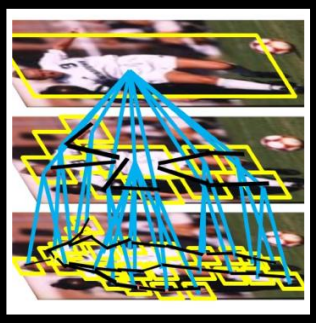
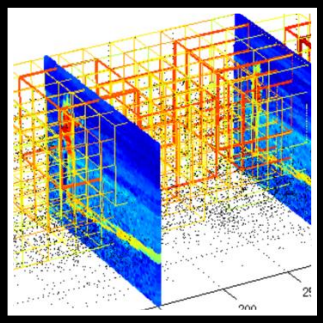
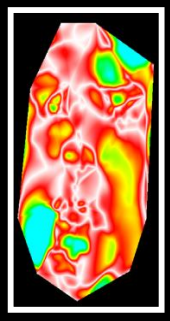
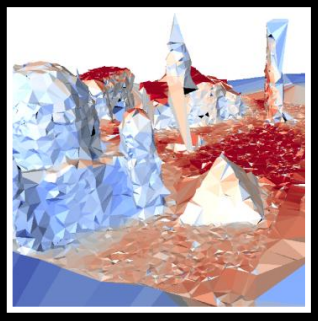
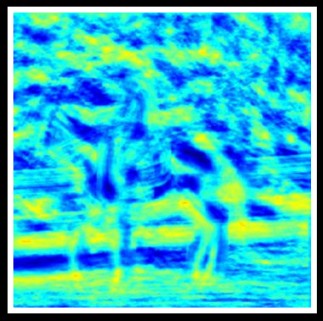




British Machine Vision Conference Programme & Abstracts



Edited by
Richard Bowden
John Collomosse
Krystian Mikolajczyk

BMVC 2012

University of Surrey

Programme overview

Monday 3rd September

13:30 –15:30	**Tutorial** Large-scale and larger-scale image search - Hervé Jégou, INRIA Rennes, France
15:30 –16:00	**Tea Break**
16:00 –18:00	**Tutorial** MAP inference in Discrete Models - Pushmeet Kohli, Microsoft Research, UK

Tuesday 4th September

08:30 –09:00	**Morning Coffee**
09:00 –09:10	**Welcome**
09:10 –10:10	**Keynote I** People in Motion - Stan Sclaroff, Boston University, MA, USA
10:10 –10:50	**Session 1: Tracking**
10:50 –11:20	**Coffee Break**
11:20 –13:00	**Session 2: People and Pose**
13:00 –14:00	**Lunch**
14:00 –15:00	**Poster Session 1**
15:00 –15:30	**Tea Break**
15:30 –16:50	**Session 3: Segmentation**
16:50 –17:30	**AGM**
19:15 –23:00	**Reception - Guildford Cathedral**

Wednesday 5th September

08:30 –09:00	**Morning Coffee**
09:00 –10:00	**Keynote II** Visual Tracking in the 21st Century - Jiří Matas, Czech Technical University, Prague
10:00 –10:40	**Session 4: Retrieval**
10:40 –11:10	**Coffee Break**
11:10 –12:50	**Session 5: Object Recognition**
12:50 –14:00	**Lunch**
14:00 –15:00	**Poster Session 2, Demos, Exhibits**
15:00 –15:30	**Tea Break**
15:30 –16:50	**Session 6: Action Recognition**
18:00 –23:30	**Banquet - Brooklands Museum**

Thursday 6th September

08:30 –09:00	**Morning Coffee**
09:00 –10:40	**Session 7: Recognition and Faces**
10:40 –11:10	**Coffee Break**
11:10 –13:00	**Session 8: Registration and Image Processing**
13:00 –13:10	**Closing Remarks**

Friday 7th September

08:30 –09:00	**Morning Coffee**
09:00 –10:00	**Workshop Keynote I** Latent Variable Models for Content-Based Image Retrieval and Structure Prediction - Ariadna Quattoni, Universitat Politecnica de Catalunya
10:00 –10:40	**Session W1**
10:40 –11:20	**Coffee Break and Posters**
11:20 –12:40	**Session W2**
12:40 –14:00	**Lunch and Posters**
14:00 –15:00	**Workshop Keynote II** Monocular SLAM and Real-Time Scene Perception - Andrew Davison, Imperial College London
15:00 –15:40	**Session W3**

Message from the chairs

Welcome to BMVC 2012, it is our pleasure to host the conference at the University of Surrey in Guildford. This is the 23rd BMVC since its inception in 1990 and the second time Surrey has organised the conference. Almost 20 years have passed since BMVC in 1993 when it was chaired by John Illingworth with the support of the Centre for Vision Speech and Signal Processing (CVSSP) at the University of Surrey. We decided it was time for BMVC to return and our thanks go to our colleagues within CVSSP and to John for his contributions to the organisation. In fact BMVC goes back even further in the form of the Alvey Vision Club and our conference T-shirt gives a list of dates and venues for BMVC going back to the early Alvey days of the 80's.

BMVC has always maintained a single track and the format has changed little over the years. As the conference has grown in popularity its quality has increased, making it a prestigious event on the vision calendar. This year we received well over 400 submissions and accepted 8% of these for oral presentation and a further 24% for posters.

For 2012, we introduced changes to the reviewing process. We increased the reviewer pool to keep the loading low, despite increased submissions, and to ensure that reviewers could invest more time in their task. At least one reviewer per paper was nominated by an area chair. We also doubled the number of area chairs so that two area chairs were allocated to each paper. Unlike other conferences which operate a buddy system, our area chairs worked anonymously and independently and every paper had 2 recommendations justified in 2 consolidation reports. ACs who gave conflicting acceptance recommendations were invited into discussion to reach a consensus. As each AC had a different batch of papers, we were looking for consistency across both reviewers and area chairs to provide a robust decision making process. Oral papers were selected based on the reviews, AC consolidation reports, ranking in ACs' batches and suitability of the content for the general audience. Feedback indicates that this process has worked well and reviewers and area chairs from all over the world have worked hard to put together the highest quality program possible.

BMVC 2012 is a truly international conference. 67.5% of accepted papers are from Europe including 29% from the UK. There is also a significant 19.5% from North as well as from South America and 13% from Asia and Australia. The quality of research in these diverse locations is exceptionally high and it is also reflected in the selection of papers for oral presentations.

Sticking with the traditional single track format, BMVC starts on Monday afternoon with two tutorials; "Large-scale and larger-scale image search" by Dr Hervé Jégou, INRIA and "MAP inference in Discrete Models" by Dr Pushmeet Kohli from Microsoft Research. We are also pleased to be able to welcome two invited speakers for the main conference; Prof Stan Sclaroff, Boston University (People in Motion: Pose, Action and Communication) and Prof. Jiří Matas, Czech Technical University (Visual Tracking in the 21st Century). On Tuesday evening there is a welcome reception hosted at Guildford Cathedral. While on Wednesday the conference banquet will be held at Brooklands Museum, the home of British Motorsport, and delegates will have exclusive access to the museum over drinks before dinner and awards in the clubhouse.

On Friday, the BMVC Student Workshop will take place. This workshop has become a regular feature of BMVC and gives students in computer vision an opportunity to network and start collaborations at an early stage in their research career. The workshop will be single track containing both oral and poster presentations with keynote talks from Ariadna Quattoni UPC, Barcelona on "Latent Variable Models for Content-Based Image Retrieval and Structure Prediction" and Andrew Davison, Imperial College London on "Monocular SLAM and Real-Time Scene Perception."

As ever, BMVC proceedings are published by the BMVA as open access with the copyright retained by the authors. Extended abstracts of all the accepted contributions appear in this book. The full papers are published electronically on a USB key provided to delegates, and on the web with associated DOI's.

Furthermore, the whole proceedings will be recorded and hosted online by videolectures.net and we are extremely grateful to the EU PASCAL2 network and video lectures for providing this service along with sponsorship of both the main conference and student workshop. We would also like to acknowledge the financial assistance of Microsoft Research, Stemmer Imaging and Toshiba all of which generously contributed to the conference as Gold sponsors. To 3DMD and Scorpion Vision as Silver sponsors and to Springer UK. We would like to thank the local organisation committee: Teofilo de Campos, John Illingworth and Fei Yan and especially Helen Cooper for taking great pleasure in doing all the organising we didn't want to do. We would also like to thank the student helpers and members of CVSSP who contributed, invited speakers and tutorial speakers and of course a huge debt to all the reviewers and area chairs for all the work they put in. Lastly many thanks to the authors for submitting the material to make this the best BMVC yet.

We sincerely hope that you all have both a rewarding and enjoyable conference.

Rich Bowden, John Collomosse, Krystian Mikolajczyk

BMVC2012 General chairs

Programme: Monday 3rd September

Tutorial I

13:30 Large-scale and larger-scale image search 1
Hervé Jégou

15:30 Tea Break

Tutorial II

16:00 MAP inference in Discrete Models 2
Pushmeet Kohli

Programme: Tuesday 4th September

08:30 Morning Coffee

09:00 Welcome

Keynote I

09:10 People in Motion: Pose, Action and Communication 3
Stan Sclaroff

Tracking

10:10 Automatic and Efficient Long Term Arm and Hand Tracking for Continuous Sign Language TV Broadcasts 4
Tomas Pfister, James Charles, Mark Everingham, Andrew Zisserman

10:30 Deformable Tracking of Textured Curvilinear Objects 5
Nicolas Padoy, Gregory Hager

10:50 Coffee Break

People and Pose

11:20 Using Richer Models for Articulated Pose Estimation of Footballers 6
Vahid Kazemi, Josephine Sullivan

11:40 Dynamical Pose Filtering for Mixtures of Gaussian Processes 7
Martin Fergie, Aphrodite Galata

12:00 Close-Range Human Detection and Tracking for Head-Mounted Cameras 8
Dennis Mitzel, Bastian Leibe

12:20 Detection and Tracking of Occluded People 9
Siyu Tang, Mykhaylo Andriluka, Bernt Schiele

12:40 Latent SVMs for Human Detection with a Locally Affine Deformation Field 10
Lúbor Ladický, Philip Torr, Andrew Zisserman

13:00 Lunch

Poster Session 1

14:00 Sparsity Potentials for Detecting Objects with the Hough Transform 11
Nima Razavi, Nima Sedaghat Alvar, Juergen Gall, Luc Van Gool

Gradient Edge Map Features for Frontal Face Recognition under Extreme Illumination Changes 12
Ognjen Arandjelović

Exemplar Driven Character Recognition in the Wild 13
Karthik Sheshadri, Santosh Divvala

Efficient Learning-based Image Enhancement: Application to Super-resolution & Compression Artifact Removal 14
Younghee Kwon, Kwang In Kim, Jin Kim, Christian Theobalt

Contour-HOG: A Stub Feature based Level Set Method for Learning Object Contour 15
Zhi Yang, Yu Kong, Yun Fu

Virtual Line Descriptor and Semi-Local Graph Matching Method for Reliable Feature Correspondence 16
Zhe Liu, Renaud Marlet

Learning Edge-Specific Kernel Functions For Pairwise Graph Matching 17
Michael Donoser, Martin Urschler, Horst Bischof

Genetic Programming-Evolved Spatio-Temporal Descriptor for Human Action Recognition 18
Li Liu, Ling Shao, Peter Rockett

Racing Bib Numbers Recognition 19
Idan Ben-Ami, Tali Basha, Shai Avidan

Learning Discriminative Chamfer Regularization 20
Pradeep Yarlagadda, Angela Eigenstetter, Björn Ommer

Feature Mining for Localised Crowd Counting 21
Ke Chen, Chen Change Loy, Shaogang Gong, Tony Xiang

Super-Resolution from Corneal Images 22
Christian Nitschke, Atsushi Nakazawa

Real-time Learning and Detection of 3D Texture-less Objects: A Scalable Approach 23
Dima Damen, Pished Bunnun, Andrew Calway, Walterio Mayol-Cuevas

15:00 Tea Break

Segmentation

Programme: Wednesday 5th September

08:30 Morning Coffee

Keynote II

Retrieval

10:40 Coffee Break

Object Recognition

12:50 Lunch

Poster Session 2

15:00 Tea Break

Action Recognition

Programme: Thursday 6th September

08:30 Morning Coffee

Recognition and Face

09:00 Scene Text Recognition using Higher Order Language Priors 127
Anand Mishra, Karteek Alahari, Cv Jawahar

09:20 Data-Driven Scene Understanding from 3D Models 128
Scott Satkin, Jason Lin, Martial Hebert

09:40 Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification 129
Thomas Berg, Peter Belhumeur

10:00 Let the Shape Speak - Discriminative Face Alignment using Conjugate Priors 130
Pedro Martins, Rui Caseiro, João Henriques, Jorge Batista

10:20 Dense Active Appearance Models Using a Bounded Diameter Minimum Spanning Tree 131
Robert Anderson, Bjorn Stenger, Roberto Cipolla

10:40 Coffee Break

Registration and Image Processing

11:10 PMBP: PatchMatch Belief Propagation for Correspondence Field Estimation 132
Frederic Besse, Carsten Rother, Andrew Fitzgibbon, Jan Kautz

11:30 Deformable 3D Reconstruction with an Object Database 133
Pablo Alcantarilla, Adrien Bartoli

11:50 Incremental Light Bundle Adjustment 134
Vadim Indelman, Richard Roberts, Chris Beall, Frank Dellaert

12:10 Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding 135
Marco Bevilacqua, Aline Roumy, Christine Guillemot, Marie-Line Alberi Morel

12:30 Fast Non-uniform Deblurring using Constrained Camera Pose Subspace 136
Zhe Hu, Ming-Hsuan Yang

12:50 Closing Remarks

Programme: Friday 7th September: Student Vision Workshop

08:30 Morning Coffee

Workshop Keynote I

09:00 Latent Variable Models for Content-Based Image Retrieval and Structure Prediction 137
Ariadna Quattoni

Papers Session W1

10:00 Invited : Person Re-identification by Attributes
Ryan Layne, Tim Hospedales, Shaogang Gong

10:20 Invited : Comparing Visual Feature Coding for Learning Disjoint Camera Dependencies
Xiatian Zhu, Shaogang Gong, Chen Change Loy

10:40 Coffee Break and Posters

Papers Session W2

11:20 Invited : Binary Pattern Analysis for 3D Facial Action Unit Detection
Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic

11:40 GEI + HOG for Action Recognition
Tenika Whytock, Alexander Belyaev, Neil Robertson

12:00 Real Time Single and Multiuser Gesture Recognition Based on Skin Colour and Optical Flow
Muhammad Raza Ali, Tim Morris

12:20 MCMC-PF Based Multiple Head Tracking in a Room Environment
Ata Ur-Rehman , Syed Mohsin Naqvi , Raphael Phan , Wenwu Wang , Jonathon Chambers

12:40 Lunch and Posters

Workshop Keynote II

14:00 Monocular SLAM and Real-Time Scene Perception 138
Andrew Davison

Papers Session W3

15:00 Invited : Fast Pedestrian Detection by Cascaded Random Forest with Dominant Orientation Templates
Danhang Tang, Yang Liu, Tae-Kyun Kim

15:20 Simultaneous Human Segmentation, Depth and Pose Estimation via Dual Decomposition
Glenn Sheasby, Jonathan Warrell, Yuhang Zhang, Nigel Crook, Philip Torr

15:40 End of conference

Tutorial

Large-scale and larger-scale image search.

Hervé Jégou

INRIA Rennes, France

The first part of this tutorial, dedicated to large-scale image retrieval, will first introduce the typical use-cases and the datasets used for evaluation of image search when considering an unsupervised framework. We will present different classes of techniques considering different trade-offs with respect to efficiency and search quality. Starting with the most costly but precise patch-based matching and spatial verification techniques, we will present the bag-of-words model, its matching interpretation and several improvements, including re-ranking techniques based on spatial verification and query expansion. Finally, the most scalable techniques based on aggregation/coding techniques and compressed-domain search will be detailed.

Hervé Jégou is a researcher employed by INRIA, in the TEXMEX team headed by Patrick Gros. He is a former student of the Ecole Normale Superieure de Cachan, holding a M.S. (2002) and PhD (2005) from University of Rennes I. During his PhD, he worked on error-resilient compression and joint source channel coding, supervised by Christine Guillemot. After that, he turned out to Computer Vision and Pattern Recognition. He joined the LEAR group (INRIA Grenoble) as a permanent researcher in 2006, and moved to INRIA Rennes in 2009. His work is mainly focused on large scale image/video/audio retrieval, and multi-dimensional indexing techniques. He has designed methods that scale from millions to billions of vectors/images while being ressource efficient (one machine, relatively low memory usage)

Tutorial

MAP inference in Discrete Models.

Pushmeet Kohli

Microsoft Research, UK

Many problems in Computer Vision are formulated in form of a random filed of discrete variables. Examples range from low-level vision such as image segmentation, optical flow and stereo reconstruction, to high-level vision such as object recognition. The goal is typically to infer the most probable values of the random variables, known as Maximum a Posteriori (MAP) estimation. This has been widely studied in several areas of Computer Science (e.g. Computer Vision, Machine Learning, Theory), and the resulting algorithms have greatly helped in obtaining accurate and reliable solutions to many problems. These algorithms are extremely efficient and can find the globally (or strong locally) optimal solutions for an important class of models in polynomial time. Hence, they have led to a significant increase in the use of random field models in computer vision and information engineering in general. This tutorial is aimed at researchers who wish to use and understand these algorithms for solving new problems in computer vision and information engineering. No prior knowledge of probabilistic models or discrete optimization will be assumed. The tutorial will answer the following questions: (a) How to formalize and solve some known vision problems using MAP inference of a random field? (b) What are the different genres of MAP inference algorithms? (c) How do they work? (d) What are the recent developments and open questions in this field?

Pushmeet Kohli is a research scientist in the Machine Learning and Perception group at Microsoft Research Cambridge, an associate of the Psychometric Centre and Trinity Hall, University of Cambridge. Pushmeet was awarded his PhD from Oxford Brookes in 2007 and was the first of Phil Torr's students to graduate from that group.

Pushmeet's research revolves around Intelligent Systems and Computational Sciences, and he publishes in the fields of Machine Learning, Computer Vision, Information Retrieval, and Game Theory. His current research interests include 'human behaviour analysis' and the 'prediction of user preferences'. Pushmeet is interested in designing autonomous and intelligent computer vision, bargaining and trading systems which learn by observing and interacting with users on social media sites such as Facebook. He is also investigating the use of new sensors such as KINECT for the problems of pose estimation, scene understanding and robotics.

Pushmeet has won a number of awards and prizes for his research. His PhD thesis, titled "Minimizing Dynamic and Higher Order Energy Functions using Graph Cuts", was the winner of the British Machine Vision Association's 'Sullivan Doctoral Thesis Award', and was a runner-up for the British Computer Society's 'Distinguished Dissertation Award'. Pushmeet's papers have appeared in Computer Vision (ICCV, CVPR, ECCV, PAMI, IJCV, CVIU, BMVC, DAGM), Machine Learning, Robotics and AI (NIPS, ICML, AISTATS, AAAI, AAMAS, UAI, ISMAR), Computer Graphics (SIGGRAPH, Eurographics), and HCI (CHI, UIST) conferences. They have won best paper awards in ICVGIP 2006, 2010, ECCV 2010 and ISMAR 2011. His research has also been the subject of a number of articles in popular media outlets such as Forbes, The Economic Times, New Scientist and MIT Technology Review. Pushmeet is a part of the Association for Computing Machinery's (ACM) Distinguished Speaker Program.

Keynote

People in Motion: Pose, Action, and Communication

Stan Sclaroff

Boston University

This talk will give an overview of some of the research in the Image and Video Computing Group at Boston University related to tracking, analysis, recognition and retrieval of images and video based on humans and their actions.

First, efficient methods for inference of human pose will be presented, where a tree-based articulated pose model incorporates higher-order constraints. In one approach, scale and rotation invariant matching is made tractable through the use of linearly augmented trees; this enables efficient optimization over continuous scale and rotation parameters. Our experiments on ground truth data and a variety of real images and videos show that the proposed method is ef?cient, accurate and reliable. In another approach, articulated pose estimation with loopy graph models is made efficient via a branch-and-bound strategy for finding the globally optimal pose; the algorithm converges rapidly in practice due to a novel method for quickly computing tree-based lower bounds.

Second, methods for learning human action models from Web images and video will be presented. The methods are unsupervised in the sense that they require no human intervention other than the action keywords to be used to form text queries to Web image and video search engines. Thus, it is easy extend the vocabulary of actions, by simply making additional search engine queries. Experiments show the benefits of this approach in two areas: improving the retrieval precision of human action images, and tagging human actions in YouTube videos. A Multiple Instance Learning framework for exploiting properties of the scene, objects, and humans in video is also proposed for action classification in video.

Third, work towards automatic recognition and retrieval of American Sign Language in video databases will be presented. The effort involves collaboration between computer scientists and linguists. The goal is to enable users to search ASL video content simply by video-recording a query sign and relying on computer-based sign-recognition for lookup. As part of this effort, methods for gesture-based retrieval of signs that can exploit phonological constraints, e.g., on start/end hand shapes in lexical signs, are being developed. The American Sign Language Lexicon Video Dataset (ASLLVD), representing more than 3,300 distinct signs, each produced by 1-6 native ASL signers, for a total of almost 9,800 tokens, is an ASL video corpus that has been gathered and linguistically annotated as part of this effort, and will soon be made available as a resource to the research community.

Collaborators in this work include (in alphabetical order): Vassilis Athitsos, Nazli Ikizler-Cinbis, Hao Jiang, He Kun, Shugao Ma, Carol Neidle, Joan Poole-Nash, Ashwin Thangali, Tai-peng Tian, and others.

Stan Sclaroff founded the Image and Video Computing research group at the University of Boston. He received the PhD degree from MIT in 1995. In 1996, he received an ONR Young Investigator Award and an NSF Faculty Early Career Development Award. Professor Sclaroff has coauthored numerous scholarly publications in the areas of tracking, video-based analysis of human motion and gesture, surveillance, deformable shape matching and recognition, as well as image/video database indexing, retrieval and data mining methods. He has served on the technical program committees of over 60 computer vision conferences and workshops. Stan Sclaroff has served as an Associate Editor for IEEE Transactions on Pattern Analysis, 2000-2004, and 2006-present. He is a Senior Member of the IEEE.

Automatic and Efficient Long Term Arm and Hand Tracking for Continuous Sign Language TV Broadcasts

Tomas Pfister[1]
tp@robots.ox.ac.uk
James Charles[2]
j.charles@leeds.ac.uk
Mark Everingham[2]
m.everingham@leeds.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] School of Computing
University of Leeds
Leeds, UK

We present a fully automatic arm and hand tracker that detects joint positions over continuous sign language video sequences of more than an hour in length.

The standard approach of Buehler *et al*. [1] for tracking arms and hands requires manual labelling of 64 frames per video, which is around three hours of manual user input per one hour of TV footage. In addition, the tracker (by detection) is based on expensive computational models and requires hundreds of seconds computation time per frame. These two factors have hindered the large scale application of this method. In this paper we describe a method for tracking joint positions (of arms and hands) without any manual annotation and, after automatic initialisation, the system runs in real-time.

Our contributions are (i) a co-segmentation algorithm that automatically separates the signer from any signed TV broadcast using a generative layered model; (ii) a method of predicting joint positions given only the segmentation and a colour model using a random forest regressor; and (iii) demonstrating that the random forest can be trained from an existing semi-automatic, but computationally expensive, tracker. Figure 1 illustrates the processing steps.

Random forests for pose estimation. In recent years there has been increasing interest in random forest/fern-based methods. In particular we are interested in the work on human pose estimation, where random forests have most recently been used to infer full body pose [2]. However, the success of these pose methods depends upon the use of depth imagery which is colour and texture invariant, while also making background subtraction much easier. Here we propose an upper body pose estimation method that exploits the efficiency and accuracy of random forests without the need for depth images, and instead use raw RGB images with only a partially known background (as described below). See Figures 4 and 5 for illustrations of this method and a comparison against ground truth.

Co-segmentation for signer extraction. Co-segmentation methods consider sets of images where the appearance of foreground and/or background share some similarities, and exploit these similarities to obtain accurate foreground-background segmentations. In our case we exploit the fact that sign language broadcasts consist of a layered model of the foreground and two separate backgrounds, one that is static throughout each video and another that changes with each frame. The signer stands partially against the static background and partially against the changing background (which they are describing).

To this end we propose a co-segmentation algorithm that automatically separates signers from any signed TV broadcast by building a generative layered model as shown in Figures 2 and 3. We use this layered model of the signer in conjunction with a foreground colour model to provide a suitable input representation for the random forest regressor, superior to using the raw input image itself, and not requiring depth data.

The method is applied to signing footage with changing background, challenging imaging conditions, and for different signers. We achieve superior joint localisation results to those obtained using the method of Buehler *et al*. [1].

[1] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Upper body detection and tracking in extended signing sequences. *IJCV*, 95(2):180–197, 2011.

[2] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. CVPR*, 2011.

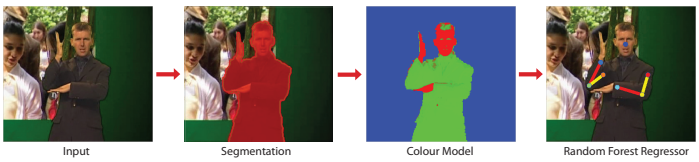


Figure 1: Arm and hand joint positions are predicted by first segmenting the signer using a layered foreground/background model, and then feeding the segmentation together with a colour model into a random forest.

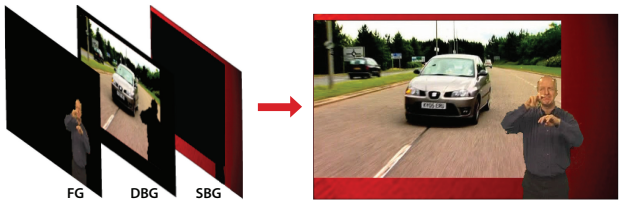


Figure 2: Generative layered model of each frame. The co-segmentation algorithm separates the signer from any signed TV broadcast by building a layered model consisting of a foreground, dynamic and static background.

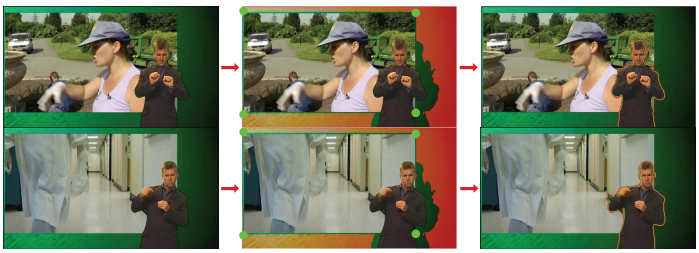


Figure 3: Co-segmentation. (a) the original frames; (b) the changing background (rectangle spanned by the green dots) and the permanently fixed background (in red); (c) the final segmentation.

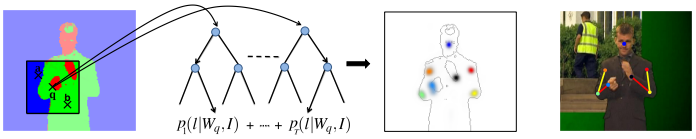


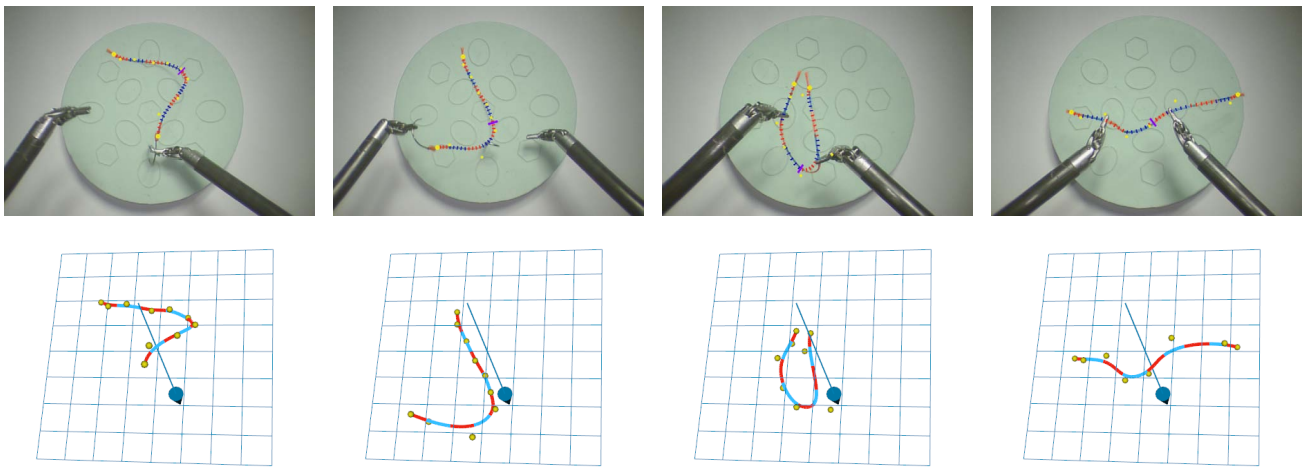
Figure 4: Estimating joint positions. (a) input colour model image; (b) random forest classifies each pixel using a sliding window; (c) probability density function of each joint location, shown in different colours per joint (more intense colour implies higher probability); (d) joint estimates, shown as small circles linked by a skeleton.

Figure 5: Joint estimation results. (a) shows a colour posterior from which we obtain probability densities of joint locations in (b) (black crosses mark maximum probability); (c) compares estimated joints (filled circles) with ground truth (open circles).

Deformable Tracking of Textured Curvilinear Objects

Nicolas Padoy
padoy@jhu.edu
Gregory Hager
hager@jhu.edu

Johns Hopkins University
Baltimore, Maryland, USA



Thread tracking illustration on two sequences: original images overlaid with spline models (top); virtual views (bottom).

Threads and wires are deformable 3-dimensional (3D) and curvilinear objects which are commonly manipulated by humans in various medical and manufacturing tasks. Several applications, including computer-assisted evaluation, augmented reality guidance, and autonomous robotic manipulation [2, 3] would benefit from the real-time estimation of the 3D shapes of these deformable objects from images. This estimation is however challenging due to multiple factors: 1) little information is available within an image to visually detect and distinguish a curvilinear object due to its thin and usually uniform appearance; 2) different 3D shapes may lead to the same visual perception, even in a stereo setting in case portions of the objects lie in an epipolar plane; and 3) the motions and deformations can be large, depending on the stiffness of the object. Additionally, a tracking approach that can consistently track specific points along the object defined by their arclength, such as the extremities or midpoint, would be particularly useful in the aforementioned applications.

