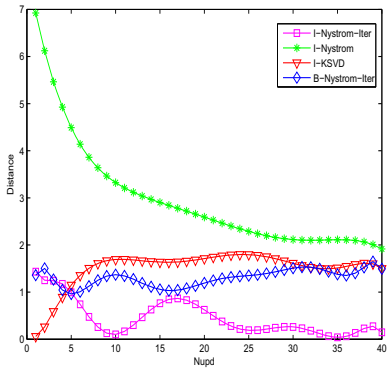
The subspace distance and time at each incremental step are plotted in Fig. 1. As shown in Fig. 1, the time cost of I-Nyström is the least in the dynamic case, however, the extracted subspace is with the biggest subspace distance. As  $N_{upd}$  increases, the performance of I-Nyström gets better. For those online applications that involve long-term update and care more about running speed, I-Nyström is a good choice. Compared to other methods, I-Nyström-Iter is the most accurate and the time cost is between I-KSVD and B-Nyström-Iter at each incremental step. Although the subspace extracted with I-Nyström-Iter is instable at each incremental step, where the subspace distance fluctuates in the range  $[0, 2.0]$  in letter dataset,  $[0, 1.5]$  in cpusmall dataset, the overall quality of the extracted subspace is reliable. The possible factor that results in the instability of I-Nyström-Iter is the convergence speed of orthogonal iteration algorithm. An alternative strategy of accelerating the convergence will be more helpful.

In addition, it is also worth noting that the computation costs of all methods obviously increase as  $N_{upd}$  becomes large. One of the feasible solutions is the reduced set, which is demonstrated in the next experiment.

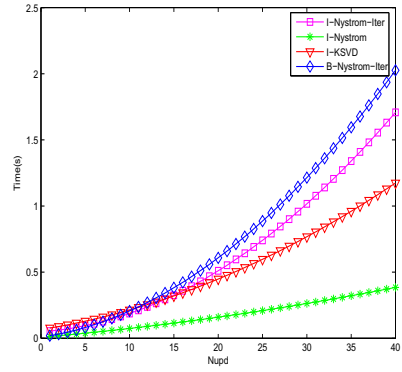
## 5.2 Visual Tracking

This section presents an application of I-Nyström-Iter-RS to visual tracking. Such application was also examined in [11], where the authors proposed a tracking method based on incremental PCA to incrementally learn the representation of a low-dimensional subspace. In this experiment, we use the I-Nyström-Iter-RS to replace incremental PCA, and do the update at every 10 frames. For each update, only the first  $p = 5$  principal eigenvectors of kernel matrix are kept, and  $\tau$  is set 3 to compress the eigenvectors. The first 10 frames are tracked according

<sup>1</sup> <http://www.csie.ntu.edu.tw/~{cjlin/libsvmtools/datasets/>

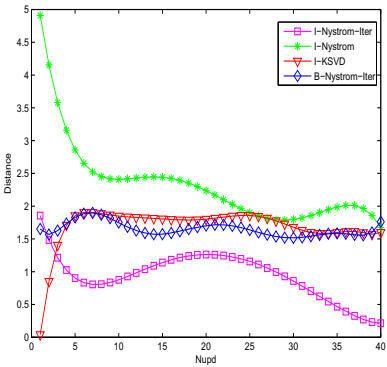


Subspace Distance

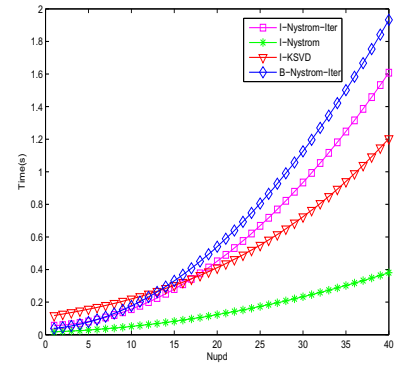


Time

(a) cpusmall

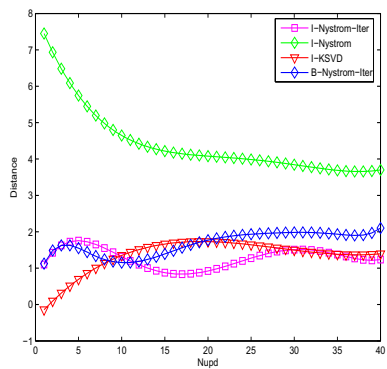


Subspace Distance

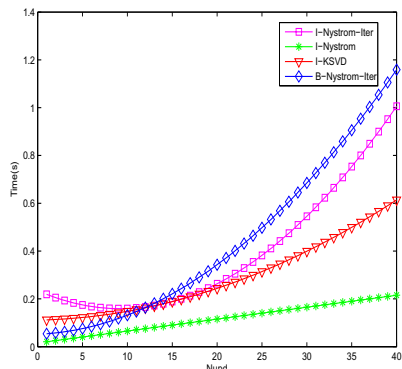


Time

(b) letter



Subspace Distance



Time

(c) EYFDB

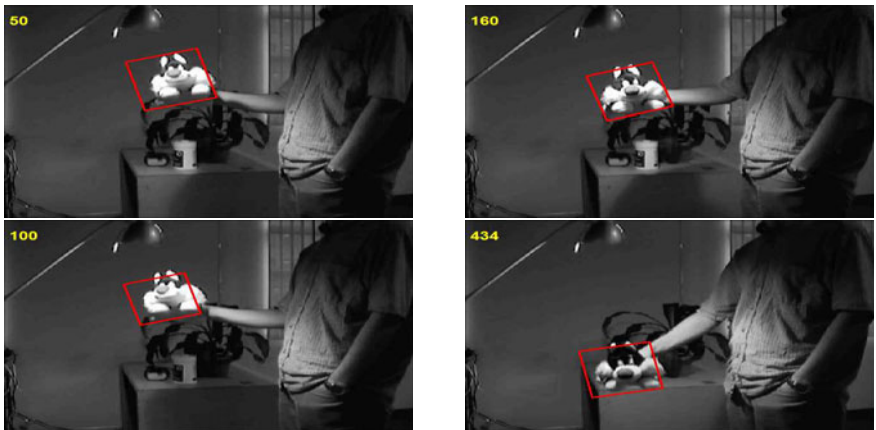
Fig. 1. Comparison of performance of each method on different datasets



(a) Car11



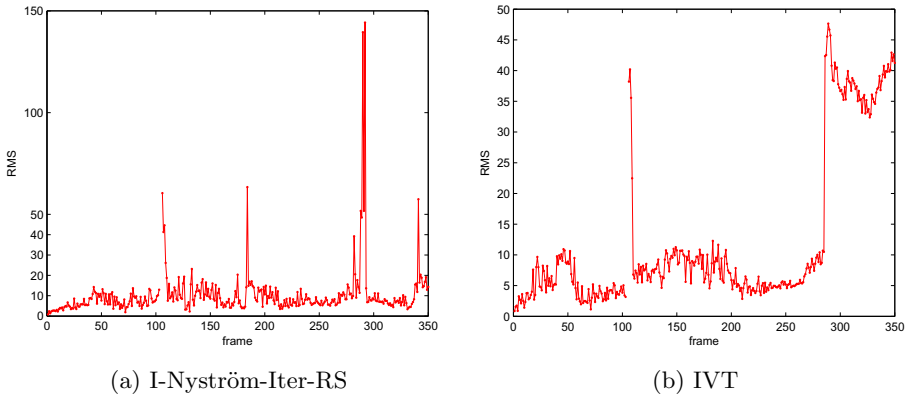
(b) Dudek



(c) Sylv

**Fig. 2.** Tracking results on three video clips





**Fig. 3.** the RMS error at tracking feature points on the "Dudek" clip

to the distance to the first frame in the input space. The number of particles generated at each frame is 100. The tracking results on three video clips<sup>2</sup> are shown in Fig. 2.

Fig. 2 shows that I-Nyström-Iter-RS can effectively update the appearance model to accommodate the conditions of low resolution and contrast (Fig. 2(a)), severe expression variation and temporary occlusion (Fig. 2(b)), and pose change (Fig. 2(c)).

In order to evaluate the tracking results quantitatively, the root mean square (RMS) error is adopted, which represents the difference between the manually-labeled facial feature points and tracking feature points. The RMS errors with I-Nyström-Iter-RS for each frame in the Dudek clip are compared with the IVT method stated in [11], as displayed in Fig. 3. From the figure, it is clear that most frames are tracked well with low RMS error by I-Nyström-Iter-RS, and the result is basically comparable to the IVT method except some abrupt changes due to the temporary occlusion or motion blur. In addition, it is worth noting that the RMS errors of IVT in frames 300 to 350 when appearance changes a lot is bigger than ours.

With regard to the time efficiency, the IVT method can achieve the speed of 24 fps, while the I-Nyström-Iter-RS only has the speed of 3 fps. However, the reason of being slow is not the update speed of I-Nyström-Iter-RS, but the high cost in evaluating the kernel matrix of particles of every frame. Therefore, I-Nyström-Iter-RS still is highly efficient in practical applications.

## 6 Conclusions

In this paper, we propose the incremental Nyström approximation method for the dynamic learning problem and employ orthogonal iteration and reduce set

<sup>2</sup> <http://www.cs.toronto.edu/dross/ivt/>

to refine the approximation results of the incremental Nyström method. Experimental results demonstrate that the proposed method can effectively preserve the whole structure of the extracted subspace and has good potential in such real applications as visual tracking.

**Acknowledgments.** This research was partially supported by Natural Science Foundation of China Grant 60903120, 863 Project 2009AA043001 and Shanghai Natural Science Foundation Grant 09ZR1434400.

## References

1. Zhang, K., Tsang, I.W., Kwok, J.T.: Improved nyström low-rank approximation and error analysis. In: ICML 2008: Proceedings of the 25th International Conference on Machine Learning, pp. 1232–1239. ACM, New York (2008)
2. Fowlkes, C., Belongie, S., Malik, J.: Efficient spatiotemporal grouping using the nystrom method. In: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001, vol. 1, pp. 231–238 (2001)
3. Fowlkes, C., Belongie, S., Chung, F., Malik, J.: Spectral grouping using the nyström method. IEEE Transactions on Pattern Analysis and Machine Intelligence 26, 214–225 (2004)
4. Williams, C., Seeger, M.: Using the nyström method to speed up kernel machines. In: Advances in Neural Information Processing Systems 13 (2001)
5. Heath, M.-T.: Scientific Computing: An Introduction Survey. The McGraw-Hill Companies, Inc., New York (2002)
6. Scholkopf, B., Smola, A.: Learning with kernels. The MIT press, Cambridge (2002)
7. Baker, C.: The numerical treatment of integral equations. Clarendon Press, Oxford (1977)
8. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. J. Mach. Learn. Res. 4, 913–931 (2003)
9. Lee, K., Ho, J., Kriegman, D.: Acquiring Linear Subspaces for Face Recognition under Variable Lighting. IEEE Trans. Pattern Anal. Mach. Intelligence 27, 684–698 (2005)
10. Georghiades, A., Belhumeur, P., Kriegman, D.: From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intelligence 23, 643–660 (2001)
11. Ross, D.A., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. Int. J. Comput. Vision 77, 125–141 (2008)

# Background Modeling via Incremental Maximum Margin Criterion

Cristina Marghes and Thierry Bouwmans

Laboratoire MIA, University of La Rochelle, 17000 La Rochelle, France

**Abstract.** Subspace learning methods are widely used in background modeling to tackle illumination changes. Their main advantage is that it doesn't need to label data during the training and running phase. Recently, White et al. [1] have shown that a supervised approach can improve significantly the robustness in background modeling. Following this idea, we propose to model the background via a supervised subspace learning called Incremental Maximum Margin Criterion (IMMC). The proposed scheme enables to initialize robustly the background and to update incrementally the eigenvectors and eigenvalues. Experimental results made on the Wallflower datasets show the pertinence of the proposed approach.

## 1 Introduction

Many background subtraction methods have been developed in video-surveillance to detect moving objects [2][3][4]. These methods have different common steps: background modeling, background initialization, background maintenance and foreground detection. The background modeling describes the kind of model used to represent the background. Once the model has been chosen, the background model is initialized during a learning step by using  $N$  frames. Then, a first foreground detection is made and consists in the classification of the pixel as a background or as a foreground pixel. Thus, the foreground mask is applied on the current frame to obtain the moving objects. After this, the background is adapted over time following the changes which have occurred in the scene and so on. The last decade witnessed very significant contributions in background modeling via unsupervised subspace learning [5] due to their robustness to illumination changes. The first approach developed by Oliver et al. [6] consists in applying Principal Component Analysis (PCA) on  $N$  images to construct a background model, which is represented by the mean image and the projection matrix comprising the first  $p$  significant eigenvectors of PCA. In this way, foreground segmentation is accomplished by computing the difference between the input image and its reconstruction. The main limitation of this method appears for the background maintenance because it is computationally intensive to perform model updating using the batch mode PCA. Moreover without a mechanism of robust analysis, the outliers or foreground objects may be absorbed into the background model. In this context, some authors proposed different algorithms

of incremental PCA. The incremental PCA proposed by Rymel et al. [7] need less computation but the background image is contaminated by the foreground object. To solve this, Li et al. [8] proposed an incremental PCA which is robust in presence of outliers. However, when keeping the background model updated incrementally, it assigned the same weights to the different frames. Thus, clean frames and frames which contain foreground objects have the same contribution. The consequence is a relative pollution of the background model. To solve this, Skocaj et al. [9] used a weighted incremental and robust. The weights are different following the frame and this method achieved a better background model. However, the weights were applied to the whole frame without considering the contribution of different image parts to building the background model. To achieve a pixel-wise precision for the weights, Zhang and Zhuang [10] proposed an adaptive weighted selection for an incremental PCA. This method performs a better model by assigning a weight to each pixel at each new frame during the update. Wang et al. [11] used a similar approach using the sequential Karhunen-Loeve algorithm. Recently, Zhang et al. [12] improved this approach with an adaptive scheme. All these incremental methods avoid the eigen-decomposition of the high dimensional covariance matrix using approximation of it and so a low decomposition is allowed at the maintenance step with less computational load. However, these incremental methods maintain the whole eigenstructure including both the eigenvalues and the exact matrix. To solve it, Li et al. [13] proposed a fast recursive and robust eigenbackground maintenance avoiding eigen-decomposition. This method achieves similar results than the incremental PCA [8] at better frames rates. In another way, Yamazaki et al. [14] and Tsai et al. [15] proposed to use the Independent Component Analysis (ICA) which is a variant of PCA in which the components are assumed to be mutually statistically independent instead of merely uncorrelated. This stronger condition allows remove the rotational invariance of PCA, i.e. ICA provides a meaningful unique bilinear decomposition of two-way data that can be considered as a linear mixture of a number of independent source signals. The ICA model was tested on traffic scenes [14] and show robustness in changing background like illumination changes. Recently, Chu et al. [16] used a Non-negative Matrix Factorization algorithm to model dynamic backgrounds and Bucak et al. [17] preferred an Incremental version of the Non-negative Matrix Factorization (INMF) which presents similar performance than the incremental PCA [8]. In order to take into account the spatial information, Li et al. [18] used an Incremental Rank-(R1,R2,R3) Tensor (IRT). Results [18] show better robustness to noise. The Table 1 shows an overview of the background modeling based on subspace learning.

However, these different approaches are unsupervised subspace learning methods. Indeed, it doesnt need to label data. Recently, White et al. [1] proved that the Gaussian Mixture Model (GMM) [19] gives better results when some coefficients are determined in a supervised way. Following this idea, we propose to use a supervised subspace learning for background modeling. Thus, the Maximum Margin Criterion (MMC) offers a nice framework. It was proposed by Li et al. [20] and it can outperform PCA and Linear Discriminant Analysis (LDA) on

many classification tasks [21]. MMC search for the projection axes on which the data points of different classes are far from each other meanwhile where data points of the same class are close to each other. As the original PCA and LDA, MMC is a batch algorithm and so it requires that the data must be known in advance and be given once altogether. Recently, Yan et al. [22] have proposed incremental version of MMC which is suitable to update online the background model.

The rest of this paper is organized as follows: In the Section 2, we firstly remind the Incremental Maximum Margin Criterion (IMMC). In the Section 3, we present our method using subspace learning via IMMC for background modeling. Then, a comparative evaluation is provided in the Section 4. Finally, the conclusion is given in Section 5.