To deal with visual ambiguities such as drift along the curve, we propose to texture the object with a coarse pattern of alternating colors and formulate the shape estimation as a deformable 1D template tracking problem. Tracking is expressed as an energy minimization over a set of control points $\mathcal{Q}$ parameterizing a 3D NURBS $\mathcal{C}^{3D}$ modeling the object:

$$\mathcal{C}^{3D}(\mathcal{Q},u) = \sum_{i=1}^{n} R_{i,d}(u)\mathcal{Q}_n, \quad u \in [0,1],$$

where $R_{i,d}$ are the rationale spline basis functions. We make use of the projective invariance properties of NURBS and, in a stereo setup, denote by $\mathcal{C}_1^{2D}$ and $\mathcal{C}_2^{2D}$ the 2D projected splines defined from the projections of $\mathcal{Q}$. The color pattern texturing the thread is represented by a general function associating the curve parameter u to its color $c(u)$:

$$c(u) : u \in [0,1] \longrightarrow S,$$

where S is a color space. For generality, we do not require the two cameras to possess the same color-calibrations, but maintain instead two representations of the texture by using two functions: c_i with $i \in \{1,2\}$ learned from the first images.

Assuming the object's in-extensibility, we propose a novel energy $E = E_{ext} + E_{int}$ based on a texture-sensitive distance map. The first term is a data term enforcing the curvilinear and texture appearances:

$$E_{ext} = \frac{1}{2(K+1)} \sum_{i=1}^{2} \sum_{k=0}^{K} \left(\alpha \mathcal{D}_i^{tub}(\mathcal{C}_i^{2D}(u_k))^2 + \beta \mathcal{D}_i^{tex}(\mathcal{C}_i^{2D}(u_k), u_k)^2 \right),$$

where $\mathcal{D}_i^{tub}(x)$ is a distance map indicating the distance to the closest ridge from position x in image I_i and $\mathcal{D}_i^{tex}(x,u)$ is a texture sensitive distance map indicating the distance to the closest pixel with color $c_i(u)$. $\mathcal{D}^{tex}$ is proposed instead of direct SSD evaluation to increase the convergence radius since the object has a thin structure. $(u_1,\ldots,u_K)$ defines a uniform sampling of the parameters and α,β are weights balancing the two terms. The second term enforces the in-extensibilty constraint for a thread with length L_{ref} and maintains an arc-length parameterization:

$$E_{int} = \frac{\gamma}{K} \sum_{k=0}^{K-1} \left(1 - \frac{\int_{u_k}^{u_{k+1}} \|\mathcal{C}^{3D'}(u)\| \mathrm{d}u}{L_{loc}} \right)^2,$$

where $L_{loc} = L_{ref}/K$ and γ is a weighting coefficient. This term also maintains consistency between the parameterizations of the texture c and of the spline $\mathcal{C}^{3D}$, thereby avoiding the re-computation of the arclength parameterization and of the corresponding spline basis functions at each time-step. Optimization is performed using Levenberg-Marquardt.

We demonstrate the benefits of this energy in synthetic and real experiments, using data illustrating the deformation and manipulation of a thread with a da Vinci robot. Usual curve distances [1, 4] are not fully suitable for thread tracking evaluation in the context of robotic manipulation, because they do not properly evaluate the tracking of specific points. We therefore use an *arclength error* based on r specific points from the thread uniquely defined by their arclength parameters $\{v_k | 1 \leq k \leq r\}$:

$$e_{\text{acl}}^{3D} = \frac{1}{r} \sum_{k=1}^{r} \|\mathcal{C}^{3D}(v_k) - \mathcal{C}_{gt}^{3D}(v_k)\|.$$

In particular, we show that the approach allows for deformable tracking in the absence of normal motion along the curve, a challenging practical situation that occurs frequently in practice when the thread is dragged by one extremity.

[1] Tim Hauke Heibel, Ben Glocker, Martin Groher, Nikos Paragios, Nikos Komodakis, and Nassir Navab. Discrete tracking of parametrized curves. In *Proc. of CVPR*, 2009.

[2] Nicolas Padoy and Gregory D. Hager. Human-machine collaborative surgery using learned models. In *Proc. of ICRA*, 2011.

[3] Jur van den Berg, Stephen Miller, Daniel Duckworth, Humphrey Hu, Andrew Wan, Xiao-Yu Fu, Ken Goldberg, and Pieter Abbeel. Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations. In *Proc. of ICRA*, 2010.

[4] Peng Wang, Terrence Chen, Ying Zhu, Wei Zhang, Shaohua Kevin Zhou, and Dorin Comaniciu. Robust guidewire tracking in fluoroscopy. In *Proc. of CVPR*, 2009.

Using Richer Models for Articulated Pose Estimation of Footballers

Vahid Kazemi
vahidk@nada.kth.se
Josephine Sullivan
sullivan@nada.kth.se

CVAP
KTH, The Royal Institute of Technology
Stockholm, Sweden

This work tackles the problem of automatically reconstructing the 3D pose of a person, in particular a football player, from multiple images taken from uncalibrated affine cameras. We adopt a bottom up approach, summarized as, localize the skeletal 2D joints in each image independently and then perform factorization with limb length constraints to estimate the 3D pose. The joint localization task is the more challenging part and is the paper's main focus.

Localization of a person's limbs in an image is very difficult for a myriad of reasons most notably the range of articulations of the person (especially true in sports footage), self-occlusion, foreshortening of limbs and motion blur. However, in recent years significant progress has been made with the introduction of pictorial structure type models using discriminatively learned parts [1, 2, 4]. These models compromise between accurate modeling of the underlying flexibility in the appearance and spatial configuration of the person's limbs and computational concerns to make the parameter learning and the inference tractable.

Despite this progress, though, the results are far from perfect in real world scenarios. Figure 1(a) shows the results from the state-of-the-art *Flexible Mixture of Parts* (FMP) model [4] on images from our football dataset. The right of figure 1(a) shows an example of a common failure. The problem is partly due to the simplifications made in the modeling. However, the main observation exploited in this paper is that while the *true configuration* might not always correspond to the global optimum of the FMP's cost function, it frequently gets a high score. One can observe this by examining figure 1(b). It shows that on our football dataset a correct configuration - all the parts are localized correctly - is in the top 1000 scoring configurations w.r.t. the FMP cost function 88% of the time, while the top scoring configuration is a correct configuration only 36% of the time.

As a correct configuration is frequently in the set of the top n scoring configurations w.r.t. the simplified (FMP) scoring function and it is straightforward to obtain these configurations [3], we only need to evaluate a more accurate and arbitrarily complex scoring/re-ranking function on this small set.

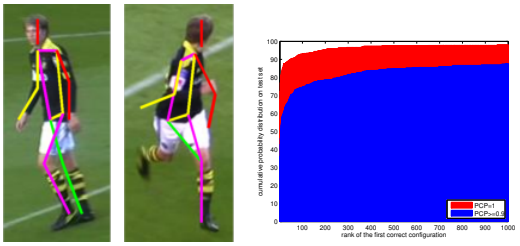


Figure 1: **(a)** Shown is the top scoring configuration returned by the FMP model and its PCP score for two images. The PCP score is the proportion of correctly localized limbs. **(b)** This is a cumulative histogram of the rank of the first correctly predicted pose by the FMP model. In 36% of the test cases the top scoring configuration has PCP=1. While 88% of the time there exists a configuration with PCP=1 in the top scoring 1000 configurations. These percentages change to 68% and 98% when the definition of a correct configuration is lowered to having PCP ≥ 0.9.

Since we only consider the n-best configurations returned by the FMP model we are at liberty to exploit more complicated and computationally expensive scoring of a configuration. Our new model re-weights appearance scores from the FMP model to prevent the double counting of evidence. We also use the colour distribution of foreground and background to penalize configurations which do not explain all the foreground. The crucial factor here is that we allow ourselves to consider the global configuration simultaneously as opposed to only considering pairs of parts at a time.

Ranking function	left/right flips **not** ignored	left/right flips ignored
Flexible Mixture of Parts	0.884	0.895
Re-ranking SVM-Rank	0.917	0.936
Oracle re-ranking	0.982	0.982

Table 1: Summary of the results on our football dataset with and without the re-ranking function. The first column of numbers displays the average PCP score of the top scoring configuration returned by the FMP model, our learnt re-ranking function and an oracle re-ranking function. The second column is the average PCP score when the left and right labels for the arms and legs are ignored.

To evaluate our method we have annotated a dataset of 771 images of football players, which includes images taken from 3 views at 257 time instances. Table 1 summarizes the results on our dataset with and without using the re-ranking function, as well as the results of picking the closest configuration to the ground truth between top 1000 configurations. In addition to the standard PCP score, we have provided the PCP scores ignoring the left/right limb assignments. It can be seen that in both case using a re-ranking function improves the performance comparing to the state of the art FMP model. The difference is much more significant if we only compare the top scoring configuration. The probability of the true configuration getting the top score based on FMP model is 36%, while this probability is increased to 51% using our model (an oracle ranking function in this case could improve the results up to 88%).

Finally, we have used the 2D estimates from our model to reconstruct the configuration of the player in 3D. With no assumptions about the pose of the player this is an extremely difficult task. However, when we have fairly good 2D estimates across all views we are able to get reasonable results. Figure 2 shows a stick figure of the 3D reconstruction of the top scoring 2D configurations, along with the back projected 2D estimates.

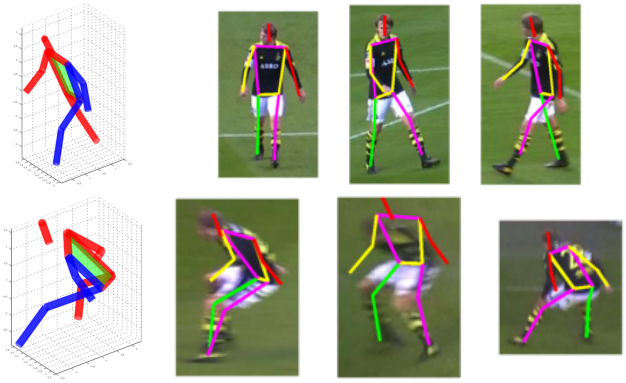


Figure 2: The result of the 3D reconstruction of the body joints computed from the top scoring 2D configurations, along with the back projected 2D estimates.

[1] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1): 55–79, 2005.

[2] O. Firschein and M. A. Fischler. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.

[3] D. Park and D. Ramanan. N-best maximal decoders for part models. In *Proceedings of the International Conference on Computer Vision*, 2011.

[4] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the Conference on Computer vision and Pattern Recognition*, 2011.

Dynamical Pose Filtering for Mixtures of Gaussian Processes

Martin Fergie
mfergie@cs.man.ac.uk
Aphrodite Galata
a.galata@cs.man.ac.uk

School of Computer Science,
University of Manchester, UK

In this paper we present a method for performing discriminative human pose estimation using a mixture of Gaussian Processes appearance model to map directly from the image features to the multi-model pose distribution. In order to obtain a pose estimate for a sequence of frames, we introduce a dynamic programming algorithm for inferring a smooth pose sequence from the multi-model distribution given by our appearance model.

1 Mixture of Gaussian Processes

A mixture of experts model gives a predictive distribution over the pose $\mathbf{y}$ conditioned on the image observation $\mathbf{x}$ as a mixture of Gaussian distributions:

$$p(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{K} p(z=i|\mathbf{x})\mathcal{N}(\mu_i(\mathbf{x}),\Sigma_i(\mathbf{x})),$$

where each μ_i and Σ_i are given as a function of $\mathbf{x}$ and $p(z=i|\mathbf{x})$ is a weight applied to each component as a function of $\mathbf{x}$ such that $\sum_i p(z=i|\mathbf{x}) = 1$ and $0 \leq p(z=i|\mathbf{x}) \leq 1$. We use a model where each expert prediction, $\mathcal{N}(\mu_i(\mathbf{x}),\Sigma_i(\mathbf{x}))$, is given by a Gaussian Process allowing each model to map a non-linear region of the dataset. To learn the model we partition the data set using an indicator variable $\mathbf{z} = \{z_n\}_1^N$ where each $z_n = i$ indicates that training point n is used to train expert i. We initialise $\mathbf{z}$ using k-means and then use a Gibbs sampling algorithm to optimise $\mathbf{z}$ with respect to the pose distribution:

$$p(z_n = i|\mathbf{z}_{/n},\mathbf{X},\mathbf{Y},\theta_i,\phi) \propto p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{X}_{\vartheta_i/n},\mathbf{Y}_{\vartheta_i/n},\theta_i)$$
$$p(z_n = i|\mathbf{z}_{/n},\mathbf{x}_n,\phi).$$

where $\mathbf{z} = \{z_n\}_{n=1}^N$, $z_n \in \{1 \ldots K\}$ indicates which expert each data point belongs to, $\vartheta_{i/n}$ is the set of indices of data points that belong to expert i, with the n^{th} point removed. Learning is performed in an *expectation-maximisation* fashion where we iterate between training the experts and resampling $\mathbf{z}$.

Figure 2 demonstrates the effect of the Gibbs sampling process. The top figure gives the predictive distribution when $\mathbf{z}$ are set by running k-means on the training pose data. The lower plot shows that the Gibbs iterations leads to a more accurate pose distribution.

2 Dynamic Pose Filtering

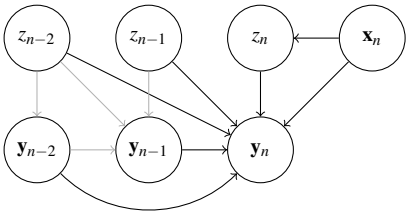


Figure 1: Graphical model for 2nd order pose filtering showing the nodes involved in computing $\mathbf{y}_n$.

The mixture of experts model gives us a Gaussian mixture model over the pose for each frame. The naive approach to estimating the pose for each frame would be to take the expectation of this distribution, getting a pose estimate by taking a weighted average of the Gaussian components. This averages out the multi-modal regions and does not utilise any temporal information resulting in a jittery tracking sequence.

We introduce a dynamic programming algorithm that incorporates a second order dynamical prediction model to infer a smooth path through the predictive distributions of each frame. Our algorithm propagates multiple predictions for each frame, where each prediction represents the observation of one appearance expert. For frame n and expert $z_n = i$, we obtain a Gaussian prediction over the pose given by:

$$\hat{\mathbf{y}}_{i,n} = p(\mathbf{y}_n, z_n = i|\mathbf{x}_{1:n}) = p(\mathbf{y}_n|\mathbf{x}_n, z_n)$$
$$\sum_{z_{n-2}}\sum_{z_{n-1}} p(\mathbf{y}_n|\hat{\mathbf{y}}_{z_{n-1},n-1},\hat{\mathbf{y}}_{z_{n-2},n-2})p(z_{n-1}|\mathbf{x}_{1:n-1})p(z_{n-2}|\mathbf{x}_{1:n-2}).$$

Where $p(\mathbf{y}_n|\mathbf{x}_n, z_n)$ is the appearance prediction for expert $z_n = i$ and $p(\mathbf{y}_n|\hat{\mathbf{y}}_{z_{n-1},n-1},\hat{\mathbf{y}}_{z_{n-2},n-2})$ is a dynamical prediction. Figure 1 shows a graphical model illustrating this process. The forward-backward algorithm is used to infer a optimal sequence of apperance experts $\mathbf{z} = \{z_n\}_n^N$ from which we obtain a smooth pose estimate $\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_n\}_n^N$.

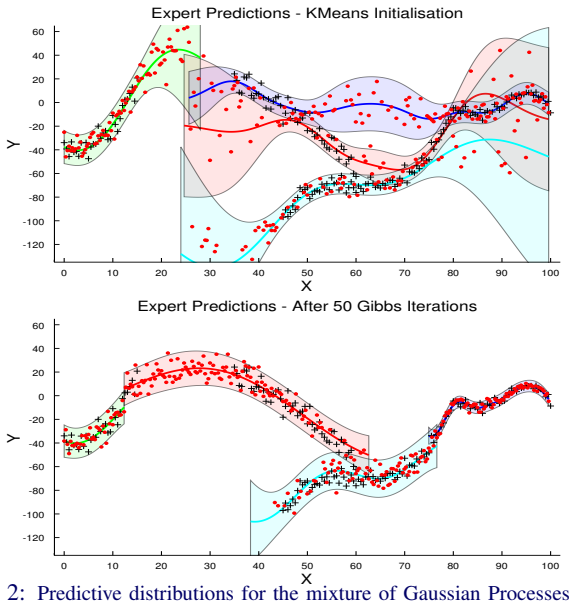


Figure 2: Predictive distributions for the mixture of Gaussian Processes model on a toy dataset. The black crosses represent the training points, the red dots are samples drawn from the predictive distribution and the coloured lines represent the predictive mean and variance of each expert.

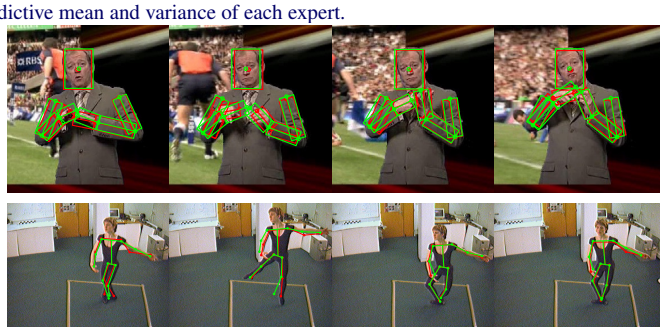


Figure 3: Tracking results for the sign language and ballet datasets showing every fifth frame of a continous sequence. Ground truth shown in red, predicted pose is shown in green.

3 Results

We evaluate our method on a 2D sign language data set taken from BBC television and a 3D Ballet dance sequence. We show visual tracking results along with quantitive results comparing our method to other state of the art methods for discriminative pose estimation.

Dataset:	Ballet		Sign Language
Model/Feature	BOW SC	HMAX	HMAX
Our Method (app)	32.52	32.68	6.88 ± 0.48
BME [1]	51.71	71.72	11.87 ± 0.92
Urtasun and Darrell [3]	36.13	38.18	8.11 ± 0.62
sKIE [2]	31.55	37.57	9.36 ± 0.52
Kernel Regression	71.68	71.71	10.65 ± 0.30

Table 1: Quantitive Results.

[1] Bo and Sminchisescu. *CVPR*, 2008.
[2] Memisevic et al. *PAMI*, 2012.
[3] Urtasun and Darrell. *CVPR*, 2008.

Close-Range Human Detection for Head-Mounted Cameras

Dennis Mitzel
mitzel@vision.rwth-aachen.de
Bastian Leibe
leibe@vision.rwth-aachen.de

Computer Vision Group
RWTH Aachen University
Aachen Germany

Motivation. Robust multi-person detection and tracking is an important prerequisite for many applications. Examples include the use of mobile service robots in busy urban settings or mobile AR applications such as Google's project Glass. In this paper, we address the problem of stereo based person detection from the perspective of a moving human observer wearing a head-mounted stereo camera system. From this viewpoint many pedestrians in a crowded scenario are only partially visible due to occlusions at the image boundaries. In such situations, standard full-body object detectors such as [3] are not well-suited, since they cannot deal with the large degree of occlusion. On the other hand, we can take advantage of the elevated viewpoint of a head-mounted camera, which typically leaves the head-shoulder region of close-by pedestrians well visible.

Taking inspiration from a recently proposed human upper body detector for Kinect RGB-Depth data by [2], this paper proposes an improved stereo depth-template based approach which can quickly and reliably detect close-by pedestrians. Similar to [1] we generate regions of interest (ROIs) based on the stereo data in order to reduce the search space of the detector. Our approach learns a continuous normalized depth template from annotations of human upper bodies and slides this template over the extracted depth ROIs at several scales in order to compute a normalized distance score. The output of this process are distance matrices whose entries represent the distance between the template and the overlayed segment of the ROI for each scale. After non-minimum suppression (NMS) in the distance matrices we obtain several detections (from different scales) for a person that are pruned to a set of final detections by a second, template-based NMS stage. We systematically evaluate this approach and characterize the effects of its parameters. In addition, we show how it can be integrated into a mobile multi-person tracking framework.

Approach. In Fig. 1 we illustrate a compact overview of our proposed detection and tracking framework. For each new frame, given the stereo pair and the corresponding depth map, we project the 3D points onto a (automatically estimated) ground plane and extract the ROIs using connected components on the ground projection image. For each extracted 3D ROI, we generate the corresponding ROI in the image plane by back-projecting the ROIs from the ground plane to the image. The 2D ROIs are passed to the detector, which slides the learned upper body template over the ROIs and computes the distance matrix by taking the Euclidean distance between the template and the overlayed, normalized depth image segment. Using a minimum filter on the distance matrix, we obtain possible bounding box hypotheses for the upper bodies. These hypotheses are further pruned to a final detection set by using a template based intersection-over-union (IoU) NMS stage, where the detection with the lowest distance is chosen first and all other detections within a certain overlap area are removed iteratively.

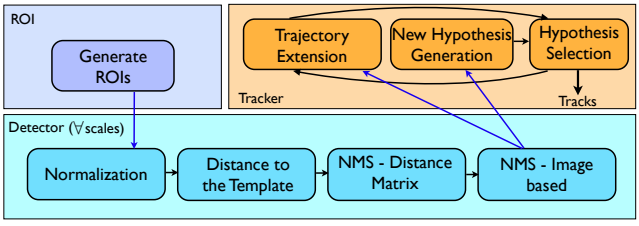


Figure 1: Overview of the different modules of the proposed approach.

Depth Template Detector. The pipeline for the detector consists of the following steps. First, for each ROI in the image plane, we discard the pixels which are not in the depth range of the ROIs in 3D by setting them to zero, as illustrated in Fig. 2. Then starting from an initial template size that is one third of the ROI height, we slide the template over the ROI. At each position, the segment of the ROI that is overlayed with the template is normalized with its median depth and then the distance between the template and the segment is computed. As a final result we obtain a distance matrix that contains for each position of the template the corresponding distance to the segment in the depth image, see Fig. 2 (bottom right). For Multi-Scale Handling we need to rescale the template and slide it over ROI again, because the initial scale estimation based on the height of the 2D ROI might not be representative for all pedestrians in the group. The multi-scale approach introduces several additional detections on a person for a number of neighboring scales, as the scale stride is usually small. To reduce this set to only one representative detection for each pedestrian, we perform an image based NMS.

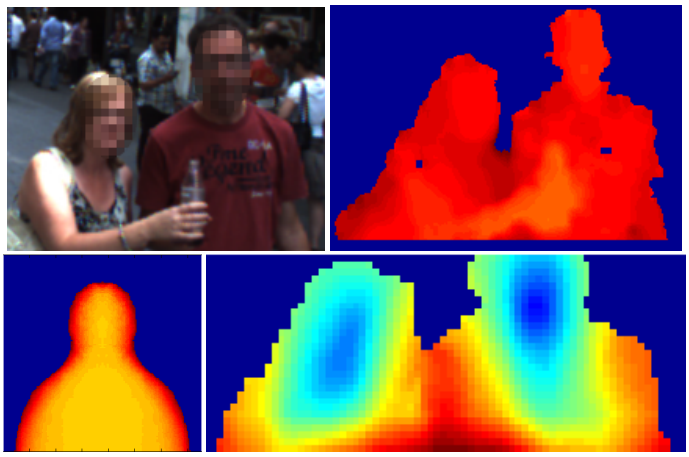


Figure 2: (upper left) Original ROI from the left image. (upper right) Input depth from the ROI. (bottom left) Depth template learned based on 600 upper body annotations. (bottom right) The resulting distance matrix for the initial scale.

Figure 3: Experimental detection and tracking results

[1] M. Bansal, S. H. Jung, B. Matei, J. Eledath, and H. S. Sawhney. A real-time pedestrian detection system based on structure and appearance classification. In *ICRA*, 2010.

[2] W. Choi, C. Pantofaru, and S. Savarese. Detecting and Tracking People using an RGB-D Camera via Multiple Detector Fusion. In *CORP*, 2011.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

Detection and Tracking of Occluded People

Siyu Tang
tang@mpi-inf.mpg.de
Mykhaylo Andriluka
andriluka@mpi-inf.mpg.de
Bernt Schiele
schiele@mpi-inf.mpg.de

Computer Vision and
Multimodal Computing
MPI Informatics
Saarbrücken, Germany

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

Figure 1: Procedure to synthetically generate training images for our double-person detector. (a) background person, (b) foreground person, (c) foreground person map, (d) generated synthetic training image.

We consider the problem of detection and tracking of multiple people in crowded street scenes. Several methods, i.e. tracking and 3D scene reasoning approaches [3, 7, 10], have been proposed to track people even in the presence of long-term occlusions. While these approaches allow to reason across potentially long-term and full occlusions they still require that each person is sufficiently visible at least for a certain number of frames. State-of-the-art approaches to people detection [5, 6] are able to reliably detect people under a variety of imaging conditions, people poses, and appearance, but their performance degrades when people become partially occluded. Careful reasoning about association of image evidence to detection hypotheses has been proposed in [4, 8, 9], but these approaches treat partial occlusion as nuisance and perform decisions based on the image evidence that corresponds to the visible part of the person, which makes them unreliable in cases of severe occlusions.

Here, we explore an alternative strategy, observing that in crowded street scenes most occlusions happen due to overlaps between people, we consider the joint evidence of both people. This is possible since overlapping people result in characteristic appearance patterns that are otherwise uncommon. Our approach builds on the powerful deformable part models (DPM [6]), which we extend in three ways. First we propose a new double-person detector that allows to predict bounding boxes of two people even when they occlude each other by 50% or more. Second, we propose a joint person detector, that is jointly trained to detect single- as well as two-people in the presence of occlusions. Last, we integrate the above joint model into a tracking approach to show its potential for people detection and tracking.

Double-person detector: We build the double-person model upon the DPM framework to detect the presence of two people and to predict the bounding boxes of both people, the occluding person as well as the occluded person. For training, we synthetically generate two-people samples (Fig. 1) based on the TUD training data [1]. The synthetic images are ideal for training as they come with accurate occlusion-level estimates. We demonstrate experimentally that our double-person detector significantly outperforms a single-person detector in the presence of severe occlusions (Fig. 2).

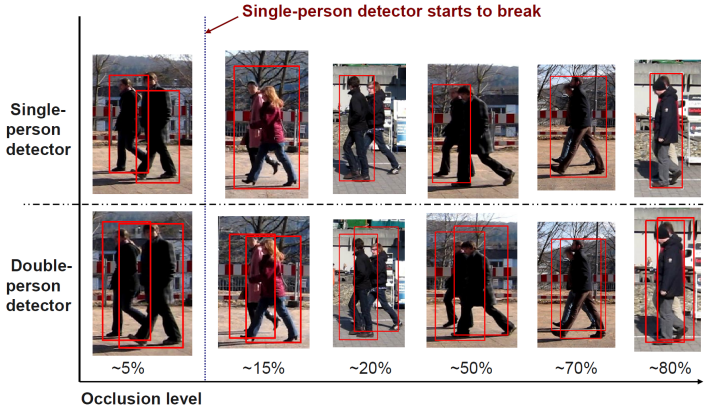


Figure 2: Qualitative comparison of single- and double-person detectors with occlusion.

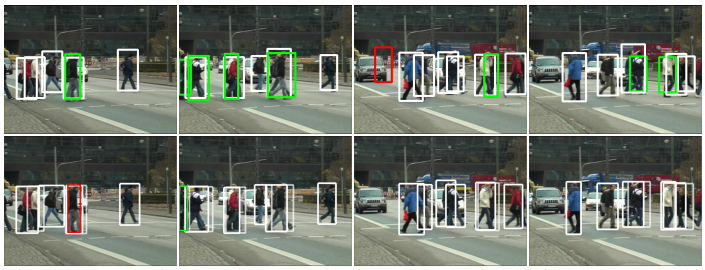


Figure 3: Detection results at equal error rate obtained with the approach of [4] (top) and our joint detector (bottom) on the TUD-Crossing [1] dataset. False-positive detections are shown in red and missing detections in green. One of the two bounding boxes predicted from the two-person detection is shown with the dotted line.

Multi-Person Detection: The joint person detector is again built upon the DPM-approach where the role of the different components is now to differentiate both between single and two people as well as between different occlusion levels among two people. Similarly to double-person detector we initialize the double-person components with training examples corresponding to gradually increasing levels of occlusion. For the single-detector components we rely on the standard initialization based on the bounding box aspect ratio. Our experiments on TUD-Crossing [1] dataset confirm the benefit of the joint detector in the realistic scenes (Fig. 3).

Multi-Person Tracking: We compare the performances of a single-person and the joint detector in the context of multiple people tracking. To that end we apply the people tracking-by-detection approach of [2] without modification both to the output of the single-person and the joint detectors on TUD-Crossing [1] dataset. The tracker based on the joint detector is able to correctly track people even in cases of severe occlusions clearly showing the potential of using our joint detector as the basis for multi-people tracking in scenes with many people and in the presence of severe occlusions.

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR'08*, .

[2] M. Andriluka, S. Roth, and B. Schiele. Monocular 3d pose estimation and tracking by detection. In *CVPR'10*, .

[3] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR'12*.

[4] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using hough transform. In *CVPR'10*.

[5] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian Detection: A Benchmark. In *CVPR'09*.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI'10*.

[7] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV'08*.

[8] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR'05*.

[9] X. Wang, T.X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV'09*.

[10] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *IJCV*, 75:247–266, November 2007.

Latent SVMs for Human Detection with a Locally Affine Deformation Field

Ľubor Ladický[1]
lubor@robots.ox.ac.uk
Philip H.S. Torr[2]
philiptorr@brookes.ac.uk
Andrew Zisserman[1]
az@robots.ox.ac.uk

[1] Department of Engineering Science
University of Oxford
Oxford, UK

[2] School of Technology
Oxford Brookes University
Oxford, UK

Human detection is typically formulated as a problem, where the objective is to find all the people within an image and enclose each one of them by a tight bounding box. Dalal and Triggs [1] introduced the histograms of oriented gradients (HOG) feature for this problem over cells composing the bounding box, efficiently matching object shape with the learnt rigid template of edge directions. This method was originally applied to pedestrian detection, but it turned out to give good performance for a wide range of object classes with distinctive shape. Intuitively, a higher dimensional template should capture more small details and should lead to a better performance. However, even under small local deformations of the data it is impossible to align the data properly and the discriminative edges often fall into the neighbouring cell. To overcome this problem, Felzenszwalb *et al*. [2] proposed a star-graph part based model allowing a predetermined number of rigid parts to change their relative location with respect to the centre of the object. Large intra-class variance was modelled by splitting training samples based on their aspect ratio and training a classifier for each set of training samples independently. This procedure works if the different aspect ratio corresponds to a different viewpoint, such as for example for a car. However, it is not very suitable for human detection, where different human poses often have the same aspect ratio and the method does not learn an independent model for each one of them.

Motivated by this work, we propose a new latent variable SVM allowing for any deformations of the template, expressed in terms of a deformation field. Rather than restrict ourselves to a star-graph model, we allow the template to deform according to a locally affine deformation field.

The classifier for our deformable template then takes the form :

$$H(\mathbf{c}) = \max_{\mathbf{d}\in\mathcal{A}} \left(\mathbf{w}^* \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c})) + b^* - R(\mathbf{d}) \right), \tag{1}$$

where $\mathbf{c}$ is the set of cells, $\mathbf{h}(D^{\mathbf{d}}(\mathbf{c}))$ are the histograms of oriented gradients on the template deformed by the deformation field $\mathbf{d}$, $R(\mathbf{d})$ is the regularisation cost taking the form of the pairwise Markov Random Field (MRF) and $\mathcal{A}$ is the set of locally affine deformation fields 1, in which any 2×2 neighbouring cells transform into a parallelogram.

The latent SVM optimisation problem for learning the weights $\mathbf{w}^*$ and the bias b^* becomes:

$$\begin{aligned} (\mathbf{w}^*, b^*) &= \arg\min_{(\mathbf{w},b)} \lambda ||\mathbf{w}||^2 + \sum_{k=1}^{M} \xi^k \\ \text{s.t. } \forall k &\in \{1,..,M\} : \\ \xi^k &\geq 0 \\ \xi^k &\geq 1 - z^k \max_{\mathbf{d}\in\mathcal{A}} \left(\mathbf{w} \cdot \mathbf{h}(D^{\mathbf{d}}(\mathbf{c}^k)) + b - R(\mathbf{d}) \right), \end{aligned} \tag{2}$$

where M is the set of training samples and $z^k \in \{-1,1\}$ is the label of the k-th training sample. This problem is non-convex, however, we can follow the same approach as [2] and iteratively estimate the weight vector $\mathbf{w}$ with the bias b, and the deformation field $\mathbf{d}$ for each training sample.

The problem of finding the optimal weight vector $\mathbf{w}$ and bias b given estimated deformation fields for each training sample is a standard SVM problem and can be solved with any standard SVM algorithm. The problem of finding the optimal deformation field given weight vector is the max-a-posteriori (MAP) estimation of the pairwise MRF problem under the additional locally affine deformation field constraints. We start with the observation that the deformation of all cells in the first row and in the first column of the deformation field fully determines the deformation of any other cell. Locally affine constraints can be satisfied for any deformations of the cells in the first row and in the first column. Thus, any locally affine deformation field can be reached by two moves - the first in which we move each row j by a deformation $\Delta^r d_j = (\Delta^r d_j^x, \Delta^r d_j^y)$ and the second

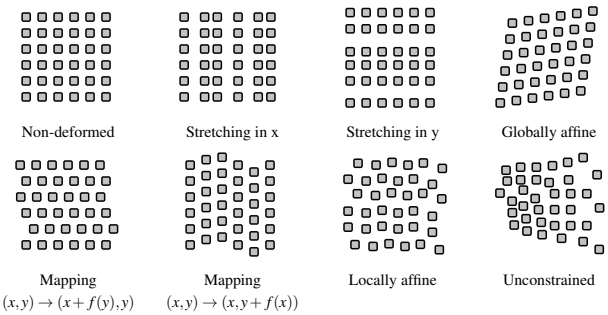


Figure 1: Expressive power of the locally affine deformation field. The locally affine constraints allow for stretching or mapping of the template in both axes, global affine transformation of the template or the combination of all of them resulting in the general locally affine transformation, in which any 2×2 neighbouring cells transform into a parallelogram.

Figure 2: Typical results on the Buffy data set. Positive detections are overlaid with the learnt HOG template of the corresponding model, deformed by the deformation field.

in which we move each column i by a deformation $\Delta^c d_i = (\Delta^c d_i^x, \Delta^c d_i^y)$. Trivially, both of these moves do not break the local affinity property and can lead to any deformation of the cells in the first row and in the first column and thus to any arbitrary locally affine deformation field. Both of these subproblems can be solved exactly using dynamic programming.

Typically, different viewpoints are modelled by splitting the positive samples based on the aspect ratio and trained independently for each aspect ratio. However, this approach could not model different poses with similar bounding boxes independently and the star-graph model (or alternatively our locally affine deformation field) could not capture this kind of deformations. We can take an advantage of our deformation field model and cluster the problem into subproblems based on the similarity of training samples, defined as their scalar product, regularised by the MRF cost of the deformation field which transforms one training sample to another.

We tested our method on the more challenging Buffy data set of [3], which consists of images with large variety of poses and truncations by the edge of the image, which makes it suitable for our clustering method. Our method significantly outperformed other state-of-the-art approaches [1, 2]. We assume, our locally affine deformation field formulation could be used in the future for other computer vision tasks, such as tracking or optical flow estimation.

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[2] P. F. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[3] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR*, 2008.

Sparsity Potentials for Detecting Objects with the Hough Transform

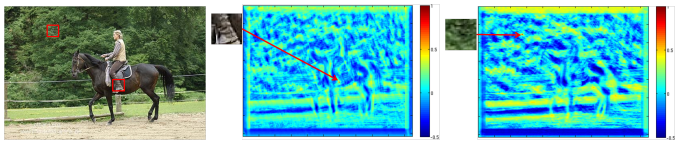
Nima Razavi[1]
nrazavi@vision.ee.ethz.ch
Nima Sedaghat Alvar[2]
n.sedaghat@gmail.com
Juergen Gall[3]
juergen.gall@tue.mpg.de
Luc van Gool[1,4]
luc.vangool@esat.kuleuven.be

[1]Computer Vision Laboratory
ETH Zurich, Switzerland
[2]Machine Vision and Intelligence Lab
Sharif University of Technology, Iran
[3]Perceiving Systems Department
MPI for Intelligent Systems, Germany
[4]IBBT/ESAT-PSI
K.U. Leuven, Belgium

Hough transform based object detectors divide an object into a number of patches and combine them using a shape model. For efficient combination of patches into the shape model, the individual patches are assumed to be independent of one another. Although this independence assumption is key for fast inference, it requires the individual patches to have a high discriminative power in predicting the class and location of objects. In this paper, we make the following two observations:

- the similarity in appearance of patches in a neighborhood of a central patch exhibit different sparsity values when the central patch appears on an object as opposed to a background region.
- the codebook entries associated with texture or simple edge patterns are consistently less sparse in their neighborhood as opposed to entries which are associated to more complex patterns (see Fig. 1).

Based on these observations, we argue that the sparsity of the appearance of a patch in its neighborhood can be a very powerful measure to increase the discriminative power of a local patch and incorporate it as a sparsity potential for object detection.



(a) (b) (c)

Figure 1: The patches in an image exhibit different sparsity values. While the self-similarity of a non-texture patch (a) to its neighboring patches is low, the patch on the tree (b) is less sparse and much more similar to its neighborhood. Base on this observation, we introduce a measure which captures the *sparseness* of a patch within its neighborhood and incorporate it as a "sparsity potential" for object detection.

We base our sparsity or distinctiveness measure on self-similarity. Let us assume that we have a metric that measures the similarity of a patch f_i with all patches in its neighborhood, $\{f_n | n \in \mathcal{N}^i\}$, e.g. by Normalized Cross Correlation as in Fig. 1. Further, we assume that the returned similarity is normalized to be in $[0,1]$ with 1 representing the most similar and 0 the most dissimilar patch. In this case, one is getting a real valued self-similarity vector $\mathbf{ss}_i = (ss_1, \ldots, ss_{|\mathcal{N}^i|})$ where each element ss_n records the normalized similarity of f_n to f_i.

The sparsity of the self-similarity vector $\mathbf{ss}_i$ can be measured in many different ways, e.g., by using entropy or various vector norms. In this work, we use the L1-norm,

$$||\mathbf{ss}_i||_1 = \sum_{n \in \mathcal{N}^i} |ss_n| \qquad (1)$$

which is both simple and fast to calculate.

For detecting objects, we incorporate the sparsity measure by training a classifier for each code-word and object class. For training the sparsity classifiers, first a set of features on the validation set, both on objects as well as background, are extracted and are assigned to one or more codebook entries ω_j. Given a neighborhood function $\mathcal{N}^i$, the sparsity measure of every feature f_i is calculated. Next, for each ω_j and class label c, these sparsity measures are collected and used to learn a simple threshold θ_{c,ω_j}. These thresholds are then used to estimate class probability $p(c|\omega_j, \mathcal{N}^i)$ as

$$p(c|\omega_j, \mathcal{N}^i) \propto \begin{cases} p(c|\omega_j) & \text{if} \quad ||\mathbf{ss}_i||_1 \leq \theta_{c,\omega_j} \\ 0 & \text{otherwise,} \end{cases} \qquad (2)$$

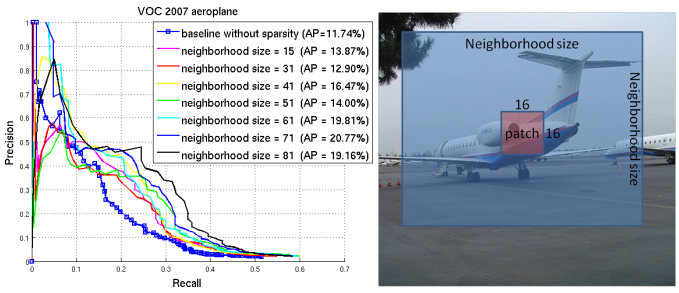


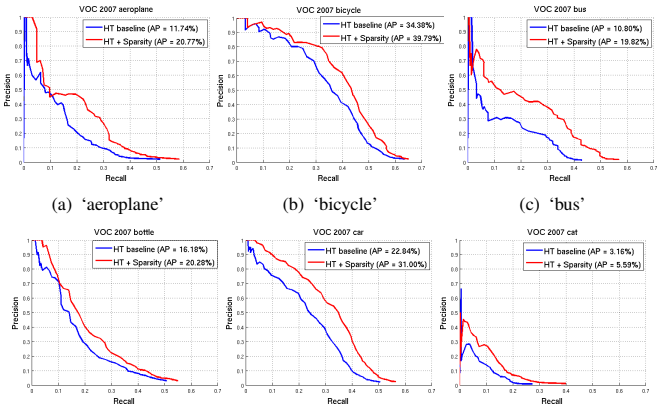
Figure 2: This figure evaluates the effect of the neighborhood size used for calculating the sparsity on the accuracy of the detector. The performance comparison of our Hough Forest baseline [1] with and without sparsity measure potentials is shown. As can be seen, the proposed sparsity potential improves the accuracy. The performance tends to increase with the window size until it saturates at around 71 pixels, almost doubling the average precision (AP) compared to the baseline.The sparsity potential is calculated on a square neighborhood of every 16×16 patch.

where $p(c|\omega_j)$ is the class probability estimated at the codebook entry ω_j.

Using the sparsity measure as a single dimensional feature, the thresholds are learned such as to separate the positive and negatives with the best classification accuracy with zero false negatives on the training data.

The evaluation is carried out on the PASCAL VOC 2007 dataset. Our experiments confirm the benefit of using the proposed sparsity potential for object detection increase the mean average precision (mAP) of our Hough transform baseline [1] from 14.82 to 20.68. Example Precision/Recall curves for some categories of the VOC'07 dataset is shown in Fig. 3.

In the future, it would be interesting to use the sparsity potentials in a multi-class setup to also discriminate classes from one another. Since the self-similar patches tend to belong to the same label, it would be also interesting to incorporate their sparsity as a higher order potential for image segmentation.



(a) 'aeroplane' (b) 'bicycle' (c) 'bus'

(d) 'bottle' (e) 'car' (f) 'cat'

Figure 3: The precision recall curves for some categories of the PASCAL VOC 2007. As can be seen, the proposed sparsity potentials substantially improve the detection performance. The average precision (AP) is calculated by the integral under the curve.

[1] J. J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *TPAMI*, 33(11):2188–2202, 2011.

Gradient Edge Map Features for Frontal Face Recognition under Extreme Illumination Changes

Ognjen Arandjelović
ognjen.arandjelovic@gmail.com
Swansea University, UK

The aim of this work is to match images of frontal faces across extreme illumination changes. This is a problem of importance in a broad range of practical scenarios. For example, the user is commonly asked to face the camera in security applications which perform authentication before granting access to a resource. Retrieval systems also often focus on nearly frontal faces because face detection is most reliable for this pose.

Discriminative information is not uniformly distributed across different parts of a face. Rather, most of it is contained in the regions which exhibit substantial variation in either geometry or albedo and which can be readily detected using direct computations on pixel intensities. One of the simplest methods of accomplishing this is by applying a 2D high pass filter. However, when applied on images acquired under extreme illuminations, the simple high pass filter fails in achieving a satisfactory result. One of the reasons can be readily observed by examining Figures 1(a) and 1(b). Notice that the discontinuities in the poorly lit, shadowed regions are less pronounced than those in well illuminated regions.

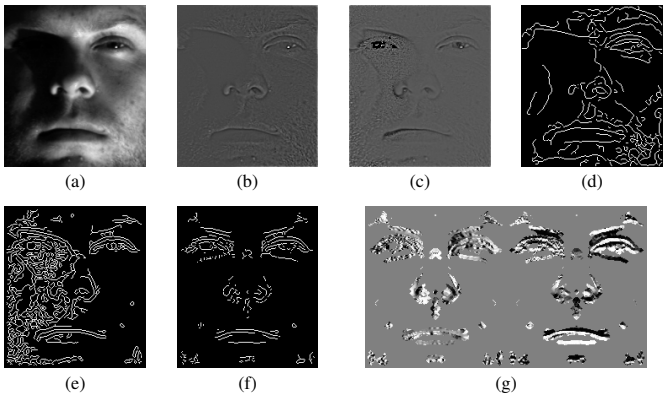


Figure 1: (a) Raw and (b) band pass filtered appearance, (c) self-quotient image, edges computed from (d) the original and (e) the self-quotient image, (f) symmetrically consistent and reliable edges, and (g) the proposed representation.

The dependence of the magnitude of the intensity discontinuities preserved by high pass filtering can be addressed by using one of the Retinex-like methods. These are in part inspired by the human visual system and the observation that humans perceive brightness in a relative rather than absolute manner. In other words, a discontinuity of a small magnitude in a dark region should have a greater effect than a discontinuity of the same magnitude in a bright region. The high pass filter can thus be modified simply by dividing pixel-wise the filtered result with the low pass filtered image which has the effect of averaging image intensity. This is a variant of the self-quotient image:

$$I_{SQI}(x,y) = I_{HP}(x,y) \;/\; I_{LP}(x,y) \tag{1}$$

The result of applying this filter is shown in Figure 1(c) which indeed appears to be an improvement over the output of the high pass filter. However, when this representation is used for matching on our data set, as discussed in detail in the paper, the error rate is increased to nearly that achieved by using raw appearance. A more detailed inspection of the resulting image reveals insight into the causes. Specifically, the noise in the poorly lit regions of the face has been amplified, as has the originally imperceptible boundary of the shadow caused by (weak) ambient illumination. The filter also causes the appearance of artefacts around interfaces between very bright and very dark image regions.

Steps 1 and 2: Provisional Edges Our method avoids the described difficulties associated with the use of absolute intensity by concentrating on the binary edge image. This is the first step of the proposed cascade. Note that we do not detect edges directly in the original image. Instead, we apply the Canny edge detector on the self-quotient image, to ensure that very weak edges in poorly illuminated regions are correctly detected. The difference between the two approaches is illustrated in Figures 1(d) and 1(e). Note that our approach results in higher automatically estimated Canny thresholds. While this has the effect of producing fewer false edges in the well lit regions of the face and more true edges in the poorly lit regions, the number of spurious edges in the poorly lit regions is also increased. This problem is addressed in the next step of our cascade.

Step 3: Spurious Edge Removal The edge map computed from the self-quotient image may contain many false edges e.g. from the amplification of noise in poorly lit regions or from the boundaries of cast shadows. Highly saturated image regions may also cause the hallucination of edges. Regardless of what the underlying cause is, false edges can decrease the matching accuracy. For example, it is straightforward to see that the left hand side of the edge map in Figure 1(e), full of densely packed false edges, will match nearly any true face edge map rather well.

We remove false edges by exploiting the vertical symmetry of frontal faces by requiring agreement between the left hand and right hand sides of the edge map. If E is the binary edge image and $\overleftrightarrow{\cdot}$ the vertical mirroring operator the edge image E_T with spurious edges removed is computed as:

$$E_T(x,y) = \begin{cases} 1 & : E(x,y) = 1 \text{ and } \overleftrightarrow{E}_{DT}(x,y) \leq 2 \\ 0 & : \text{otherwise} \end{cases} \tag{2}$$

where E_{DT} is the distance transformed edge image. In other words, we remove all edge segments which are not within ≈ 2 pixels from the corresponding mirrored edges, see Figure 1(f).

Steps 4 and 5: Edge Reliability Refinement After spurious edges are removed in the previous step of the cascade, the resulting binary image E_T is not necessarily vertically symmetric. We interpret this lack of symmetry as arising from true but unreliable edges. To ensure that the final representation contains only those true edges which are repeatably detectable, we again exploit the vertical symmetry of frontal faces. We first dilate the edge image E_T using a 4×4 pixel solid circle structuring element S and then combine the dilated edge information from the left hand and right hand sides of the face:

$$E_R(x,y) = \min\left[E_T(x,y) \oplus S, \overleftrightarrow{E}_T(x,y) \oplus S\right] \tag{3}$$

Steps 6 and 7: Merging Edge and Gradient Information In the last step of the proposed cascade, we incorporate into our representation further discriminative information. The specific limitation of the edge map that we wish to overcome is its limited ability to robustly capture shape. This is a consequence of the observation that each edge map pixel by itself only contains information about whether an edge passes through it or not. Edge pixels carry no additional information about the directionality of the corresponding edge. We demonstrate that a highly discriminative representation can be obtained by combining the dilated reliable edges map and the corresponding gradient phase. This is achieved by computing a 3D image comprising two "stacked" 2D images which contain horizontal and vertical gradients at the dilated edges:

$$E_{GM}(x,y,1) = \begin{cases} \frac{\partial I}{\partial x}/|\nabla I| & : E_R(x,y) > 0 \\ 0 & : E_R(x,y) = 0 \end{cases} \tag{4}$$

$$E_{GM}(x,y,2) = \begin{cases} \frac{\partial I}{\partial y}/|\nabla I| & : E_R(x,y) > 0 \\ 0 & : E_R(x,y) = 0 \end{cases} \tag{5}$$

The normalization by magnitude is performed to account for the unreliability of absolute or even relative image intensity across different illuminations. The directionality of the gradient, on the other hand, is preserved well in the vicinity of strong discontinuities (but not necessarily elsewhere). The proposed representation is illustrated in Figure 1(g), displayed as the two stacked images side by side.

The effectiveness of the proposed representation was demonstrated on the notoriously challenging YaleB data set, which covers a wide range of illumination conditions, many of which are extreme (rear lateral, overhead). Unlike most of the previous work we used only a single image per person for training and a single probe image as test, and did not eliminate any of the images from the evaluation. Our gradient edge map achieved outstanding results, incorrectly recognizing in only 0.8% of the cases and exhibiting nearly perfect receiver-operator characteristic behaviour. This performance vastly exceeds that reported previously in the literature on this data set and using the same evaluation methodology.

Exemplar Driven Character Recognition in the Wild

Karthik Sheshadri
sheshadri@cmu.edu
Santosh K. Divvala
santosh@ri.cmu.edu

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania
USA

Character recognition in natural scenes continues to represent a formidable challenge in computer vision. Traditional optical character recognition (OCR) methods fail to perform well on characters from scene text owing to a variety of difficulties in background clutter, binarisation, and arbitrary skew. Further, English characters group into only 62 classes whereas many of the world's languages have several hundred classes. In particular, most Indic script languages such as Kannada exhibit large intra class diversity, while the only difference between two classes may be in a minor contour above or below the character. These considerations motivate an exemplar approach to classification; one which does not seek intra class commonality among extreme examples which are essentially sub classes of their own. Exemplar SVM's have been recently introduced in the object recognition context. The essence of the exemplar approach is that rather than seeking to establish commonality within classes, a separate classifier is learnt for each exemplar in the dataset. To make individual classification simple, linear SVM's are used and each classifier is hence an exemplar specific weight vector. Each exemplar in the dataset is resized to standard dimensions, and thence HOG features are densely extracted to create a rigid template x_E. A set of negative samples N_E are created by the same process from classes not corresponding to the exemplar. Each classifier (w_E, b_E) maximizes the separation between x_E and every window in N_E. This is equivalent to optimizing the convex objective[4]:

$$\Omega_E(w,b) = \| w \|^2 + C_1 h(w^T x_E + b) + C_2 \sum_{x \in N_E} h(-w^T x - b), \quad (1)$$

where $h(.)$ indicates the hinge loss function, and C_1, C_2 are constants.

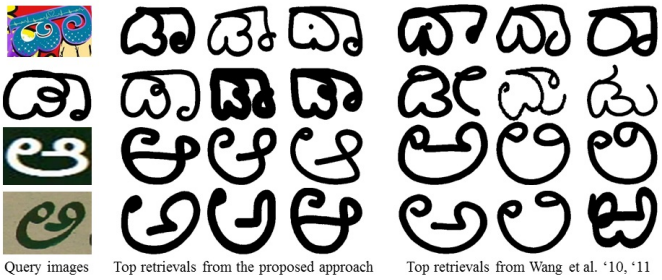


Figure 1: The problem of fine grained character recognition in unconstrained visual scenes is addressed in this paper.

Table 1: Our results (English) on chars74k and ICDAR-CH, and comparison to baseline methods.

Model	Chars74k-5	Chars74k-15	ICDAR-CH
HOG+ESVM+AFF	**48.43 ± 2.40**	**69.66**	**70.53**
HOG+ESVM+on fly calib	27.76 ± 1.74	60.00	66.67
HOG+NN+AFF	47.61 ±0.81	64.22	63.59
HOG+ESVM	16.33 ± 2.33	44.68	41.44
HOG+NN[2]	45.33 ± 0.99	57.50	52
NATIVE+FERNS[3]	−−	54	64
MKL[1]	−−	55.26	−−

1 Calibrating Exemplar SVM's for Character Recognition

In return for simpler classification at the level of each exemplar, we must now deal with the problem of decision calibration: combining decisions from independently trained and hence non compatible classifiers. In this work, we explore the following two calibration methods.

1.1 Calibration based on SVM scores

In the spirit of [4], we adopt an "on the fly" calibration method, using positives selected by each exemplar based on SVM scores. Exemplars which achieve low scores on ground truth labelled query images from the validation set are suppressed by moving the decision boundary in their requisite classifier towards the exemplar and well performing exemplars are boosted by moving the decision boundary in their classifier away from the dataset. Given a detection 'x' and the learned sigmoid parameters α_E, β_E, the calibrated detection score for each exemplar E is as follows: $f(x|\alpha_E, \beta_E, w_E) = \frac{1}{1+e^{-\alpha_E(w_E^T x - \beta_E)}}$. This rescaling and shifting of the decision boundary conditions each classifier to fire only on visually similar examples, and underlines the explicit correspondence offered by the exemplar SVM based approach.

1.2 Calibration based on affine motion estimation

This calibration approach is based on a simple observation: variations in font and shape essentially constitute small affine transformations. Characters from visual scenes are often affine warps of characters from normal text: they are oriented differently, different character contours are irregularly shaped, and are of different sizes, etc. Hence on thinned character images, one could compute affine motion between train and test characters, and minimize the sum of absolute differences to refine candidate choices obtained by simple max voting of the exemplar SVM's. Our proposed approach is summarized as follows: (i) count the number of positive votes in favour of each class, computed based on a preselected threshold (ii) extract the top k of these classes, and perform affine motion estimation $M_{E_C,Q}$ between the thinned binarized query image Q and every exemplar E_C, C being the class of the exemplar, in the training subset corresponding to the top k classes (iii) recognize the character as that class which minimizes the sum of absolute differences (SAD) between the test character and any exemplar in the training subset corresponding to top k classes. Equation (2) illustrates the approach:

$$B = _{C \in C_K} \{E_C - M_{E_C,Q} Q\} \quad (2)$$

where B is the computed belief class of query image Q, and $M_{E_C,Q}$ is the affine transformation matrix which warps Q with respect to E_C.

The proposed approach beats the existing state of the art on the chars74k and ICDAR datasets by over 10% for English, and around 24% for Kannada. Motivated by the performance on two languages ranging from conventional to extremely complex, we argue that leveraging fine grained categorization and generic object recognition approaches is a promising research direction for character recognition unconstrained by language or setting.

[1] T. de Campos, B. Babu, and M. Varma. Character recognition in natural images. In: VISAPP 2009.

[2] K. Wang and S. Belongie. Word Spotting in the Wild. In: ECCV 2010.

[3] Kai Wang, Boris Babenko, and Serge Belongie. End to End Scene Text Recognition. In: ICCV 2011.

[4] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond. In: ICCV 2011.

Efficient Learning-based Image Enhancement: Application to Super-resolution and Compression Artifact Removal

Younghee Kwon[1]

Kwang In Kim[2]
http://www.mpi-inf.mpg.de/~kkim

Jin Hyung Kim[3]
http://ai.kaist.ac.kr/~jkim/

Christian Theobalt[2]
http://www.mpi-inf.mpg.de/~theobalt

[1] Google Inc.
Mountain View, CA, USA

[2] Max-Planck-Institut für Informatik
Saarbrücken, Germany

[3] KAIST
Daejeon, Korea

Many widely used imaging operations lead to specific degradations of images with respect to the ground truth. The removal of these degradations is one of the most important tasks in computer vision, image processing, and computational photography. For instance, image encoding deficiencies such as block artifacts have to be removed frequently. Deterioration and information loss due to the limitations of the optical system, such as limited sensor resolution or defocusing, should also be erased.

This paper presents an algorithm for learning-based image enhancement. At each pixel in the given degraded image, a small sub-window encompassing that pixel (patch) is extracted and the corresponding desired patch is estimated based on Gaussian process (GP) regression. As the output patches (i.e., the predictive means) overlap with their neighbors, the result of the regression step constitutes a set of candidates for each pixel location. The final pixel-valued output is synthesized by combining the candidates based on the corresponding predictive variances and trading the consistency with them with a global image prior as a regularizer [1].

While GP regression has been shown to be competitive on a wide range of small-scale applications, its application to large-scale problems is limited due to its unfavorable scaling behavior. A standard approximate approach to overcome this limitation is to introduce a small set of *inducing variables* $\mathbf{f}_{\mathcal{U}} = \{f(\mathbf{u}_1), \ldots, f(\mathbf{u}_m)\}$ (corresponding to *inducing inputs* $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$) through which the conditional independence of the training ($\mathbf{f}$) and target ($\mathbf{f}_*$) latent variables is assumed in the approximation of the joint prior (cf. the unified framework of [2]):

$$p(\mathbf{f}_*, \mathbf{f}) \approx q(\mathbf{f}_*, \mathbf{f}) = \int q(\mathbf{f}_*|\mathbf{f}_{\mathcal{U}}) q(\mathbf{f}|\mathbf{f}_{\mathcal{U}}) p(\mathbf{f}_{\mathcal{U}}) d\mathbf{f}_{\mathcal{U}}. \quad (1)$$

The *training conditional* $q(\mathbf{f}|\mathbf{f}_{\mathcal{U}})$ is approximated subsequently. This leads to a set of approximations which are referred to as *sparse* GPs where the inference is carried out through $\mathbf{f}_{\mathcal{U}}$ summarizing l training data points.

In existing sparse GP algorithms, once identified, the inducing inputs $\mathcal{U}$ are fixed throughout the entire test set. The problem is then cast into an optimization where one constructs $\mathcal{U}$ based on a certain measure of approximation quality (e.g., marginal likelihood and information gain). The performance of a sparse approximation depends heavily on the inducing inputs $\mathcal{U}$. However, usually the corresponding optimization problem is non-convex and accordingly is not easy to solve.

In this paper, we present a simple alternative to these *off-line* approaches: We build a sparse GP which is specially tailored for a given test input $\mathbf{x}_*$ (i.e., $\mathcal{U} \equiv \mathcal{U}_*$ is chosen depending on $\mathbf{x}_*$; The corresponding GP model is constructed only when it is presented with a test point $\mathbf{x}_*$). An important advantage of this *on-line* approach is that it naturally leads to an extremely simple strategy for identifying $\mathcal{U}_*$: If we introduce a spatial Markov assumption on $\{f_*, \mathbf{f}\}$

$$p(f_*|\mathbf{f}, \mathcal{N}(f_*)) \approx q(f_*|\mathcal{N}(f_*)), \quad (2)$$

where $\mathcal{N}(f_*)$ denotes the values of the latent function f for the inputs in the spatial neighborhoods $\mathcal{N}(\mathbf{x}_*)$ (of $\mathbf{x}_*$), the decomposition (1) becomes exact once we use $\mathcal{N}(\mathbf{x}_*)$ for $\mathcal{U}_*$.

This approximation dramatically reduces the computation time during training. Actually, the only training component is building a data structure for nearest neighbor (NN)-search, which facilitates identifying $\mathcal{N}(f_*)$. However, for large scale problems ($l \approx 2 * 10^5$ in the current applications), this approximation might be still impractical. The second step of our approximation is to introduce an additional Markov-like assumption directly on the observations:

$$p(f_*|\mathcal{Y}, \mathcal{N}_1(\mathbf{y}_*)) \approx q(f_*|\mathcal{N}_1(\mathbf{y}_*)), \quad (3)$$

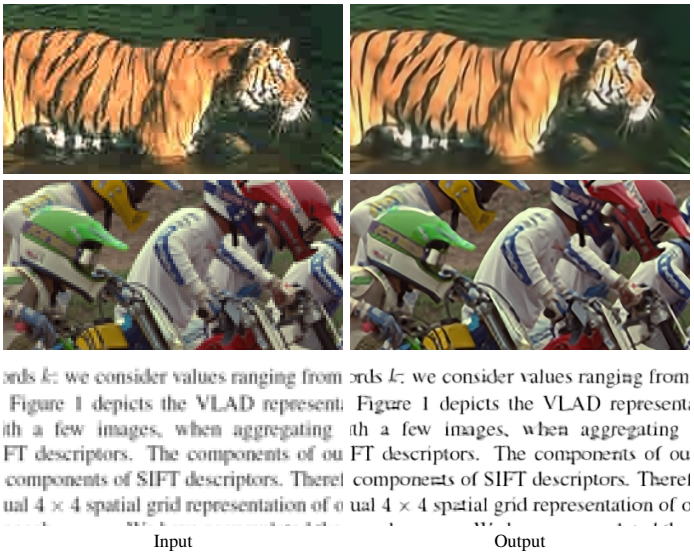


Figure 1: Examples of image enhancement: (top) JPEG artifact removal, (middle) (generic) single-image super-resolution, and (bottom) (document-specific) single-image super-resolution.

where $\mathcal{Y}$ is the set of training labels and $\mathcal{N}_1(\mathbf{y}_*)$ denotes the observed training target values in the spatial neighborhood $\mathcal{N}_1(\mathbf{x}_*)$ of $\mathbf{x}_*$. To guarantee that the resulting GPs are non-locally regularized, we set $\mathcal{N}(\mathbf{x}_*) \subset\subset \mathcal{N}_1(\mathbf{x}_*)$. The spatial Markov assumption (2) is fairly natural and has proven to be effective in many different applications while the second approximation step (3) is motivated by the large-scale behavior of full GPs: For large l, the predictive distribution $p(f_*|\mathcal{Y})$ of a full GP is not affected by the data points which are sufficiently distinct from $\mathbf{x}_*$ [3].