**Table 1.** Subspace Learning for background modeling: An Overview

Subspace Learning - Methods	Authors - Dates
<b>Principal Components Analysis</b>	
Batch PCA	Oliver et al. (1999) [6]
Incremental PCA	Rymel et al. (2004) [7]
Incremental and Robust PCA	Li et al. (2003) [8]
Weighted Incremental and Robust PCA	Skocaj et al. (2003) [9]
Adaptive Weighted Incremental and Robust PCA	Zhang and Zhuang (2007) [10]
<b>Independent Component Analysis</b>	
Batch ICA	Yamazaki et al. (2006) [14]
Incremental ICA	Tsai and Lai (2009) [15]
<b>Independent Component Analysis</b>	
Batch NMF	Chu et al. (2010) [16]
Incremental NMF	Bucak et al. (2007) [17]
<b>Independent Component Analysis</b>	
Incremental Rank-(R1,R2,R3) Tensor	Li et al. (2008) [18]

## 2 Incremental Maximum Margin Criterion (IMMC)

This section reminds briefly the principle of IMMC developed in [22]. Suppose the data sample points  $u(1), u(2), \dots, u(N)$  are  $d$ -dimensional vectors, and  $U$  is the sample matrix with  $u(i)$  as its  $i^{th}$  column. MMC [20] projects the data onto a lower-dimensional vector space such that the ratio of the inter-class distance to the intra-class distance is maximized. The goal is to achieve maximum discrimination and the new low-dimensional vector can be computed as  $y = W^T u$  where  $W \in \mathbf{R}^{d \times p}$  is the projection matrix from the original space of dimension  $d$  to the low dimensional space of dimension  $p$ . So, MMC [20] aims to maximize the criterion:

$$J(W) = W^T(S_b - S_w)W \tag{1}$$

where

$$S_b = \sum_{i=1}^c p_i(m_i - m)(m_i - m)^T \tag{2}$$

$$S_w = \sum_{i=1}^c p_i E(u_i - m_i)(u_i - m_i)^T \quad (3)$$

are called respectively the inter-class scatter matrix and the intra-class scatter matrix and  $c$  is the number of classes,  $m$  is the mean of all samples,  $m_i$  is the mean of the samples belonging to class  $i$  and  $p_i$  is the prior probability for a sample belonging to class  $i$ . The projection matrix  $W$  can be obtained by solving:

$$(S_b - S_w)w = \lambda w \quad (4)$$

To incrementally maximize the MMC criterion, Yan et al. [22] constraint  $W$  to unit vectors, i.e.  $W = [w_1, w_2, \dots, w_p]$  and  $w_k^T w_k = 1$ . Thus the optimization problem of  $J(W)$  is transformed to:

$$\max \sum_{k=1}^p w_k^T (S_b - S_w) w_k \quad (5)$$

subject to  $w_k^T w_k = 1$  with  $k = 1, 2, \dots, p$ .  $W$  is the first  $k$  leading eigenvectors of the matrix  $S_b - S_w$  and the column vectors of  $W$  are orthogonal to each other. Thus, the problem is learning the  $p$  leading eigenvector of  $S_b - S_w$  incrementally.

## 2.1 Updating Incrementally Leading Eigenvectors

Let  $C = S_b + S_w$  be the covariance matrix, then we have  $J(W) = W^T (2S_b - C)W$ ,  $W \in \mathbf{R}^{d \times p}$ . Then maximizing  $J(W)$  means to find the  $p$  leading eigenvectors of  $2S_b - C$ .

The inter-class scatter matrix of step  $n$  after learning from the first  $n$  samples can be written as below,

$$S_b(n) = \sum_{j=1}^c p_j(n) (m_j - m(n))(m_j(n) - m(n))^T \quad (6)$$

and

$$S_b = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n S_b(i) \quad (7)$$

On the other hand,

$$C = E(u(n) - m)(u(n) - m)^T \quad (8)$$

$$= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n (u(n) - m(n))(u(n) - m(n))^T \quad (9)$$

$2S_b - C$  should have the same eigenvectors as  $2S_b - C + \theta I$  where  $\theta$  is a positive real number and  $I \in \mathbf{R}^{d \times d}$ . From (7) and (9) we have the following equation:

$$2S_b - C + \theta I = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n A(i) = A \quad (10)$$

where  $A(i) = 2S_b(i) - (u(i) - m(i))(u(i) - m(i))^T + \theta I$ ,  $A = 2S_b - C + \theta I$ .

The general eigenvector form is  $Ax = \lambda x$ , where  $x$  is the eigenvector of matrix  $A$  corresponding to the eigenvalue  $\lambda$ . By replacing matrix  $A$  with the MMC matrix at step  $n$ , an approximate iterative eigenvector computation formulation is obtained with  $\nu = \lambda x$ .

$$\begin{aligned} \nu(n) = & \frac{1}{n} \sum_{i=1}^n (2 \sum_{j=1}^c p_j(i) \Phi_j(i) \Phi_j(i)^T \\ & - (u(i) - m(i))(u(i) - m(i))^T + \theta I) x(i) \end{aligned} \tag{11}$$

where  $\Phi_j(i) = m_j(i) - m(i)$ ,  $\nu(n)$  is the  $n$  step estimation of  $\nu$  and  $x(n)$  is the  $n$  step estimation of  $x$ . Once the estimation of  $\nu$  is obtained, eigenvector  $x$  can be directly computed as  $x = \nu / \|\nu\|$ . Let  $x(i) = \nu(i - 1) / \|\nu(i - 1)\|$ , then the incremental formulation is the following:

$$\begin{aligned} \nu(n) = & \frac{n - 1}{n} \nu(n - 1) \\ & + \frac{1}{n} (2 \sum_{j=1}^c p_j(n) \alpha_j(n) \Phi_j(n) \\ & - \beta(u(n) - m(n)) + \theta \nu(n - 1)) / \|\nu(n - 1)\| \end{aligned} \tag{12}$$

where  $\alpha_j(n) = \phi_j(n)^T \nu(n - 1)$  and  $\beta(n) = (u(n) - m(n))^T \nu(n - 1)$ ,  $j = 1, 2, \dots, c$ . For initialization,  $\nu(0)$  is equal to the first data sample.

### 2.2 Updating Incrementally the Other Eigenvectors

To compute the  $(j + 1)^{th}$  eigenvector, its projection is subtracted on the estimated  $j^{th}$  eigenvector from the data,

$$u_{1_n}^{j+1}(n) = u_{1_n}^j(n) - (u_{1_n}^j(n)^T \nu^j(n)) \nu^j(n) \tag{13}$$

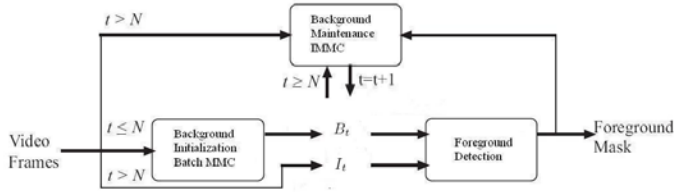
where  $u_{1_n}^1(n) = u_{1_n}(n)$ . The same method is used to update  $m_i^j(n)$  and  $m^j(n)$ ,  $i = 1, 2, \dots, c$ . Since  $m_i^j(n)$  and  $m^j(n)$  are linear combinations of  $x_{l_i}^j(i)$ , where  $i = 1, 2, \dots, k$ , and  $l_i \in \mathbf{1}, \mathbf{2}, \dots, \mathbf{C}$ .  $\Phi_i$  are linear combination of  $m_i$  and  $m$ , for convenience, only  $\Phi$  is updated at each iteration step by:

$$\Phi_{l_n}^{j+1}(n) = \Phi_{l_n}^j(n) - (\Phi_{l_n}^j(n)^T \nu^j(n)) \nu^j(n) \tag{14}$$

In this way, the time-consuming orthonormalization is avoided and the orthogonal is always enforced when the convergence is reached.

## 3 Application to Background Modeling

The Figure [1](#) shows an overview of the proposed approach. The background modeling framework based on IMMC includes the following stages: (1) Background initialization via MMC using  $N$  frames ( $N = 30$  practically) (2) Foreground detection (3) Background maintenance using IMMC. The steps (2) and (3) are executed repeatedly as time progresses.



**Fig. 1.** Overview of the proposed approach

Denote the training video sequences  $S = \{I_1, \dots, I_N\}$  where  $I_t$  is the frame at time  $t$ . Let each pixel  $(x, y)$  be characterized by its intensity in the grey scale and assume that we have the ground truth corresponding to this training video sequences, i.e we know for each pixel its class label which can be foreground or background. Thus, we have:

$$S_b = \sum_{i=1}^c p_i (m_i - m)(m_i - m)^T \quad (15)$$

$$S_w = \sum_{i=1}^c p_i E(u_i - m_i)(u_i - m_i)^T \quad (16)$$

where  $c = 2$ ,  $m$  is the mean of the intensity of the pixel  $x, y$  over the training video and  $m_i$  is the mean of samples belonging to class  $i$  and  $p_i$  is the prior probability for a sample belonging to class  $i$  with  $i \in \{Background, Foreground\}$ . Then, we can apply the batch MMC to obtain the first leading eigenvectors which correspond to the background. The corresponding eigenvalues are contained in the matrix  $L_M$  and the leading eigenvectors in the matrix  $\Phi_M$ . Once the leading eigenbackground images stored in the matrix  $\Phi_M$  are obtained and the mean  $\mu_B$  too, the input image  $I_t$  can be approximated by the mean background and weighted sum of the leading eigenbackgrounds  $\Phi_M$ .

So, the coordinate in leading eigenbackground space of input image  $I_t$  can be computed as follows:

$$w_t = (I_t - \mu_B)^T \Phi_M \quad (17)$$

When  $w_t$  is back projected onto the image space, a reconstructed background image is created as follows:

$$B_t = \Phi_M w_t^T + \mu_B \quad (18)$$

Then, the foreground object detection is made as follows:

$$|I_t - B_t| > T \quad (19)$$

where  $T$  is a constant threshold.



**Table 2.** Performance Evaluation on Wallflower dataset [23]

Algorithm	Error Type	Problem Type							Total Errors (TE)
		MO	TD	LS	WT	C	B	FA	
SG Wren et al. [24]	False neg	0	949	1857	3110	4101	2215	3464	35133
	False pos	0	535	15123	357	2040	92	1290	
MOG Stauffer et al. [25]	False neg	0	1008	1633	1323	398	1874	2442	27053
	False pos	0	20	14169	341	3098	217	530	
KDE Elgammal et al. [26]	False neg	0	1298	760	170	238	1755	2413	26450
	False pos	0	125	14153	589	3392	933	624	
PCA Oliver et al. [6]	False neg	0	879	962	1027	350	304	2441	17677
	False pos	1065	16	362	2057	1548	6129	537	
INMF Bucak et al. [17]	False neg	0	724	1593	3317	6626	1401	3412	19098
	False pos	0	481	303	652	234	190	165	
IRT Li et al. [18]	False neg	0	1282	2822	4525	1491	1734	2438	17053
	False pos	0	159	389	7	114	2080	12	
IMMC Proposed method	False neg	0	1336	2707	4307	1169	2677	2640	15714
	False pos	0	11	16	6	136	506	203	
































































Once the first foreground detection is made, we apply the IMMC to update the background model using (12) and (14). The class label for each pixel is obtained using the foreground mask.

**Remark:** Note that the IMMC can be applied directly at time  $t=1$  but it is less robust than to use firstly the batch algorithm on  $N$  frames and then to apply the IMMC to update the background.

## 4 Experimental Results

For the performance evaluation, we have compared our supervised approach with the unsupervised subspace learning methods PCA, INMF and IRT using the Wallflower dataset provided by Toyama et al. [23]. This dataset consists in a set of images sequences where each sequence presents a different type of difficulty that a practical task may meet: Moved Object (MO), Time of Day (TD), Light Switch (LS), Waving Trees (WT), Camouflage (C), Bootstrapping (B) and Foreground Aperture (F). The performance is evaluated against hand-segmented ground truth. Three terms are used in evaluation: False Positive (FP) is the number of background pixels that are wrongly marked as foreground; False Negative (FN) is the number of foreground pixels that are wrongly marked as background; Total Error (TE) is the sum of FP and FN. The Table 2 shows the performance in term of FP, FN and TE for each algorithm. The corresponding results are shown in Table 3. As we can see, the IMMC gives the lowest TE followed by the IRT, the INMF and the PCA. Secondly, we have compared our supervised approach with the state of the art algorithms: SG [24], MOG [25] and KDE [26]. As we can see on the Table 2 and Table 3, our algorithm gives

**Table 3.** Results on Wallflower dataset [23]

Sequence	MO	TD	LS	WT	C	B	FA
Frame	Frame 985	Frame 1850	Frame 1865	Frame 247	Frame 251	Frame 299	Frame 449
Test Image							
Ground Truth							
SG [24]							
MOG [25]							
KDE [26]							
PCA [6]							
INMF [17]							
IRT [18]							
IMMC							

better results particularly in the case of illumination changes. The results for SG, MOG and PCA comes from [27]. The results for the INMF was provided by their authors [17]. The KDE was implemented in Microsoft Visual C++ and the IRT and IMMC was implemented in Matlab.

## 5 Conclusion

In this paper, we have proposed to model the background using a supervised subspace learning called Incremental Maximum Criterion. This approach allows to initialize robustly the background and to upate incrementally the eigenvectors and eigenvalues. Experimental results made on the Wallflower datasets show the pertinence of the proposed approach. Indeed, IMMC outperforms the supervised PCA, INMF and IRT. For future investigations, supervised subspace learning methods such as Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA) seem to be very interesting approaches. For example, LDA exists in several incremental versions as incremental LDA using fixed point method [28] or sufficient spanning set approximations [29]. In the same way, Partial Least Squares (PLS) methods [30] give a nice perspective to model robustly the background.

## References

1. White, B., Shah, M.: Automatically tuning background subtraction parameters using particle swarm optimization. In: ICME 2007, pp. 1826–1829 (2007)
2. Elhabian, S., El-Sayed, K., Ahmed, S.: Moving object detection in spatial domain using background removal techniques - state-of-art. In: RPCS, vol. 1, pp. 32–54 (January 2008)
3. Bouwmans, T., Baf, F.E., Vachon, B.: Statistical background modeling for foreground detection: A survey. In: Handbook of Pattern Recognition and Computer Vision, vol. 4, pp. 181–189. World Scientific Publishing (2010)
4. Bouwmans, T., Baf, F.E., Vachon, B.: Background modeling using mixture of gaussians for foreground detection: A survey. In: RPCS, vol. 1, pp. 219–237 (November 2008)
5. Bouwmans, T.: Subspace learning for background modeling: A survey. In: RPCS, vol. 2, pp. 223–234 (November 2009)
6. Oliver, N., Rosario, B., Pentland, A.: A bayesian computer vision system for modeling human interactions. In: ICVS 1999 (January 1999)
7. Rymel, J., Renno, J., Greenhill, D., Orwell, J., Jones, G.: Adaptive eigen-backgrounds for object detection. In: ICIP 2004, pp. 1847–1850 (October 2004)
8. Li, Y., Xu, L., Morphett, J., Jacobs, R.: An integrated algorithm of incremental and robust pca. In: ICIP 2003, pp. 245–248 (September 2003)
9. Skocaj, D., Leonardis, A.: Weighted and robust incremental method for subspace learning. In: ICCV 2003, pp. 1494–1501 (2003)
10. Zhang, J., Zhuang, Y.: Adaptive weight selection for incremental eigen-background modeling. In: ICME 2007, pp. 851–854 (July 2007)
11. Wang, L., Wang, L., Zhuo, Q., Xiao, H., Wang, W.: Adaptive eigenbackground for dynamic background modeling. LNCS, vol. 2006, pp. 670–675 (2006)
12. Zhang, J., Tian, Y., Yang, Y., Zhu, C.: Robust foreground segmentation using subspace based background model. In: APCIP 2009, vol. 2, pp. 214–217 (July 2009)
13. Li, R., Chen, Y., Zhang, X.: Fast robust eigen-background updating for foreground detection. In: ICIP 2006, pp. 1833–1836 (2006)
14. Yamazaki, M., Xu, G., Chen, Y.: Detection of moving objects by independent component analysis. In: Narayanan, P.J., Nayar, S.K., Shum, H.-Y. (eds.) ACCV 2006. LNCS, vol. 3852, pp. 467–478. Springer, Heidelberg (2006)
15. Tsai, D., Lai, C.: Independent component analysis-based background subtraction for indoor surveillance. IEEE Transactions on Image Processing, IP 2009 8, 158–167 (2009)
16. Chu, Y., Wu, X., Sun, W., Liu, T.: A basis-background subtraction method using non-negative matrix factorization. In: International Conference on Digital Image Processing, ICDIP 2010 (2010)
17. Bucak, S., Günsel, B.: Incremental subspace learning and generating sparse representations via non-negative matrix factorization. Pattern Recognition 42, 788–797 (2009)
18. Li, X., Hu, W., Zhang, Z., Zhang, X.: Robust foreground segmentation based on two effective background models. In: MIR 2008, pp. 223–228 (October 2008)
19. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR 1999, pp. 246–252 (1999)
20. Li, H., Jiang, T., Zhang, K.: Efficient and robust feature extraction by maximum margin criterion. Advances in Neural Information Processing Systems vol. 16 (2004)

21. Wang, F., Zhang, C.: Feature extraction by maximizing the average neighborhood margin. In: CVPR 2007, pp. 1–8 (2007)
22. Yan, J., Zhang, B., Yan, S., Yang, Q., Li, H., Chen, Z.: Immc: incremental maximum margin criterion. In: KDD 2004, pp. 725–730 (August 2004)
23. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. In: ICCV 1999, pp. 255–261 (September 1999)
24. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfunder: Real-time tracking of the human body. *IEEE Transactions on PAMI* 19, 780–785 (1997)
25. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: CVPR 1999, pp. 246–252 (1999)
26. Elgammal, A., Harwood, D., Davis, L.: Non-parametric model for background subtraction. In: ECCV 2000, pp. 751–767 (June 2000)
27. Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. In: International Conference on Computer Vision, pp. 255–261 (September 1999)
28. Chen, D., Zhang, L.: An incremental linear discriminant analysis using fixed point method. In: Wang, J., Yi, Z., Žurada, J.M., Lu, B.-L., Yin, H. (eds.) *ISNN 2006*. LNCS, vol. 3971, pp. 1334–1339. Springer, Heidelberg (2006)
29. Kim, T., Wong, S., Stenger, B., Kittler, J., Cipolla, R.: Incremental linear discriminant analysis using sufficient spanning set approximations. In: CVPR, pp. 1–8 (June 2007)
30. Rosipal, R., Krämer, N.C.: Overview and recent advances in partial least squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) *SLSFS 2005*. LNCS, vol. 3940, pp. 34–51. Springer, Heidelberg (2006)

# Trace Norm Regularization and Application to Tensor Based Feature Extraction

Yoshikazu Washizawa

Brain Science Institute, Riken

**Abstract.** The trace norm regularization has an interesting property that is rank of a matrix is reduced according to its continuous regularization parameter. We propose a new efficient algorithm for a kind of trace norm regularization problems. Since the algorithm is not gradient-based approach, its computational complexity does not depend on initial states or learning rate. We also apply the proposed algorithm to a tensor based feature extraction method, that is an extension of the trace norm regularized feature extraction.