Since the only training component of the new approximation is building a data structure for NN-search, the off-line processing is very fast. Therefore, the resulting image enhancement system is very flexible as the it can be easily adapted to the distribution of a specific (non-generic) class of images. This is important especially when *a priori* knowledge of the problem is available in terms of a class-specific set of example images. For instance, if it is known that the image of interest to be processed is representing documents (whose statistical properties might be distinct from those of general images), one could quickly generate examples from this specific class of images on which the system is trained. While this leads to much better results (see the last low of Fig. 1), it is infeasible in conventional sparse GPs due to their high complexity in training (which includes the identification of inducing inputs).

We demonstrate the utility of our algorithm in two example image enhancement applications that can benefit from the high efficiency of our approximation (both in training and in testing): suppression of compression artifacts in JPEG images and single-image super-resolution (Fig. 1).

[1] P. V. Gehler and M. Welling. Product of "edge-perts". In *NIPS*, Cambridge, MA, 2005. MIT Press.

[2] J. Quińonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *JMLR*, 2005.

[3] P. Sollich and C. K. I. Williams. Using the equivalent kernel to understand Gaussian process regression. In *NIPS*, 2005.

Contour-HOG: A Stub Feature based Level Set Method for Learning Object Contour

Zhi Yang
zhiyang@buffalo.edu
Yu Kong
yukong7@buffalo.edu
Yun Fu
raymondyunfu@gmail.com

Department of ECE and College of CIS
Northeastern University
Boston, MA, USA
Department of CSE
State University of New York
Buffalo, NY, USA

An object can be effectively characterized by its contour. Caselles *et al*. [1] introduced the concept of geodesic active contours, which applies the energy reducing form to acquire contours. Shape priors are great helpful to obtaining more accurate contours. Leventon [6] utilized the curvature prior as the shape prior for different classes of objects to guide contour evolution. Etyngier *et al*. [3] proposed a non-linear manifold learning method for learning shape prior. Another line of work uses edges to describe objects which provide local perspective of an object and are robust when part of the object is occluded. However, since the global perspective is missing, the arrangement of the edge features, such as the pairwise interactions between edge features [2, 5] or the relative positions of edge features with respect to the centroid of the shape [7], is exploited to improve the edge based models.

We propose a novel edge-based method for learning objects. Given an image, our method first detects edgelet feature as a rough contour for an object. Edgelet feature indicates potential positions for the contour and may stop curve evolution. These positions are referred to as *stub features*. Object contour is adaptively refined by the level set method. The evaluation criteria for contour evolution is defined by the similarity between the evolving contour and the target contour computed by their HOG features. Therefore the curve evolution method is referred to as the *Contour-HOG* method. We formulate the joint distribution of the edgelet feature, the HOG feature and the curvature of the evolved contour in a probabilistic model, and perform classification by computing the posterior of the evolved contour conditioned on the three types of features. Compared with previous methods, our method uses stub features to roughly localize a target object. This allows us to accurately capture the contour of the object. Moreover, the method fuses both local and global features to better describe the contour and thus improves the recognition accuracy.

Our method begins by detecting edgelet feature [8]. We use this feature to roughly find an object in an image. With the detected stub feature, we compute their similarities with the stub feature in training data under a predefined edge mask $M_{k,s}$. The similarity is computed as

$$p(x,s,k) = p_o(x,s,k)p_g(x,s,k), \tag{1}$$

where x is the 2D coordinate for a stub feature, $p_g(x,s,k)$ represents the likelihood of a local gradient and a mask having the similar magnitude, and $p_o(x,s,k)$ is the distribution of a gradient sharing the same orientation with the mask. We define $p_g(x,s,k)$ as a Gaussian distribution:

$$p_g(x,s,k) = \frac{1}{\sqrt{2\pi}|\Sigma_g|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}[g(x)-\mu]^T\Sigma_g^{-1}[g(x)-\mu]\right). \tag{2}$$

Here, μ and Σ_g are the mean and standard deviation of gradient magnitude, respectively. $g(x)$ is the magnitude of gradient vector derived by operating a mask $M_{k,s}$ on an input image I. The similarity of two orientation vectors is computed by

$$p_o(x,s,k) = \frac{\|v(I_q(M_{k,s}(x)), I_t(M_{k,s}))\|_2^2}{NT^2}, \tag{3}$$

where $v(o_1,o_2)$ is the distance between two quantized directions, N denotes the number of pixels selected by the mask, T represents the number of orientations, $I_q(M_{k,s}(x))$ and $I_t(M_{k,s})$ are the orientation vectors for a testing image and a training image, respectively.

After stub feature detection, we run a contour evolution method to obtain the contour of an object. We adopt the Elliptic Fourier descriptor (EFD) [4] to represent the object contour.

Our evolution model utilizes curvature force to guide contour evolution. The curvature force is defined as the ratio between the curvature of the shape prior κ_p and the curvature of the evolving contour κ_c: $f_\kappa = \frac{\kappa_p}{\kappa_c}$.

To obtain the global perspective of the evolving contour, we compute the HOG feature of an evolving contour and measure its similarity with the contours of classes. The using of the global similarity measure of contours allows us to accurately obtain an object contour.

In our work, an object is classified by the similarity between the evolved contour of the object and the contour of a target object in the training dataset. The similarity is evaluated based on their curvatures, stub features and HOG features. We compute the similarity as

$$p(\kappa_c,\kappa_t,S_{\text{hog}},S_{\text{stub}}) = p(\kappa_c|\kappa_t,S_{\text{hog}},S_{\text{stub}})p(\kappa_t,S_{\text{hog}},S_{\text{stub}}), \tag{4}$$

where κ_c and κ_t denote the curvature of an evolved contour and a target contour, respectively. S_{hog} is the affinity of contour-HOG features between an evolved contour and a target contour, and S_{stub} is the affinity of stub features between them. In our work, we assume κ_t and S_{hog} are independent of S_{stub}. Then Eq.(4) can be given by

$$p(\kappa_c,\kappa_t,S_{\text{hog}},S_{\text{stub}}) = p(\kappa_c|\kappa_t,S_{\text{hog}},S_{\text{stub}})p(\kappa_t,S_{\text{hog}})p(S_{\text{stub}}). \tag{5}$$

Here, we define $p(\kappa_c|\kappa_t,S_{\text{hog}},S_{\text{stub}})$ as a Gaussian distribution:

$$p(\kappa_c|\kappa_t,S_{\text{hog}},S_{\text{stub}}) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{||\kappa_c-\kappa_t||_2^2}{2\sigma^2}), \tag{6}$$

where σ is the standard deviation. In Eq.(5), distribution $p(\kappa_t,S_{\text{hog}})$ is the similarity of HOG features between an evolved contour and a contour of a class. $p(S_{\text{stub}})$ is the similarity of an edge segment and a predefined edge mask computed in Eq.(1).

We use a likelihood function $\Lambda(Y;\theta)$ to measure the likelihood of a particular model with N training samples. We define the joint likelihood function as a Gaussian mixture model (GMM):

$$\Lambda(Y;\theta) = \prod_{n=1}^{N}\sum_{k=1}^{K} w_k G(y;\mu_k,\delta_k). \tag{7}$$

Our model is learned by maximizing likelihood estimation (MLE):

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\{\Lambda(Y;\theta)\}. \tag{8}$$

[1] V. Caselles, F. Catte, T. Coll, and F. Dibos. A geometric model for active contours in image processing. *Numerische Mathematik*, 66(1): 1–31, 1993.

[2] S. M. Ali Eslami, Nicolas Heess, and John Winn. The shape boltzmann machine: a strong model of object shape. In *CVPR*, 2012.

[3] Patrick Etyngier, Florent Ségonne, and Renaud Keriven. Active-contour-based image segmentation using machine learning techniques. In *MICCAI*, pages 891–899, 2007.

[4] F. P. Kuhl and C. R. Giardina. Elliptic fourier features of a closed contour. *Computer Graphics and Image Processing*, 18(3):236–258, 1982.

[5] M. Leordeanu, M. Hebert, and R. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *CVPR*, 2007.

[6] Michael E. Leventon, W. Eric L. Grimson, Olivier Faugeras, and William M. Wells III. Level set based segmentation with intensity and curvature priors. In *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 4–11, 2000.

[7] A. Opelt, A. Pinz, and A. Zisserman. A boundary-fragment-model for object detection. In *ECCV*, 2006.

[8] Bo Wu and Ram Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, pages 90–97, 2005.

Virtual Line Descriptor and Semi-Local Matching Method for Reliable Feature Correspondence

(Extended abstract — see details in BMVC 2012 full paper)

Zhe Liu
zhe.liu@enpc.fr
Renaud Marlet
renaud.marlet@enpc.fr

University Paris-Est, LIGM (UMR CNRS),
Center for Visual Computing,
Ecole des Ponts ParisTech,
6-8 av. Blaise Pascal, 77455 Marne-la-Vallée, France

Matching detected features in two images based on the similarity of their descriptor often provides good correspondences (inliers). But it often also includes false matches (outliers). Eliminating those false matches while preserving true correspondences remains challenging for images with numerous ambiguities or strong transformations, e.g., due to strong occlusions. In these cases, individual feature matching is not enough; global methods are required, such as RANSAC or graph matching.

However, RANSAC-like methods hardly treat low inlier rates (less than 10%) and, when estimating a fundamental matrix, they cannot eliminate outliers corresponding to points that have matches near their epipolar line. As for graph matchers, they can cope with higher-order constraints (involving more than one match) and optimize a global consistency score. However, most of them are based on geometric constraints rather than photometry. Besides, they are not well suited for a high outlier rate, and their time and space complexity grows exponentially with the order, which limits in practice applications to a few hundred points.

In this paper, we define a 2nd-order photometric descriptor for virtual lines joining two neighbouring feature points. We show it can be used in existing graph matchers to significantly improve their accuracy. We also define a scalable, semi-local matching method based on this descriptor. We show that it is robust to strong transformations and more accurate than existing graph matchers for scenes with significant occlusions, including for very low inlier rates. If used as a preprocessor, it also significantly improves the robustness of RANSAC and reduces camera calibration errors.

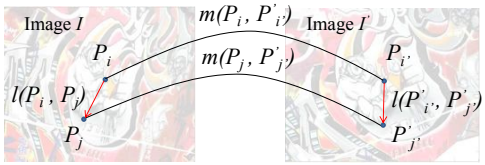


Figure 1: Lines (P_i,P_j) and $(P'_{i'},P'_{j'})$ are unlikely to be similar unless matches $(P_i,P'_{i'})$ and $(P_j,P'_{j'})$ are correct.

Virtual line descriptor (VLD). Our approach is based on the fact that, for any two points P_i,P_j in image I, and any two points $P'_{i'},P'_{j'}$ in image I', it is unlikely to find similar photometric information around lines (P_i,P_j) and $(P'_{i'},P'_{j'})$ unless both $(P_i,P'_{i'})$ and $(P_j,P'_{j'})$ are correct matches (see Fig. 1). To measure this property, we define a virtual line descriptor (VLD) that captures photometric information between any two points.

We consider a regular disk covering, with overlap, of an image strip between P_i and P_j, and use a SIFT-like descriptor to represent each disk. The covering consists of U disks D_u of radius $r=\frac{dist(P_i,P_j)}{U+1}$ (see Fig. 2). Each disk is described at image scale $\max(r/r_{\min},1)$ where $r_{\min}$ is a minimum description radius. Scales are actually discretized and precomputed. The disk descriptor includes a small-size histogram of gradients and an orientation (see Fig. 2). The global, virtual line descriptor is the concatenation of all disk descriptors. Although it requires less dimensions, VLD inherits its SIFT descriptor's robustness to noise and changes of scale, orientation and illumination. Last we define a distance between VLDs that can be used in the pairwise score of a graph matcher.

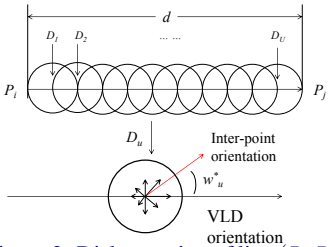


Figure 2: Disk covering of line (P_i,P_j) and histogram of gradient orientation.

K-VLD matching method. We introduce K-VLD, a novel semi-local, 2nd-order matching method. It relies both on geometric and photometric consistency (based on VLD). Two matches $(P_i,P_{i'})$ and $(P_j,P_{j'})$ are considered *VLD-consistent* iff virtual lines (P_i,P_j) in I and $(P_{i'},P_{j'})$ in I' have a VLD-distance less than a given threshold. In the paper, we also define a geometric consistency measure for two matches based on the scale and main orientation of feature points, assuming the local transformation is close to a similitude; if this measure is under a given threshold, then the matches are considered *geometry-consistent*. Finally, two matches are *gVLD-consistent* iff they are both geometry- and VLD-consistent.

The basic idea of K-VLD relies on the fact that, given a potential match $(P_i,P'_{i'})$, if there are in the neighborhood of P_i and $P'_{i'}$ at least K other matches $(P_{j_k},P'_{j'_k})_{k\in\{1,\dots,K\}}$ that are gVLD-consistent with $(P_i,P'_{i'})$, then $(P_i,P'_{i'})$ is likely to be a correct match. Given a match m among a set of matches M, the paper defines a notion of neighborhood $\mathcal{N}_M(m)$ whose size adapts to the density of feature points. Experimentally, requiring that good matches have at least K gVLD-consistent neighbors eliminates many outliers, but some may still remain, especially in ambiguous scenes. We found that adding an extra constraint on the proportion of geometry-consistent neighbors and on their average measure of geometric consistency helped in removing many of these remaining outliers.

The K-VLD algorithm starts with all the potential matches and iteratively removes matches that have less than K gVLD-consistent neighbors and matches that do not satisfy the extra geometric constraint, until no match is removed. Ambiguous matches are solved too, based on a preference for matches with many gVLD-consistent neighbors or, when the number of such neighbors is equal, a preference for matches with lowest average VLD-distance. Additional optimizations and heuristics ensure a performance that is in practice quasi linear in the number of matches.

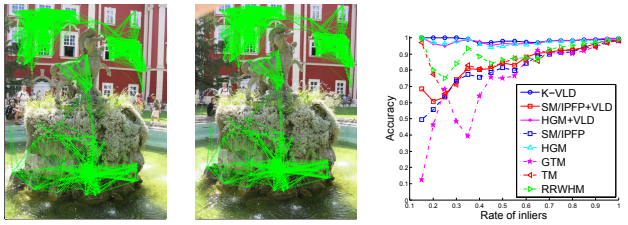


Figure 3: Dětenice fountain: K-VLD clusters & average accuracy.

Evaluation. We experimented with various matching methods: probabilistic hypergraph matching (HGM), tensor matching, hypergraph matching via reweighted random walks, spectral matching (SM) / integer projected fixed point, and game-theoretic matching. We also augmented methods SM and HGM with our VLD. And we compared with K-VLD.

We evaluated matching accuracy w.r.t. changing imaging conditions with Mikolajczyk's dataset. K-VLD often outperforms other methods and VLD significantly improves existing methods, especially for scenes with viewpoint or scale changes. We also evaluated the case of strong occlusion with Dětenice fountain's dataset (see Fig. 3). K-VLD creates clusters of consistent matches despite occlusions, mostly outperforming other methods. Last we tested K-VLD as a pre-filter to RANSAC-based calibration (ORSA) using Strecha's castle dataset (see Fig.). It considerably improves the quality of match selection. As it can eliminate false matches near epipolar lines, it greatly improves precision, as well as stability. It also substantially reduces the number of iterations, which improves speed.

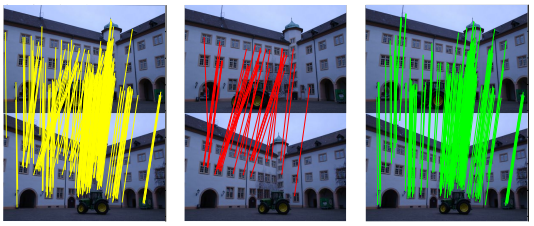


Figure 4: Left: inliers by ORSA. Middle: false matches near epipolar lines by ORSA, rejected by K-VLD. Right: inliers by K-VLD + ORSA.

Learning Edge-Specific Kernel Functions For Pairwise Graph Matching

Michael Donoser[1]
donoser@icg.tugraz.at
Martin Urschler[2]
martin.urschler@cfi.lbg.ac.at
Horst Bischof[1]
bischof@icg.tugraz.at

[1] Institute for Computer Graphics and Vison
Graz University of Technology
Austria
[2] Ludwig Boltzmann Institute for
Clinical Forensic Imaging
Graz, Austria

Motivation Graph matching has become widely used in several computer vision applications including tracking, shape matching or object detection. Many different approaches are available for solving the NP-hard problem in an approximated manner, e. g. based on spectral techniques, probabilistic methods or graduated assignments. Surprisingly, only few papers focused on the important graph potentials themselves, which have a tremendous influence on the quality of the obtainable results. For example, it was shown in [1] that solving a linear assignment problem using well chosen potentials even improves over related state-of-the-art quadratic assignment solutions.

One important challenge of using powerful potentials in graph matching is their right parametrization, which is mostly done manually. Only a few papers focused on the problem of choosing the right parameters. Caetano et al. [1] showed how to learn optimal parameters for the features used in the potentials from manually labeled reference data sets and Leordeanu et al. [2] extended this idea to an unsupervised setting. Both approaches strongly agree on the fact that learning the parameters is important for improving the matching performance.

In this paper we follow the idea of learning optimal parameters for the task of graph matching, but instead of learning fixed parameters for the features used as done in [1, 2], we directly learn edge-specific kernel functions for each node pair, assuming that the setting of graph matching is a-priori known. Such a-priori knowledge is indeed available in several important computer vision applications like automated face alignment, model fitting and object localization.

Method Our approach is divided into two main steps. First, in the training step, we learn a statistical shape model from labeled training images, obtaining a model of the location uncertainties of the graph nodes. Our model is then defined by edge-specific kernel functions for every pair of nodes. Second, during testing, our method is an extension of standard graph matching formulated as quadratic assignment problem. As the main difference to standard graph matching solutions, we exploit the learned kernel functions for improving matching quality.

We define our kernel functions $\mathcal{K}_{ij}$ to relate an edge connecting points i and j in our reference graph (consisting of N_1 nodes) to an edge connecting points a and b in the query graph (N_2 nodes) by deriving statistics of the point location distributions within a labeled training set. Thus, we assume that we have given a set of training images, with the same number of labeled points in each image, where we require the labeled points to be corresponding over the training set. We register all labeled points of the training set to each other using Procrustes Analysis, which then allows to describe the spatial distribution of each point over the training set by a Gaussian as it is visualized in Figure 1.

Figure 1: Building a statistical shape model from labeled training data. Unaligned point sets (left), Procrustes aligned sets (middle) and obtained location uncertainties (right) are shown.

The goal of graph matching is to find a one-to-one mapping between two graphs, which is defined by a binary assignment vector $\mathbf{x}^* \in \mathbb{R}^{N_1 N_2}$, where $x^*_{ia} = 1$ if node i of the reference graph matches to node a of the query graph and $x^*_{ia} = 0$ otherwise and $\sum_i x^*_{ia} = 1, \sum_a x^*_{ia} = 1$. Such standard quadratic assignment problems (QAP) are solved by

$$\mathbf{x}^* = \underset{\mathbf{x}}{argmax} \left(\mathbf{x}^T \mathbf{A}\, \mathbf{x} \right) = \underset{x}{argmax} \sum A_{ia,jb}\, x_{ia}\, x_{jb}\,, \qquad (1)$$

where $\mathbf{A}$ is a provided $N_1 N_2 \times N_1 N_2$ affinity matrix describing how well a pair of nodes in the reference (i, j) agrees in terms of local descriptors and geometry with a pair of nodes in the query (a, b).

Similar to related methods we use shape context ($\mathbf{s}_i$) as local descriptor for each node i, but replace the standard analysis of the differences in edge lengths by our learned edge-specific kernel functions. For this reason, the affinity matrix entries are adapted to

$$A_{ia,jb} = exp - \left(\mathbf{w}_1^T |\mathbf{s}_i - \mathbf{s}_a| + \mathbf{w}_2^T |\mathbf{s}_j - \mathbf{s}_b| + \mathbf{w}_3^T\, \mathcal{K}_{ij}(ab) \right), \qquad (2)$$

where $\mathcal{K}_{ij}(ab)$ is the learned pairwise kernel function. Thus, in our setting deviations from the reference graph geometry are penalized depending on the location uncertainty as learned in the statistical shape model.

We relax the integer optimization of Equation 1 into the continuous domain and solve it using Replicator Dynamics [3], an evolutionary algorithm from the field of game theory. These dynamics iteratively update the assignment vector $\mathbf{x}$ using

$$x_i^{t+1} = x_i^t \frac{(\mathbf{A}\, \mathbf{x}^t)_i}{\mathbf{x}^{tT} \mathbf{A}\, \mathbf{x}^t}, \qquad (3)$$

where $\mathbf{x}^t$ is the assignment vector at time t. As a necessary additional constraint $\mathbf{x}$ has to lie on the simplex $\left(\mathbf{x} \in \mathbb{R}^{N_1 N_2} : x_i \geq 0 \text{ and } \mathbf{1}^T \mathbf{x} = 1 \right)$. Replicator dynamics return an optimal assignment vector $\mathbf{x}^*$, which is a local (!) maximum of the optimization problem shown in Equation 1.

Experiments In a first experiment we use our method to align a set of face images using the IMM and AR face data sets. We used the mean point set obtained from the training graphs (*Mean*) or each of them (*All*) as reference graph and compared it to all point sets of the remaining test data. Table 1 shows the average percentage of correct assignments, comparing our proposed, learned potentials to standard ones. As can be seen, using our learned kernel function clearly improves results by up to **25%**.

GM	*AR data set*			*IMM data set*		
	Orig.	*Learned*	*Impr.*	*Orig.*	*Learned*	*Impr.*
All	88.1	98.5	+10.5	56.3	81.3	+25.1
Mean	95.6	98.7	+3.1	69.2	80.5	+11.3

Table 1: Percentage of correct assignments for matching to the mean point model (*Mean*) or each model of the training data (*All*) using standard (*Orig.*) and our learned potentials (*Learned*).

More experiments, e. g. on evaluating the influence of the number of training samples on the matching quality and an application for feature point based localization of previously unseen category instances in images, are provided in the main paper.

[1] L. Caetano, L. Cheng, Q. Le, and A.J. Smola. Learning graph matching. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2007.

[2] M. Leordeanu and M. Hebert. Unsupervised learning for graph matching. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[3] M. Pelillo. Replicator equations, maximal cliques, and graph isomorphism. *Neural Computation*, 11(8):1933–1955, 1999.

Genetic Programming-Evolved Spatio-Temporal Descriptor for Human Action Recognition

Li Liu

Ling Shao

Peter Rockett

Department of Electronic and Electrical Engineering,
University of Sheffield

Human action recognition has attracted a great deal of attention due to its potential usage in areas such as: video search and retrieval, intelligent surveillance systems and human-computer interaction.

In this paper, we propose a novel method by using Genetic Programming (GP) [2] to automatically generate a highly-performing low-level spatio-temporal descriptor for high-level human action recognition tasks. Our method is outlined in Fig. 1. For a given group of 3D sequence processing operators, GP first randomly assembles them into a variety of descriptors as the initialized population. The population is then continually evolved/evaluated by calculating the recognition error rate to evolve, hopefully, better-performing individuals into the next generation. Finally, one best-so-far individual can be selected as the final spatio-temporal descriptor. Genetic Programming (GP), as an evolutionary computation methodology, allows the computer to solve pre-defined tasks without requiring users to specify the form or structure of the solution in advance. GP can also escape local optima which may trap deterministic methods. Because of this, the use of GP is not limited to any research domain and can create relatively generalized solutions for target tasks. The basic Genetic Programming flow is shown in Algorithm. 1.

Algorithm 1 Genetic Programming

Start

Initialization Randomly create an initial population of computer programs from the available primitives (terminal set & function set).

Repeat

(1) Execute each program and evaluate its fitness.

(2) Choose programs from the population with a particular probability based on the fitness to involve genetic operations

(3) Create new generation programs by applying genetic operations.

If An acceptable solution is found or reach the maximum number of generations defined by user.

Stop

Return The best-so-far solution selected by Genetic Programming.

End

In our GP architecture, we have pre-defined three significant components as follows:

Terminal set: We flatten the 3D action sequences into 1D vectors as the programs' external inputs for GP.

Function set: We apply 12 unary 3D operators (*i.e.* Gaussian pyramid filters, Laplacian pyramid filters, Wavelet pyramid filers, *etc.*) and 4 basic binary arithmetic functions (*i.e.* Add, Subtraction, Multiply, Absolute subtraction) as our function set.

Fitness function: We use the classification error E_r for evaluating the performance of candidate descriptors. A support-vector machine (SVM) is adopted as the classifier to compute corresponding error rate. To achieve a fairer and more accurate result, ten-fold cross-validation is simultaneously employed on our dataset. We define the fitness function as follows:

$$E_r = (1 - (\sum_{i=1}^{n} (SVM[acu_i])/n)) \times 100\% \quad (1)$$

where $SVM[acu_i]$ denotes the classification accuracy of the fold i by the SVM, n indicates the total number of cross-validation folds, here, $n = 10$.

We test our GP architecture on a mixed dataset combining the KTH dataset[1] [3] with the Weizmann dataset[2] [1] to obtain a promising spatio-temporal descriptor. The parameter settings for our GP running are listed in Table 1.

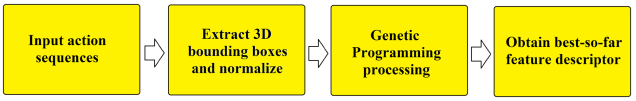
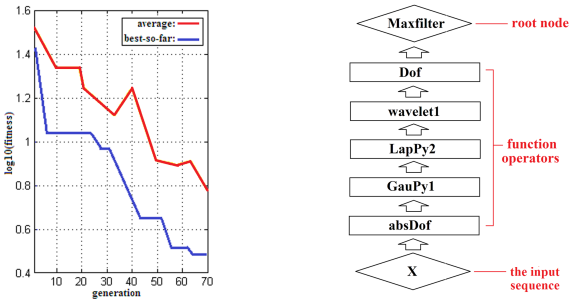


Figure 1: The outline of our proposed method



(a) (b)

Figure 2: (a) Evolutional average and best-so-far values of fitness (b) Tree-based genomic structure for the best-so-far program

We calculate the average error rate as the fitness value and obtain an accuracy of 96.9% for the GP-generated spatio-temporal descriptor. Fig. 2(a) shows the evolution of the average and best-so-far fitnesses. The final descriptor is illustrated in Fig. 2(b).

To demonstrate the generalizability of our method, the descriptor has further been tested on the more challenging IXMAS dataset [4] (composed of eleven human daily actions performed by ten actors and recorded from five different viewpoints.) The accuracy is 93.6% for multi-view fusion. We observe that our method achieves improvements and significantly outperforms previous work. The details can be seen in our paper.

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *International Conference on Computer Vision*, volume 2, pages 1395–1402, Beijing, China, 2005.

[2] R. Poli, W.B. Langdon, and N.F. McPhee. *A field guide to genetic programming*. Published via `http://lulu.com` and freely available at `http://www.gp-field-guide.org.uk`, 2008.

[3] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *International Conference on Pattern Recognition*, volume 3, pages 32–36, Cambridge, UK, 2004.

[4] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3D exemplars. In *IEEE 11th International Conference on Computer Vision*, pages 1–7, Rio de Janeiro, Brazil, 2007.

[1]The KTH dataset contains six types of human action examples (boxing, handwaving, handclapping, jogging, running and walking) performed by 25 different subjects with four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. From http://www.nada.kth.se/cvap/actions/.

[2]The Weizmann dataset contains ten actions types (bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2) performed by nine different subjects

Table 1: The parameter settings for GP

Population Size: 100	Generation: 70	Crossover Rate: 90% Mutation Rate: 10%
Selection for Reproduction: 'Lexictour'	Survival Method: 'Keepbest'	Stopping Conditions: Equal or lower than 2% of error rate

Racing Bib Numbers Recognition

Idan Ben-Ami
idan.benami@gmail.com
Tali Basha
talib@eng.tau.ac.il
Shai Avidan
avidan@eng.tau.ac.il

School of Electrical Engineering,
Tel Aviv University,
Tel Aviv 69978, Israel

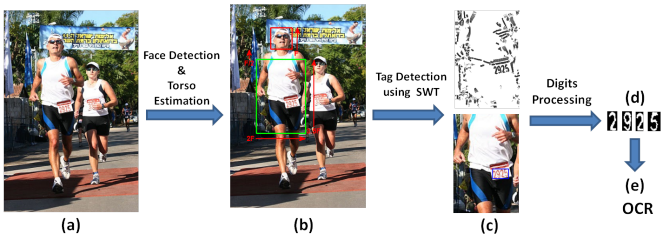


Figure 1: **Method Outline: (a)** the input image; **(b)** face detection results in red; the hypothesis estimated region of the RBN in green; **(c)** the stroke-width map of the hypothesis region (top) and the detected tag (bottom); **(d)** the detected tag after processing is fed to the OCR **(e)**.

Running races, such as marathons, are broadly covered by professional as well as amateur photographers. This leads to a constantly growing number of photos covering a race, making the process of identifying a particular runner in such datasets difficult. Today, such identification is often done manually. In running races, each competitor has an identification number, called the Racing Bib Number (RBN), used to identify that competitor during the race. RBNs are usually printed on a paper or cardboard tag and pined onto the competitor's T-shirt during the race. We introduce an automatic system that receives a set of natural images taken in running sports events and outputs the participants' RBN.

This specific application can be studied in the wider context of detecting and recognizing text in natural images of unstructured scenes. Existing methods that fall into this category fail to reliably recognize RBNs (as demonstrated in our experiments), due to the large variability in their appearance, size, and the deformations they undergo. The RBNs usually cover only a small portion of the image and are surrounded by complex backgrounds. Moreover, the images often contain irrelevant text such as sponsor billboards, signs, or text printed on people's clothes. Therefore, text detection methods are expected to be inefficient and to produce many false detections. This is demonstrated in Fig. 2(a)-(c), showing the results of applying SWT [1] on the entire image.

In this paper, we propose a method specifically designed for RBN recognition. Our method can be applied without any adjustments to images taken at various running races by different photographers. We show that by using prior information - the spatial relation between a person's face and the tag he/she wears, we obtain an effective RBNR system that outperforms text detection methods and state of-the-art commercial LPR software.

Method Outline: The input to our method is a collection of images covering a running race. The images are generally different in size, resolution, and viewpoint, and may be taken using different cameras. Each image is assumed to capture one or more participants. The RBN tags are allowed to have any color, font and scale. The only assumption is that the RBN tag is located on the front torso of the participant. The outline of our method is shown in Fig. 1. First, we use a face detector [2] to generate hypotheses regarding the RBN location and scale. We then adapt and enhance the stroke width transform (SWT) [1] to detect the location of the tag, which is then processed and fed to a standard optical character recognition (OCR) engine [3].

Results in a Glance: To test our method, we collected images from running races found on the Web. Our database includes 217 color images divided into three sets, each taken from a different race. The tag dimensions vary between 13x28–120x182 pixels while digit stroke widths vary from 13 pixels to as few as 2 pixels in the smallest tags. To verify and compare our results, we manually generated the ground truth RBNs.

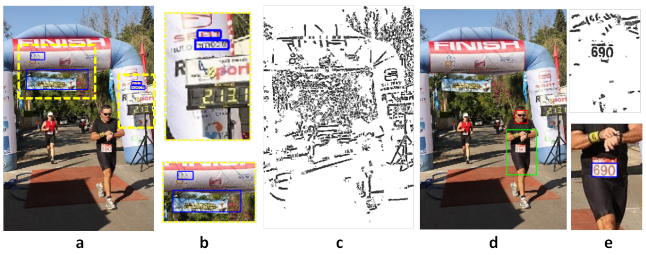


Figure 2: **(a)** The result of applying SWT on the entire image; the detected text regions are marked in blue inside the yellow dashed regions; **(b)** zoom-in of the yellow dashed regions; **(c)** the stroke width map computed on the entire image; **(d)** the detected face in red, and the hypothesis tag region in green;

Comparison to Conventional Approaches: In this experiment, we evaluated the contribution of each component in our pipeline. To do so, we compared our full pipeline (**"Face Detector+E-SWT+OCR"**) to the following two sub-systems: (1) **"SWT+OCR":** the SWT is applied on the entire image to locate the RBNs, followed by standard OCR on the detected text regions; (2) **"Face Detector+E-SWT+OCR":** the face detector is added to generate hypothesis search regions. The SWT combined with our enhancements (noted by E-SWT) is applied on the hypothesis regions, followed by standard OCR. Fig. 3 presents the results of the two sub-systems compared to our full pipeline on one of the datasets.

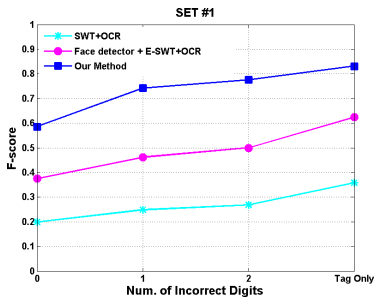


Figure 3: The graph show the computed F-score for each dataset, for the cases of perfect detection, one wrong digit, two wrong digits, and only tag (the tag area is correctly detected but with more than 2 wrong digits).

Comparison to LPR: We compared our performance to the CARMEN FreeFlow LPR system. CARMEN is a leading, commercial, general-purpose LPR system, that provides high rate plate recognitions in a large variety of image scenes and plate types. To adapt CARMEN for our purpose, a special parameter adjustment was required. The performance of CARMEN on our datasets is shown in Table 1. For each dataset, the precision, recall and F-score are measured for perfectly correct recognition (i.e., all digits are correct). The results indicate that our system achieved higher F-score than CARMEN (we achieved higher recall than CARMEN and similar precision rate).

[1] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. *CVPR*, 2010.

[2] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900. Ieee, 2002.

[3] R. Smith. An overview of the tesseract ocr engine. *ICDAR*, 2007.

	Set #1			Set #2			Set #3		
	Prec.	Rec.	F	Prec.	Rec.	F	Prec.	Rec.	F
Our method	**0.66**	**0.50**	**0.57**	**0.75**	**0.45**	**0.56**	**0.65**	**0.62**	**0.63**
CARMEN	0.67	0.37	0.48	0.68	0.38	0.49	0.73	0.47	0.57

Table 1: The computed precision, recall and F-score (w.r.t the ground truth) of our results compared with the results of CARMEN LPR system.

Learning Discriminative Chamfer Regularization

Pradeep Yarlagadda *
pradeep.yarlaga@iwr.uni-heidelberg.de
Angela Eigenstetter *
aeigenst@iwr.uni-heidelberg.de
Björn Ommer
ommer@uni-heidelberg.de

Interdisciplinary Center for Scientific Computing (IWR)
University of Heidelberg
Germany

Chamfer matching is an effective and widely used technique for detecting objects or parts thereof by their shape. However, a serious limitation is its susceptibility to background clutter. The primary reason for this is that the presence of individual model points in a query image is measured independently. A match with the object model is then represented by the sum of all the individual model point distance transformations. Consequently, i) all object pixels are treated as being independent and equally relevant, and ii) the model contour (the foreground) is prone to accidental matches with background clutter.

As demonstrated by Attneave [1], and various experiments on illusionary contours, object boundary pixels are not all equally important due to their statistical interdependence. Moreover, in dense background clutter the points on the model have a high likelihood to find good spurious matches [1, 3]. However, any arbitrary model would match to such a cluttered region, which consequently gives rise to matches with high accidentalness. Chamfer matching only matches the template contour and thus fails to discount the matching score by the accidentalness, i.e., the likelihood that this is a spurious match.

We take account of the fact that boundary pixels are not all equally important by applying a discriminative approach to chamfer distance computation, thereby increasing its robustness. Let $T = \{\mathbf{t}_i\}$ and $Q = \{\mathbf{q}_j\}$ be the sets of template and query edge map respectively. Let $\phi(\mathbf{t}_i)$ denote the edge orientation of the edge point $\mathbf{t}_i$. For a given location $\mathbf{x}$ of the template in the query image, directional chamfer matching [2] finds the best $\mathbf{q}_j \in Q$ for each $\mathbf{t}_i \in T$, thus resulting in a matching cost $p_i^{(T,Q)}(\mathbf{x})$.

$$p_i^{(T,Q)}(\mathbf{x}) = \min_{\mathbf{q}_j \in Q} |(\mathbf{t}_i + \mathbf{x}) - \mathbf{q}_j| + \lambda |\phi(\mathbf{t}_i + \mathbf{x}) - \phi(\mathbf{q}_j)| \quad (1)$$

Adjacent template pixels are statistically dependent and, thus, we do average (1) over the direct neighbors of pixel i. The resulting $\bar{p}_i$ are then used to learn the importance of contour pixels.

While learning the weights for individual pixels improves the robustness of template matching, chamfer matching is still prone to accidental responses in spurious background clutter. To estimate the accidentalness of a match, a small dictionary of simple background contours T_{bg} is utilized. Rather than placing background contours at a fixed single location, i.e., at the center of the model contour as in [3], background elements are trained to focus at locations where, relative to the foreground, typically accidental matches occur.

Let $d_{DCM}^{(T,Q)}(\mathbf{x})$ denote the directional chamfer distance between Q and T with a relative displacement $\mathbf{x}$. To measure where clutter typically interferes with the model contour we compute $d_{DCM}^{(T_{bg},T)}$ between each background contour T_{bg} and the object template T. We consider placements of the background contour with better (lower) chamfer matching score to be more important since they occur on or close to the model contour. In order to weight these matching locations higher we create a mask $M^{(T_{bg},T)}(\mathbf{x})$

$$M^{(T_{bg},T)}(\mathbf{x}) = 1 - d_{DCM}^{(T_{bg},T)}(\mathbf{x}) \quad (2)$$

To describe the background matching costs for a hypothesis in a robust way we build weighted histograms over chamfer matching scores $d_{DCM}^{(T_{bg},Q)}$ obtained from matching a background contour T_{bg} with the query image Q. Let $B(\bar{\mathbf{x}})$ be the bounding box region with center $\bar{\mathbf{x}}$ for a specific placement of the foreground template T in the query image Q. For each foreground hypothesis we build weighted histograms $h^{(T_{bg},Q)}$ over the directional chamfer matching scores $d_{DCM}^{(T_{bg},Q)}$ in the corresponding bounding box region. The weights introduced in (2) are used to weight the histogram votes. Therefore chamfer matching scores $d_{DCM}^{(T_{bg},Q)}$ are weighted

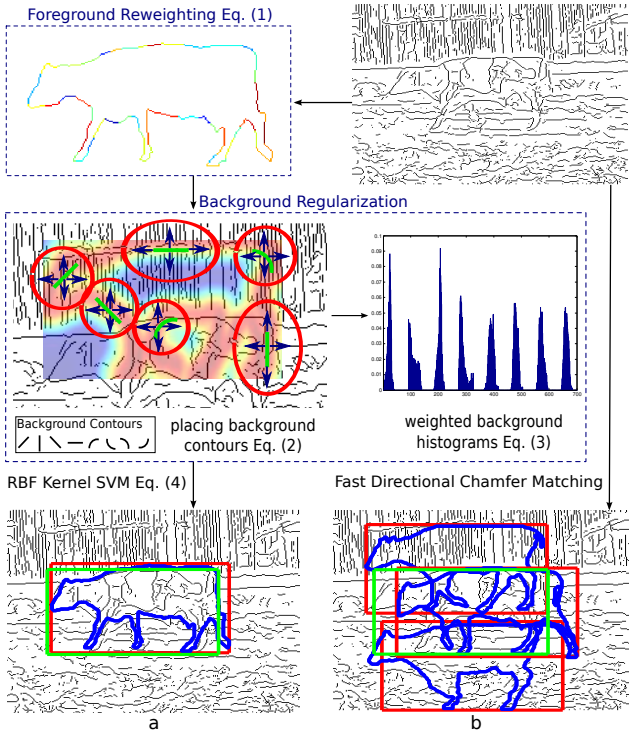


Figure 1: Comparison of a) reguarized chamfer matching with b) directional chamfer matching.

according to their position relative to the foreground template. Each histogram consists of K bins where $\mathcal{M}_k$ is the range of the kth bin and $k = 1, ..., K$. A histogram bin $h_k^{(T_{bg},Q)}$ is defined as

$$h_k^{(T_{bg},Q)} = \sum_{\substack{\mathbf{x} \in B(\bar{\mathbf{x}}) \\ d_{DCM}^{(T_{bg},Q)}(\mathbf{x}) \in \mathcal{M}_k}} M^{(T_{bg},T)}(\mathbf{x}), \quad (3)$$

for each background contour T_{bg} on a certain position of the foreground template T in the query image Q.

For each object hypothesis we build a feature vector $f_i = [\bar{p}_1 \ldots \bar{p}_L\, h_1 \ldots h_$ consisting of the average pixel cost $\bar{p}_i$ and the corresponding background histograms h_i, where L is the number of template edge pixels and G is the number of background contours.

Finally, a max-margin classifier is employed to learn the co-placement of all background contours and the foreground template. This classifier yields a regularized distance function d_{RDCM}

$$d_{RDCM}^{(T,Q)}(\mathbf{x}) = 1 - \Big(\sum_i \alpha_i \mathcal{K}(f_j, S_i) + b\Big). \quad (4)$$

$\mathcal{K}$ denotes the kernel used in the SVM. b denotes the offset. S_i, α_i denotes the support vectors and their respective coefficients.

Our approach is easily integrated into an off-the-shelf directional chamfer matching approach and it shows significant improvements over state-of-the-art chamfer matching on standard benchmark datasets. The qualitative and quantitative results are detailed in the paper.

[1] F. Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183–193, 1954.

[2] M. Liu, O. Tuzel, A.Veeraraghavan, and R. Chellappa. Fast directional chamfer matching. *CVPR*, 2010.

[3] T. Ma, X. Yang, and L.Latecki. Boosting chamfer matching by learning chamfer distance normalization. *ECCV*, 2010.

* Both authors contributed equally to this work.

Feature Mining for Localised Crowd Counting

Ke Chen[1]
cory@eecs.qmul.ac.uk
Chen Change Loy[2]
ccloy@visionsemantics.com
Shaogang Gong[1]
sgg@eecs.qmul.ac.uk
Tao Xiang[1]
txiang@eecs.qmul.ac.uk

[1] School of Electronic Engineering and Computer Science
Queen Mary, University of London
London E1 4NS, UK

[2] Vision Semantics
London E1 4NS, UK

Crowd counting in public places has a wide spectrum of applications especially in crowd control, public space design, and pedestrian behaviour profiling. Existing counting by regression methods, which aim to learn a direct mapping between low-level features and people count without segregation or tracking of individuals, can be categorised into either global approaches or local approaches. Global approaches [1, 3, 4] learn a single regression function between image features extracted globally from the entire image space and the total people count in that image space. Since spatial information is lost when computing global features, such a model assumes implicitly that a feature should be weighted the same regardless where in the scene it is extracted. However, this assumption is largely invalid in real-world scenarios. To overcome these limitations of a global approach, local models [5, 7] aim to relax the global assumption to certain extent by dividing the image space into cell regions, each of which modelled by a separate regression function. However, existing local methods suffer a scalability issue due to the need to learn multiple regression models, the number of which can become very large. In addition, an inherent drawback of existing local models is that no information is shared across spatially localised regions in order to provide a more context-aware feature selection for more accurate crowd counting.

We consider that *localised feature importance mining* and *information sharing among regions* are two key factors for accurate and robust crowd counting, which are missing in all existing techniques. To this end, we propose a single multi-output model for joint localised crowd counting based on ridge regression [6], which takes inter-dependent local features from local spatial regions as input and people count from individual regions as multi-dimensional structured output. Unlike global regression methods, our model relaxes the one-to-one mapping assumption by learning spatially localised regression functions jointly in a single model for all the individual cell regions in a scene, as such our model can capture feature importance locally. Unlike existing approaches to building multiple local regression models, our single model is learned by joint optimisation to enforce dependencies among cell regions. Therefore information from all local spatial regions can be shared to achieve more reliable count prediction.

Figure 1 gives an overview of our framework: (Step-1) We first infer a perspective normalisation map using the method described in [2]. (Step-2) Given a set of training images, we extract low-level imagery features, including local foreground, edges and texture features, from each cell region. (Step-3) Local features from each cell are used to construct a local intermediate feature vector before all local intermediate feature vectors are concatenated into a single ordered (location-aware) feature vector. (Step-4) A multi-output regression model based on multivariant ridge regression is trained using the single concatenated feature vector and the vector, each element being actual count in each region, as a training pair. Given a new test frame, features are extracted and mapped to the learned regression model for generating a structured output that estimates the crowd count in each local region simultaneously.

For a training video frame i, where $i = 1,2\ldots N$ and N denotes the total number of training frames, we first partition the frame into K cell regions (see Step-3 in Figure 1). We then extract low-level imagery features $\mathbf{z}_i^j$ from each cell region j and combine them into an intermediate feature vector $\mathbf{x}_i \in \mathbb{R}^d$. We also concatenate the localised labelled ground truth u_i^j from each cell region into a multi-dimensional output vector, $\mathbf{y}_i \in \mathbb{R}^m, i = 1,2\ldots N$

$$\mathbf{x}_i = [\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^{K-1}, \mathbf{z}_i^K], \quad \mathbf{y}_i = [u_i^1, u_i^2, \ldots, u_i^{K-1}, u_i^K].$$

Let $(\mathbf{x}_i, \mathbf{y}_i)$ be the observation and target vectors, multivariate ridge regression can be presented as follows

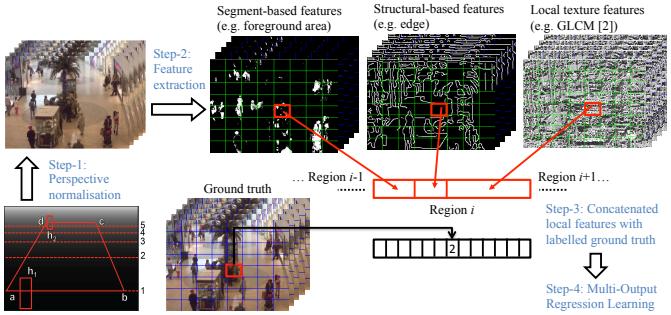


Figure 1: A multi-output regression framework for localised crowd counting by feature mining.

$$\min \quad \frac{1}{2}\|\mathbf{W}\|_F^2 + C\sum_{i=1}^{N}\|\mathbf{y}_i^T - \mathbf{x}_i^T\mathbf{W} - \mathbf{b}\|_F^2, \qquad (1)$$

where $\mathbf{W} \in \mathbb{R}^{d\times K}$ and $\mathbf{b} \in \mathbb{R}^{1\times K}$ denote a weight matrix and a bias vector respectively. The $\|\cdot\|_F$ denotes the Frobenius-norm, and C is a parameter that controls the trade-off between the penalty and the fit. The weight matrix $\mathbf{W}$ plays an important role in capturing the local feature importance and facilitating the sharing of features. In particular, for each localised cell, we formulate our model to jointly weigh the features extracted from both the corresponding localised cell and other cell regions in the image. Owing to its inbuilt ability for feature mining according to changing crowd conditions presented in different local spatial cell regions in the scene, our model outperforms multiple localised regressors and also compares favourably against existing single global regressor based crowd counting models on existing UCSD benchmark dataset and a new more challenging shopping mall dataset.

[1] A.B. Chan and N. Vasconcelos. Counting people with low-level features and Bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177, 2012.

[2] A.B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: counting people without people models or tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[3] S.Y. Cho, T.W.S. Chow, and C.T. Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(4):535–541, 1999.

[4] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In *British Machine Vision Conference*, 2005.

[5] W. Ma, L. Huang, and C. Liu. Crowd density analysis using co-occurrence texture features. In *International Conference on Computer Sciences and Convergence Information Technology*, pages 170–175, 2010.

[6] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *International Conference on Machine Learning*, pages 515–521, 1998.

[7] X. Wu, G. Liang, K.K. Lee, and Y. Xu. Crowd density estimation using texture analysis and learning. In *IEEE International Conference on Robotics and Biomimetics*, pages 214–219, 2006.

Super-Resolution from Corneal Images

Christian Nitschke[1]
christian.nitschke@cmc.osaka-u.ac.jp
Atsushi Nakazawa[1,2]
nakazawa@cmc.osaka-u.ac.jp

[1] Cybermedia Center, Osaka University
Toyonaka, Osaka, Japan

[2] PRESTO, Japan Science and Technology Agency (JST)
Kawaguchi, Saitama, Japan

The cornea of the human eye reflects the light from a person's environment. Modeling corneal reflections from an image of the eye enables a number of applications, including the computation of scene panorama and 3D model, together with the person's field of view and point of gaze [4]. The obtained environment map enables general applications in vision and graphics, such as face reconstruction, relighting [3] and recognition [5]. In reality, however, even if we use a carefully-adjusted high-resolution camera in front of the eye, the quality of corneal reflections is limited due to low resolution and contrast, iris texture and geometric distortion.

This paper introduces an approach to overcome these issues through a super-resolution (SR) [6] strategy for corneal imaging that reconstructs a high-resolution (HR) scene image from a series of lower resolution (LR) corneal images such as occurring in surveillance or personal videos. The process comprises (1) single image environment map recovery, (2) multiple image registration, and (3) HR image reconstruction. This is also the first non-central catadioptric approach for multiple image SR.

Corneal reflection modeling. We apply a common geometric eye model, where eyeball and cornea (Figure 1 (a)) are approximated as two overlapping spherical surfaces. A simple strategy assuming weak perspective projection recovers the pose of the model by reconstructing the pose of the circular iris from its elliptical projected contour (Figure 1 (b)).

A corneal image is transformed into a spherical environment map by calculating the intersection and reflection at the corneal surface. Since the eye model only approximates the true corneal geometry, it is not possible to obtain an accurate registration for the whole environment map. Instead, we assume spherical curvature for a user-defined region of interest, where we project the environment map to a local tangent plane (Figure 1 (c)).

Registration further requires the forward projection from the tangent plane into the image. As common iterative methods are not feasible to handle the large number of re-projections, we apply a recent analytic method that requires solving a 4th-order polynomial equation (for the case of a spherical mirror), that is calculated in closed form [1].

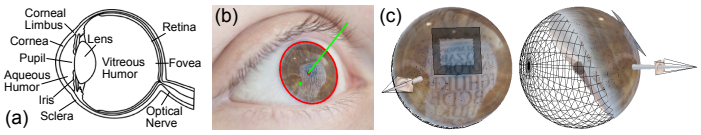


Figure 1: Modeling. (a) Cross-section of the eye. (b) Cropped eye image with iris contour, center (red), cornea center and gaze direction (green). (c) Environment map and local tangent plane at region of interest.

Registration. Regarding the small change of corneal sphere locations in continuous video frames, it is feasible to assume the cornea to be centered at the world origin, where the task of alignment amounts to finding the pose of the camera w.r.t. the world frame. This is achieved through a multiple-step iterative process: (1) Coarse alignment is carried out using at least two feature correspondences for each LR image (Figure 2 (b)-(d)). The transformation between the environment maps is a rotation around the origin that we estimate by minimizing the deviation of backprojected feature directions. To compensate for the error in eye pose estimation we continue adjusting corneal sphere locations through a pairwise and bundle registration. (2) Fine alignment is carried out through image matching in the local tangent plane (Figure 3 (d)), by minimizing the sum of absolute differences (SAD) at uniform sampling points using forward projection lookup. Finally, the remaining misalignment is corrected through a 2D subpixel rigid registration in the plane.

Super-resolution. Using the registration parameters we back project the region of interest for all corneal images and apply the rigid alignment. The obtained (LR) points represent non-uniform samples (observations) of an unknown HR image. The image is estimated through a MAP (maximum a posteriori) based SR approach using gradient descent to minimize the error between the observed and synthesized LR images under a Gaussian PSF assumption. As image priors, we evaluate the norm of the HR image filtered by either a Laplacian of Gaussian filter (LoG), a bilateral filter residual (BL) [7] or a bilateral total variation filter (BTV) [2].

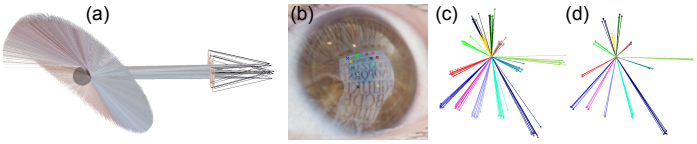


Figure 2: Alignment. (a) Cameras, image planes, corneal sphere (world origin), and limbus backprojection rays for 10 images. (b) Single corneal image with 13 feature correspondences. (c) Back-projected features for all images before alignment, and (d) after coarse alignment.

Experiments. In a number of experiments for indoor and outdoor scenes we confirmed that the strategy using MAP-BL and -BTV performs best and recovers high-frequency textures (with a quality high enough to recognize small characters, human faces and fine structures) that are lost in the source images (Figure 3). We also confirmed this for a spherical mirror, suggesting applicability to other non-central catadioptric systems such as specular and liquid surfaces in everyday environments.

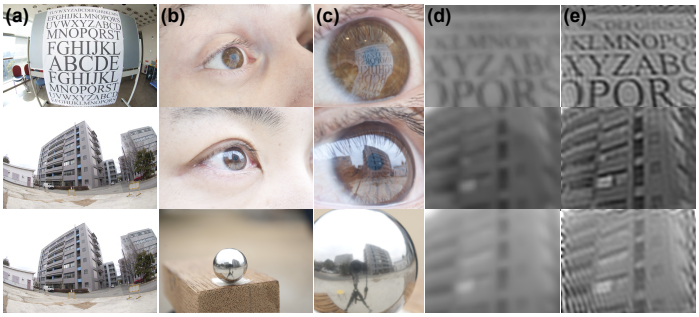


Figure 3: SR result for two scenes. (a) Fisheye scene image. (b) Single LR image. (c) Cropped eye/mirror region. (d) Local tangent plane projection. (e) Multiple image SR result using MAP with BTV prior.

Since this solves the quality degradation problem in corneal imaging techniques, we believe our contribution can become a foundation for future applications in this research category. The obtained information about a person and the environment has the potential to enable novel applications, *e.g.*, for surveillance systems, personal video, human-computer interaction, and upcoming head-mounted cameras.

[1] A. Agrawal, Y. Taguchi, and S. Ramalingam. Beyond Alhazen's problem: Analytical projection model for non-central catadioptric cameras with quadric mirrors. In *Proc. CVPR*, pages 2993–3000, 2011.

[2] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Trans. Img. Proc.*, 13(10):1327–1344, 2004.

[3] K. Nishino and S. K. Nayar. Eyes for relighting. In *Proc. SIGGRAPH*, pages 704–711, 2004.

[4] K. Nishino and S. K. Nayar. Corneal imaging system: Environment from eyes. *Int. J. Comput. Vision*, 70(1):23–40, 2006.

[5] K. Nishino, P. N. Belhumeur, and S. K. Nayar. Using eye reflections for face recognition under varying illumination. In *Proc. ICCV*, pages 519–526, 2005.

[6] J. Tian and K.-K. Ma. A survey on super-resolution imaging. *Signal Image Video Process.*, 5(3):329–342, 2011.

[7] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. ICCV*, pages 839–846, 1998.

Real-time Learning and Detection of 3D Texture-less Objects: A Scalable Approach

Dima Damen[1]
damen@cs.bris.ac.uk
Pished Bunnun[2]
pished.bunnun@nectec.or.th
Andrew Calway[1]
andrew@cs.bris.ac.uk
Walterio Mayol-Cuevas[1]
wmayol@cs.bris.ac.uk

[1] University of Bristol
Bristol, UK
[2] National Electronics and Computer Technology Center
Bangkok, Thailand

We present a method for the learning and detection of multiple rigid texture-less 3D objects intended to operate at frame rate speeds for video input. The method is geared for fast and scalable learning and detection by combining tractable extraction of edgelet constellations with library lookup based on rotation- and scale-invariant descriptors. Most shape-based detectors either require offline training or scale linearly as more objects are being searched for, or commonly both. To address speed and scalability in learning and testing, this paper proposes the use of predefined paths that specify the relative direction between edgelets, and importantly, make the search tractable for real-time operation. The traced edgelets are represented by a simple to compute transformation invariant descriptor, that is used as an index to similarly stored descriptors, in a way that revisits geometric hashing. The approach learns object views in real-time, and is generative - enabling more objects to be learnt without the need for re-training. During testing, a random sample of edgelet constellations is tested for the presence of known objects.

A key and distinguishing element of the method is the use of *path tracing* for both training and testing (Fig. 1). Each path defines the relative direction between the constellation's constituent edgelets. This introduction of paths is critical; as it limits the number of possible constellations and allows tractable generation of a library of descriptors. For a constellation of n edgelets, a **path** Θ is a sequence of angles $\Theta = (\theta_0, .., \theta_{n-2})$. From any starting edgelet, the base angle θ_0 specifies the direction of a tracing vector v_1, initially with unspecified length, relative to the orientation of the starting edgelet. If this tracing vector intersects with another edgelet in the edge map, then the edgelet is added to the constellation.

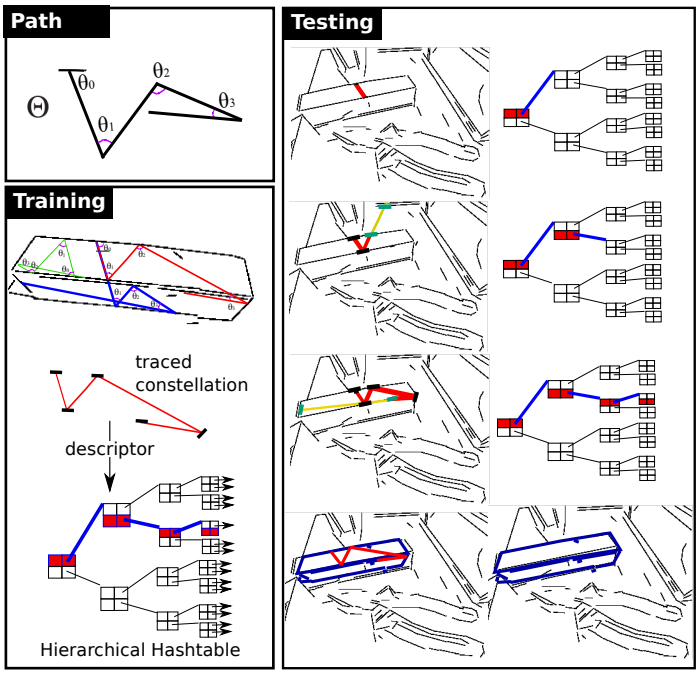


Figure 1: For a given tracing path, constellations of edgelets are traced out from training views exhaustively, and an affine-invariant descriptor for each constellation is inserted into a quantised hierarchical hash table. For a test image, constellations are traced out using the same path. Candidate detections are found by comparing the descriptor to the hash table, tested using all the training edgelets, and refined using iterative closest edgelet.

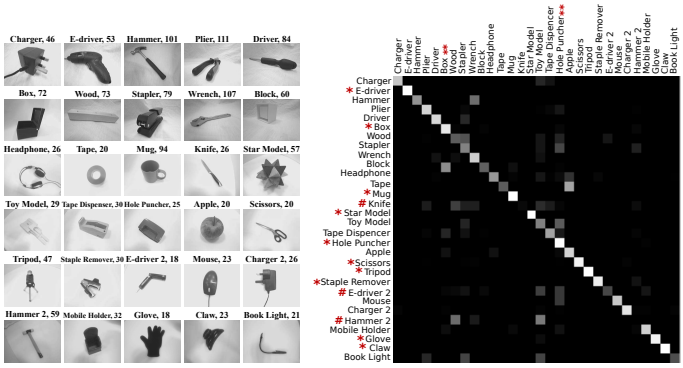


Figure 2: Thirty texture-less objects in the dataset along with the confusion matrix from 10 runs. Ten objects achieved recall > 90% (*), two of them with precision < 90% (**). Three objects are difficult to detect with recall < 30% (#).

The next tracing vector v_2 then has the direction θ_1 relative to v_1, i.e. $cos(\theta_1) = (v_1 \cdot v_2)/(|v_1||v_2|)$. This process continues until the constellation has n edgelets.

For a traced constellation, the descriptor specifies the relative orientations and distances between the consecutive edgelets in the constellation's tuple. By keeping a comprehensive library of descriptors for all constellations guided by one path Θ from all starting edgelets, it is sufficient to extract one constellation using the same path from the object in the test image to produce a candidate detection that is verified using the rest of the view edgelets. Several paths are used and a separate library is built for each chosen path. The choice of paths is discussed in the paper.