Computational simulations show that the proposed algorithm provides an accurate solution in less time than conventional methods. The proposed trace based feature extraction method show almost that same performance as Multilinear PCA.

## 1 Introduction

The regularization has been researched in mathematics and computer science, especially machine learning and pattern recognition. The regularization is known to help to avoid the over-fitting (over-learning) problems, e.g. the ridge regression or the support vector machines (SVMs). Furthermore, the regularization itself can be used for feature extraction [1], and the  $l_1$  norm regularization is known to provide a sparse solution.

Recently, the trace norm (or nuclear norm) regularization has been utilized for matrix optimization problems, and applied to several applications such as multi-task learning or recommender systems [2,3,4,5,6,7],

$$\underset{\mathbf{X}}{\text{minimize}} \quad f(\mathbf{X}) + \mu \|\mathbf{X}\|_T, \quad (1)$$

where  $\|\mathbf{X}\|_T = \text{Trace}[(\mathbf{X}^\top \mathbf{X})^{1/2}]$  is the trace norm, and  $\mu \geq 0$  is the regularization parameter. For a symmetric matrix  $\mathbf{B}$ ,  $\mathbf{B}^{1/2}$  is a symmetric matrix that satisfies  $\mathbf{B}^{1/2} \mathbf{B}^{1/2} = \mathbf{B}$ . The most advantage of using trace norm is that the rank of the solution  $\mathbf{X}^*$  is usually reduced, and hence for input vector  $\mathbf{b}$ , multiplication,  $\mathbf{X}^* \mathbf{b}$ , is in the subspace. The number of dimension of the subspace (rank of  $\mathbf{X}^*$ ) depends on  $\mu$ .

Let  $\mathbf{A}$  be a given matrix. If  $f(\mathbf{X})$  has forms  $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{A} - \mathbf{X}\|_F^2$  or  $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2$ , the problem has closed form solutions [4,1], where  $\|\cdot\|_F$  denotes the Frobenius norm. If  $f(\mathbf{X})$  is a convex and Lipschitz continuous function,

several gradient-based approaches and the semi-definite programming (SDP) approaches have been proposed [5,6]. However, SDP is known to have high computational cost [2], and gradient-based approaches depend on initial conditions or learning rates.

In this paper, we study the case  $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{A} - \mathbf{X}\mathbf{B}\|_F^2$ , where  $\mathbf{A}$  and  $\mathbf{B}$  are given matrices. We propose an efficient and direct algorithm to the problem. Since our method is not gradient based approach, it does not depend on initialization, and faster than the other methods. Actually the method does not output the strict optimal solution, but a good approximation of the solution. Therefore, if we want to have an accurate solution, we can use the output of the proposed method as an initial value of gradient based approaches.

Furthermore, we apply the method to the tensor based feature extraction. Tensor (multi-linear) approaches are accompanied by the appearance of powerful computer environments. In vector or matrix analysis, data is often transformed to a set of vectors or matrices, then vector/matrix based approaches are applied. In tensor analysis, we can directly treat the structure of the data. For example, bio-medical data has many indices such as subjects, time (date), trials, sensor channels, etc. Multi-linear PCA (MPCA) is an extension of PCA [8]. In this paper, we extend the trace constrained feature extraction method [1] to tensor analysis using the proposed algorithm in a similar way with MPCA.

We summarize the problem and introduce a feature extraction using the trace norm regularization in Section 2. In Section 3, we introduce an efficient algorithm for the problem. Section 4 discusses the tensor feature extraction using the trace norm regularization. We show the experimental results in Section 5, and conclude in Section 6.

## 2 Properties of Trace Norm Regularization and Previous Works

### 2.1 Why Trace Norm Induces Rank Reduction

Suppose that the singular value decomposition (SVD) of  $\mathbf{X}$  be  $\mathbf{X} = \sum_{i=1}^{\text{rank}(\mathbf{X})} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ . Since there exists an ambiguity with respect to the sign, we usually assume that the singular values  $\sigma_i$  are non-negative. We, here, do not assume  $\sigma_i \geq 0$ , then the trace norm of  $\mathbf{X}$  is  $\|\mathbf{X}\|_T = \sum_{i=1}^{\text{rank}(\mathbf{X})} |\sigma_i| = \|\boldsymbol{\sigma}\|_1$ , where  $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \dots, \sigma_{\text{rank}(\mathbf{X})}]^\top$ , and  $\|\cdot\|_1$  is the  $l_1$  norm. It is well-known that the  $l_1$  norm minimization induces the sparse solution, therefore the singular values of the solution matrix  $\mathbf{X}^*$  is sparse, and the rank of  $\mathbf{X}^*$  is reduced. Recently several approaches or analyses have been reported [1,2,3,4,5,6,7].

### 2.2 Trace Norm Regularization in Vector Approximation Problems

We, here, review the feature extraction method using the trace norm [1]. We propose a tensor based feature extraction method based on this method in Section 4.2.

Let  $\mathbf{x}_1, \dots, \mathbf{x}_n$  be  $d$ -dimensional samples. Then approximation problem with the trace norm regularization is given by

$$\underset{\mathbf{X}}{\text{minimize}} \quad \frac{1}{2n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{X}\mathbf{x}_i\|^2 + \mu \|\mathbf{X}\|_T = \frac{1}{2} \|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2 + \mu \|\mathbf{X}\|_T, \quad (2)$$

where  $\mathbf{A} = \frac{1}{\sqrt{n}}[\mathbf{x}_1, \dots, \mathbf{x}_n]$ . The first term minimizes the averaged squared error between  $\mathbf{x}_i$  and  $\mathbf{X}\mathbf{x}_i$ , and the second term minimizes the trace norm of  $\mathbf{X}$ . This problem has a form  $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{A} - \mathbf{X}\mathbf{A}\|_F^2$  in eq. (1).

If  $\mu = 0$ ,  $\mathbf{X}^* = \mathbf{I}$  (identity matrix) extracts all features of  $\mathbf{A}$ . The regularization term  $\mu \|\mathbf{X}\|_T$  limits the degree of freedom of  $\mathbf{X}$ . Thus  $\mathbf{X}^*$  extracts intrinsic feature in the samples. The Frobenius norm could be useful too, however, Frobenius norm does not give lower rank solutions [1]. For classification problems, we obtain  $\mathbf{X}^*$  for each class. Let  $\mathbf{X}_c^*$  be the matrix of the class  $c$ . An unknown input pattern  $\mathbf{x}$  is classified to the class  $k$  that is  $\text{argmin}_k \|\mathbf{x} - \mathbf{X}_k^* \mathbf{x}\|$ . In subspace method,  $\mathbf{X}_c^*$  is a projection onto the subspace of the class  $c$ . The one advantage of the trace norm is that the parameter  $\mu$  is a real number (continuous) whereas the parameter of the subspace methods, the dimension, is a natural number which is smaller than original dimension.

### 3 Algorithm for the Case $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{A} - \mathbf{X}\mathbf{B}\|_F^2$

#### 3.1 Minimization Algorithm

We here consider the case  $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{A} - \mathbf{X}\mathbf{B}\|_F^2$ , then the problem is

$$\underset{\mathbf{X}}{\text{minimize}} \quad J(\mathbf{X}) = \frac{1}{2} \|\mathbf{A} - \mathbf{X}\mathbf{B}\|_F^2 + \mu \|\mathbf{X}\|_T, \quad (3)$$

where  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times L}$  are given matrix.

We provide an efficient and direct algorithm for the problem (3). Since the derivation of the algorithm is rather complex, we put its details in the Appendix. We, here, show only outline of the derivation and the procedure of the algorithm in Algorithm 1.

Let  $\mathbf{R}_B = \mathbf{B}\mathbf{B}^\top$  and  $\mathbf{R}_{AB} = \mathbf{A}\mathbf{B}^\top$ . Assume that  $\mathbf{R}_B$  and  $\mathbf{R}_{AB}$  are full-rank. If they are not full-rank, we can reduce the dimension using SVD [2]. Suppose that SVD of  $\mathbf{X}$  be  $\mathbf{X} = \mathbf{P}_1 \boldsymbol{\Sigma} \mathbf{P}_2^\top$  ( $(\boldsymbol{\Sigma})_{ii} > 0$ ). Since the trace norm  $\|\mathbf{X}\|_T$  is not differentiable, we use the sub-gradient of  $J(\mathbf{X})$  that is

$$\delta J(\mathbf{X}) = \mathbf{X}\mathbf{R}_B - \mathbf{R}_{AB} + \mu \mathbf{P}_1 \mathbf{P}_2^\top + \mu \mathbf{S}, \quad (4)$$

where  $\mathbf{S}$  is a matrix that satisfies  $\mathbf{P}_1^\top \mathbf{S} = 0$ ,  $\mathbf{S}\mathbf{P}_2 = 0$ , and  $\|\mathbf{S}\|_2 \leq 1$ .  $\|\cdot\|_2$  denotes the spectral norm [7]. We attempt to find the solution that satisfies  $\delta J(\mathbf{X}) = 0$ . From  $\delta J(\mathbf{X}) = 0$ , we have following propositions.

**Proposition 1.** *Suppose that SVD of  $\mathbf{R}_{AB}$  is  $\mathbf{R}_{AB} = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$ . If  $\delta J(\mathbf{X}) = 0$ , there exists  $C_1 \subset \{1, \dots, d\}$ , and i)  $\mathbf{P}_1 \mathbf{P}_1^\top = \sum_{i \in C_1} \mathbf{u}_i \mathbf{u}_i^\top$  and ii)  $\mathbf{P}_2 \mathbf{P}_2^\top = \sum_{i \in C_1} \mathbf{v}_i \mathbf{v}_i^\top$  are satisfied.*

**Algorithm 1.** Algorithm to minimize problem (3)

---

```

1: Input:  $\mathbf{A}, \mathbf{B}, \mu$ 
2: Output:  $\mathbf{X}^*$ 
3: Obtain SVD  $\mathbf{R}_{AB} = \mathbf{A}\mathbf{B}^\top = \sum_{i=1}^d \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$ . Suppose that eigenvalues  $\lambda_i$  are sorted
   in descending order.
4: Set  $l_1$  be the maximum index  $i$  that satisfy  $\lambda_i > \mu$ , and set  $l_2 = d$ .
5: repeat
6:   Calculate  $l = \lceil (l_1 + l_2)/2 \rceil$ ,  $\mathbf{V}_1 = [\mathbf{v}_1, \dots, \mathbf{v}_l]$ ,  $\mathbf{K} = \mathbf{V}_1^\top \mathbf{B}\mathbf{B}^\top \mathbf{V}_1$ ,  $\mathbf{A}_1 =$ 
    $\text{diag}(\lambda_1, \dots, \lambda_l)$ ,  $\mathbf{E} = (\mathbf{K}^{-1} \mathbf{A}_1^2 \mathbf{K}^{-1}) - \mu \mathbf{K}^{-1}$ 
7:   if  $\mathbf{E}$  is positive-definite then
8:      $l_1 = \lceil (l_1 + l_2)/2 \rceil$ 
9:   else
10:     $l_2 = \lceil (l_1 + l_2)/2 \rceil$ 
11:   end if
12: until  $l_2 = \lceil (l_1 + l_2)/2 \rceil$ 
13: Calculate  $\mathbf{V}_1 = [\mathbf{v}_1, \dots, \mathbf{v}_l]$ ,  $\mathbf{K} = \mathbf{V}_1^\top \mathbf{B}\mathbf{B}^\top \mathbf{V}_1$ ,  $\mathbf{A}_1 = \text{diag}(\lambda_1, \dots, \lambda_l)$ ,  $\mathbf{E} =$ 
    $(\mathbf{K}^{-1} \mathbf{A}_1^2 \mathbf{K}^{-1}) - \mu \mathbf{K}^{-1}$ 
14: Calculate EVD  $\mathbf{E} = \mathbf{V}_X \boldsymbol{\Sigma} \mathbf{V}_X^\top$ .
15: Calculate  $\mathbf{U}_X = \mathbf{A}_1^{-1} (\mathbf{V}_1^\top \mathbf{B}\mathbf{B}^\top \mathbf{V}_1 \mathbf{V}_X \boldsymbol{\Sigma} + \mu \mathbf{V}_X)$ ,  $\mathbf{U}_1 = [\mathbf{u}_1, \dots, \mathbf{u}_{l_1}]$ ,
16: Output  $\mathbf{X}^* = \mathbf{U}_1 \mathbf{U}_X \boldsymbol{\Sigma} \mathbf{V}_X^\top \mathbf{V}_1^\top$ 

```

---

**Proposition 2.** If  $\lambda_i > \mu$ , the index  $i$  is in  $C_1$ .

Therefore, the optimal  $\mathbf{P}_1$  and  $\mathbf{P}_2$  are limited to the space that is spanned by  $\{\mathbf{u}_i\}_{i \in C_1}$  and  $\{\mathbf{v}_i\}_{i \in C_1}$  respectively. Suppose that  $\mathbf{U}_1$  and  $\mathbf{V}_1$  be matrices whose column vectors are  $\{\mathbf{u}_i\}_{i \in C_1}$  and  $\{\mathbf{v}_i\}_{i \in C_1}$  respectively. Then we only have to find the set  $C_1$  and the unitary matrices  $\mathbf{U}_X \in \mathbb{R}^{|C_1| \times |C_1|}$  and  $\mathbf{V}_X \in \mathbb{R}^{|C_1| \times |C_1|}$  that is  $\mathbf{P}_1 = \mathbf{U}_1 \mathbf{U}_X$  and  $\mathbf{P}_2 = \mathbf{V}_1 \mathbf{V}_X$ . We also provide a method to obtain  $\mathbf{U}_X$  and  $\mathbf{V}_X$  that is described in Appendix. Thus we only have to determine  $C_1$ .

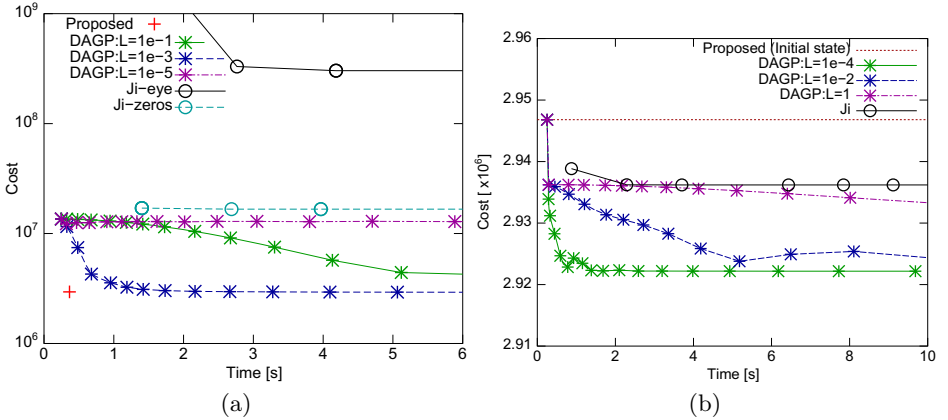
Strictly speaking, we have to seek all possibilities  $i \in C_1$  and  $i \notin C_1$  for each  $i$  such that  $\lambda_i \leq \mu$ . However, we found that if we only check positive-definiteness of a matrix  $\mathbf{E}$  (defined in Algorithm 1) according to the descending order of the singular value  $\lambda_i$ , a good approximation of the solution can be found.

### 3.2 Comparison with Other Optimization Methods

We compared our method with gradient based approaches, the dual accelerated gradient-projection (DAGP) [2], and [5]. [6] also treats the same problem (1), however, they did not use  $\mu$  directly but the other parameter  $t$  (in [6]). Actually there exists corresponding  $t$  for every  $\mu$ , however, we cannot obtain  $t$  from  $\mu$  directly. Therefore we cannot compare our method with [6].

We generated matrices  $\mathbf{A} = \mathbf{A}_1 \mathbf{A}_2$  and  $\mathbf{B} = \mathbf{B}_1 \mathbf{A} + \mathbf{B}_2$ , where  $\mathbf{A}_1, \mathbf{B}_1 \in \mathbb{R}^{200 \times 200}$  are random matrix whose elements follow the uniform distribution  $[0, 1]$ , and  $\mathbf{A}_2, \mathbf{B}_2 \in \mathbb{R}^{200 \times 10000}$  are also random matrix whose elements follow the standard Gaussian distribution. We set  $\mu = 50000$ . Each algorithm was tested 10 times using different  $\mathbf{A}_2$  and  $\mathbf{B}_2$ , and we obtain mean values of the cost function and runtime. We coded in GNU Octave compiled with Intel Math





**Fig. 1.** Cost and runtime, (a): Proposed method is a left-bottom cross. “DAGP” is the dual accelerated gradient-projection [2]. “Ji” is a method in [5]. (b) is the case that the output of our method is used for initial states.

Kernel Library, and conducted the simulation on PC with Intel i7 X980 3.33GHz (we used only one core for the simulation). Since runtime depends on coding skills, we open our source codes<sup>1</sup>.