The method is tested on a dataset of 30 texture-less objects. It trades recall for speed, testing a sample of edgelet constellations in each processed frame. The method is tested at frame rates varying from 1 to 17 fps. At 7fps, recall of 50% (precision = 74%) was achieved when 30 objects were learnt (1433 views - Fig. 2). As the number of objects in the library increases from 1 to 30, the increase of detection time is dependent on the shape's ambiguity rather than the number of objects.

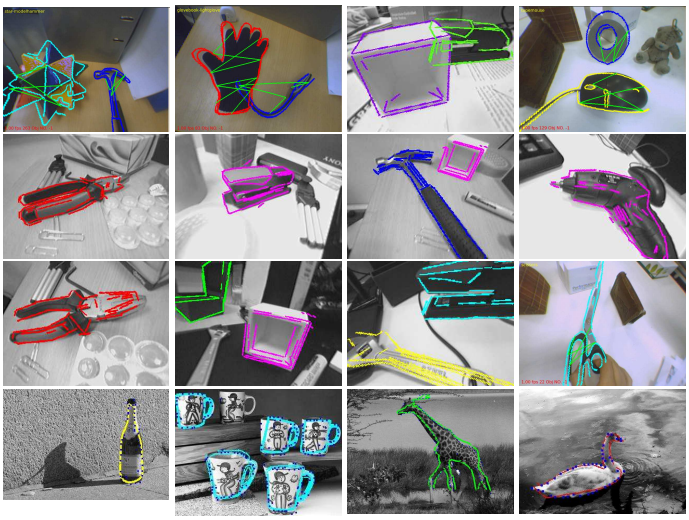


Figure 3: Sample set of results on the tools and ETHZ datasets.

Person Re-identification by Attributes

Ryan Layne
rlayne@eecs.qmul.ac.uk
Timothy Hospedales
tmh@eecs.qmul.ac.uk
Shaogang Gong
sgg@eecs.qmul.ac.uk

Queen Mary Vision Laboratory,
School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London, E1 4NS, U.K.

Automatic re-identification of a human candidate from public space CCTV video is challenging due to spatiotemporal visual feature variations and strong visual similarity between different people, low-resolution and poor quality video data. In this work, we propose a novel method for re-identification that learns a selection and weighting of mid-level semantic attributes to describe people. The model learns an attribute-centric, parts-based feature representation which differs from and complements existing low-level features for re-identification that rely purely on bottom-up statistics for feature selection.

We are motivated by the operating procedures of human experts and recent research in attribute learning to introduce a new class of mid-level *attribute* features. When performing person re-identification, human experts tend to seek and rely upon matching attributes appearance or functional attributes that are unambiguous in interpretation, such as hair-style, shoe-type or clothing-style. Some of these mid-level attributes can be measured reasonably reliably with modern computer-vision techniques, and hence provide a valuable additional class of features which has thus far not been exploited for re-identification. These attributes provide a very different source of information – effectively a separate modality – to the typical low-level features used.

We make three main contributions: (i) We introduce and evaluate an ontology of useful attributes from the subset of attributes used by human experts which can also be relatively easily measured by bottom-up low-level features computed using established computer vision methods. (ii) We show how to select those attributes that are most effective for re-identification and how to fuse the attribute-level information with standard low-level features. (iii) We show how the resulting synergistic approach – Attribute Interpreted Re-identification (AIR) – obtains state of the art re-identification performance on two standard benchmark datasets.

We first extract a low-level colour and texture feature vector denoted $\mathbf{x}$ from each person image I following the method in [3]. This consists of 464-dimensional feature vectors extracted from the image. Each vector uses 8 colour channels (RGB, HSV and YCbCr) and 21 texture filters (Gabor, Schmid) derived from the luminance channel.

We train Support Vector Machines (SVM) to detect attributes and select the Intersection kernel as it compares closely with χ^2 but can be trained much faster. For each attribute, we perform cross validation to select values for SVM slack parameter C from the set $C \in \{-2,\ldots,5\}$ with increments of $\varepsilon = 1$. The SVM scores are probability mapped, so each attribute detector i outputs a posterior $p(a_i|\mathbf{x})$ for that attribute.

Given the learned bank of attribute detectors, any person image can now be represented in a semantic attribute space by an N_a dimensional vector: $A(\mathbf{x}) = [p(a_1|\mathbf{x}_i^+),\ldots,p(a_{N_a}|\mathbf{x}_{N_a}^+)]^T$.

We investigate how attributes can be fused to enhance performance and we choose to build on *Symmetry-Driven Accumulation of Local Features* (SDALF), introduced by Farenzena *et al.* [1]. SDALF provides a low-level feature and Nearest Neighbour (NN) matching strategy giving state-of-the-art performance for a non-learning NN approach, as well as a compatible fusion method capable of admitting additional sources of information. Farenzena *et al.* introduce features from which seperate distance metrics can be constructed. These distance metrics are combined in order to obtain the distance d between two particular person images I_p and I_q. Within this nearest neighbour strategy, we can integrate our attribute-based distance d_{ATTR} as follows:

$$d(I_p,I_q) = \beta_{WH}\cdot d_{WH}(WH(I_p),WH(I_q)) \quad (1)$$
$$+ \beta_{MSCR}\cdot d_{MSCR}(MSCR(I_p),MSCR(I_q)) \quad (2)$$
$$+ \beta_{RHSP}\cdot d_{RHSP}(RHSP(I_p),RHSP(I_q)) \quad (3)$$
$$+ \beta_{ATTR}\cdot d_{ATTR}(ATTR(I_p),ATTR(I_q)). \quad (4)$$

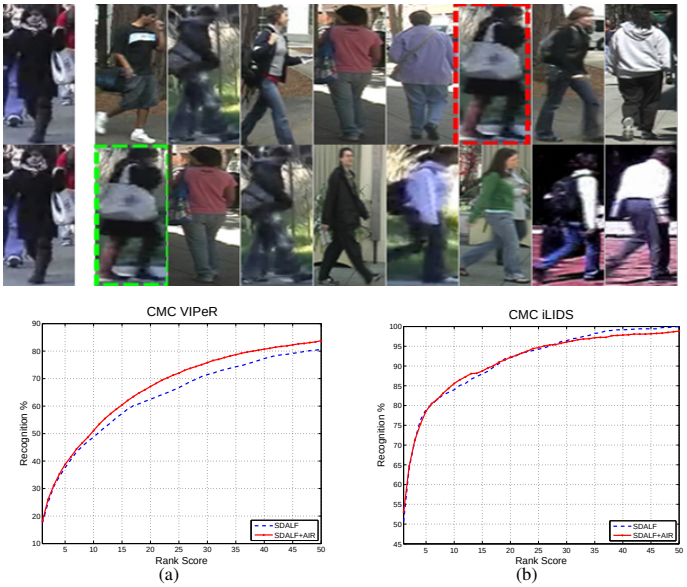


Figure 1: Above: Illustrative result for AIR (green) and SDALF (red); Below: CMC curves for (a) VIPeR ($p = 250$) ; and (b) i-LIDS ($p = 60$).

Here Eqs. (1-3) correspond to the three SDALF distance measures and Eq. (4) fuses our attribute-based distance distance metric. WH, $MSCR$ and $RHSP$ represent the metrics calculated for each of the separate SDALF features using Bhattacharyya.

To compare images' semantic attribute representation, we learn an $L2$-norm distance metric, d_{ATTR}. For diagonal weight matrix W:

$$d_{ATTR}(I_p,I_q) = \left(A(\mathbf{x}_p)-A(\mathbf{x}_q)\right)^T W \left(A(\mathbf{x}_p)-A(\mathbf{x}_q)\right), \quad (5)$$
$$= \sqrt{\sum_i w_i \left(p(a_i|\mathbf{x}_{p,i}^+) - p(a_q|\mathbf{x}_{q,i}^+)\right)^2}. \quad (6)$$

Searching the N_a dimensional space of weights directly to determine attribute selection and weighting is computationally intractable. We therefore employ a greedy search which selects and weighs attributes to maximally improve the re-identification rate.

The re-identification performance of our complete system is summarised in Figure 1. In each case, our AIR outperforms vanilla SDALF [1], (which in turn decisively outperforms [2]). At the important rank 1 (perfect match), we obtain a relative improvement over SDALF of 3.2% and 4.8% for VIPeR and iLIDS respectively.

The proposed attribute-centric re-identification model provides an important contribution and novel research direction for practical re-identificatior both by providing a complementary and informative mid-level cue, as well as by opening up completely new applications via the interpretable semantic representation. As a novel application, consider how semantic attributes could potentially be used to constrain or permute a search for a particular person, for example by specifying invariance to whether or not they have removed or added a hat.

[1] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2010.

[2] Douglas Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *European Conference on Computer Vision*, pages 262–275, 2008.

[3] Bryan Prosser, Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Person Re-Identification by Support Vector Ranking. In *British Machine Vision Conference*, 2010.

A Closed Form Solution for the Self-Calibration of Heterogeneous Sensors

Alessio Del Bue
alessio.delbue@iit.it
Marco Crocco
marco.crocco@iit.it
Igor Barros Barbosa
igorbb@gmail.com
Vittorio Murino
vittorio.murino@iit.it

Pattern Analysis & Computer Vision - PAVIS
Istituto Italiano di Tecnologia - IIT
Via Morego, 30, 16163 Genova, Italy

We present a novel closed-form solution for the joint self-calibration of video and range sensors solely from measurements as shown in Fig. 1. The approach single assumption is the availability of synchronous time of flight (i.e., range distances) measurements and visual position of the target on images acquired by a set of cameras. In such case, we make explicit a rank constraint that is valid for both image and range data. This rank property is used to find an initial and affine solution via bilinear factorization, which is then corrected by enforcing the metric constraints characteristic for both sensor modalities (i.e., camera and anchors constraints). The output of the algorithm is the identification of the target/range sensor 3D position and the calibration of the cameras.

Let us consider m range sensors and n point-like targets laying in a 3D space. Assuming no measurement errors, the following equations hold:

$$s_{i1}^2+s_{i2}^2+s_{i3}^2+t_{j1}^2+t_{j2}^2+t_{j3}^2-2s_{i1}t_{j1}-2s_{i2}t_{j2}-2s_{i3}t_{j3}=d_{i,j}^2 \quad (1)$$

for $i=1...m$, $j=1...n$, where s_{il}, t_{jl} and $d_{i,j}$ denote respectively the sensor and target coordinates and the measured distance among them. By centering the sensors and target coordinates to the first sensor and the first target, the six quadratic terms in (1) disappear and a bilinear form can be obtained [1]:

$$-2\tilde{\mathtt{S}}\tilde{\mathtt{T}}=\tilde{\mathtt{D}}. \quad (2)$$

where $\tilde{\mathtt{S}}$, $\tilde{\mathtt{T}}$ and $\tilde{\mathtt{D}}$ matrices have dimension $(m-1)\times 3$, $3\times(n-1)$ and $(m-1)\times(n-1)$ respectively. Analogously, let us consider c affine cameras displaced in 3D space. Assuming an ideal projection of the n targets in the cameras frames, the following equations hold:

$$\mathbf{g}_{jk}=\begin{pmatrix} u_{jk} \\ v_{jk}\end{pmatrix}=\left[\begin{array}{c|c} \mathtt{R}_k & \mathbf{z}_k\end{array}\right]\begin{pmatrix} t_{j1}\\ t_{j2}\\ t_{j3}\\ 1\end{pmatrix}=\mathtt{G}_k\begin{pmatrix} t_{j1}\\ t_{j2}\\ t_{j3}\\ 1\end{pmatrix} \quad (3)$$

for $k=1...c$, $j=1...n$, where u_{kj} and v_{kj} represents the two image coordinates of the target j as seen by camera k. The 2×3 matrix $\mathtt{R}_k$ and the 2-vector $\mathbf{z}_k$ are the parameters of the cameras. By centering the target coordinates to the first target, Eq.(3) can be expressed in a matrix form as:

$$\tilde{\mathtt{G}}=\tilde{\mathtt{C}}\tilde{\mathtt{T}} \quad (4)$$

where matrices $\tilde{\mathtt{G}}$ and $\tilde{\mathtt{C}}$ have dimension $2c\times(n-1)$ and $2c\times 3$ respectively.

The common property for solving jointly the self-calibration problem is that both measured data sussist on a common subspace as defined by the target positions $\tilde{\mathtt{T}}$. The consequence is that the fusion of the modalities is for the first time strictly geometrical, in the sense that the data is now explicitly linked by the metric position of the targets. This leads to the possibility of computing a joint closed form solution using the range-visual constraints of the heterogeneous sensors. In particular Equations (2) and (4) can be merged together obtaining:

$$\mathtt{Y}=\begin{bmatrix}\tilde{\mathtt{D}}\\ \tilde{\mathtt{G}}\end{bmatrix}=\begin{bmatrix}-2\tilde{\mathtt{S}}\\ \tilde{\mathtt{C}}\end{bmatrix}\tilde{\mathtt{T}}. \quad (5)$$

The joint measurement matrix $\mathtt{Y}$ of size $(m+2c-1)\times(n-1)$ has rank equal to three since it is a product between a $(m+2c-1)\times 3$ matrix and a $3\times(n-1)$. If we apply a SVD to $\tilde{\mathtt{Y}}$ we have, in case of no noise, that the singular values after the third are equal to zero. Thus we can truncate these SVD components such as:

$$\mathtt{UVW}=\tilde{\mathtt{Y}}, \quad (6)$$

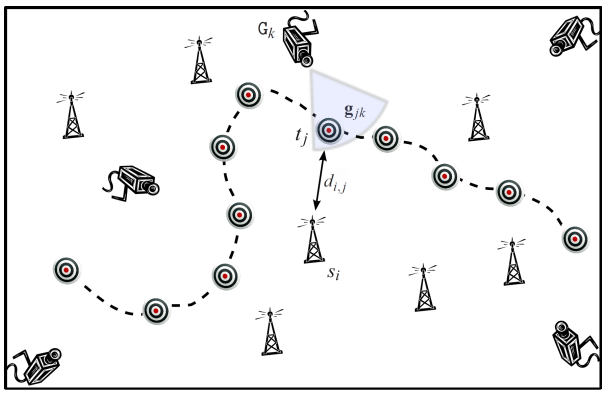


Figure 1: An example of the self-calibration problem for a heterogeneous sensor network. A target with 3D position $\mathbf{t}_j$ is measured by both a range sensor $\mathbf{s}_j$ and a video camera $\mathtt{G}_k$. Using just the scalar range distance d_{ij} from the sensors and the image coordinates of the target $\mathbf{g}_{jk}$ from the cameras, our algorithm recovers the 3D locations of the targets, sensors and it simultaneously calibrates each camera.

where $\mathtt{U}$ is an $(m-1)\times 3$ matrix, $\mathtt{V}$ is a 3×3 diagonal matrix and $\mathtt{W}$ is a $3\times(n-1)$ matrix. In a practical situation, in presence of measurement noise, the rank of $\tilde{\mathtt{Y}}$ will be higher than three: in this case only the three biggest singular values in $\mathtt{V}$ will be considered reducing the size of $\mathtt{U}$, $\mathtt{V}$ and $\mathtt{W}$ according to the noise-free case. From (5) and (6), for every invertible 3×3 matrix $\mathtt{C}$, the following relationships hold:

$$\begin{bmatrix}-2\tilde{\mathtt{S}}\\ \tilde{\mathtt{C}}\end{bmatrix}=\mathtt{UQ}_j \quad \text{and} \quad \tilde{\mathtt{T}}=\mathtt{Q}_j^{-1}\mathtt{VW}.$$

The matrix $\mathtt{Q}_j$ is called the "mixing matrix" since it mixes the components obtained from the SVD in order to obtain the correct solution given the original sensors localization problem. The matrix $\mathtt{Q}_j$ can be found exploiting the linear constraints given by the a priori known positions of a subset of range sensors, named anchors, as well as the quadratic constraints inherent to the affine camera model. We show in the paper that such constraints can be merged together, finding $\mathtt{Q}_j$ as a Cholesky decomposition of a matrix $\mathtt{H}$, whose entries are found through a linear least squares procedure.

The application extent of our approach is broad and versatile. In fact, with the same framework, we can deal with, but not restricted to, two very different applications. The first is aimed at calibrating cameras and microphones deployed in an unknown environment. The second uses a RGB-D device to reconstruct the 3D position of a set of keypoints using the camera and depth map images. Synthetic and real tests show the algorithm performance under different levels of noise and configurations of target locations, number of sensors and cameras. Though geometrical approaches for self calibration of range and video sensors are present in literature as two distinct problems, to the authors knowledge, for the first time we have presented a new geometrical constraint for the fusion of information acquired from video cameras and range sensors.

[1] M. Crocco, A. Del Bue, and V. Murino. A bilinear approach to the position self-calibration of multiple sensors. *IEEE Transactions on Signal Processing*, 60(2):660–673, February 2012.

On Cross-Spectral Stereo Matching using Dense Gradient Features

Peter Pinggera[123]
pinggera@alumni.tugraz.at
Toby Breckon[1]
toby.breckon@cranfield.ac.uk
Horst Bischof[2]
bischof@icg.tugraz.at

[1] School of Engineering, Cranfield University, Bedfordshire, UK.
[2] Institute for Computer Graphics and Vision, TU Graz, Austria.

Here we address the problem of scene depth recovery within cross-spectral stereo imagery (each image sensed over a differing spectral range). We compare several robust matching techniques which are able to capture local similarities between the structure of cross-spectral images and a range of stereo optimisation techniques for the computation of valid dense depth estimates for this case.

As the performance of standard optical camera systems can be severely affected by environmental conditions the use of combined sensing systems operating in differing parts of the electromagnetic spectrum is increasingly common [5]. As a result, an attractive solution is the combination of both optical and thermal images in many sensing and surveillance scenarios as the complementary nature of both modalities can be exploited and the individual drawbacks largely compensated. Despite the inherent stereo setup of this common two sensor deployment, in practical scenarios it is rarely exploited. Here, we specifically deal with the recovery of dense depth information from thermal (far infrared spectrum) and optical (visible spectrum) image pairs where large differences in the characteristics of image pairs make this task significantly more challenging than the common stereo case (Figure 1A).

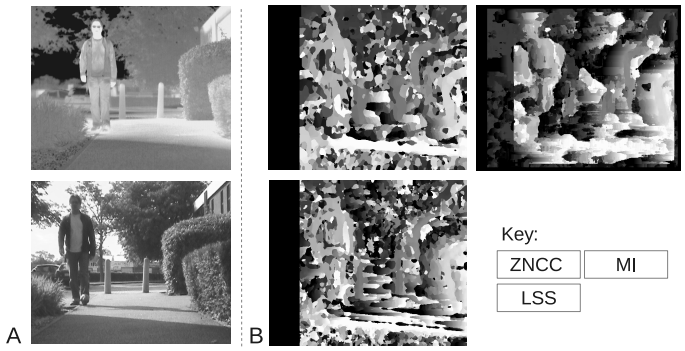


Figure 1: Performance of ZNCC, MI and LSS approaches

Prior work on cross-spectral stereo is weak and either recovers depth from isolated scene objects (Local Self-Similarity (LSS) features, [9]) or relies on an evaluation based on simulated cross-spectral imagery (Mutual Information (MI), [2, 3, 4]). Related work on the problem of radiometric differences in stereo image pairs [6] uses (amongst others) Zero Mean Normalised Cross Correlation (ZNCC). The poor performance of these prior techniques on an example cross-spectral stereo pair (Figure 1A) is shown in Figure 1B.

By contrast, we show cross-spectral stereo matching can be achieved, by using dense gradient features combined with strong optimisation criteria, to produce a scene depth image usable for further scene analysis and understanding (Figure 2). Our approach facilitates full scene depth recovery comparable in quality to standard optical stereo techniques under identical scene conditions.

This extends prior work which is limited to simulated cross-spectral results [2, 3, 4], or isolated object depth recovery [7, 8, 9]. We illustrate that dense gradient feature approaches outperform methods based on prior work using Mutual Information (MI) [2, 3, 4] and Local Self-Similarity (LSS) features [9]. Furthermore, we show that prior results on radiometric image differences [6] or simulated imagery [2, 3, 4] do not readily transfer to the true cross-spectral case.

The prevalence of dense gradient approaches, notably dense unsigned Histograms of Oriented Gradient (HOG) features [1], is shown over a range of disparity optimisation approaches with improved results under strong optimisation criteria of Graph Cuts (GC) and Semi-Global Matching (SGM). Although the results remain somewhat coarse in comparison to contemporary work in optical stereo [6], this work illustrates both :-

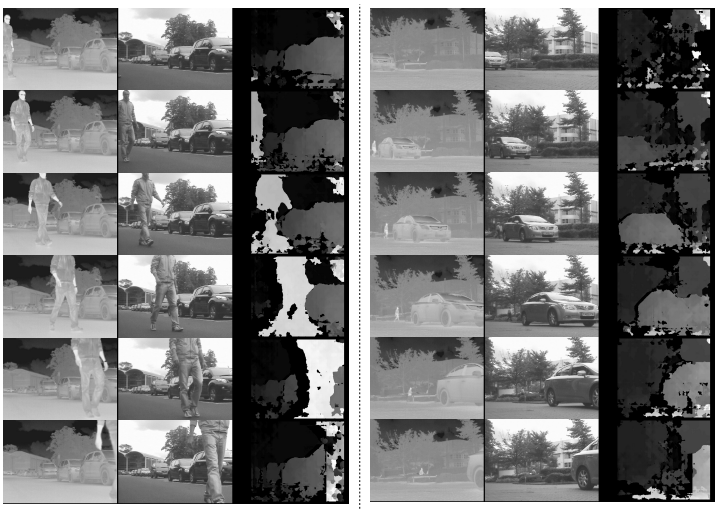


Figure 2: Two cross-spectral stereo sequences obtained using HOG+SGM depth recovery without explicit temporal consistency constraints

a) the additional challenge of cross-spectral stereo in comparison to other stereo matching cases (e.g. radiometric differences [6]) and b) that results suitable for further scene analysis and understanding are achievable via a dense gradient feature approach (Figure 2).

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 886–893. IEEE, 2005.

[2] G. Egnal. Mutual information as a stereo correspondence measure. Technical report, University of Pennsylvania, 2000.

[3] C. Fookes and S. Sridharan. Investigation & comparison of robust stereo image matching using mutual information & hierarchical prior probabilities. In *Proc. Second Int. Conf. on Signal Processing and Communication Systems*, pages 1–10. IEEE, 2008.

[4] C. Fookes, A. Maeder, S. Sridharan, and J. Cook. Multi-spectral stereo image matching using mutual information. In *Proc. Second Int. Symposium on 3D Data Processing, Visualization and Transmission*, pages 961–968. IEEE, 2004.

[5] A. Gaszczak, T.P. Breckon, and J.W. Han. Real-time people and vehicle detection from UAV imagery. In *Proc. SPIE Conf. Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques*, volume 7878, 2011.

[6] H. Hirschmüller and D. Scharstein. Evaluation of stereo matching costs on images with radiometric differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1582–1599, 2009.

[7] S. Krotosky and M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *Computer Vision and Image Understanding*, 106(2-3):270–287, 2007.

[8] S. Krotosky and M. Trivedi. Registering multimodal imagery with occluding objects using mutual information: application to stereo tracking of humans. In R.I. Hammoud, editor, *Augmented Vision Perception in Infrared: Algorithms and Applied Systems*, chapter 14, pages 321–347. Springer, 2009.

[9] A. Torabi and G.-A. Bilodeau. Local self-similarity as a dense stereo correspondence measure for thermal-visible video registration. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 61–67, 2011.

[3] Peter Pinggera is now with Daimler Research and Development, Germany.

Exploiting Relationship between Attributes for Improved Face Verification

Fengyi Song
f.song@nuaa.edu.cn
Xiaoyang Tan
x.tan@nuaa.edu.cn
Songcan Chen
s.chen@nuaa.edu.cn

Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, P.R. China

Recently work has shown the advantages of using attribute representation over low-level feature descriptors in face verification [6], due to its capability to explicitly encode high-level semantic meaning with economical coding bits. These advantages allow us to explicitly investigate the similarity relationship between attributes and to see how such relationship could be exploited to improve the performance of face verification. Actually, research in the field of cognitive discovery has shown that infants as young as 3 months of age gain the capability to encode the relations among object features, and use such feature configuration for general object recognition [2]. Indeed, despite of the partial success of using attribute descriptors by treating them statistically independent to each other [4, 6], recent work has shown that it is beneficial to exploit the relationship between attributes under various contexts [7][3][5]

In this paper we proposed a novel method to model the relationship between attributes and exploit such information to improve face verification. In particular, we first represent the meaning of each attribute as a high-dimensional vector in the subject space, which enable us to conveniently construct the corresponding attribute-relationship graph based on the distribution of attributes in that space (c.f., Fig.1). The resulting attribute-relationship encode the pairwise closeness relationship between any two attributes, which are further integrated into a linear classifier to constrain the searching space of its parameters, based on the idea that similar attributes should have similar weights. This is helpful to avoid overfitting and improve the generalization capability of the learned classifier. We also extend the model to handle uncertainty in attribute responses.

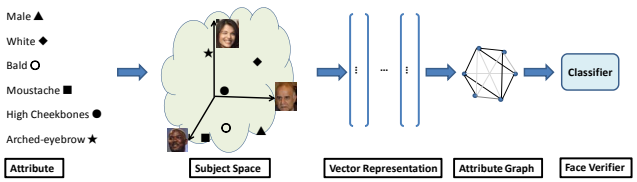


Figure 1: The overall pipeline of the proposed algorithm. Each attribute descriptor is first projected into a common subject space to obtain a high-dimensional vector representation, which are then used to construct an attribute graph. The graph is finally exploited to regularize the objective of a linear SVM-based face verifier.

The overall pipeline of our algorithm is presented in Fig.1, with the following major steps:

Subject space projection. Although the meaning of each attribute is clear to human beings, the way to represent each attribute as a real number is too simple to estimate their correlation. Therefore, we need to represent each attribute in a richer manner to support more advanced inference. Our method is inspired by the vector representation of words in the literature of text categorization. Assuming that we are given a set of M attribute descriptors $A = \{A_i \in R\}_{i=1}^{M}$ for each face image. We use the subjects available in the training set and call the space spanned by these subjects subject space (see Fig.1). Hence for K subjects, we have a subject space with K-dimensions and the meaning of each attribute is represented as a high-dimensional vector in the subject space, with each entry representing whether the corresponding subject owns such attribute.

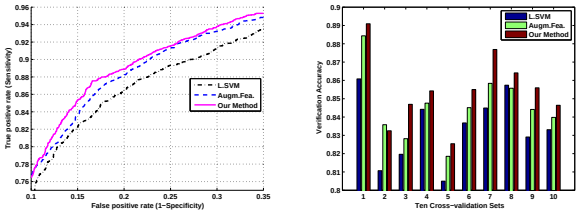
Attribute graph building. Through projecting all the attributes into the subject space, we may model their relationship based on the distribution of each attribute in an information theory framework. In particular, we first compute the point-wise mutual information $I(A_i, y_j)$ of each attribute A_i with each subject with label y_j. After this, correlated information encoded by M attributes and K subjects is organized as the following matrix, based on which, the attribute graph can be constructed by treating each row as a node.

$$\begin{pmatrix} & y_1 & \cdots & y_j & \cdots & y_K \\ A_1 & I(A_1,y_1) & \cdots & I(A_1,y_j) & \cdots & I(A_1,y_K) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_i & I(A_i,y_1) & \cdots & I(A_i,y_j) & \cdots & I(A_i,y_K) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ A_M & I(A_M,y_1) & \cdots & I(A_M,y_j) & \cdots & I(A_M,y_K) \end{pmatrix} \quad (1)$$

Attribute graph constraints. After the weighted attribute graph is built, its Laplacian L is then constructed. We add this as an extra regulariser into the standard SVM objective function and use it to regularize the identity prediction for face images, as follows,

$$\min_{w} \sum_{i=1}^{N} \max\{0, 1 - y_i(w^T x_i + b)\} + \frac{\lambda_1}{2} w^T w + \frac{\lambda_2}{2} w^T L w \quad (2)$$

The above formulation is similar to that of Laplacian SVM [1], but instead of constructing an instance-graph, we build an attribute-relationship graph. One advantage of attribute-graph is that its complexity is controllable since its size will not grow with the number of instances as in [1] but only with the number of attributes. Furthermore, our graph is not meant to constrain the output space of instances but the searching space of model parameters, based on the simple idea that similar attributes should play similar roles in the learnt classifier.



(a) (b)

Figure 2: Comparison of our method with [6][7] on the LFW dataset: a) overall ROC curve, b) detailed results on 10 cross-validation test sets.

Detailed implementations are described in the paper. Fig.2 gives the major experimental results, which indicates that the performance of the attribute-based face verification method can be improved by regularizing it with attribute relationships graph induced from subject space.

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[2] R.S. Bhatt and C. Rovee-Collier. Infants' forgetting of correlated attributes and object recognition. *Child development*, 67(1):172–187, 1996.

[3] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Describing people: Poselet-based attribute classification. In *ICCV'11*, 2011.

[4] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR'09*, pages 1778–1785, 2009.

[5] V. Ferrari and A. Zisserman. Learning visual attributes. In *Advances in Neural Information Processing Systems*, 2007.

[6] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *PAMI*, (99):1–1, 2011.

[7] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. *ECCV'10*, pages 155–168, 2010.

Single Image Segmentation with Estimated Depth

Ryo Yonetani[1]
yonetani@vision.kuee.kyoto-u.ac.jp
Akisato Kimura[2]
akisato@ieee.org
Hitoshi Sakano[2]
sakano.hitoshi@lab.ntt.co.jp
Ken Fukuchi[3]
k2.fukuchi@jaist.ac.jp

[1] Kyoto University
Japan
[2] NTT Communication Science Labs.
Japan
[3] Japan Advanced Institute of Science and Technology
Japan

Object segmentation is a fundamental problem in computer vision. Although many segmentation methods have been proposed, most of them still rely on the appearances of images (i.e., colors or textures) [1, 2, 3, 4, 6, 8]. Consequently, they have a difficulty in distinguishing an object from the background with a similar appearance to the object.

To overcome this difficulty, **we employ a depth map of an input image as an additional cue to the object segmentation**. The main contribution of this work is to introduce a novel segmentation framework that utilizes the depth map combined with a color image to describe the features of objects and backgrounds, where the depth map is estimated from the color image. While a depth map has great potential for use in segmentation, finding a way of integrating two completely different physical quantities, namely the color and depth, has remained unclear. We introduce an integration of the color and depth likelihood on objectness and backgroundness, which simply and effectively extends a traditional segmentation framework based on the Markov random fields (MRF) [2]. By refining the likelihood with the depth information, our proposed method can suppress the incorrect detection of misleading backgrounds.

A single image is expressed by K, where K includes color information $\mathcal{C} = \{C_x \in \mathbb{R}^3\}_{x\in\Omega}$, and in our case, depth information $\mathcal{Z} = \{Z_x \in \mathbb{R}\}_{x\in\Omega}$ (x is a position in the image domain $\Omega \subset \mathbb{N}^2$). Object segmentation is the problem of assigning the label $\mathcal{A} = \{A_x\}_{x\in\Omega}$, which gives a label $A_x = \{0,1\}$ to each pixel, where the labels 1 and 0 at x respectively correspond to the object and background. The statistical relationship between K and $\mathcal{A}$ can be described by an MRF, and the appropriate configuration of the labels can be derived by minimizing the following energy function E:

$$E = \sum_{x\in\Omega} \left\{ \phi_{\mathrm{D}}(K \mid A_x) + \xi_{\mathrm{D}}(A_x) + \sum_{y\in N_x} \left(\phi_{\mathrm{S}}(K \mid A_x, A_y) + \xi_{\mathrm{S}}(A_x, A_y) \right) \right\},$$

where N_x is a 4-neighborhood system of the position x. The data prior term $\xi_{\mathrm{D}}(A_x)$ evaluates how likely to an object the position is for all the pixels in an image, and the smoothness prior term $\xi_{\mathrm{S}}(A_x, A_y)$ is given by the Kronecker delta to ensure the spatial continuity of labels. The data likelihood term $\phi_{\mathrm{D}}(K \mid A_x)$ is modeled by the negative log likelihood of the data value conditioned by the labels: traditionally $\phi_{\mathrm{D}}(K \mid A_x) \propto -\log p(C_x \mid A_x)$. On the other hand, the smoothness likelihood term gives the difference of intensities where labels are spatially discontinuous.

An integration of the color and depth cue is a central part of this study. For introducing depth information into the segmentation framework, we here present a key observation of a structural difference of depth maps from color images. As shown in Figure 1, depth-map structures are quite different from those of color images. In particular, the spatial discontinuities between the pixel values of objects and backgrounds in depth maps do not always agree with those in color images (e.g., an object and the floor on which the object is placed). As a result, **a consideration of depth continuities prevents us from distinguishing objects from backgrounds, which implies that depth information is inappropriate to the smoothness term** ϕ_{S}. On the other hand, Figure 1 also demonstrates that the averages in depth distributions appear at different values between the object and background, while the corresponding intensity distributions look like each other. From these observation, **we decide to introduce depth information into the data term** ϕ_{D}. Specifically, we take particular note of "foregroundness" (nearness) Z_x besides a color value C_x, and fuse them as a weighted sum of likelihood values to derive the data likelihood term $\phi_{\mathrm{D}}(K \mid A_x)$. Consequently, $\phi_{\mathrm{D}}(K \mid A_x)$ is modified as follows:

$$\phi_{\mathrm{D}}(K \mid A_x = i) \propto -\log p(C_x \mid A_x = i) - \alpha_i \log p(Z_x \mid A_x = i) \ (i = 0, 1),$$

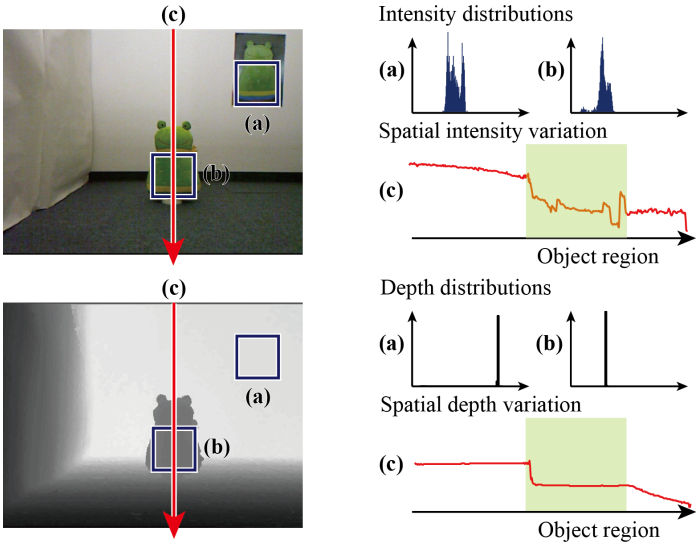


Figure 1: Distributions (blue) and spatial variations (red) of the data in color images and depth maps. The green rectangles describe the region on which the object is displayed.

where α_i is a scale factor of the depth likelihood individually set for both $A_x = 1$ and $A_x = 0$. Note that depth and color distributions may take a different variation because of the difference in the possible range of themselves. We determine α_i by cross validation in the experiments.

Our implementation of the proposed framework comprises not only the integration of colors and depths but automatic computation of the prior term $\xi_{\mathrm{D}}(A_x)$ using a visual attention models [1, 5], and single-image depth estimation via supervised learning [7]. That is, **the proposed method performs automatic object segmentation from a single image, which requires no actual depth maps corresponding to input images.**

[1] K. Akamine, K. Fukuchi, A. Kimura, and S. Takagi. Fully automatic extraction of salient objects from videos in near real time. *The Computer Journal*, 55(1):3–14, 2012.

[2] Y. Boykov and G. Funka-Lea. Graph Cuts and Efficient N-D Image Segmentation. *IJCV*, 70(2):109–131, 2006.

[3] K. Fukuda, T. Takiguchi, and Y. Ariki. Graph Cuts by Using Local Texture Features of Wavelet Coefficient for Image Segmentation. In *ICME*, 2008.

[4] Z. Kato and T. Pong. A Markov Random Field Image Segmentation Model for Color Textured Images. *Image and Vision Computing*, 24 (10):1103–1114, 2006.

[5] D. Pang, A. Kimura, T. Takeuchi, J. Yamato, and K. Kashino. A Stochastic Model of Selective Visual Attention with a Dynamic Bayesian Network. In *ICME*, 2008.

[6] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM Trans. on Graphics*, 23(3):309–314, 2004.

[7] A. Saxena, S. Chung, and A. Ng. Learning Depth from Single Monocular Images. In *NIPS*, 2006.

[8] S. Vicente, V. Kolmogorov, and C. Rother. Joint Optimization of Segmentation and Appearance Models. In *ICCV*, 2009.

Teaching Stereo Perception to YOUR Robot

Marcus Wallenberg
http://users.isy.liu.se/cvl/wallenberg
Per-Erik Forssén
http://users.isy.liu.se/cvl/perfo

Computer Vision Laboratory
Linköping University
Linköping, Sweden

This paper describes a method for generation of dense stereo ground-truth using a consumer depth sensor such as the Microsoft Kinect. Such ground-truth allows adaptation of stereo algorithms to a specific setting. The method uses a novel residual weighting based on error propagation from image plane measurements to 3D. We use this ground-truth in wide-angle stereo learning by automatically tuning a novel extension of the best-first-propagation (BFP) dense correspondence algorithm. We extend BFP by adding a coarse-to-fine scheme, and a structure measure that limits its propagation along linear structures and flat areas. The tuned correspondence algorithm is evaluated in terms of accuracy, robustness, and ability to generalise. Both the tuning cost function, and the evaluation are designed to balance the accuracy-robustness trade-off inherent in patch-based methods such as BFP.

Wide-angle stereo provides an overview of a scene, even at very short range. The large *field of view* (FoV) also ensures that the visual fields from different points of view have a high degree of overlap. For these reasons, wide-angle lenses are popular in navigation, mapping and visual object search on robot platforms. However, the radial distortion caused by these lenses complicates the application of traditional stereo algorithms. A common approach to wide-angle stereo is to first attempt to remove radial distortion and then apply a descriptor-based wide-baseline stereo algorithm. An alternative approach is to use simpler matching metrics, and instead leverage *correspondence propagation*. One such algorithm is the *best-first propagation* (BFP) algorithm [2], and a recent addition is the *generalised PatchMatch algorithm* (GPM) [1]. Though these algorithms are more general than stereo algorithms, they have previously been applied to the stereo problem.

Since we make use of both the inverse depth and pixel coordinates in our calibration procedure, the effect of errors in these measurements must be taken into account during calibration. We therefore propagate error variances from these measurements into both 3D reconstructions and resulting 2D projections. We then fuse multiple Kinect range scans in a reference view coincident with the left stereo camera. This is done by estimating a disparity distribution for each pixel in this image, and using *mean-shift* to find the visible surface closest to the camera. We have imaged three indoor scenes, and for each of them calculated 51 full-resolution wide-angle disparity maps. Examples of individual range scans, an intermediate point cloud and the final disparity maps are shown in figure 2 (top row).

We use these to tune our extension of the BFP algorithm, which we call *coarse-to-fine best-first propagation* (CtF-BFP). Novelties are the use of multiple scales, a structure threshold that limits propagation along linear structures, and a sub-pixel refinement step. These novelties add a multitude of parameters, which make manual tuning difficult. We therefore use an automatic tuning procedure, that minimises an objective function that balances on accuracy, coverage and robustness.

When measuring performance of the stereo algorithm, we denote the estimated disparity map to be evaluated by $\mathbf{D}(u,v)$ defined on the domain $\mathcal{V}$ (pixels where disparities have been estimated). Similarly, the ground-truth disparity image is $\mathbf{D}^*(u,v)$, and the set of valid ground-truth pixels is $\mathcal{V}^*$.

To find a useful set of parameters, we minimise an objective function

$$J(t_a,t_r) = \lambda r(t_r) - (1-\lambda)\int_0^{t_a} a(t)dt. \quad (1)$$

based on the acceptance rate $a(t_a)$ (the relative portion of disparities with an error below a threshold t_a) and rejection rate $r(t_r)$ (the relative portion of estimated pixels that differ from ground-truth by more than a threshold t_r). The parameter $\lambda \in [0,1]$ allows us to control the relative weight of the rejection and acceptance terms. Different values of λ produce different behaviour of $J(t_a,t_r)$ by influencing the trade-off between accuracy, coverage and robustness. Results of the automatic tuning procedure are shown in figure 3.

Figure 1: Left: In wide-angle images, angular resolution is near uniform. Middle: If they are rectified to preserve straight lines, most of the image is spent representing the periphery.
Right: Pan-tilt stereo rig with Kinect. (A) - SLP projector, (B) - RGB camera, (C) - NIR camera, (D) - Left wide-angle camera, (E) - Right wide-angle camera, (F) - Diffusor (raised).

Figure 2: Top: Example of ground-truth generation. Left: Individual Kinect range scans. Right: several range scans projected into 3D.
Bottom: Examples of views for two scenes. Columns left to right: Left wide-angle image, right wide-angle image, magnitude of disparities deemed reliably reconstructed. The final column shows magnitude of disparity estimate obtained using tuned CtF-BFP.

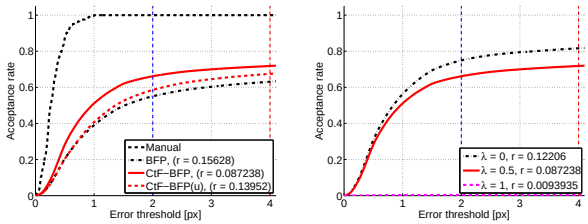


Figure 3: Left: Average acceptance curves over all data sets for automatically tuned CtF-BFP with $\lambda = 0.5$, BFP with original parameters, CtF-BFP(u) (before tuning). Errors on manually selected correspondences included as a best case. Right: Average acceptance curves over all data sets for parameters tuned using $\lambda = 0, 0.5, 1$.

[1] Connelly Barnes, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. The generalized PatchMatch correspondence algorithm. In *European Conference on Computer Vision*, LNCS, September 2010.

[2] Maxime Lhuillier and Long Quan. Match propagation for image-based modelling and rendering. *IEEE TPAMI*, 24(8):1140–1146, 2002.

Recognizing activities with cluster-trees of tracklets

Adrien Gaidon
http://lear.inrialpes.fr/people/gaidon
Zaid Harchaoui
http://lear.inrialpes.fr/people/harchaoui
Cordelia Schmid
http://lear.inrialpes.fr/people/schmid

LEAR - INRIA Grenoble, LJK
655, avenue de l'Europe
38330 Montbonnot, France

Although the structure of simple actions can be captured by rigid grids [3] or by sequences of short temporal parts [2], *activities* are composed of a variable number of sub-events connected by more complex spatio-temporal relations. In this paper, we learn how to automatically represent activities as a hierarchy of mid-level motion components in order to improve activity classification in real-world videos. This hierarchy is a video-specific, data-driven decomposition obtained by clustering *tracklets*, *i.e.* local point trajectories of a fixed small duration.

Our first contribution is a hierarchical spectral clustering algorithm, based on top-down recursive bi-partitioning. We propose to robustly threshold tracklet projections on an approximate spectral embedding by minimizing a spatio-temporal *connectivity* cost. This allows for clusters of arbitrary shape and an efficient greedy splitting strategy that automatically determines the number of clusters. The resulting hierarchical decomposition provides structural information relating motion parts together.

Our second contribution is the use of this entire tree structure, called *cluster-tree* (*c.f.* Figure 1), in order to build a hierarchical model of the motion content of a video. We introduce a corresponding tree representation of actions, called *BOF-tree*. The BOF-tree of a video has the same structure as its cluster-tree and each node is modeled by a bag-of-features (BOF) over the MBH descriptors [6] of its constitutive tracklets. Efficiently using this structural information is challenging as cluster-trees have a variable number of nodes and a structure specific to each video. Furthermore, there is no natural left-to-right ordering of the two children of a parent node. Therefore, we introduce an efficient positive definite kernel — called the "All Tree Edge Pairs" (ATEP) kernel — that computes the structural and visual similarity of two hierarchical decompositions by relying on models of their parent-child relations as described in the following. Let $\mathcal{T}_i = (\mathcal{V}_i, \mathcal{E}_i)$, $i \in \{1,2\}$, be two BOF-trees, defined from their set of vertices (nodes) $\mathcal{V}_i$ and directed edges (parent-child relations) $\mathcal{E}_i$. Each node $v \in \mathcal{V}_i$ is represented by a BOF — noted $b[v]$ — over its constitutive tracklets. We model a directed edge $e = (v_p, v_c) \in \mathcal{E}_i$ by the concatenation — noted $b[e] = (b[v_p], b[v_c])$ — of the BOF of the child node v_c with the BOF of its parent node v_p. Let h be a kernel between BOF, $r_i \in \mathcal{V}_i$ be the root of $\mathcal{T}_i$, and $w_r \in (0,1)$ a cross-validated parameter encoding a prior on the importance of the root-to-root comparisons. Our ATEP kernel is defined as:

$$k(\mathcal{T}_1, \mathcal{T}_2) = w_r \cdot h(b[r_1], b[r_2]) + \frac{1 - w_r}{|\mathcal{E}_1||\mathcal{E}_2|} \cdot \sum_{\substack{e_1 \in \mathcal{E}_1 \\ e_2 \in \mathcal{E}_2}} h(b[e_1], b[e_2])$$

As described in the paper, this kernel can be seen as a weighted similarity between all sub-trees of two BOF-trees.

We use our ATEP kernel in conjunction with SVM classifiers on videos represented by BOF-trees, *i.e.* hierarchically structured sets of motion components. We present experimental results on two recent challenging benchmarks focusing on complex activities: the Olympics Sports dataset [4] and the human-human interactions of the High Five dataset [5]. Table 1 reports performance comparisons between baselines (described in the legend), the state of the art and our method. Our hierarchical ATEP kernel on SDT BOF-trees improves over the unstructured BOF baselines. This confirms the importance of leveraging structure information to recognize complex activities. Note, however, that only our method yields clear performance improvements, whereas other structured baselines are less accurate. This shows that properly decomposing activities is a challenging problem that is critical for performance. In particular, using hierarchical relations between motion components with our ATEP kernel consistently improves over the "flat" baselines relying on unrelated sets of clusters such as the leaves of cluster-trees or clusters obtained by k-means. Table 1 also shows that the BOF-trees produced by our SDT algorithm yield more powerful models than the SDKM ones obtained by bi-partitioning k-means. Finally, our approach outperforms the state of the art on both datasets, including latent part models [4] (+10.6%), complex graphical models resulting from video segmentation [1] (+5.4%), and interaction-specific structured learning [5] (+22.8%).

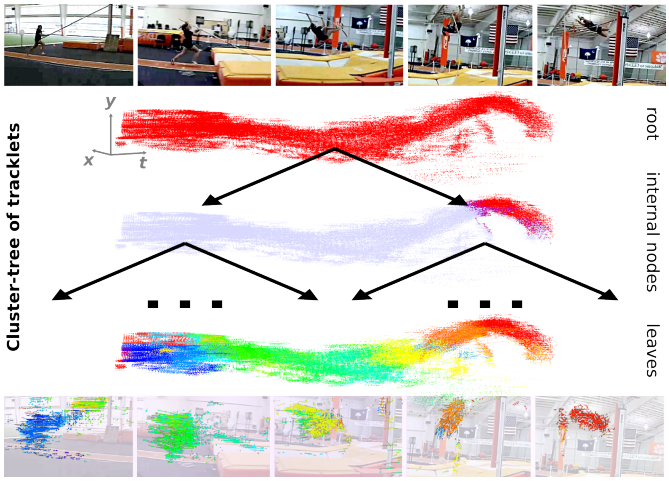


Figure 1: Example of a hierarchical motion decomposition obtained by our divisive clustering algorithm on dense tracklets. Our approach uses the whole cluster-tree to compare videos.

SDT trees with ATEP	**82.7**
SDKM trees with ATEP	76.7
SDT leaves	77.9
SDKM leaves	72.2
spectral	71.7
kmeans	70.8
Wang *et al.* [6]	75.9
Laptev *et al.* [3]	61.3
Brendel and Todorovic [1]	77.3
Niebles *et al.* [4]	72.1

(a) Olympics Sports (Accuracy in %)

SDT trees with ATEP	**55.6**
SDKM trees with ATEP	53.8
SDT leaves	49.5
SDKM leaves	48.6
spectral	48.9
kmeans	50.1
Wang *et al.* [6]	53.4
Laptev *et al.* [3]	36.9
Patron-Perez *et al.* [5]	32.8

(b) High Five (AP in %)

Table 1: Performance on the Olympics Sports [4] and High Five [5] datasets. Results with our Spectral Divisive Thresholding algorithm are noted "SDT". We compare with the state of the art, BOF baselines [3, 6], "flat" decompositions obtained by k-means and spectral clustering, as well as hierarchical decompositions obtained by a baseline spectral divisive bi-partitioning k-means algorithm noted "SDKM".

Acknowledgments This work was partially funded by the MSR/INRIA joint project, the European project AXES and the PASCAL 2 Network of Excellence.

[1] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, 2011.

[2] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom sequence models for efficient action detection. In *CVPR*, 2011.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[4] J.C. Niebles, CW. Chen, , and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV*, 2010.

[5] A. Patron-Perez, M. Marszałek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in TV shows. In *BMVC*, 2010.

[6] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *CVPR*, 2011.

A Videography Analysis Framework for Video Retrieval and Summarization

Kang Li*[1]
kangli@buffalo.edu
Sangmin Oh*[2]
sangmin.oh@kitware.com
A. G. Amitha Perera[2]
amitha.perera@kitware.com
Yun Fu[3]
raymondyunfu@gmail.com

[1] Department of CSE
State University of New York
Buffalo, NY, USA

[2] Kitware, Inc.
Clifton Park, NY, USA

[3] Department of ECE and College of CIS
Northeastern University
Boston, MA, USA

Overview: In this work, we focus on developing features and approaches to represent and analyze videography styles in unconstrained videos. By unconstrained videos, we mean typical consumer videos with significant content complexity and diverse editing artifacts, mostly with long duration. We present an approach for *unsupervised videography analysis* for unconstrained videos. Intuitively, each videography can be understood as a camera director's direction on a movie script, *e.g.*, "capture the running actress by panning the camera, to have her face appear at 20 percent size of the video". The idea is that different classes of video content will have different styles—the videography style of a wedding video should be different from a sports video—and so, the videography style should provide a valuable signal for automated content analysis.

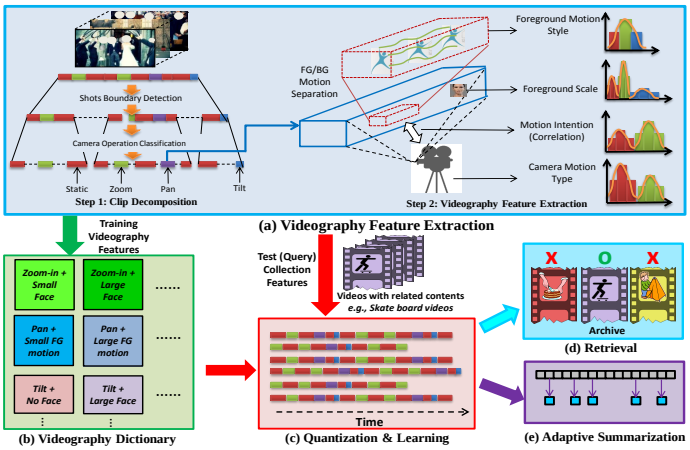


Figure 1: Framework for videography analysis and applications.

Videography Analysis: The overall framework of our approach is illustrated in Fig. 1(a). First, a two-level motion analysis is conducted to decompose long clips into sequences of segments with coherent motion types (S/P/T/Z). Second, multiple features related to motion and scale patterns are measured from every segment, which are used to characterize videography. Throughout this work, we utilize densely computed KLT tracks over the entire clips as main basis for the derived features.

We assume that there are diverse videography styles in unconstrained videos, which are discovered as a *videography dictionary* via unsupervised clustering on proposed features. Then, a video clip can be represented as a series of segments with varying videography words. For the underlying videography features, we extend conventional features such as camera motion and foreground (FG) object motion (e.g., [1]) by incorporating two novel features: *motion correlation* and *scale* information.

Once videography features are obtained from segments, they are used to build *videography dictionary* (VD) shown in Fig. 1(b). The computed VD will be used to quantize video clips into sequences of videography words (VWs), as shown in Fig. 1(c). Our analysis shows that there are regularized patterns in the videography used in the unconstrained Internet videos, and correlations between the exhibited videography styles and video contents. Such observation on discriminative correlations suggests

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

* Indicates equal contributions.

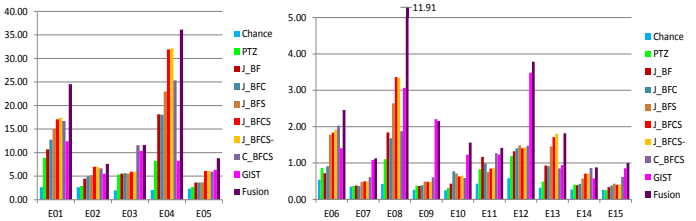


Figure 2: Average Precision (%) of video retrieval results on MED corpus, for 15 events: ***(E01) Board trick***, *(E02) Feeding animal, (E03) Fishing,* ***(E04) Wedding***, *(E05) Working wood project,* ***(E06) Birthday party***, *(E07) Change vehicle tire,* ***(E08) Flash mob***, *(E09) Getting vehicle unstuck, (E10) Groom animal, (E11) Make sandwich,* ***(E12) Parade***, ***(E13) Parkour***, *(E14), Repair appliance, and (E15) Sewing project.*

that videography analysis can actually be used for challenging tasks such as content-based retrieval and content summarization.

Video Retrieval: For retrieval, we computed bag-of-word representations based on the videography word sequences and employed them as the basis for content-based video retrieval tasks. We have conducted experiments on a large TRECVID '11 MED dataset where we tried diverse variations of the proposed approach as well as using more conventional features such as GIST. Our results indicate that the proposed videography features effectively improve the retrieval performance and are complimentary to traditional appearance features such as GIST, improving performance further when both features are used jointly. Figure 2 shows the list of video event classes and the extent of conducted retrieval experiments as well as summarized performance profiles. Event classes that show the most benefits by videography-based analysis are marked in bold.

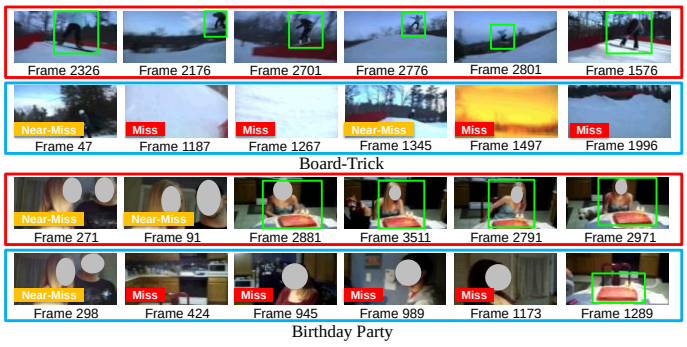


Figure 3: Videography-aware adaptive summarization. Three summarization results by this work (red rows) and baseline (blue rows). Detected FG regions (green) and human judgements on relevance of key frames (good:none, near-miss: yellow, miss: red) are marked on each image.

Video Summarization: We also show that the proposed videography analysis can be used to provide videography-aware adaptive summarization method. For example, Fig. 3 shows example summarization results for different events where the segments with distinctive videography styles for particular events are highlighted in the summaries, e.g., board tricks during snowboarding and candle blowing during a birthday party. Summarization produced by our proposed approach is shown in red and results by baseline approaches of using color histogram changes are shown in blue.

[1] Xingquan Zhu, Ahmed K. Elmagarmid, Xiangyang Xue, Lide Wu, and Ann Christine Catlin. InsightVideo: Towards hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, 7(4):648–666, 2005.

Efficient Point Feature Tracking based on Self-aware Distance Transform

Min-Gyu Park
mpark@gist.ac.kr
Kuk-Jin Yoon
kjyoon@gist.ac.kr

Computer Vision Lab.
School of Information and Communications,
Gwangju Institute of Science and Technology (GIST),
Gwangju, Republic of Korea
http://cvl.gist.ac.kr

The tracking of point features is an essential problem in computer vision because the acquisiton of feature correspondence from successive frames is a front-end step in many problems. We divide feature tracking algorithm rougly into three categories, i.e., tracking-by-detection [2, 6], tracking-by-template matching [3, 5, 8], and tracking-by-Lucas-Kanade-Tomasi (KLT) [1, 4] tracker. In this paper, we focus on improving the second approach that inherits the intrinsic characteristics of the template matching problem. Previous studies focused mainly on increasing the speed of matching [3], improving the computational search efficiency [8], and developting robust similarity measures. However, the size of search region is still retained as a predefined parameter, although the size of a search region affects the performance of an algorithm significantly.

To tackle this problem, we propose a Self-aware Distance Transform (SDT) with an efficient feature-tracking method. The aim of the SDT is to estimate the optimal search region size based on the autocorrelation with a template in the initial frame. We use the spatial relationship of the cross-correlation coefficients relative to the best match as the function of a coefficient in the predicted position of a feature. After extracting a feature [7] and its corresponding template, we immediately perform autocorrelation of the image and the template; then generate a set of groups based on the distance to the best match as follow:

$$F = \{F_1, F_2, ..., F_{M-1}, F_M\},$$
$$F_k = \{C(\mathbf{p}) | round(\sqrt{p_x^2 + p_y^2}) = k\} \text{ for } 1 \leq k \leq M, \quad (1)$$

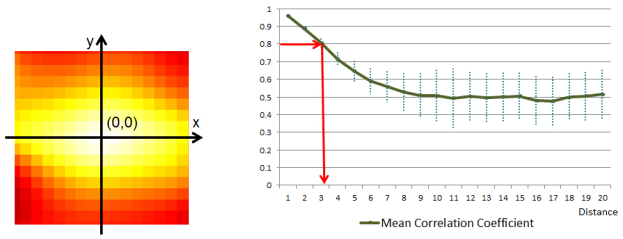
where F_k is a set of correlation coefficients with the same distance from the best match, $\mathbf{p} = [p_x \; p_y]^T$ is a relative position vector centered on the best match $(0,0)$, $C(\mathbf{p})$ is the autocorrelation coefficient at $\mathbf{p}$, and k ranges from 1 to the predefined constant M. This constant is tuned automatically during the last step, so the selection of this value is not a significant problem. Rather than using a continuous distance, we discretize the distance values for the group of pixels and this distance is computed using a Chamfer distance transform. Next, we compute the mean and variance of each group as follows:

$$\mu_k = \frac{1}{|F_k|} \sum_{C(\mathbf{p}) \in F_k} C(\mathbf{p}), \; \sigma_k^2 = \frac{1}{|F_k|} \sum_{C(\mathbf{p}) \in F_k} (C(\mathbf{p}) - \mu_k)^2, \quad (2)$$

where μ_k and σ_k indicate the mean and standard deviation of the autocorrelation coefficients at the distance k, while $|F_k|$ represents the cardinality of a set of correlation coefficients. These two statistics are the essence of SDT because they are used to the compute the optimal size of a search region. Figure 1 shows the autocorrelation result and the relationship between the mean and distance values. This relationship is used as a function of an NCC coefficient, which allows the size of a search region to be determined automatically at each prediction step. For example, Fig. 1 shows that the best match is probably within 3 pixels if the correlation value is 0.8. Finally, the SDT is defined as a function of a real valued vector (the position of a feature), which yields a positive integer value as follows:

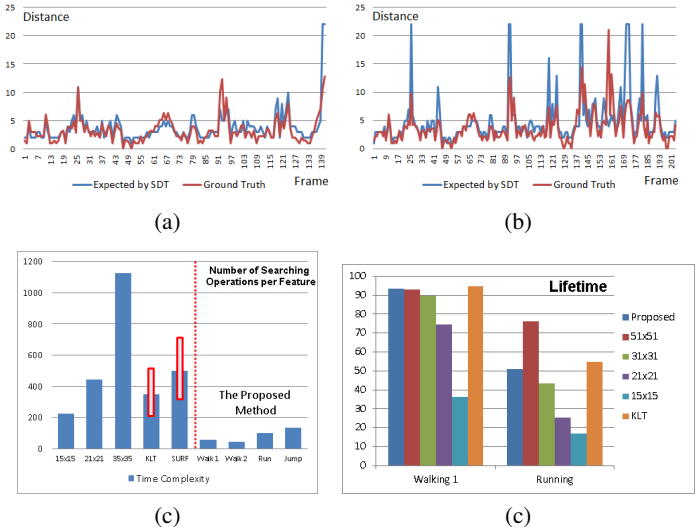
$$SDT : \Re^2 \rightarrow N^+ \text{ s.t.}$$
$$\hat{d}_{t+1} = arg\min_{1 \leq k \leq M} |\mu_k - C_{t+1}(\hat{\mathbf{x}}_{t+1})|, \quad (3)$$

where $C_{t+1}(\hat{\mathbf{x}}_{t+1})$ indicates the NCC coefficient of a predicted position (we use a subscript to indicate that this computes the NCC between the template and a successive image at time $t+1$) while the expected distance is computed by minimizing the difference between the mean value and the NCC coefficients. Indeed, the SDT can be used for any other prediction models; we use the constant velocity model for the prediction. The expected distance contains uncertainty that is proportional to corresponding variance σ_k where k equals $\hat{d}_{t+1}$. To avoid unreliable estimation of expected distance, therefore, we restrict the range of the valid expected distance, $(0, d_{max}]$ is determined by thresholding larger variance values than a predefined threshold. For the experiment, we computed both the ground truth displacement and the predicted distance for the SDT evaluation. As shown in Fig. 2 (a–b), the SDT well approximates the actual displacement of features; thus, the size of a search region can be adaptively chosen. As a consequence, the time complexity of the proposed feature tracking method reduced significantly compared to other methods while maintaining a certain level of robustness against abrupt motion of features, as shown in Fig. 2 (c–d), respectively.