We show the relation between runtime and value of the objective function (cost) in Figure 1 (a). The left-bottom cross is the result of the proposed method. The proposed method provides a good solution in less time.  $L$  is a learning coefficient for DAGP. When  $L = 1e - 3$ , DAGP converges to almost the same value as the proposed method. However, it takes more than two seconds whereas our method achieves in 0.4 sec. Moreover the performance of DAGP highly depends on the learning coefficient  $L$ . “Ji” stands for a method in [5], and “eye” and “zeros” stand for the initial  $\mathbf{X}$ , that are the identity matrix and the zero matrix respectively. This method does not converge to the optimal solution, and performance depends on the initial state.

Since the problem (3) is convex, gradient-based methods guarantee to converge to the global solution. We also tried the case that the proposed method is used for the initial state of the other methods (Figure 1 (b)). Both DAGP and Ji achieve smaller cost than the proposed method, however, the difference is smaller than 0.8%. The proposed method provides almost the optimal solution.

Although our algorithm does not guarantee the optimal solution, it outputs very good approximation. We think the reason is that the optimal solution is almost determined by the several largest singular values and vectors. Therefore, even if we only evaluate larger singular values and vectors, the solution is close to the optimal solution.

<sup>1</sup> Due to the double-blind review, we will open source codes after the acceptance decision.

## 4 Application to Tensor Based Feature Extraction

In this section, we apply the proposed algorithm to the tensor based feature extraction method that is an extension of the method in Section 2.2. Multilinear PCA (MPCA) is an extension of PCA that treats tensors [8]. We extend the trace regularization method in a similar way. We, first, describe tensor algebra and MPCA, then introduce the proposed method.

### 4.1 Tensor Algebra and Multi-linear PCA

Due to the limitation of the space, we enumerate notations, terms and operations of tensor briefly.

- **Tensor** - denoted by a calligraphic large letter e.g.,  $\mathcal{A}$  or  $\mathcal{A}(i_1, i_2, \dots, i_N)$  ( $i_1 = 1, 2, \dots, I_1, \dots, i_N = 1, 2, \dots, I_N$ ), where  $N$  is the number of modes. Suppose that each index of a mode if positive integer, and the maximum number of the index is called the dimension of the mode.
- **Fiber** - an  $I_j$ -dimensional vector obtained by fixing all modes except the  $j$ th mode, is called the fiber of the  $j$ th mode.
- **Unfolding matrices** -  $(I_j)$  by  $(\prod_{k \neq j} I_k)$  matrix laying all possible fibers of the  $j$ th mode is called the unfolding matrix of the  $j$ th mode. We denote the unfolding matrix of the  $j$ th mode by  $A_{(j)}$ . The inverse operation of the unfolding is called the folding. We denote the unfold and the fold operator of the  $j$ th mode by  $\text{Unfold}_j(\cdot)$  and  $\text{Fold}_j(\cdot)$  respectively ( $A_{(j)} = \text{Unfold}_j(\mathcal{A})$ ,  $\mathcal{A} = \text{Fold}_j(A_{(j)})$ ).
- **Tensor-matrix multiplication** - suppose  $\mathbf{a}$  is a fiber of a tensor  $\mathcal{A}$ . Given  $m$  by  $I_j$  matrix  $B$ , the  $j$ th multiplication  $\mathcal{A} \times_j B$  is done by replacing all possible fibers  $\mathbf{a}$  by  $B\mathbf{a}$ . The dimension of the  $j$ th mode of  $(\mathcal{A} \times_j B)$  is  $m$ . By using folding,  $\mathcal{A} \times_j B = \text{Fold}_j(BA_{(j)})$ .
- We denote  $\mathcal{A} \times_1 B_1 \times_2 B_2 \cdots \times_N B_N = \prod_{i=1}^N \mathcal{A} \times_i B_i$ .
- **Frobenius norm** of a tensor  $\mathcal{A}$  is defined by

$$\|\mathcal{A}\|_F^2 = \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_N=1}^{I_N} [\mathcal{A}(i_1, i_2, \dots, i_N)]^2 = \|\text{Unfold}_j(\mathcal{A})\|_F^2 \quad \text{for any } j.$$

Let  $\mathcal{A}_1, \dots, \mathcal{A}_L \in \mathbb{R}^{I_1 \times \dots \times \text{times} I_N}$  are given tensors, and suppose that the mean tensor is zero. MPCA is defined by an optimization problem

$$\underset{\mathbf{U}_1, \dots, \mathbf{U}_N}{\text{maximize}} \quad \sum_{l=1}^L \left\| \prod_{i=1}^N \mathcal{A}_l \times_i \mathbf{U}_i \right\|_F^2, \quad \text{subject to} \quad \mathbf{U}_i^\top \mathbf{U}_i = I, \mathbf{U}_i \in \mathbb{R}^{r_i \times I_i}. \quad (5)$$

$r_1, \dots, r_N$  are the parameters that specify dimensions. Since there is no way to obtain  $\mathbf{U}_1, \dots, \mathbf{U}_N$  directly, [8] adopts the alternating least square (ALS) method. If we fix  $\mathbf{U}_i$ , ( $i \in C_j = \{1, \dots, N\} \setminus \{j\}$ ), and obtain  $\mathbf{U}_j$ , the problem is reduced to the sub-problem,

$$\underset{\mathbf{U}_j}{\text{max}} \quad \sum_{l=1}^L \left\| \mathbf{U}_j \text{Unfold}_j \left( \prod_{i \in C_j} \mathcal{A}_l \times_i \mathbf{U}_i \right) \right\|_F^2, \quad \text{subject to} \quad \mathbf{U}_j^\top \mathbf{U}_j = I, \mathbf{U}_j \in \mathbb{R}^{r_j \times I_j}.$$

Since  $\text{Unfold}_j(\prod_{i \in C_j} \mathcal{A} \times_i \mathbf{U}_i)$  is a matrix, this problem can be solved by standard PCA approach. For each  $j = 1, \dots, L$ , the sub-problem is solved in rotation. This procedure monotonically decreases the cost function of the main problem (5).

## 4.2 Tensor Based Feature Extraction Using Trace Norm Regularization

We extend the trace norm regularization technique (2).

$$\underset{\mathbf{U}_1, \dots, \mathbf{U}_N}{\text{minimize}} \quad \sum_{l=1}^L \|\mathcal{A}_l - (\prod_{i=1}^N \mathcal{A}_l \times_i \mathbf{U}_i)\|_F^2 + \sum_{i=1}^N \mu_i \|\mathbf{U}_i\|_T, \quad (6)$$

where  $\mu_1, \dots, \mu_N$  are regularization parameters. This is natural extension of the problem (2), replacing vectors  $\mathbf{x}_i$  to tensor  $\mathcal{A}_l$ .

Since this problem is also difficult to obtain  $\mathbf{U}_1, \dots, \mathbf{U}_N$  simultaneously, we adopt the ALS strategy. If we fix  $\mathbf{U}_i$ , ( $i \in C_j = \{1, \dots, N\} \setminus \{j\}$ ), and obtain  $\mathbf{U}_j$ , the problem is reduced to

$$\underset{\mathbf{U}_j}{\text{minimize}} \quad \sum_{l=1}^L \|\text{Unfold}_j(\mathcal{A}_l) - \mathbf{U}_j \text{Unfold}_j(\prod_{i \in C_j} \mathcal{A}_l \times_i \mathbf{U}_i)\|_F^2 + \mu_j \|\mathbf{U}_j\|_T,$$

If we let  $\mathbf{A} = [\text{Unfold}_j(\mathcal{A}_1) \dots \text{Unfold}_j(\mathcal{A}_L)]$ ,  $\mathbf{B} = [\text{Unfold}_j(\prod_{i \in C_j} \mathcal{A}_1 \times_i \mathbf{U}_i) \dots \text{Unfold}_j(\prod_{i \in C_j} \mathcal{A}_L \times_i \mathbf{U}_i)]$ , the problem has the form (3), and can be solved by the algorithm proposed in Section 3.1.

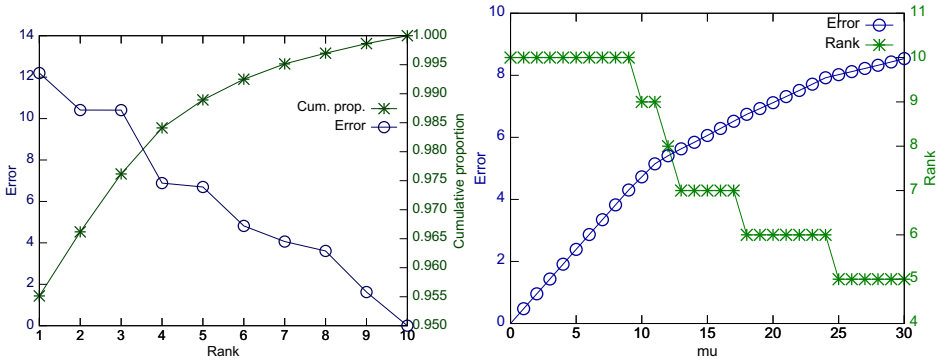
Let  $\mathcal{X}$  be an input tensor. Then transformed tensor is given by  $\prod_{i=1}^N \mathcal{X} \times_N \mathbf{U}_N$ , and dissimilarity  $d(\mathcal{X})$  is given by  $d(\mathcal{X}) = \|\mathcal{X} - \prod_{i=1}^N \mathcal{X} \times_i \mathbf{U}_i\|_F^2$ .

## 5 Experimental Results

### 5.1 Eigenface vs. Traceface

We, here, show example of trace norm regularization. Although this example is not adopt tensor approach, the result shows interesting property of the trace norm regularization. Eigenface has been used for face feature extraction [9]. However, if the number of samples (faces) is not sufficient, the residual error draws a discontinuous line because its parameter, the number of dimension, has to be a natural number. Even if the number of samples is sufficient, sometimes eigenvalues of the covariance matrix are discontinuous. Therefore, we sometimes cannot tune optimal parameter that provides right approximation. On the other hand, trace norm regularization draws a continuous line because the parameter  $\mu$  is a real number. Furthermore, the dimension is reduced according to  $\mu$ .

We show an example using the first person of the Olivetti Research Laboratory (ORL) face database [10]. We obtained PCA and trace regularized feature extractor from ten images, then we transformed the first one image  $\mathbf{x}$ . The residual error was obtained by  $\|\mathbf{x} - \mathbf{X}\mathbf{x}\|$ , where  $\mathbf{X}$  is PCA or the trace regularized



**Fig. 2.** Eigenface (left) and traceface (right): Residual error and the parameter



**Fig. 3.** Eigenface (top) and traceface (bottom)

feature extractor. We show the result in Figure 2. The residual error of eigenface has discontinuous line whereas that of traceface is smooth. Moreover rank of the traceface is reduced when  $\mu$  is large. Figure 3 shows images of the methods. The variation of traceface is more smooth than that of eigenface.

## 5.2 Handwritten Digit Recognition by Tensor Feature Extraction

We used USPS handwritten digit database. Each image is gray-scaled 16x16 pixel image, and has a label '0'-'9'. The database has 7291 samples for training, and 2007 samples for testing. Each image is described in 256-dimensional vector. We added indices that are *scaling*, *rotation*, *horizontal shift*, and *vertical shift*. We generated three images for each index, therefore, one sample is a five-mode tensor in  $\mathbb{R}^{256 \times 3 \times 3 \times 3 \times 3}$ . *Scaling* and *rotation* were done by commands `imreshape` and `imrotate` of GNU octave, respectively.

We used randomly selected 10% of the training samples (729 samples) for training, and the remaining samples are used for validation. We obtained tensor based feature extractor,  $\mathbf{U}_1, \dots, \mathbf{U}_5$  (in Section 4.2) for each class, using the training samples. Then we calculated the dissimilarity between validation samples  $\mathcal{X}$  and the class  $c$ , and classify the sample to the class whose dissimilarity is minimum.

**Table 1.** Result of handwritten digit classification

Method	Proposed	MPCA	CLAFIC	TQC
Error rate [%]	$4.23 \pm 0.75$	$4.20 \pm 0.63$	$5.28 \pm 0.58$	$4.47 \pm 0.30$

Table 1 shows the experimental result, the mean values and standard deviation of three trials. For the class feature information compression (CLAFIC) and the trace constrained quadratic classifier (TQC) [1], original 256-dimensional vectors are used. For MPCA and CLAFIC, all possible sets of parameters were evaluated, and obtained the lowest error rate. For the proposed method and TQC, we sought optimal parameter from several picks. The proposed method shows slightly worse performance than MPCA. However, as TQC outperforms CLAFIC, the trace norm approach is promising.

## 6 Conclusion

We proposed a new efficient algorithm for the trace regularized problem (3). Although the proposed algorithm does not guarantee the optimal solution, it provides a very good approximation or an initial point in very low computational cost. Our simulation in Section 3.2 demonstrated our algorithm.

Furthermore, we applied the proposed algorithm to the tensor based feature extraction using trace norm regularization. The sub-problem of the optimization problem (6) can be solved by the proposed algorithm.

Experimental results of the traceface showed the advantage of the trace norm regularization. In the handwritten digit classification problem, by introducing tensor method, the proposed method shows almost the same performance as MPCA.

As we mentioned, the proposed algorithm to minimize the trace regularization problem does not guarantee the optimal solution, it has to be verified in various problems. Moreover, estimated error from the optimal solution also should be discussed in future research.

## References

1. Washizawa, Y.: Feature extraction using constrained approximation and suppression. *IEEE Trans. Neural Networks* 21, 201–210 (2010)
2. Pong, T.K., Tseng, P., Ji, S., Ye, J.: Trace norm regularization: reformulations and multi-task learning (2009), [http://www.public.asu.edu/~sji03/papers/pdf/Pong\\_tnr\\_mtl.pdf](http://www.public.asu.edu/~sji03/papers/pdf/Pong_tnr_mtl.pdf)
3. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *Computer* 42, 30–37 (2009)
4. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. *SIAM J. Optim.* 20, 1956–1982 (2010)
5. Ji, S., Ye, J.: An accelerated gradient method for trace norm minimization. In: *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, pp. 457–464 (2009)

6. Jaggi, M., Sulovský, M.: A simple algorithm for nuclear norm regularized problems. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pp. 471–478 (2010)
7. Bach, F.R.: Consistency of trace norm minimization. Journal of Machine Learning Research 8, 1019–1048 (2008)
8. Lu, H., Plataniotis, K.N., Venetsanopoulos, A.N.: MPCA: multilinear principal component analysis of tensor objects. IEEE Trans. Neural Networks 19, 18–39 (2008)
9. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3, 71–86 (1991)
10. A & T Laboratories Cambridge: Orl database (2002), <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>

## Appendix – Derivation of the Algorithm

We attempt to find  $\mathbf{A}$  that gives  $\delta J(\mathbf{A}) = 0$ . From  $\mathbf{P}_1^\top \mathbf{S} = 0$ ,  $\mathbf{S}\mathbf{P}_2 = 0$ ,  $\mathbf{S}$  can be parameterized by using a matrix  $\mathbf{B}$ , like  $\mathbf{S} = (\mathbf{I} - \mathbf{P}_1\mathbf{P}_1^\top)\mathbf{B}(\mathbf{I} - \mathbf{P}_2\mathbf{P}_2^\top)$ . Substituting this to  $\delta J(\mathbf{A})$ , and multiplying  $(\mathbf{I} - \mathbf{P}_1\mathbf{P}_1^\top)$  and  $(\mathbf{I} - \mathbf{P}_2\mathbf{P}_2^\top)$  from left and right hand side respectively, we have  $\mathbf{B} = \frac{1}{\mu}(\mathbf{I} - \mathbf{P}_1\mathbf{P}_1^\top)\mathbf{R}_X(\mathbf{I} - \mathbf{P}_2\mathbf{P}_2^\top) + \mathbf{P}_1\mathbf{C}\mathbf{P}_2^\top$ , where the second term will be vanished when we substitute. Then we have  $\partial J(\mathbf{A}) = \mathbf{P}_1\boldsymbol{\Sigma}\mathbf{P}_2^\top\mathbf{R}_B - \mathbf{R}_{AB} - \mu\mathbf{P}_1\mathbf{P}_2^\top + (\mathbf{I} - \mathbf{P}_1\mathbf{P}_1^\top)\mathbf{R}_{AB}(\mathbf{I} - \mathbf{P}_2\mathbf{P}_2^\top)$ . By multiplying  $(\mathbf{I} - \mathbf{P}_1\mathbf{P}_1^\top)$  from left-hand side,  $\mathbf{P}_2\mathbf{P}_2^\top$  from right-hand side, we obtain  $(\mathbf{I} - \mathbf{P}_1\mathbf{P}_1^\top)\mathbf{R}_{AB}\mathbf{P}_2\mathbf{P}_2^\top = 0$ . Hence, Proposition [1](#) is derived. Suppose that a set  $C_1$  in Proposition [1](#) is known. Suppose that SVD of  $\mathbf{R}_{AB}$  is  $\mathbf{R}_{AB} = \sum_i \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$ . Let  $\mathbf{A}_1$  be a diagonal matrix of  $\{\lambda_i\}_{i \in C_1}$ ,  $\mathbf{U}_1$  and  $\mathbf{V}_1$  be matrices whose column vectors are  $\{\mathbf{u}_i\}_{i \in C_1}$  and  $\{\mathbf{v}_i\}_{i \in C_1}$ . Then  $\mathbf{P}_1$  and  $\mathbf{P}_2$  can be expressed by  $\mathbf{P}_1 = \mathbf{U}_1\mathbf{U}_X$ , and  $\mathbf{P}_2 = \mathbf{V}_1\mathbf{V}_X$ , where  $\mathbf{U}_X$  and  $\mathbf{V}_X$  are unitary matrices.

Then  $\delta J(\mathbf{A})$  is given by  $\delta J(\mathbf{A}) = \mathbf{U}_1(\mathbf{U}_X\boldsymbol{\Sigma}\mathbf{V}_X^\top\mathbf{V}_1^\top\mathbf{R}_B - \mathbf{A}_1\mathbf{V}_1^\top + \mu\mathbf{U}_X\mathbf{V}_X^\top\mathbf{V}_1^\top)$ . Since  $\mathbf{U}_1$  is the matrix above,  $\delta J(\mathbf{A}) = 0$  when inside of the bracket is zero. Then we have the relation,  $\mathbf{U}_X = \mathbf{A}_1^{-1}\mathbf{V}_1^\top\mathbf{R}_B\mathbf{V}_1\mathbf{V}_X\boldsymbol{\Sigma} + \mu\mathbf{A}_1^{-1}\mathbf{V}_X$ . Let  $\mathbf{K} = \mathbf{V}_1^\top\mathbf{R}_B\mathbf{V}_1$ ,  $\mathbf{K}$  is positive-definite. From  $\mathbf{U}_X^\top\mathbf{U}_X = \mathbf{I}$ , we also have a relation,  $(\mathbf{V}_X\boldsymbol{\Sigma}\mathbf{V}_X^\top + \mu\mathbf{K}^{-1})\mathbf{K}\mathbf{A}_1^{-2}\mathbf{K}(\mathbf{V}_X\boldsymbol{\Sigma}\mathbf{V}_X^\top + \mu\mathbf{K}^{-1}) = \mathbf{I}$ . Both  $(\mathbf{V}_X\boldsymbol{\Sigma}\mathbf{V}_X^\top + \mu\mathbf{K}^{-1})$  and  $\mathbf{K}\mathbf{A}_1^{-2}\mathbf{K}$  are symmetric, the equation hold when  $\mathbf{V}_X\boldsymbol{\Sigma}\mathbf{V}_X^\top = (\mathbf{K}\mathbf{A}_1^{-2}\mathbf{K})^{1/2} - \mu\mathbf{K}^{-1}$ . Since  $\mathbf{V}_X$  is an unitary matrix and  $\boldsymbol{\Sigma}$  is a diagonal-matrix, the left-hand side should be eigenvalue decomposition (EVD) of the right-hand side. Consequently, if we obtain  $C_1$ , matrices  $\mathbf{V}_X$  and  $\mathbf{U}_X$  are obtained, and the solution  $\mathbf{X}^*$  is given.

Here,  $\mathbf{S}$  is given by  $\mathbf{S} = \frac{1}{\mu}\mathbf{U}_2\mathbf{A}_2\mathbf{V}_2^\top$ , where  $\mathbf{A}_2$  is a diagonal matrix of  $\{\lambda_i\}_{i \notin C_1}$ ,  $\mathbf{U}_2$  and  $\mathbf{V}_2$  are matrices whose column vectors are  $\{\mathbf{u}_i\}_{i \notin C_1}$  and  $\{\mathbf{v}_i\}_{i \notin C_1}$ . From  $\|\mathbf{S}\|_2 \leq 1$ , if  $\lambda_i/\mu > 1$ , the index  $i$  should be in  $C_1$ , hence Proposition 2 is derived.

# Fast and Robust Face Recognition for Incremental Data

I. Gede Pasek Suta Wijaya<sup>1,2</sup>, Keiichi Uchimura<sup>1</sup>, and Gou Koutaki<sup>1</sup>

<sup>1</sup>Computer Science and Electrical Engineering of GSST, Kumamoto University,  
Kurokami 2-39-1, Kumamoto Shi, Japan

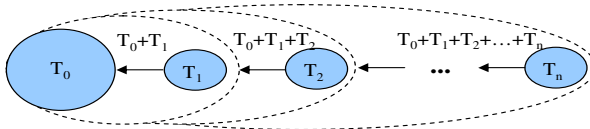
<sup>2</sup>Electrical Engineering Department, Faculty of Engineering, Mataram University,  
Jl. Majapahit 62 Mataram, West Nusa Tenggara, Indonesia  
gdepasek@navi.cs.kumamoto-u.ac.jp,  
{uchimura,koutaki}@cs.kumamoto-u.ac.jp

**Abstract.** This paper proposes fast and robust face recognition system for incremental data, which come continuously into the system. Fast and robust mean that the face recognition performs rapidly both of training and querying process and steadily recognize face images, which have large lighting variations. The fast training and querying can be performed by implementing compact face features as dimensional reduction of face image and predictive LDA (PDLDA) as face classifier. The PDLDA performs rapidly the features cluster process because the PDLDA does not require to recalculate the between class scatter,  $S_b$ , when a new class data is registered into the training data set. In order to get the robust face recognition achievement, we develop the lighting compensation, which works based on neighbor analysis and is integrated to the PDLDA based face recognition.

## 1 Introduction

Face recognition is one to many matches which compare a query face features against all training face features to determine the identity of a query face. It remains hard to be done because variations in a single face can be very large, while the variations between different faces can be quite small. In addition, face variability also depends on ethnicity and registration technique (i.e., capture method, lighting condition, and devices).

PCA [1], LDA [2], and their variations [3,4,5] based face recognition are most popular approach because of their uncomplicated processing. However, the main problem of them is that they have to retrain all of the samples to get the optimum projection matrix ( $W$ ) when new data come continuously into the system (as shown in Fig. 1 with  $T_i$  representing  $i$ -th incremental data). Recently, two methods have been proposed to address to this problem, as described in Refs. [6,7]. Ref. [6] algorithm redefined within class scatter ( $S_w$ ) formulation and made simplification of calculating the global mean. However, the between class scatter ( $S_b$ ) still depends on the global mean and the  $W$  has to be determined, as done by the LDA algorithm which requires  $O(n^3)$  computational complexity. In other



**Fig. 1.** The illustration incremental data

side, the Ref. [7] proposed another strategy to this problem called as generalized singular value decomposition-incremental LDA (GSVD-ILDA) which determined  $W$  of incremental data using SVD which has less time computation than that of Ref. [6]. However, the GSVD-ILDA still has to recalculate the global mean for constructing the  $H_b = [\sqrt{p_1}\mu_1 - \mu_a, \dots, \sqrt{p_L}\mu_L - \mu_a]$  and requires QR decomposition, twice SVD, and twice inverse to obtain  $W$  for each incremental data. As known, the inverse matrix needs  $(O(n^3))$  computational complexity, where  $n$  is size of matrix.

This paper proposes fast and robust face recognition system for incremental data. This system is another strategy to overcome retraining problem, which can performs rapidly both of training and querying process and steadily recognize face images with large lighting variations. It can be realized by: implementing simple lighting compensation algorithm which provides better performance than that of existed methods, redefining the  $S_b$  using predictive and constant global means, and implementing compact face features as dimensional reduction of face image. The redefined  $S_b$  has the same characteristic as the original one in terms of its symmetrical and separable and has much less computational complexity than that of the original one for incremental data. In addition, this paper is much difference with the Ref. [10] in term of lighting compensation, the effect of  $S_b$  on the recognition rate, and the incremental data processing.

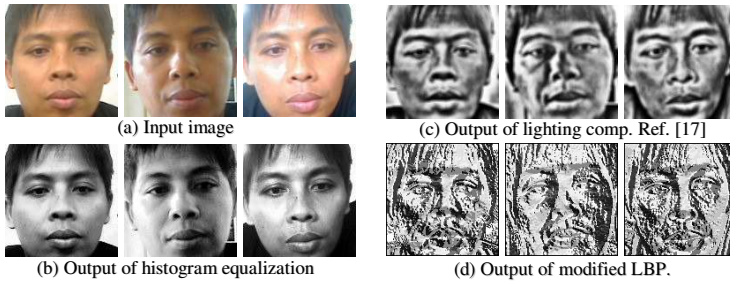
## 2 The Proposed Algorithm

Our proposed algorithm mainly consists of three processes: pre-processing consist of lighting compensation and data normalization; holistic features extraction that is used to get the specifics and powerful information of face image; and features classifier which is used to obtain most separable projected features cluster and to determine the similarity score between the query projected features and the training projected features set.

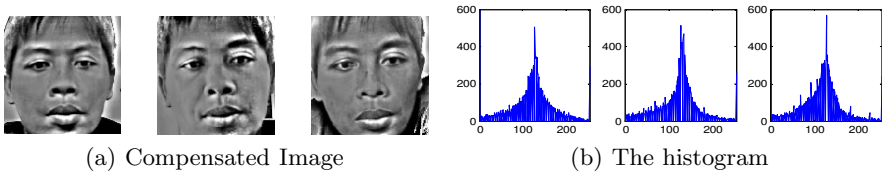
### 2.1 Pre-processing

Histogram equalization is commonly used to remove non-uniform lighting effect when face image is captured. However, this method did not work at all when input images have large lighting variations (outdoor and indoors), as shown in Fig. 2(a). J. Ruiz-de-Solar et al. [8] has been performed a comparative study of different pre-processing approach to illumination compensation. Based on this project, the self-quotient image (SQI) and the modified linear binary pattern (mLBP) approaches are the most suitable algorithm to achieve illumination





**Fig. 2.** The example of lighting compensation



**Fig. 3.** The output of our illumination compensation

compensation for Eigenface face recognition system. In addition, Kurita et al. [9] proposed robust pre-processing for illumination compensation of face image which is based on low pass filter with providing robust result over the SQI. However, its algorithm is much the same as the SQI algorithm (see Ref. [8]) and both of them are not easy to know what kind of low pass filter that is suitable for this process.

In this paper, we adopt the SQI and Ref. [9] based methods to develop simple illumination compensation algorithm which main goal is to provide better achievement than that of recent best existed methods such as mLBP and Wijaya et al. (Ref. [10]) methods. The developed algorithm consists of four steps, as follows:

1. Color space transformation, which transforms the face image in RGB color space to YCbCr because the RGB is not necessarily. Then, the compensation just performs in the intensity (Y) component because the lighting just affects the contrast and brightness of the image which is placed on the Y component.
2. Illuminance definition, which is determined by dividing the input image (i.e. Y component) into  $N$ -by- $N$  blocks, then computing the mean of each block, and finally resizing the result into the input image size. This process will get the non-uniform lighting effect on the face image. By using trial and error, the best block size ( $N$ ) for this process is 4.
3. Dividing the original image ( $I(x,y)$ ) that represents the input stimulus with the result of brightness definition in point 2 ( $L(x,y)$ ) that represent the illuminance or perception using  $g$ :  $R(x,y) = \{I(x,y)/L(x,y)\} \cdot \alpha$ , where  $\alpha$  is constant coefficient for making centering the image intensity.
4. Normalizing the output of the point 3 process using stretching algorithm to get the uniform contrast and brightness of the input face image.

The output of the lighting compensation can be seen in Fig. 3. It shows that all of the images have almost the same brightness and contrast, which is shown by almost identical histogram data (Fig. 3(b)) and remaining to provide good local facial features such as clearer eyes, mouth, nose, and face outline (see Fig. 3(a)). It means the proposed lighting compensation tends to overcome the large variations of face images due to the lighting variations. When the Fig. 3(a) is compared with the compensated image of existed methods, as shown in the Fig. 2(c and d), our method tends to provide better achievement which keeps the most significant information such as local facial features after the compensation. Consequently, it tends to give robust performance for large variation of the illumination data, such as YaleB database.

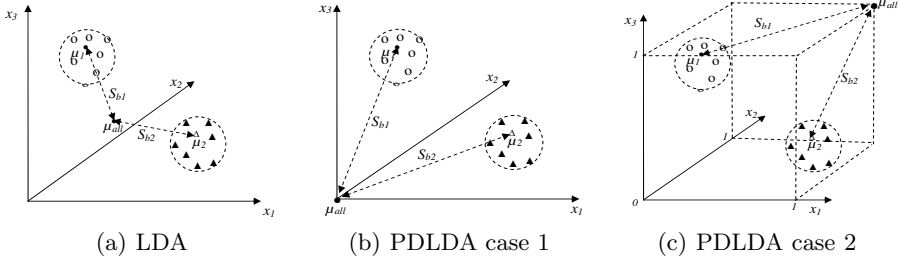
## 2.2 Holistic Features Extraction

In this research, a compact holistic features (HF) of face image, which is a set dominant frequency content and moment information of entire face, is implemented as dimensional reduction instead of raw face image. This concept has been reported effectively for dimensional reduction which compressed by about 99% of original size with providing the good enough achievement [10,11].

The compact HF is created using three steps: firstly, convert the DCT decomposition coefficients of face image to a vector using row ordering technique; secondly, sort the vector descending using quick sort algorithm, and finally truncate  $m$  first vector elements (i.e., less than 100 elements). From the dominant frequency content, if they are reconstructed into the face images, the reconstructed face images are different. However, we can still understand that they are the face images, as shown and described clearly in Refs. [10,11]. It means that the dominant frequency content existing in low-frequency components is sufficient for face image representation, which can be implemented as part of HF of face image that has powerful discriminant information and small dimension. In order to get robust HF of any face pose variations, the moment information that provides invariant measure of face images shape is considered. The moment information is obtained using invariant moment analysis, which is derived from central moment analysis [12]. The invariant moment set is invariant to translation, scale change, and rotation, therefore this concept can be employed to get the holistic information of any face pose variations. The strengths of the proposed features are that it has higher discriminant power (DP) and provides better performances than that of without moment information, as reported by Wijaya et al. [10].

## 2.3 Features Classifier

Suppose, we have the three-dimensional data cluster of two classes which is normalized in the range [0-1], shown in Fig. 4(a). By expanding this illustration to  $n$ -dimensional data which have  $L$  classes and each class ( $k$ -th) has  $N_k$  samples, then the optimum projection matrix ( $W$ ), which has to satisfy the Fisher criterion (Eq. 1), can be determined by eigen analysis of  $S_w^{-1}S_b$  and then select  $m$



**Fig. 4.** The illustration PDLDA in three-dimensional data

orthonormal eigenvectors corresponding to the largest eigenvalues (i.e.  $m < n$ ). Where, the  $S_b = \frac{1}{L} \sum_{k=1}^L P(x^k)(\mu_k - \mu_a)(\mu_k - \mu_a)^T$  is between-class scatter matrix,  $S_w = \frac{1}{N} \sum_{k=1}^L \sum_{i=1}^{N_k} (x_i^k - \mu_k)(x_i^k - \mu_k)^T$  is the within-class scatter matrix,  $\mu_k$  is mean features vector of  $k$ -th class, and  $\mu_a$  is mean of all samples.

$$J_{LDA}(W) = \underset{W}{arg\ max} \frac{|W^T S_b W|}{|W^T S_w W|} \tag{1}$$

This LDA algorithm has been implemented successfully as face recognition with providing good and stable performance in both small and large sample size data, as explained in Ref. [2][10]. However, it has to retrain all data samples to obtain optimum projection matrix when new data samples enter into the system. The retraining has to be done because the  $S_b$  depends on the global means, which has to recalculated when new data sample comes. In order to avoid this problem and to decrease its computational load, we develop a predictive LDA (PDLDA), which is derived by defining the global mean  $\mu_a$  as a constant value for all samples, as shown in Fig. 4(a and b).

From this illustration, if the global mean,  $\mu_a$ , is set-up as a constant vector ( $\mu_p$ ) by moving the  $\mu_a$  to the origin point (see Fig. 4(b)) or to maximum value of the range (see Fig. 4(c)), it will make  $S_b$  not only require much less computational complexity but also have the same basic structure as the original one  $S_b^{org}$  in terms of separable scatter and symmetrical matrix. If a new data class,  $x^{new}$ , incrementally comes into the system, the predictive  $S_b$  can be updated using the following equation.

$$\begin{aligned} S_b^p &= \sum_{k=1}^L P(x^k)(\mu_k - \mu_p)(\mu_k - \mu_p)^T + P(x^{new})(\mu_{new} - \mu_p)(\mu_{new} - \mu_p)^T \\ &= S_b^{old} + S_b^{new} \end{aligned} \tag{2}$$

Hereafter, By substituting the  $S_b$  with the  $S_b^p$  of LDA eigen analysis, we will get the optimum projection matrix called as PDLDA projection matrix ( $W_{PDLDA}$ ). As note,  $S_w$  is determined by the same way as done in the original one. By using this  $W_{PDLDA}$ , the projected features of the both training and querying data set can be performed by the following equation:

$$Y_i^k = W_{PDLDA}^T X_i^k \quad (3)$$

Finally, the Euclidean distance based on nearest neighbor rule is implemented for face classification.

## 2.4 The Effect of $S_b^p$ to the Discrimination

In order to prove that  $S_b^p$  has the same separability as the original one ( $S_b$ ), we calculate the discrimination power (DP) [2], which represents the ability of features separation, using the following equation.

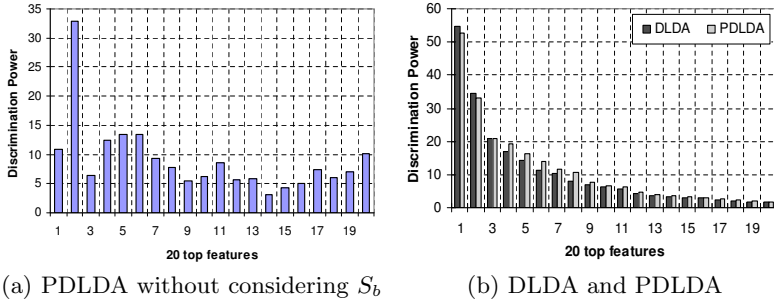
$$J(W) = sep(W) = trace(S_W^{-1} S_b) \quad (4)$$

In this case, we examine the DP of the PDLDA projected data using the procedure below on two conditions: when the  $S_b$  is unconsidered and considered for determining the optimum  $W_{PDLDA}$ . The procedure consists of three steps:

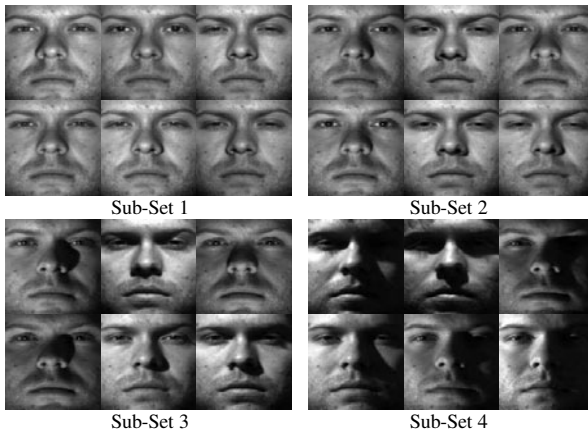
1. Determine the projected data ( $Y_i^k$ ) using the Eq. 3.
2. Determine the within and between class scatter of projected data ( $Y_i^k$ ) called as  $S_w^{Pro}$  and  $S_b^{Pro}$  as the same as done by LDA respectively.
3. Calculate the DP of the projected data that is done by substituting  $S_w$  and  $S_b$  of Eq. 4 with  $S_w^{Pro}$  and  $S_b^{Pro}$  respectively.

In the first condition, we substitute the  $S_b$  with identity matrix ( $I$ ) of the eigen analysis ( $S_w^{-1} S_b$ ) to obtain optimum  $W_{PDLDA}$ . By using this optimum  $W_{PDLDA}$ , we examine the DP using the above procedures in well-known ORL database. The examination result shows that the classification information of face image is not placed in few top discriminant vectors but spreading to all over features vector, as shown in Fig. 5(a). It means to get the better recognition rate the more discriminant vectors have to be considered. Consequently, the more discriminant vectors are, the larger the face features size will be, which affect to the memory space requirements. While in the last condition, we perform the same process as the first condition except on substituting  $S_b$  of the eigen analysis with the  $S_b^p$ . The results of the last condition examination show that the PDLDA (see Fig. 5.b) have closely the same DP as that of the Direct LDA (DLDA [4]) algorithm and have higher DP than that of the first condition (Fig. 5(a)). It means that the PDLDA have the same characteristic as the original one in terms the ability of features separation or in other word, the result prove that  $|W^T S_b^p W|$  has much the same value as the  $|W^T S_b^{org} W|$  which make the data cluster of the PDLDA be the same separable as that of the original one. The higher DP tends to provide the higher recognition rate which will be proven using experimental data in the next section.

Regarding to time complexity of recalculating  $S_b$  using Eq. (2), it requires:  $(n + n^2)$  multiplication and  $n$  addition operations. However, the original one requires  $(L + 1)(n + n^2)$  multiplication and  $(L + 1)n$  addition operations, where  $L + 1$  is total class member of data training and  $n$  is the dimensional size of features vector.



**Fig. 5.** The discrimination power of our proposed methods



**Fig. 6.** Example of face with large lighting variations

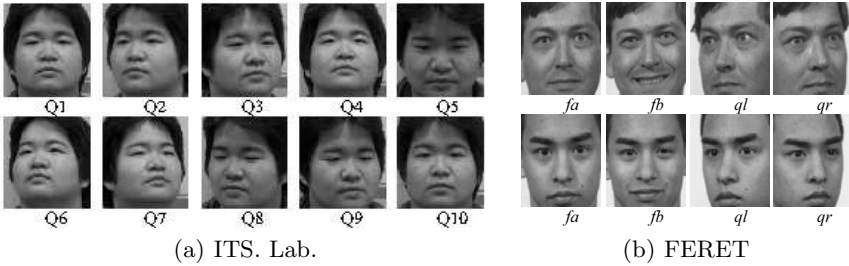
### 3 Experimental Setup and Results

The experiments were carried out using several challenge face databases: YALEB database (YAL) [14], ITS-Lab. Kumamoto University database (ITS) [10], and FERET database (FER) [13]. Each database has special characteristics. The tests were performed using PC with specification: Core-Duo Processor 1.7 GHz and 2 GB RAM.

The first experiment, which was carried out on the YaleB database investigated the robustness of our proposed lighting compensation to any variations of lighting condition compared with the established method, such as histogram equalization (HE), modified Linear Binary Pattern (mLBP [8]), method in Ref. [10]. The example of face image with lighting variations of the YALEB database is divided by four set, as shown in Fig. 6. In this case, the sub-set 1 was chosen as training and the remaining sub-sets were selected as testing. From the experimental results, our lighting compensation can improve the existed methods significantly, such as mLBP and method of Ref. [10] by about 5.45% and 4.88% respectively, as shown in Table 1. The significant improvement of

**Table 1.** The comparison of the recognition rate of the proposed lighting compensation to established algorithms.

No	Methods	Recognition Rate (%)			
		1 vs 2	1 vs 3	1 vs 4	Average
1	HE	95.39	60.13	13.69	56.39
2	mLBP	100	100	78.71	92.90
3	Method in Ref. [10]	100	100	80.40	93.47
4	Our Method	<b>100</b>	<b>100</b>	<b>95.06</b>	<b>98.35</b>

**Fig. 7.** Pose example of ITS and FERET face database

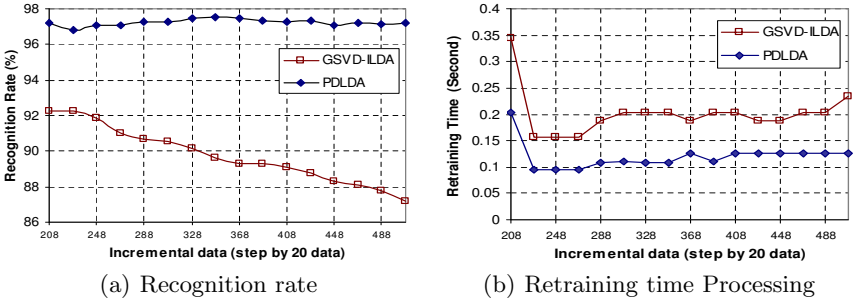
recognition rate is given by sub-set 4, because face images on this sub-set contains large lighting variations. It can be achieved because our lighting compensation provides better compensated face image with almost identical histogram data than that of mLBP and method of Ref. [10], as shown in Fig. 3. It means any lighting condition of face images are compensated into almost the same contrast and brightness face images by our proposed lighting compensation method.

The second experiment was carried out with ITS. Lab database (consisting of 100 classes) and FERET database (consisting of 508 classes) which represents small and large size database respectively. This test investigated effectiveness of the integration of our lighting compensation and PDLDA against to the established methods. The example of face pose variation of the databases can be shown in Fig. 7. From these data, half of the samples were selected as the training sample and remaining as test samples. In addition, the recent existing GSVD-ILDA, which has been reported to provided better achievement than GSVD-LDA and IDR/QR methods (in detail, see Ref. [7]), is used as comparison. The experimental results (Table 2) show that the proposed method tends to provide better achievement than those of the established methods and event than that of GSVD-ILDA. It can be achieved because the PDLDA almost has the same discrimination power as the DLDA, see Fig. 5 while the GSVD-ILDA just working using the approximation projection matrix. This achievement also proves that the  $S_b$  of PDLDA has the same structure and satisfy the same optimum criterion as that of the DLDA.

In order to show that our proposed method requires less time processing for retraining, the next experiment was performed. It was done on FERET face database with incremental data scenario: firstly, it was trained 208 face classes

**Table 2.** The comparison of the recognition rate of the lighting compensation+PDLDA to established methods

No	Methods	Features Dimension	Recognition Rate (%)		
			TTS. Lab.	FERET	Average
1	HF+2DLDA	8x8	93.69	91.66	92.68
2	HF+(2D) <sup>2</sup> LDA	8x8	95.56	91.24	93.40
3	HF+(2D) <sup>2</sup> PCALDA	8x8	93.16	89.99	91.58
4	HF+DLDA	24	98.76	96.94	97.85
5	HF+GSVD-ILDA [7]	24	95.47	95.79	95.63
6	HF+PDLDA	24	<b>98.77</b>	<b>97.24</b>	<b>98.01</b>

**Fig. 8.** The performance of the incremental testing

and then added gradually 20 new face classes into the system until 508 face classes. The experimental results were plotted in Fig. 8. It shows that the recognition rate for incremental data is much robust (Fig. 8(a)) with less retraining time (Fig. 8(b)) than those of GSVD-ILDA. It can be achieved because the PDLDA is identical with the original direct LDA which is the best variation of LDA while GSVD-ILDA is the approximation of the original one. These results match with the result of GSVD-ILDA as reported in Ref. [7] which provided less recognition rate than GSVD-LDA. In terms of retraining time, the PDLDA takes less retraining time than that of DLDA:  $n + n^2$  multiplication and  $n$  addition for PDLDA and  $(L + 1)(n + n^2)$  multiplication and  $(L + 1)n$  addition for DLDA as described in section (4.c). In other side, the GSVD-ILDA required QR decomposition, twice SVD, and twice inverse to obtain  $W$  for each incremental data. As known, the inverse matrix need large computational cost by about  $(O(n^3))$ , where  $n$  size of matrix which make the GSVD-ILDA take longer retraining time than our method.

## 4 Conclusion and Future Works

The proposed lighting compensation, which is employed as pre-processing of face image, provides robust performance in terms of recognition rate compared to



that of existed methods. In detail, by implementing the proposed illumination compensation significant improvement can be achieved by more than 20% of the mLBP and 15% of the method in Ref. [10], which does not effect much time processing the PDLDA base face recognition. In terms of retraining for incremental data our proposed method can achieve robust result for both small and large database size with less time processing.

In order to get more precise verification result, we will consider more local features analysis involving eyes, nose, mouth, and context information of the face image. In addition, we will investigate false acceptance and rejection rate in order to know real-time performances.

## References

1. Turk, M., Pentland, A.: Eigenfaces for Recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
2. Etemad, K., Chellappa, R.: Discriminant analysis for Recognition of Human Face Images. *J. opt. Soc. Am. A* 14(8), 1724–1733 (1997)
3. Chen, W., Meng, J.-E., Wu, S.: PCA and LDA in DCT Domain. *Pattern Recognition Letter* (26), 2474–2482 (2005)
4. Yu, H., Yang, J.: A Direct LDA algorithm for High-Dimensional Data-with Application to Face Recognition. *Pattern Recognition* 34, 2067–2070 (2001)
5. Nousath, S., Kumar, G.H., Shivakumara, P.:  $(2D)^2LDA$ : An Efficient Approach for Face Recognition. *Pattern Recognition* 39, 1396–1400 (2006)
6. Pang, S., Ozawa, S., Kasabov, N.: Incremental Linear Discriminant Analysis for Classification of Data Streams. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 35(5), 905–914 (2005)
7. Zhao, H., Yuen, P.C.: Incremental Linear Discriminant Analysis for Face Recognition. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 38(1), 210–221 (2008)
8. Ruiz-del-Solar, J., Quinteros, J.: Illumination compensation and normalization in eigenspace-based face recognition: A comparative study of different pre-processing approaches. *Pattern Recognition Letter* 29(14), 1966–1979 (2008)
9. Kurita, S., Tomikawa, T.: Study On Robust Pre-Processing For Face Recognition Under Illumination Variations. In: *The Workshop of Image Electronics and Visual Computing 2010, Nice France (March 2010)*, (CDROM)
10. IGPS, Wijaya, Uchimura, K., Hu, Z.: Improving the PDLDA Based Face Recognition Using Lighting Compensation. In: *The Workshop of Image Electronics and Visual Computing 2010, Nice France (March 2010)*, (CDROM)
11. IGPS, Wijaya, Uchimura, K., Hu, Z.: Pose Invariant Face Recognition Based on Hybrid Dominant Frequency Features. *IEICE Transactions on Information and Systems* 91-D(8), 2153–2162 (2008)
12. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn., pp. 839–842. Pearson Prentice Hall, USA (2008)
13. Philips, P.J., Moon, H., Risvi, S.A., Rauss, P.J.: The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Anal. Machine Intell.* 22(10), 1090–1104 (2000)
14. <http://cvc.yale.edu/projects/yalefacesB/yalefacesB>



# Extracting Scene-Dependent Discriminant Features for Enhancing Face Recognition under Severe Conditions

Rui Ishiyama and Nobuyuki Yasukawa

Information and Media Processing Research Laboratories, NEC Corporation  
1753, Shimonumabe, Nakahara-Ku, Kawasaki 211-8666 Japan

**Abstract.** This paper proposes a new method to compare similarities of candidate models that are fitted to different areas of a query image. This method extracts the discriminant features that are changed due to the varying pose/lighting condition of given query image, and the confidence of each model-fitting is evaluated based on how much of the discriminant features is captured in each foreground. The confidence is fused with the similarity to enhance the face-identification performance. In an experiment using 7,000 images of 200 subjects taken under largely varying pose and lighting conditions, our proposed method reduced the recognition errors by more than 25% compared to the conventional method.

## 1 Introduction

Face recognition is now successfully used in some real applications such as gate control and photo search. [1,10] Many conventional studies have shown that pattern-recognition techniques are powerful tools for face recognition under constrained imaging conditions; namely, the query image is captured in a near-frontal pose under moderate lighting variations. [5]

In applications such as video surveillance and non-interactive interfaces, however, the query image is captured under un-constrained conditions. Facial pose and lighting conditions change to a huge extent. It is impossible to collect a complete set of training data that covers their infinite variations. Furthermore, there is no feature that is invariant across any conditions. [3] Useful features for face identification are changed due to the conditions: some features are distinctive under one condition but become invisible under a different condition. Consequently, the recognition algorithms using the fixed features that are predetermined using the training data works well only under limited conditions, and its recognition performance tends to degrade significantly under severe conditions.

The subspace methods are widely-used and powerful tools for modeling the infinite variations due to the lighting conditions. [2,11] If 3D shape and albedo are acquired in the enrollment process, the parametric model is constructed that can synthesize facial images under arbitrary pose and lighting conditions. Such face recognition methods have been proposed that construct the 3D models for each subject in the database and recognize 2D query image. [2,4,6,9] In

those methods, the individual models are fitted to the query image, and the images of the enrolled subjects under the same condition as the query image are reconstructed. Then the similarities between the query and reconstructed images are compared (Fig. 1). Since the pose/lighting conditions of all images are compensated to be the same as the query, recognition is successfully conducted without being affected by the variations.

Such model-based methods are commonly confronted with the following problem: which pixels of the query image belong to the foreground that the model should be fitted to? As its pose and subject differs, the foreground significantly changes for each query image. It is unpractical to assume that the accurate foreground is given by the face detection process. Therefore, the model-fitting starts with a rough estimation and adjusts the pose parameters so as to maximize the similarity. Since the fitting is individually conducted for each model, each pose (and the foreground determined by it) differs. Thus the methods have to compare the similarities of facial images that have different foregrounds. In the conventional methods, the similarity of each model is calculated as the reconstruction error averaged over each foreground. [2,6]

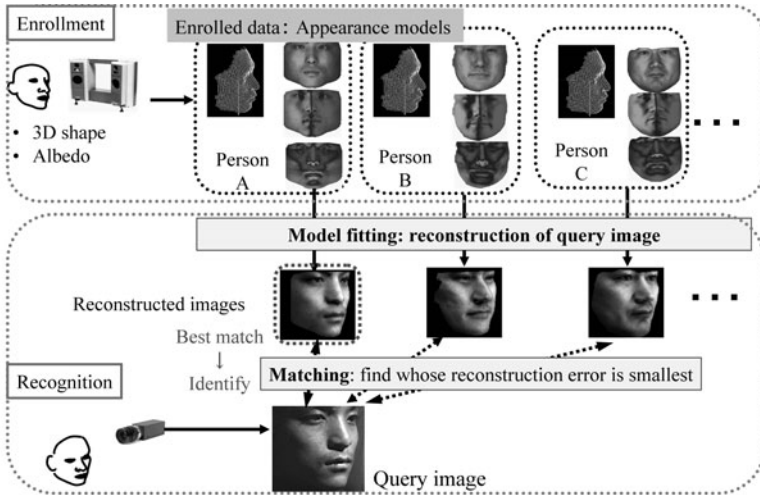
The problem is that an incorrect model can be fit to an invalid foreground and produce 'better' similarity score than the correct model fitted to the valid foreground. Such "poor fitting" is inevitable because any model can perfectly fit to such pixels covered with dark shadows or having no textures. An example is illustrated in Fig. 2. The pose may be adjusted to make the foreground to contain more such pixels that can be fitted with little error. Although several methods using edge features to avoid the poor fitting have been proposed [8], they cannot solve the problem perfectly because it is difficult to determine which edges belong to the true foreground. A new method is required to evaluate the confidence of model-fitting and to reject the poor fitting in the recognition.

The main contribution of this paper is to propose a new metric that evaluates the confidence of model-fitting. The confidence is evaluated based on how much of the discriminant features is captured in its foreground. The proposed method determines the scene-dependent discriminant features, i.e. the pixels that are important for face recognition under the particular pose/lighting condition of given query. Since the discriminant features significantly differ depending on the condition, our proposed method determines those features online. Another contribution is to propose a method to reject "poor fitting" by fusing the confidence into the similarity score and to enhance face recognition performance. The experiments using a large set of the images captured under hugely varying pose/lighting conditions show the efficacy of our proposed method.

## 2 Proposed Algorithm

### 2.1 Problem: Matching Facial Images with Different Shapes

Face recognition based on fitting appearance models is illustrated in Fig. 1. Numerous methods [2,4,6,9] for constructing the appearance model and conducting face recognition using the model fitting have been proposed. [8,11] A

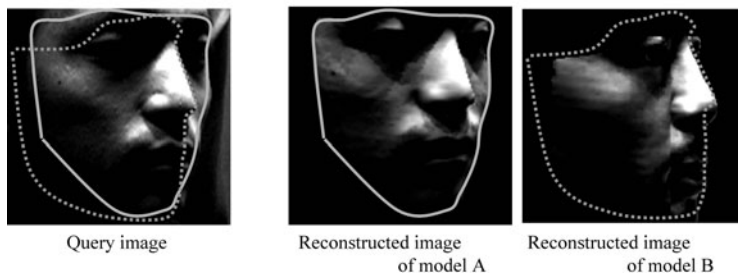


**Fig. 1.** Example of face-recognition algorithm using 3D appearance models. In the enrollment step, 3D appearance models of each subject are registered in the database. In the recognition step, each model is fitted to a query image, and the similarities of the reconstructed images are compared in order to identify the correct face.

typical method uses 3D shape and albedo to calculate the illumination bases that can predict any variations in appearance due to lighting conditions. 3D shape is also used to predict pose variations. The appearance model is fitted to the query image by estimating pose and illumination parameters. The methods proposed in [2,6] warp the illumination bases calculated on the texture space into the query image frame. The method then reconstructs the lighting condition by least-squares fitting of a linear combination of the warped illumination bases to the query image. Starting from a given rough estimate, the pose is updated by detecting optical flow between the reconstructed image and the query image. [7,8] The reconstruction and updating pose is repeated until the pose is converged.

When the query image is taken under large variations in pose and lighting conditions, it is quite difficult to determine the precise facial area to which the appearance model should be fitted. That is because face detection and image segmentation techniques often fail for images that are non-frontal and contain many shadows. Even without specifying a precise facial area, the existing algorithms can fit 3D face models to the query image. Only a rough estimation of facial pose is required to initiate the model-fitting. Since the shapes of each subject's face differ, the reconstructed images have different face areas. Consequently, the algorithm has to compare the similarities of the reconstructed images, which are defined in different foreground areas.

The conventional methods simply calculate the average error per pixel over each face area and identify the face by choosing the one having the smallest error. It seems enough for the query images captured under moderate conditions;



**Fig. 2.** Example of "poor fitting result" that affect non-frontal face recognition under severe condition. Although model B is incorrect subject, reconstruction error inside its foreground (surrounded by dotted line) is smaller than that of model A (correct subject, its foreground is surrounded by solid line).

however, it becomes problematic for severe conditions, an example of which is shown in Fig. 2. The areas to which the two example models are fitted are slightly different (surrounded by the solid line and the dotted line). Unfortunately in this case, model B (incorrect subject) has less error than model A (correct subject).

In this example shown in Fig.2, any model exhibits little error in the cheek and chin, because there is no discriminant feature due to shadows. Foreground of model B (surrounded by the dotted line) seems less confident than that of model A (surrounded by the solid line), because the foreground of model A captures the left part of the face, which is considered to be more significant for face recognition in this condition. Our problem is how to extract these discriminant features in a given query image.

In the following sections, a new method is proposed to extract significant features for face identification. Then the confidence of model-fitting is evaluated, and it will be combined into matching score to enhance face recognition performance.

## 2.2 Solution: Adaptive Extraction of Discriminant Features

In this section, a new method for extracting the discriminant feature that changes due to the imaging condition of given query image is proposed. Our method analyzes the reconstructed images by fitting the appearance model to extract the discriminant features in accordance with given query image.

In the enrollment step of the conventional 2D-3D face-recognition methods [24,6,9], 3D appearance models of each subject are acquired and registered in the matching database. The appearance model describes the face images of subject  $i$  under any pose  $\mathbf{p}$  and lighting condition  $\mathbf{l}$  by parametric model  $f(\mathbf{p}, \mathbf{l})$ . There are numerous methods to construct such appearance models. They acquire 3D shape and albedo by 3D scans [2,6] or estimation [4,9], and they calculate the illumination bases by using spherical harmonics or applying PCA to images synthesized by varying lighting conditions.

In the recognition step, appearance model  $f_i(\mathbf{p}, \mathbf{l})$  of subject  $i$  is fitted to the query image  $\mathbf{Q}$ , thus the reconstructed image  $\mathbf{R}_i$  is obtained. Pose  $\mathbf{p}_i$  and lighting parameters  $\mathbf{l}_i$  are estimated by minimizing the reconstruction errors over the pixels inside foreground  $F_i$ , which is defined as the face area of model  $i$  in the fitted pose.

$$(\mathbf{p}_i, \mathbf{l}_i) = \arg \min \sum_{\mathbf{x} \in F_i} |\mathbf{Q}(\mathbf{x}) - \mathbf{f}_i(\mathbf{x}; \mathbf{p}_i, \mathbf{l}_i)| \quad (1)$$

Reconstructed image  $\mathbf{R}_i$  reproduces the estimated pose/lighting conditions that optimally fit model  $i$  to the query image. Let us denote the mask image which indicates  $F_i$  by  $\delta_i(\mathbf{x})$ . Note that  $F_i$  and  $\delta_i(\mathbf{x})$  differ for each subject.

$$\mathbf{R}_i(\mathbf{x}) = \mathbf{f}_i(\mathbf{x}; \mathbf{p}_i, \mathbf{l}_i) \quad (2)$$

$$\delta_i(\mathbf{x}) = 1 \text{ if } \mathbf{x} \in F_i, \text{ otherwise } 0 \quad (3)$$

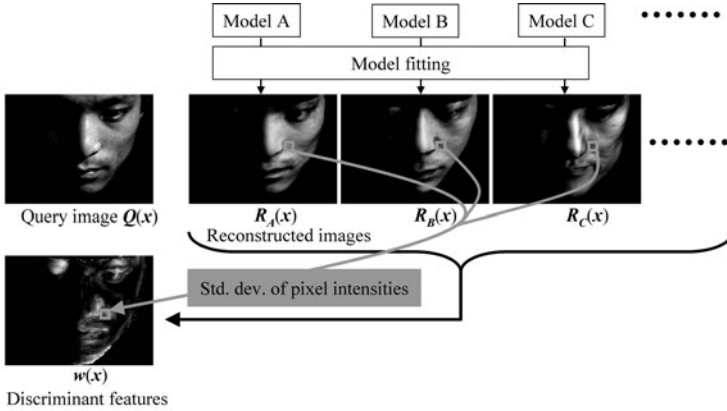
Here, a new method for extracting the discriminant features adaptively to the imaging condition of given query image is proposed. The image features (i.e. pixel intensities) that have large variations across reconstructed images  $\mathbf{R}_i$  are considered to be discriminant features and useful for face identification. Otherwise, those features can be fitted by any model with little error. With our proposed method, the importance  $w(\mathbf{x})$  of the pixel  $\mathbf{x}$  for face identification is evaluated from the standard deviations of  $\{\mathbf{R}_i(\mathbf{x})\}$  as follows:

$$w(\mathbf{x}) = \left( \sum_i \{ \delta_i(\mathbf{x}) |R_i(\mathbf{x}) - R_{avg}(\mathbf{x})|^2 \} \right)^{1/2} / \sum_i \delta_i(\mathbf{x}), \quad (4)$$

$$R_{avg}(\mathbf{x}) = \sum_i \{ \delta_i(\mathbf{x}) R_i(\mathbf{x}) \} / \sum_i \delta_i(\mathbf{x}) \quad (5)$$

Figure 3 presents a result obtained by our proposed method. First, the reconstructed images  $\mathbf{R}_i$  are obtained by fitting the enrolled models to the query image  $\mathbf{Q}$ . The discriminant feature  $w(\mathbf{x})$ , which is extracted from these images by equation (4), is shown on the lower left. This result shows that our proposed method adaptively extracts the discriminant features in the condition of given query image. Under this condition, the right half of the face is covered with strong shadows; thus, those areas have no discriminant features and are fitted by any model with little error. In contrast, the left half of the face has discriminant features in the edges. Note that  $w(\mathbf{x})$  has large values in the right eyelid. Although this area is shadowed in the query image, the area is not shadowed for some subjects (see reconstructed image of model B) and has discriminant features that appear in this particular lighting condition. Our proposed method can also extract such features.

Figure 4 shows examples of applying our proposed method for a variety of test images. The results in Fig. 4 show that our proposed method can extract the discriminant features that change according to the pose and lighting condition.



**Fig. 3.** Outline of our method for determining discriminant features of query image online

### 2.3 Fusing Fitting Confidence with Similarity for Recognition

In this section, a new method is proposed to evaluate the confidence of the model-fitting based on the discriminant features extracted by the method proposed in the previous section. The confidence will be fused with the appearance similarity to conduct face identification.

The conventional method calculates mean absolute differences of the image features over each model’s foreground  $F_i$ , and its inverse is used as the similarity score  $S_i$  to identify the face. The query image is identified to person  $i$  if  $S_i$  has the largest value.

$$S_i = 1 - \sum_{\mathbf{x} \in F_i} |Q(\mathbf{x}) - R_i(\mathbf{x})| / \sum_{\mathbf{x} \in F_i} I \tag{6}$$

where  $I$  indicates maximum value of pixel intensities.  $S_i$  is normalized to have a value from 0 to 1 so that  $S_i$  represents the similarity.

Since each subject’s facial shape is different, foreground  $F_i$  differs for each subject  $i$ . As discussed in the earlier sections, some sub-regions of the face are easy to fit by any model, but discriminant areas are not. If  $F_i$  contains more of the former and less of the latter, it becomes likely to have high similarity score. However, the confidence of similarity  $S_i$  becomes doubtful. Here, a method is proposed for evaluating the confidence  $C_i$  of similarity score  $S_i$  according to how much of discriminant feature is captured by its foreground  $F_i$ :

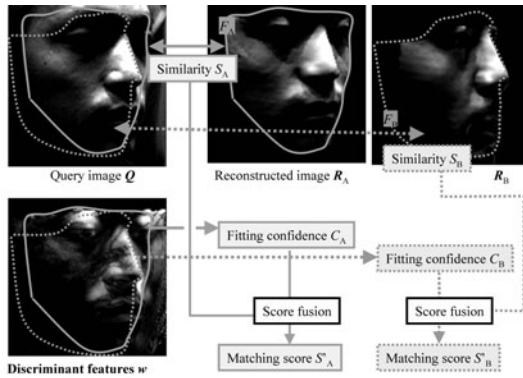
$$C_i = \sum_{\mathbf{x} \in F_i} w(\mathbf{x}) / \sum_{\mathbf{x}} w(\mathbf{x}) \tag{7}$$

$C_i$  is normalized to have value from 0 to 1, then this criterion is fused with similarity  $S_i$  as follows:

$$S'_i = S_i C_i^\gamma \tag{8}$$



**Fig. 4.** Examples of a query and its discriminant features extracted by our proposed method



**Fig. 5.** Outline of our proposed method of evaluating confidence of model-fitting and fusing appearance similarity with the confidence into matching score used for face recognition.

The resultant fused score  $S'_i$  is used for face identification. Here,  $\gamma$  is a parameter that balances the weights on the appearance similarity and the matching confidence.

Figure 5 illustrates the flow of our proposed method. Model A (correct subject) and B (incorrect subject) were fitted to areas  $F_A$  (surrounded by a solid line) and  $F_B$  (dotted line), respectively. In this case, model B has comparably high similarity scores ( $S_B > S_A$ ). Although the shape of model B is not aligned to right chin and mouth areas, it does not produce errors due to strong shadows.

Our method proposed in the previous section determines where discriminant features appear under this particular pose and illumination condition (shown on the lower left of Fig. 5). Many discriminant features are detected in the left half of the face, but no feature in the right cheek and chin areas. Since  $F_A$  covers the

former area (i.e. much more discriminant features than  $F_B$ ), a higher confidence score  $C_A$  is given to model A. The query image is thereby successfully identified, because  $S'_A > S'_B$ .

### 3 Experiments

In this section, the efficacy of our proposed method is evaluated by applying it to face recognition under large variations in pose and illumination. Our proposed method is combined with a conventional face recognition method of [3], which is selected as an example of many existing methods that use 3D appearance models [5,7]. The recognition performances of the original conventional method and that combined with our method are compared experimentally.

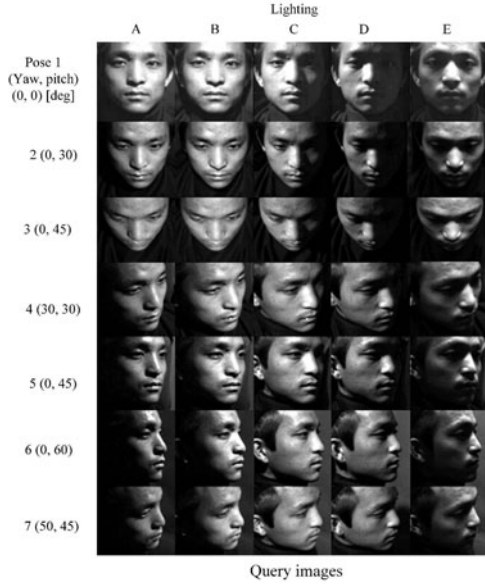
As for the query images, 7,000 images of 200 subjects were collected by setting up 30 conditions of drastically changed pose and lighting conditions (see Fig. 6). The seven pose variations included a maximum of 95 degrees in total rotation in depth, and the illumination direction was changed by up to 90 degrees from the front. For the enrolled data, three-dimensional scans of 200 subjects were also collected by using a 3D scanner.

In the experiments, it is assumed that the face-detection process provides the rough estimate of the facial pose but that the precise facial area is not determined. To simulate the rough pose estimates, locations of 12 feature points (left and right corners of eyes, nose and mouth, pupils, centers of nose and mouth) were manually annotated to the test images and 3D shape data. The pose estimation algorithm calculates the facial pose using the point correspondences between the feature points on the test image and the 3D shape. Since the locations contain some errors, the estimated pose becomes rough. The pose estimate is used to initiate model-fitting.

The recognition performances of the conventional method and that combined with our proposed method are compared. The conventional method uses only the similarity scores between the reconstructed image and the query image, which is calculated as mean absolute differences of pixel intensities in foreground. [2,6] The model with the smallest error is chosen as the matched subject. Our proposed method additionally calculates the confidence of model fitting, as described in section 2, and the fused score is used for face identification. The parameter  $\gamma$  in equation (8) is tuned to 0.01 by a preliminary experiment using a data set different from the test data.

The performances of the face identification from the 200 enrolled subjects were compared in the experiments. The error rates of the previous method and our proposed method are shown in Table 1 and 2, respectively. Bold numbers indicates the severe conditions for the previous method where error rates are higher than the average. The error rate averaged over all conditions was reduced from 1.5% to 1.1% against 7,000 test images. The recognition errors induced by the conventional method were reduced more than 25% by using our proposed method.





**Fig. 6.** Pose and lighting conditions of test images used in face identification experiments. 35 test images were taken for each of 200 subjects to be identified. 3D scans of each subject were also collected and enrolled in matching database.

**Table 1.** Error rates of face identification by conventional method using only similarity (model-fitting errors)

Conventional method: similarity only					
	Light A	B	C	D	E
Pose 1	0.0%	0.0%	0.5%	<b>3.0%</b>	<b>2.5%</b>
2	0.0%	0.0%	0.5%	0.5%	3.0%
3	<b>2.0%</b>	<b>2.0%</b>	<b>3.0%</b>	<b>5.0%</b>	<b>5.0%</b>
4	0.0%	0.0%	0.5%	0.5%	<b>2.5%</b>
5	0.0%	0.0%	0.5%	0.5%	1.5%
6	0.5%	0.5%	0.5%	<b>2.0%</b>	<b>4.5%</b>
7	1.5%	0.0%	0.5%	0.5%	<b>10.0%</b>
Average					1.5%

**Table 2.** Error rates of face identification by our proposed method using confidence of model-fitting in addition to similarity

Proposed method: similarity + confidence					
	Light A	B	C	D	E
Pose 1	0.0%	0.0%	0.5%	<b>2.5%</b>	<b>3.0%</b>
2	0.0%	0.0%	0.5%	0.5%	2.5%
3	<b>1.0%</b>	<b>1.5%</b>	<b>1.5%</b>	<b>3.0%</b>	<b>4.5%</b>
4	0.0%	0.0%	0.5%	0.5%	<b>2.0%</b>
5	0.0%	0.0%	0.5%	1.0%	2.0%
6	0.5%	0.5%	0.5%	<b>2.0%</b>	<b>3.5%</b>
7	0.5%	0.0%	0.5%	0.5%	<b>4.0%</b>
Average					1.1%

Note that our method is designed to enhance the performance under the severe conditions. Performance was especially improved under severe conditions containing large occlusions and strong cast shadows, namely, light E and poses 3, 6 and 7.

## 4 Conclusions

In this paper, a new method for enhancing face-recognition performance by evaluating the confidence of model-fitting was proposed. Our proposed method extracts the discriminant features that are useful for face identification in a particular condition of a given query image, and the confidence of the similarity score is evaluated according to how much of the discriminant feature is captured in the foreground. The similarity and confidence scores are fused into a matching score that is used to identify the face. The experiments using a large set of the images taken under drastically varying pose and lighting conditions showed the efficacy of our proposed method. The errors induced by the previous method are reduced more than 25% by using our method.

In this study, the naive image feature, i.e. pixel intensity, was used and the discriminant features are detected as the weights for the pixels. This simply showed the efficacy of our basic idea, and our method can be applied to more sophisticated features extracted by various filters. Our future work is to extend our proposed method for such features, and to construct larger database to test the algorithms.

## References

1. Abate, F., Nappi, M., Riccio, D., Sabatino, G.: 2D and 3D face recognition: A survey. *Pattern Recognition Letters* 28(14), 1885–1906 (2007)
2. Basri, R., Jacobs, D.: Lambertian Reflectance and Linear Subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(2), 218–233 (2003)
3. Chen, H.F., Belhumeur, P.N., Jacobs, D.W.: In Search of Illumination Invariants. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, vol. 1, p. 1254 (2000)
4. Georghiadis, 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(6), 643–660 (2001)
5. Phillips, P.J., Grother, P., Michaels, R.J., Blackburn, D.M., Tabassi, E., Bone, M.: *Face Recognition Vendor Test 2002: Evaluation Report NISTIR 6965*. Nat'l Inst. of Standards and Technology (2003)
6. Ishiyama, R., Hamanaka, M., Sakamoto, S.: An Appearance Model Constructed on 3D Surface for Robust Face Recognition against Pose and Illumination Variations. *IEEE Trans. Systems, Man, and Cybernetics-Part C* 35(3), 326–334 (2005)
7. Ishiyama, R., Sakamoto, S.: Fast and Accurate Facial Pose Estimation by Aligning a 3D Appearance Model. In: *Proceedings of 17th International Conference on Pattern Recognition (ICPR 2004)*, vol. 4, pp. 388–391 (2004)
8. Romdhani, S., Ho, J., Vetter, T., Kriegman, D.J.: Face Recognition Using 3-D Models: Pose and Illumination. *Proceedings of the IEEE* 94(11), 1977–1999 (2006)
9. Zhang, L., Samaras, D.: Face Recognition from a Single Training Image under Arbitrary Unknown Lighting Using Spherical Harmonics. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(3), 351–363 (2006)
10. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.: Face recognition: A literature survey. *ACM Computing Surveys* 35(4), 399–458 (2003)
11. Zhao, W., Chellappa, R.: *Face Processing*. Academic Press, London (2006)

# A Brief History of the Subspace Methods

Hitoshi Sakano

NTT Communication Science Lab.

2-4, Hikaridai, Seika-cho, gKeihanna Science Cityh Kyoto 619-0237 Japan  
sakano.hitoshi@lab.ntt.co.jp

## Abstract

I hope to start from one question. “Is the eigenface[1] a subspace method?”

Answer is weakly YES and strongly NO. In wide meaning in Subspace method of pattern recognition is that uses subspace. In this meaning the answer is YES. However in narrow meaning the term “Subspace method” means pattern recognition techniques that represent class featuring information with subspace of original feature space[2]. The eigenface subspace represent common feature of trained faces, that is differ from class information. Thus in this meaning the answer is NO[1].

For understanding the term of “Subspace method”, we shall trace back to a Subspace method root. In this article I try to clarify the meaning of Subspace method through the historical study. To this goal we trace histories of Subspace methods from their birth at 1960s to 21c. We studied the history both side of theory and applications, because sometimes new theory is inspired by new application and new theory extend applicability of Subspace methods.

The history of Subspace method is classified in three epochs.

First epoch is the birth of Subspace methods, from '60th to '70th. Subspace method was originated by two Japanese researcher Prof. Taizo Iijima and Prof. Satoshi Watanabe independently. Prof. Iijima try to formulate an observation theory of object that include scale space methods[4]. Prof. Watanabe started from the information theory and the theory of probabilistic logics[5]. Interestingly they reached same goal from other start points. Their results are “categories or class information is represented by subspaces”.

Second epoch is the age of the application to character recognition and discriminative Subspace methods. Main issue of pattern recognition research in this age is character recognition[6]. Especially Japanese Kanji recognition problem was very important industrial problem in Japan. For obtaining high recognition accuracy, many discriminative Subspace methods were proposed[7]

Third epoch was starting from Yamaguch et. al [8]. They demonstrate the effectiveness of mutual Subspace method for object recognition problem. From their paper, Subspace method is defined important technology of object recognition problem, and many improvement and extension were proposing[9,10,11,12]. Many other applications were proposed[13] in this epoch.

---

<sup>1</sup> The technology of eigenface is rediscovery of SELFIC[3].

From this historical study, we try to discuss current status and future issue of Subspace method.

## References

1. Turk, M., Pentland, A.: Eigenface for recognition. *Journal of Cognitive Neuroscience* 3(1), 71–86 (1991)
2. Watanabe, S.: *Knowing and Guessing. A Quantative study of Inference and Information.* John Wiley and Sons, West Sussex (1969)
3. Watanabe, S.: Karhunen - Loeve expansion and factor analysis - Theoretical remarks and application. In: *Proc. 4th Prague Conf. Information Theory* (1965)
4. Iijima, T.: *Theory of pattern recognition, Morikita* (1989) (in Japanese)
5. Watanabe, S., Lambert, P.F., Kulikowski, C.A., Buxton, J.L., Walker, R.: gEvaluation and selection of variables in pattern recognition. In: Tou, J. (ed.), *Computer and Information Sciences.* vol. 2, pp. 91–122. Academic Press, New York (1967)
6. Mori, S., Suen, C.Y., Yamamoto, K.: *Historical Review of OCR Research and Development.* *Proceedings of the IEEE* 80(7), 1029–1058 (1992)
7. Oja, E.: *Subspace Methods of Pattern Recognition.* Research Studies Press (1983)
8. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequence. In: *Proc. of IEEE 4th Int'l. Conf. on Face and Gesture Recognition*, pp. 318–323 (1998)
9. Fukui, K., Stenger, B., Yamaguchi, O.: A framework for 3D object recognition using the kernel constrained mutual subspace method. In: *Proceedings of Asian Conference on Computer Vision*, pp. 315–324 (2006)
10. Fukui, K., Yamaguchi, O.: The kernel orthogonal mutual subspace method and its application to 3D object recognition. In: Yagi, Y., Kang, S.B., Kweon, I.S., Zha, H. (eds.) *ACCV 2007, Part II. LNCS*, vol. 4844, pp. 467–476. Springer, Heidelberg (2007)
11. Kim, T., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. PAMI* 29(6), 1005–1018 (2007)
12. Sakano, H., Mukawa, N.: Kernel mutual subspace method for robust facial image recognition. In: *Proc. of 4th Int'l. Conf. on Knowledge based Engineering System*, vol. 1, pp. 245–248, Brighton, Sakano, H., Mukawa, N., Nakamura, T. (eds.) *Kernel mutual subspace method and its application for object recognition. Electronics and Communications in Japan.* vol. E88(6), pp. 45–53 (Journal paper version) (2000)
13. Bagan, H., Takeuchi, W., Yamagata, Y., Wang, X., Yasuoka, Y.: gExtended Averaged Learning Subspace Method for Hyperspectral Data Classification. *Sensors* 9(6), 4247–4270 (2009)

# Author Index

- Achard, Catherine I-123  
Aizawa, Kiyoharu I-410  
Akahane, Katsuhito II-324  
Akihiro, Naoki II-374  
Al-Hamadi, Ayoub I-307, I-370  
Andaroodi, Elham II-286  
Aramvith, Supavadee I-359  
Ariki, Yasuo I-400  
Avgerinakis, Konstantinos I-54
- Benedek, Csaba I-74  
Benfold, Ben I-380  
Birchfield, Stan T. I-390  
Bouwmans, Thierry II-394  
Briassouli, Alexia I-54  
Bujnak, Martin II-184
- Cai, Yinghao I-205  
Calway, Andrew II-31  
Cameron, Jonathan I-297  
Carque, Bernd II-296  
Castleman, Shannon L. II-112  
Chen, Liu II-21  
Chen, Xilin I-450  
Chi, Chen I-450  
Chihara, Kunihiro II-42  
Chu, Xiu-Qin I-256  
Chung, Ronald I-23  
Constable, Martin II-142  
Cooper, David B. I-246
- Dadgostar, Farhad I-236  
Decker, Peter II-11  
Deguchi, Daisuke I-175, II-163, II-204  
De Leo, Carter I-94  
Di Stefano, Luigi I-1, I-43  
Doman, Keisuke II-204  
Duchowski, Andrew T. I-390
- Ellis, Liam I-338
- Feixas, Miquel II-122  
Feldmann, Tobias I-113  
Förstner, Wolfgang II-334  
Fukui, Kazuhiro II-374  
Futragoon, Natchapon II-286
- Gabard, Christophe I-123  
Gee, Andrew P. II-31  
Geiger, Andreas II-235  
Gevers, Theo I-349  
Gong, Lujin II-82  
Goras, Bogdan Tudor I-349  
Grabner, Helmut I-133  
Grindinger, Thomas J. I-390  
Guillot, C. I-33, I-123
- Haberjahn, Mathias II-225  
Hahn, Hern-soo I-246  
Han, Dongjin I-246  
Han, Zhenjun I-216  
Hanbury, Allan I-349  
Harle, Robert I-297  
Hasegawa, Tsutomu I-287  
Hermann, Simon II-245  
Hiraki, Kazuo I-420  
Holzer, Peter I-195  
Hong, Jen-Shin II-132  
Hoque, Mohammed Moshui I-430  
Hori, Maiya II-62  
Hotta, Seiji II-364  
Hu, Weiming I-184  
Huang, Xiangsheng II-82  
Hui, Tak-Wai I-23  
Hwang, Jae I-246
- Ichikari, Ryosuke II-1  
Ide, Ichiro I-175, II-163, II-204  
Ikeda, Sei II-42  
Ikeuchi, Katsushi II-306  
Imiya, Atsushi II-344  
Ishiyama, Rui II-424  
Islam, Md. Zahidul I-226  
Isshiki, Masaharu II-324  
Itoh, Hayato II-344
- Jiang, Ruyi II-214  
Jiao, Jianbin I-216  
Jie, Liu II-21
- Kakarala, Ramakrishna II-112  
Kanbara, Masayuki II-62

- Kaneko, Toru II-92  
 Kawahara, Tomokazu II-364  
 Kim, Kangsoo II-276  
 Kitamoto, Asanobu II-286  
 Kittipanya-ngam, Panachit I-143  
 Klette, Reinhard II-102, II-152, II-174,  
 II-214, II-245  
 Kobayashi, Yoshinori I-430  
 Kodama, Sachiko I-430  
 Kojima, Yoshiko II-163  
 Kompatsiaris, Ioannis I-54  
 Kono, Yuki I-175  
 Koutaki, Gou II-414  
 Kukulova, Zuzana II-184  
 Kuno, Yoshinori I-430  
 Kurazume, Ryo I-287  
 Kurita, Takio I-277  
 Kurz, Julian II-235
- Lalos, Constantinos I-133  
 Lanza, Alessandro I-1, I-43  
 Lanza, Piergiorgio II-255  
 Lasenby, Joan I-297  
 Lavest, J.M. I-33  
 Lee, Chil-Woo I-226  
 Leithy, Alaa I-153  
 Li, Hongyu II-384  
 Li, Li I-216  
 Li, Wei I-184  
 Li, Yanli II-52  
 Li, Yu-Shan I-256  
 Ling, Haibin I-184  
 Liu, Jen-Chang II-132  
 Lo, Pei-Yu II-132  
 Lovell, Brian C. I-236  
 Lu, Max I-164  
 Lu, Min II-306  
 Lucat, Laurent I-123  
 Lung, Eng How I-143  
 Luo, Wenhan I-184
- Manabe, Yoshitsugu II-42  
 Manjunath, B.S. I-94  
 Marcer, Alessandra II-255  
 Marghes, Cristina II-394  
 Matini, Mohammad Reza II-286  
 Matsukawa, Tetsu I-277  
 Matsuyama, Takashi II-1  
 Mayol-Cuevas, Walterio II-31  
 Miao, Jun I-450
- Michaelis, Bernd I-307, I-370  
 Mikami, Toshiaki I-440  
 Miyaki, Rie II-92  
 Monroy, Antonio II-296  
 Morales, Sandino II-152, II-245  
 Mori, Taketoshi I-318  
 Moustafa, Mohamed N. I-153  
 Murali, Vidya N. I-390  
 Murase, Hiroshi I-175, II-163, II-204
- Naito, Takashi II-163  
 Nakaguchi, Toshiya I-440  
 Nishino, Ko II-306  
 Nobuhara, Shohei II-1  
 Noceti, Nicoletta I-84  
 Noda, Masafumi II-163  
 Noguchi, Yoshihiro I-277  
 Nomura, Kazuyoshi II-316  
 Nordberg, Klas I-338
- Odashima, Shigeyuki I-318  
 Odone, Francesca I-84, II-255  
 Ogawara, Koichi I-287  
 Oh, Chi-Min I-226  
 Okabe, Takahiro I-420  
 Ommer, Björn II-296  
 Ono, Kinji II-286  
 Onuki, Tomami I-430  
 Orero, Pilar I-390  
 Ouda, Osama I-440  
 Ozturk, Ovgu I-410
- Pajdla, Tomas II-184  
 Pan, Guodong II-194  
 Park, Jong-II II-276  
 Pathan, Saira Saleem I-370  
 Patron, Alonso I-380  
 Paulus, Dietrich II-11  
 Pham, Q.C. I-33  
 Pietikäinen, Matti I-205  
 Pinz, Axel I-195  
 Pöntiz, Thomas I-349
- Qing, Laiyun I-450  
 Qiu, Jia-Tao I-256
- Rajan, Deepu II-112  
 Rashid, Omer I-307  
 Reid, Ian I-380  
 Reulke, Ralf II-225

- Rigau, Jaume II-122  
 Roscher, Ribana II-334  
 Roser, Martin II-235  
 Roy-Chowdhury, Amit K. I-328  
  
 Sachs, Todd S. II-112  
 Sakai, Tomoya II-344  
 Sakano, Hitoshi II-364, II-434  
 Salti, Samuele I-43  
 Sato, Takayuki I-430  
 Sato, Tomokazu II-1, II-265  
 Sato, Tomomasa I-318  
 Sato, Yoichi I-420  
 Sayd, Patrick I-33, I-123  
 Sbert, Mateu II-122  
 Schauwecker, Konstantin II-174  
 Schindler, Falko II-334  
 Sebe, Nicu I-349  
 Senior, Andrew W. I-164  
 Seo, Byung-Kuk II-276  
 Sethi, Ricky J. I-328  
 Shih, Sheng-Wen II-132  
 Shimad, Atsushi I-12  
 Shimada, Keiji I-277  
 Shirazi, Sareh Abolahrari I-236  
 Siddhichai, Supakorn I-359  
 Soatto, Stefano I-266  
 Sommerlade, Eric I-380  
 Song, Li I-64  
 Stöttinger, Julian I-349  
 Sugano, Yusuke I-420  
 Sugimoto, Akihiro I-420  
  
 Takahashi, Hideyuki II-62  
 Takahashi, Tomokazu I-175, II-163  
 Takamatsu, Jun II-306  
 Takatani, Manabu I-400  
 Taketomi, Takafumi II-265  
 Takiguchi, Tetsuya I-400  
 Tamura, Hideyuki II-1  
 Tanabe, Yasufumi I-287  
 Tanaka, Hiromi T. II-316, II-324  
 Tanaka, Masayuki II-354  
 Taniguchi, Rin-ichiro I-12  
 Taron, M. I-33  
 Tetreault, Stephen I-390  
 Tian, YingLi I-164  
 Tilmant, C. I-33  
 Tombari, Federico I-1  
 Torii, Akihiko II-184  
  
 Tsuburaya, Emi I-430  
 Tsumura, Norimichi I-440  
  
 Uchimura, Keiichi II-414  
 Utasi, Ákos I-74  
  
 Valente, Crystal II-102  
 Van Gool, Luc I-133  
 Varvarigou, Theodora I-133  
 Vaudrey, Tobi II-245  
 Verri, Alessandro II-255  
  
 Wahba, Ayman I-153  
 Wakita, Wataru II-316, II-324  
 Wang, Jia I-64  
 Wang, Junqiu II-72  
 Wang, Shigang II-214  
 Washizawa, Yoshikazu II-354, II-404  
 Watanabe, Toshinori I-104  
 Watcharapinchai, Nattachai I-359  
 Wijaya, I. Gede Pasek Suta II-414  
 Wnuk, Kamil I-266  
 Wong, Kwan-Yee Kenneth II-194  
 Wu, Wei II-52  
  
 Xiaohui, Liang II-21  
 Xie, Nianhua I-184  
 Xu, Yi I-64  
 Xu, Zhenchao I-64  
  
 Yagi, Yasushi II-72  
 Yamada, Kentaro I-420  
 Yamaguchi, Osamu II-364  
 Yamasaki, Toshihiko I-410  
 Yamashita, Atsushi II-92  
 Yamauchi, Takuya I-440  
 Yarlagadda, Pradeep II-296  
 Yasukawa, Nobuyuki II-424  
 Ye, Qixiang I-216  
 Yiming, Yue II-21  
 Yokoya, Naokazu II-1, II-62, II-265  
 Yokoyama, Takanori I-104  
 Yoshinaga, Satoshi I-12  
  
 Zhang, Lin II-384  
 Zhang, Xiaoqin I-184  
 Zheng, Bo II-306  
 Zhou, Zhong II-52  
 Zini, Luca II-255  
 Zografos, Vasileios I-338