(a) (b)

Figure 1: Illustration of the SDT; (a) the result of autocorrelation and (b) the relationship between the mean values of the autocorrelation coefficients and distance. The dotted vertical line indicates the variance of the autocorrelation coefficients at the same distance.



(a) (b)

(c) (c)

Figure 2: Comparison of the expected distances (blue lines) and the ground truth distances (red lines) for features in the walking 1 sequence (a–b), and the evaluation of the proposed method in terms of time complexity (c) and the lifetime of features (d) compared to other tracking methods.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0065038).

[1] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework: Part 1. Technical Report CMU-RI-TR-02-16, Robotics Institute, Pittsburgh, PA, July 2002.

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417, 2006.

[3] Yu-Wen Huang, Ching-Yeh Chen, Chen-Han Tsai, Chun-Fu Shen, and Liang-Gee Chen. Survey on block matching motion estimation algorithms and architectures with new results. *J. VLSI Signal Process. Syst.*, 42(3):297–320, March 2006.

[4] Myung Hwangbo, Jun-Sik Kim, and T. Kanade. Inertial-aided klt feature tracking for a moving camera. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 1909 –1916, oct. 2009.

[5] J. P. Lewis. Fast normalized cross-correlation, 1995.

[6] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, November 2004. ISSN 0920-5691.

[7] Jianbo Shi and Carlo Tomasi. Good features to track. In *Computer Vision and Pattern Recognition (CVPR)*, Ithaca, NY, USA, 1993. Cornell University.

[8] Steven L Tanimoto. Template matching in pyramids. *Computer Graphics and Image Processing*, 16(4):356 – 369, 1981.

Doo-Sabin Surface Models with Biomechanical Constraints for Kalman Filter Based Endocardial Wall Tracking in 3D+T Echocardiography

Engin Dikici[1]
engin.dikici@ntnu.no
Fredrik Orderud[2]
fredrik.orderud@ge.com
Gabriel Kiss[1]
gabriel.kiss@ntnu.no
Anders Thorstensen[1]
anders.thorstensen@ntnu.no
Hans Torp[1]
hans.torp@ntnu.no

[1] Norwegian University of Science and Technology
Trondheim, Norway
[2] GE Vingmed Ultrasound
Oslo, Norway

3D+T echocardiography is a valuable tool for assessing cardiac function, as it enables real-time, non-invasive and low cost acquisition of volumetric images of the heart. The automated tracking of heart chambers in 3D+T echocardiography remains a challenging task due to reasons including speckle noise, shadowing, and the existence of intra-cavity structures [6]. Furthermore, the real-time detection of endocardial borders might be desirable for the invasive procedures and intensive care applications. State-space analysis using Kalman filtering can be employed for the detection of left ventricle (LV) structures in time-dependent recordings. Orderud et al. proposed a Kalman tracking framework for the real-time detection of LV structures in 3D+T echocardiography [5]. The study took advantage of compact Doo-Sabin model representations for rapid tracking, but it did not utilize physical properties to constrain model deformations. Liu et al. introduced a biomechanical-model constrained state-space analysis framework for the tracking of short-axis 2D+T echocardiography recordings [4]. Their study used *dense* Delaunay triangulated models and employed basic tri-nodal linear elements during the finite element analysis (FEA). Due to triangulated high resolution model representations, it offered a computationally expensive solution.

This paper proposes an approach to combine the compact model representations with biomechanical constraints for rapid and accurate tracking. We extend the real-time Kalman tracking framework described in [5] by employing biomechanically constrained state transitions. First, FEA for the tracked Doo-Sabin surface model is performed using the isoparametric method introduced in [3]. This step enables the computation of a stiffness matrix $\mathbf{K}$ for a given endocardial model using shell elements without changing the model geometry. However, the computed model might lead to inaccurate deformation modes due to hypothesized model shape and FEA parameters (e.g. Young's modulus, Poisson's ratio). Accordingly, we improve the model shape and stiffness matrix using statistical information collected from a training data via Control Point Distribution Models (CPDM) [2]. During the improvement stage, (1) the model shape is updated to the population mean, (2) the stiffness matrix for the updated model shape is computed as $\mathbf{K}'$ (see Figure 1), and (3) $\mathbf{K}'$ is further modified to $\mathbf{K}_{opt}$ to produce similar modes of deformation as the statistically observed ones using Baruch and Bar-Itzhack direct matrix modifications (BBDMM) [1]. Finally, the state prediction stage of the Kalman tracking framework is formulated to perform biomechanically constrained tracking.

In the results section, endocardial surface tracking quality is compared among (1) Doo-Sabin surface models with different control node resolutions, (2) biomechanically constrained and non-constrained state transitions, and (3) the systems employing statistically improved and not improved Doo-Sabin models (see Figure 2). Our analyses showed that

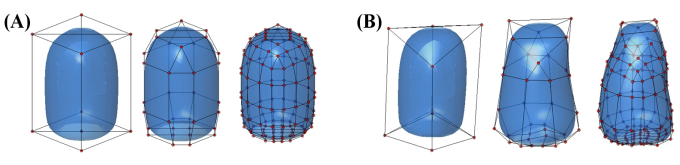


Figure 1: (A) Not-refined, refined and double-refined endocardial Doo-Sabin surface models, and (B) the same surface models after the model shape updates.

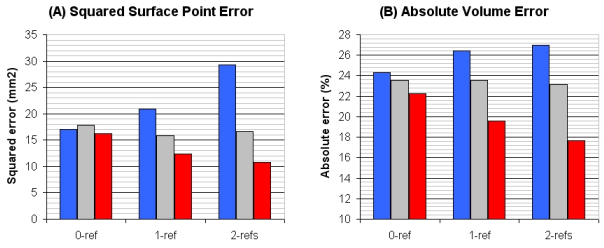


Figure 2: (A) Squared surface point error (in mm^2), and (B) absolute volume error (in percentages) for the Kalman tracking framework using no biomechanical constraints (blue), biomechanical constraints but no statistical improvements (gray), and biomechanical constraints and statistical improvements (red) for non-refined (0-ref), refined (1-ref) and double-refined (2-ref) Doo-Sabin model tracking.

the biomechanical constraints are necessary especially when the tracked model has a high control node resolution. This is due to the fact that as the model complexity increases the tracker can benefit more from a spatial regularization, which is provided by biomechanical constraints. The statistical model improvements take advantage of higher model resolution levels as (1) the model node updates provide a more realistic model shape to perform tracking, and (2) deformation modes learned from CPDM improve the stiffness matrix accuracy. The tracking framework is implemented in C++, and processes each frame in 2*ms* with not-refined (9 control nodes), 3.4*ms* with refined (34 control nodes) and 30.6*ms* with double-refined (136 control nodes) models when executed on a 2.80 GHz Intel Core 2 Duo CPU. The introduced method is (1) practical; the computed models can be directly used in a Kalman tracking framework by implementing a few modifications in the state prediction stage, (2) useful since it improves the tracking accuracy without introducing additional run-time complexity, (3) yet novel as the biomechanically constrained subdivision surfaces have not been employed in a Kalman tracker prior to our study.

[1] M. Baruch and I. Y. Bar Itzhack. Optimal weighted orthogonalization of measured modes. *AIAA Journal*, 16(4):346–351, 1978.

[2] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Training models of shape from sets of examples. In *In Proc. British Machine Vision Conference*, pages 9–18, 1992.

[3] E. Dikici, S. R. Snare, and F. Orderud. Isoparametric finite element analysis for doo-sabin subdivision models. In *Proceedings of Graphics Interface 2012*, GI '12, Toronto, Ont., Canada, Canada, 2012. Canadian Information Processing Society.

[4] Ning Lin, Weichuan Yu, and James S. Duncan. Combinative multi-scale level set framework for echocardiographic image segmentation. *Medical Image Analysis*, 7:529–537, 2002.

[5] F. Orderud and S. I. Rabben. Real-time 3d segmentation of the left ventricle using deformable subdivision surfaces. In *CVPR*, 2008.

[6] S. K. Setarehdan and John J. Soraghan. *Segmentation in echocardiographic images*, pages 64–130. Springer-Verlag New York, Inc., New York, NY, USA, 2002. ISBN 1-85233-389-8.

Efficient and Scalable Depthmap Fusion

Enliang Zheng
http://www.cs.unc.edu/~ezheng
Enrique Dunn
http://www.cs.unc.edu/~dunn
Rahul Raguram
http://www.cs.unc.edu/~rraguram
Jan-Michael Frahm
http://www.cs.unc.edu/~jmf

Department of Computer Science
University of North Carolina
Chapel Hill, NC USA

The estimation of a complete 3D model from a set of depthmaps is a data intensive task aimed at mitigating measurement noise in the input data by leveraging the inherent redundancy in overlapping multi-view observations. In this paper we propose an efficient depthmap fusion approach that reduces the memory complexity associated with volumetric scene representations. By virtue of reducing the memory footprint we are able to process an increased reconstruction volume with greater spatial resolution. Our approach also improves upon state of the art fusion techniques by approaching the problem in an incremental online setting instead of batch mode processing. In this way, are able to handle an arbitrary number of input images at high pixel resolution and facilitate a streaming 3D processing pipeline.

Our proposal builds upon recently proposed heightmap representations [2, 3] and introduces a wavelet based compression mechanism for every column in the heightmap in order to attain a reduced parametric representation of each column. The heightmap fusion method takes a set of depthmaps as input, along with external and internal camera parameters. From the external camera parameters, we can determine a ground plane for the scene [4], which serves as the lower $x-y$ boundary of the fusion volume. Following this, the $x-y$-plane (ground plane) is partitioned into cells, where the x, y size of the cell matches the desired resolution of the 3D-model. For each cell we can define a column representing the volume above the cell. One of the computational advantages of the original heightmap fusion algorithm, in contrast to earlier volumetric fusion methods, is that the fusion is solved independently within each column, allowing for easy parallelization of the fusion process. There are two main observations that we leverage: (a) in the basic heightmap formulation, columns of the 3D volume are processed independently of each other and (b) when considering a single column , it is typically the case that the occupancy function function values change smoothly within a segment, with sharp transitions between adjacent segments. These sharp transitions correspond to the change from occupied to empty space in the physical world. Accordingly, wavelet based compression techniques are very well suited for modeling such volumetric data, as the transitional intervals of the input data signal can be accurately represented using a small number of wavelet coefficients [5].

The goal of our multi-layer heightmap estimation module is to transform the input occupancy function values for all the voxels belonging to a given column into an output binary segmented set of occupancy layers. We model the associated segmentation problem as a *directed acyclic graph* traversal problem and efficiently find the optimal solution through dynamic programming (DP). Our multi-layer approach differs from [3] in the following:

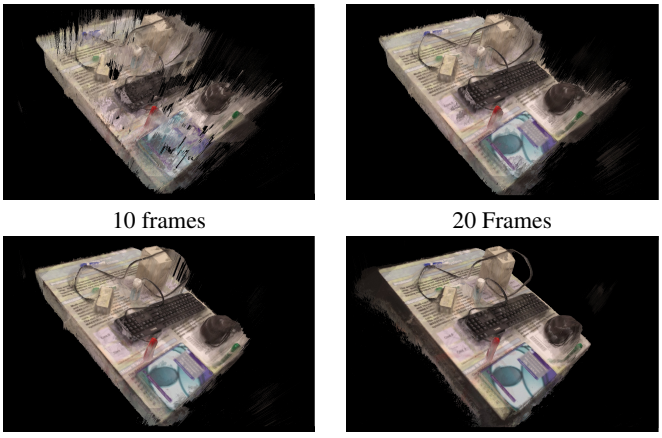
1. We deploy a general graph structure that is not restricted in regard to the parity of the number of layers nor in the topology of the assigned layers. In [3] strong assumptions on the observed scene needed to be made to define the graph structure.

2. We perform the dynamic programming in a reduced search space. We identify and quantify segments of homogeneous occupancy classification and use them as atomic elements to be labeled by the DP procedure. The importance of this pre-processing is that it enables more efficient processing at finer voxel resolutions while mitigating the memory requirements of DP search.

In our experiment, camera poses and depthmaps are computed using the pipeline of [1]. We compared our depthmap fusion approach against the one proposed in [3]. Fig. 1(a) (b) illustrate an improved robustness to noise provided by our method. We additionally tested our fusion method with the output of a SfM pipeline for close range scene reconstruction. We utilized a 16MP DSLR camera to obtain a total of 100 images (we name this data set randObject). Figure 2 depicts the sequential process of incrementally building the 3D model through online depthmap fusion. Our depthmap fusion runs at 33 Hz on one NVIDIA Tesla C2050/C2070 graphics card. Computing a 9-layer heightmap on a volume of 100x100x100 given one image takes 24.9 ms, within which the procedure of compression and decompression only takes 1.6 ms.

(a) Our Results (b) Results from [3]

Figure 1: (a) and (b) are comparative results for crowd sourced data. Both images show 3D models of the Brandenburg Gate.



10 frames 20 Frames

40 Frames 100 Frames

Figure 2: Incremental Online Depthmap Fusion. The 3D model is improved as more images are added.

[1] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building rome on a cloudless day. In *ECCV*, 2010.

[2] David Gallup, Jan-Michael Frahm, and Marc Pollefeys. A heightmap model for efficient 3d reconstruction from street-level video. In *3DPVT*, 2010.

[3] David Gallup, Marc Pollefeys, and Jan-Michael Frahm. 3d reconstruction using an n-layer heightmap. In *DAGM*, 2010.

[4] Richard Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Computer Vision*, 2: 1–104, 2006.

[5] Wikipedia. Wavelet transform, 2011. URL "`http://en.wikipedia.org//Wavelet_transform`".

Recovery of Slice Rotations with the Stack Alignment Transform in Cardiac MR Series

Constantine Zakkaroff[1]
mnkz@leeds.ac.uk
Aleksandra Radjenovic[2]
a.radjenovic@leeds.ac.uk
John P. Greenwood[3]
j.greenwood@leeds.ac.uk
Derek R. Magee[1]
d.r.magee@leeds.ac.uk

[1] School of Computing
University of Leeds, UK
[2] National Institute for Health Research,
Leeds Musculoskeletal Biomedical Research Unit
University of Leeds, UK
[3] Leeds Institute of Genetics, Health and Therapeutics
University of Leeds, UK

Displacement of individual slices in image stacks relative to each other is known as stack misalignment and it can occur in imaging modalities with discontinuous acquisition protocols. Stack misalignment in cardiac Magnetic Resonance (MR) cine series is caused by inconsistencies in breath-hold positions between slice acquisitions and depends on the ability of the patient to follow the instructions of the attending technician during the scan. Figure 1 shows a typical example of cine stack before and after misalignment correction.

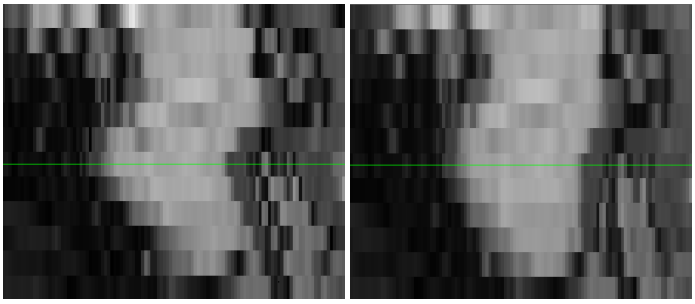


Figure 1: An example of misalignment and its correction: left ventricle (LV) in a cine stack reformatted to long-axis (LA) orientation before (left) and after (right) correction.

As with most registration applications the solutions to the problem of slice misalignment in image stacks rely on a reference image [2, 4]. However, as the authors in [1] observe, slice-to-volume registration can become an under-constrained problem due to the method's weakness which lies in the use of out-of-plane rotations. Slice displacements in cardiac cine stacks typically include translational and rotational components. Recovery of rotation in medial down to apical slices with the basic slice-to-volume registration may provide unreliable results because of rotational symmetry of she short-axis (SA) view of the LV.

This paper describes a novel method for correction of stack misalignment in cardiac cine series with recovery of the rotational component. The core of the presented method is a custom spatial transform, which improves the reliability of misalignment correction over the slice-by-slice correction approach because the image similarity metric for every iteration during optimisation is calculated on the whole image stack at once with all slice correction parameters contributing to the result.

The stack alignment transform parametrises separately the in-plane translation along the X and Y dimensions and rotation around a user-supplied centre of rotation for the individual slices independent of each other. In addition, the transform includes a parameter for global translation along the Z direction. For example, if the cine series consist of N slices, the transform is parametrised as $T = \{\{\theta_1, C_{x1}, C_{y1}, T_{x1}, T_{y1}\}, \ldots, \{\theta_N, C_{xN}, C_{yN}, T_{xN}, T_{yN}\}, T_z\}$. The transform also includes a set of fixed parameters, which are used to calculate the relevant slice number and the corresponding subset of transform parameters for a point in 3D space. Fixed parameters include image origin, size and voxel spacing. Given a point $P = [x, y, z]$ in 3D space, the transformed point $P' = [x', y', z']$ is calculated in two steps as follows:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta_n & -\sin\theta_n \\ \sin\theta_n & \cos\theta_n \end{bmatrix} \cdot \begin{bmatrix} x - C_{xn} \\ y - C_{yn} \end{bmatrix} + \begin{bmatrix} T_{xn} + C_{xn} \\ T_{yn} + C_{yn} \end{bmatrix}, \; z' = z + T_z \quad (1)$$

where n is an index into the Z dimension on the image grid, identifying the slice number, which point P falls into. The stack alignment transform deliberately avoids out-of-plane translations and rotations which otherwise pose a problem for series reconstruction: both out-of-plane translations and rotations may result in the "holes" after the reconstruction.

Misalignment correction was evaluated for convergence reliability and accuracy. First, the transform was evaluated with a protocol of registration uncertainty measurement described in [5]. The protocol is designed for assessing the stability of registration in a large number of runs each with a known introduced misalignment; in this evaluation method the mean recovered transform is used as a pseudo-gold standard based on the assumption of a zero mean distribution of errors. The comparison of misalignment correction with stack alignment transform against the basic slice-by-slice correction provides evidence that the transform improves the robustness of misalignment correction. The second part of evaluation involved the measurement of registration accuracy on the basis of manually-defined contours treated as the gold-standard. In addition, this part of the evaluation was aimed at providing evidence to support the case for recovering the rotational component of slice displacement, as is shown in Figure 2; in our study about 25% of datasets contained visually-detectable rotation. The accuracy of contour match was calculated as the Hausdorff distance metric, which is used for determining the degree of similarity between two objects when superimposed on one another [3].

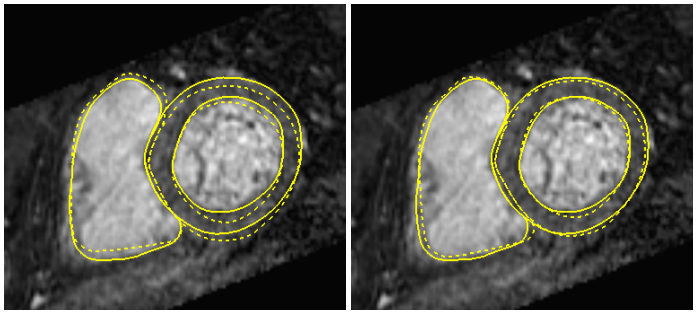


Figure 2: Improvement of correction accuracy gained through the recovery of slice rotation; moving contours (dashed line) after translation-only correction, superimposed on the fixed contours (solid line) shown in the context of the reference volume (left); moving contours after rotation and translation, superimposed on fixed contours (right).

[1] W Birkfellner, M Figl, J Kettenbach, J Hummel, P Homolka, R Schernthaner, T Nau, and H Bergmann. Rigid 2D/3D Slice-to-volume Registration and its Application on Fluoroscopic CT Images. *Medical Physics*, 34(1):246–255, 2007.

[2] A Chandler, R Pinder, T Netsch, J A Schnabel, D J Hawkes, D L G Hill, and R Razavi. Correction of Misaligned Slices in Multi-slice Cardiovascular Magnetic Resonance Using Slice-to-volume Registration. *Magnetic Resonance Imaging*, 10(1):13, 2008.

[3] D P Huttenlocher, G A Klanderman, and W J Rucklidge. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[4] J Lötjönen, M Pollari, S Kivistö, and K Lauerma. Correction of Movement Artefacts from 4-D Cardiac Short- and Long-axis MR Data. In *Medical Image Computing and Computer Assisted Intervention*, 2004.

[5] J R Sykes, D S Brettle, D R Magee, and D I Thwaites. Investigation of Uncertainties in Image Registration of Cone Beam CT to CT on an Image-guided Radiotherapy System. *Physics in Medicine and Biology*, 54(24):7263–83, 2009.

Fine-Grained Categorization for 3D Scene Understanding

Michael Stark[1]
mst@cs.stanford.edu
Jonathan Krause[1]
jkrause@cs.stanford.edu
Bojan Pepik[2]
bpepikj@mpi-inf.mpg.de
David Meger[3]
dpmeger@cs.ubc.ca
James J. Little[3]
little@cs.ubc.ca
Bernt Schiele[2]
schiele@mpi-inf.mpg.de
Daphne Koller[1]
koller@cs.stanford.edu

[1] Computer Science Department
Stanford University
Stanford, CA, USA

[2] Max Planck Institute for Informatics
Saarbrücken, Germany

[3] Computer Science Department
University of British Columbia
Vancouver, BC, Canada

Basic-level object category recognition has made remarkable progress over the last decade, both in image-level categorization and bounding box localization settings [3]. More recently, the recognition of finer-grained, subordinate categories is receiving increased attention [1, 2, 4, 7, 8, 11, 12]. It is deemed challenging due to the need to capture subtle appearance differences between categories while at the same time maintaining robustness to intra-category variations induced by changes in pose and viewpoint. As a consequence, the focus of previous work has been mostly on object categories *and* methods that favor discrimination by strong local appearance cues (such as random color image patches for birds [12]) or global image statistics (such as color histograms for flowers [8]).

Our paper goes beyond previous work on fine-grained categorization in two ways. First, in addition to exploring the task of fine-grained categorization itself, we suggest the use of fine-grained category predictions as an input for higher-level reasoning. This is based on the observation that fine-grained categories can encode, among other aspects, information about metric object sizes, which can in turn provide geometric constraints for scene-level reasoning. Accordingly, we focus our attention on rigid, geometric objects that can provide, if correctly categorized, reliable metric size estimates, and introduce a novel dataset [1] of fine-grained car types as a test bed for our approach (Fig. 1). This data set is annotated with 2D bounding boxes, viewpoint estimates, car types, and additionally includes metric object sizes (length, width, and height) for geometric reasoning.

Secondly, we design a fine-grained object class representation that captures variations in object shape and geometry rather than appearance [8, 12], in order to match the object class of interest. To that end, we introduce two different variants of utilizing part detections as indicators of object geometry, of varying complexity. Both are based on the best-performing object class detector to date, the DPM [5].

Novel car-types data set. We introduce a novel data set of fine-grained *car-types*, consisting of 1904 images of cars from 14 different categories (Fig. 1), annotated with category labels, 2D bounding boxes, and a viewpoint estimate (azimuth angle binned to 5 degrees).

Fine-grained categorization. Our approach follows the intuition that object geometry, and hence, category affiliation, can be encoded in the layout of its constituent parts. We thus design two different models that capture part layout, both building upon the DPM [5]: i) *part-bank*, a feature derived from response maps of a basic-level object class detector, similar in spirit to object-bank [6], and ii) *structDPM*, a multi-class variant of the DPM [10] that directly optimizes for fine-grained categorization. Our experiments show that both models outperform state-of-the-art classifiers by significant margins in fine-grained categorization. *structDPM* outperforms *part-bank*, at the cost of higher computational complexity.

3D Geometric reasoning. We demonstrate the potential of fine-grained category predictions to aid 3D geometric reasoning in a first, idealized experiment: the task is to predict the depth of a given object (its distance from the calibrated camera) from a single view, based on its 2D bounding box and metric size information derived from its predicted fine-grained category (Fig. 2). Our experiments on a public data set [9] confirm the benefit of these predictions over a baseline in the high precision domain.

[1] The data set will be publicly available under https://www.d2.mpi-inf.mpg.de/datasets.

Figure 1: Example images from our novel *car-types* data set with fine-grained category and viewpoint (azimuth angle) annotations.

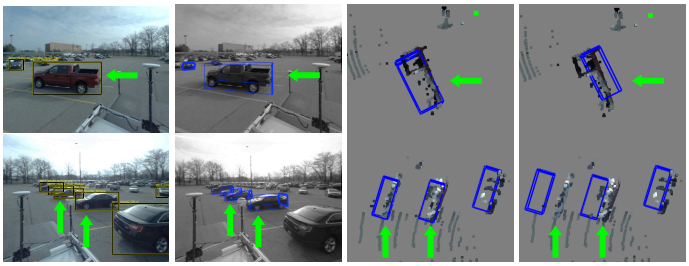


Figure 2: Depth estimation results. (1) 2D GT BB's with predicted fine-grained categories, (2) estimated 3D BBs for fine-grained categories, (3) point cloud top view for fine-grained, (4) for baseline. Green arrows: improvement. Best viewed in the electronic version, with magnification.

Acknowledgements. This material is based upon work supported by the Max Planck Center for Visual Computing and Communication and the Defense Advanced Research Projects Agency under Contract No. FA8650-10-C-7020.

[1] A. Bar-Hillel and D. Weinshall. Subordinate class recognition using relational object models. In *NIPS*, 2006.

[2] S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[4] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *ICCV*, 2011.

[5] P. F. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[6] L.-J. Li, Hao Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[7] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011.

[8] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[9] G. Pandey, J. R. McBride, and R. M. Eustice. Ford campus vision and lidar data set. *International Journal of Robotics Research*, 30(13):1543–1552, November 2011.

[10] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[11] C. Wah, S. Branson, P. Perona, and S. Belongie. Multiclass recognition and part localization with humans in the loop. In *ICCV*, 2011.

[12] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

Probabilistic Correspondence Matching using Random Walk with Restart

Changjae Oh
ocj1211@yonsei.ac.kr
Bumsub Ham
mimo@yonsei.ac.kr
Kwanghoon Sohn
khsohn@yonsei.ac.kr

School of Electrical and Electronic Engineering
Yonsei University
Seoul, Korea

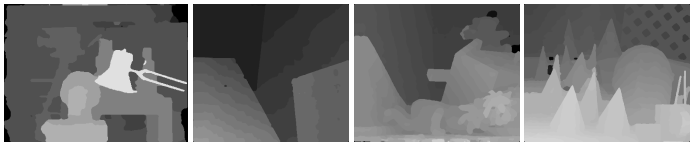


Figure 1: Disparity estimation results of the proposed method.

Introduction - Local correspondence matching methods are mainly composed of three steps: matching cost computation, cost aggregation, and disparity computation. Let us assume that a truncated absolute difference is used in matching cost computation. Each step can be represented as following probability inferring problem.

$$p_0(\mathbf{x},\mathbf{d}) = \max\left(\sigma - \|I_R(\mathbf{x}) - I_T(\mathbf{x},\mathbf{d})\|_1, 0\right) \quad (1)$$

$$p_{n+1}(\mathbf{x},\mathbf{d}) = \frac{\sum_{\mathbf{y}\in\mathcal{N}} w(\mathbf{x},\mathbf{y}) p_n(\mathbf{y},\mathbf{d})}{\sum_{\mathbf{y}\in\mathcal{N}} w(\mathbf{x},\mathbf{y})} \quad (2)$$

$$\mathbf{d}(\mathbf{x}) = \underset{\mathbf{d}\in[\mathbf{d}_1,\ldots,\mathbf{d}_D]}{\arg\max}\ p_N(\mathbf{x},\mathbf{d}) \quad (3)$$

where $\mathbf{x} = [x,y]^T$ and $\mathbf{d} = [d,0]^T$ represent the position and disparity vector, respectively. The weight between $\mathbf{x}$ and $\mathbf{y}$ is defined as $w(\mathbf{x},\mathbf{y})$. First, a matching probability $p_0(\mathbf{x},\mathbf{d})$ is computed by an absolute difference between points on the reference image $I_R(\mathbf{x})$ and on the '$\mathbf{d}$'-shifted target image $I_T(\mathbf{x},\mathbf{d})$ with the threshold σ as in Equation 1. Then, in Equation 2, the probability $p_n(\mathbf{x},\mathbf{d})$ is iteratively aggregated with $\mathbf{y}\in\mathcal{N}$ where $\mathcal{N}$ is the neighborhood of $\mathbf{x}$. Finally, an optimal disparity $\mathbf{d}(\mathbf{x})$ is selected within a search range D for the aggregated probability $p_N(\mathbf{x},\mathbf{d})$ after the maximum iteration N, by winner-takes-all (WTA) as in Equation 3.

Proposed Method - Since correspondence matching method can be regarded as probability inferring problem, the probability optimization can be used as cost aggregation. In this paper, we present the cost plane optimization using RWR framework.

Consider the cost plane as an undirected weighted graph. We present steady-state probability computation by the RWR with the given graph. Let us denote the initial cost plane as $\mathbf{P}_n^k = [p_n(\mathbf{x}_i,\mathbf{d}_k)]_{M\times 1}$ and the adjacency matrix as $\mathbf{W} = [w_{ij}]_{M\times M}$, where M and w are the size of reference image and the weight between $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. The RWR can be formulated in an iterative manner as follows:

$$\begin{aligned}\mathbf{P}_{n+1}^k &= (1-\alpha)\mathbf{D}^{-1}\mathbf{W}\mathbf{P}_n^k + \alpha\mathbf{P}_0^k \\ &= (1-\alpha)\overline{\mathbf{W}}\mathbf{P}_n^k + \alpha\mathbf{P}_0^k\end{aligned} \quad (4)$$

where n denotes the number of iterations. The initial cost $\mathbf{P}_0^k$ is returned with the probability α at each iteration. The adjacency matrix $\mathbf{W}$ is normalized as $\overline{\mathbf{W}} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D} = diag(D_1,\ldots,D_M)$, and $D_i = \sum_{j=1}^{M} w_{ij}$. When the solution reaches to the steady-state, $\mathbf{P}_n^k$ and $\mathbf{P}_{n+1}^k$ become identical, *i.e.*, the energy transition with respect to time approaches 0. Therefore, Equation 4 can be reformulated as follows:

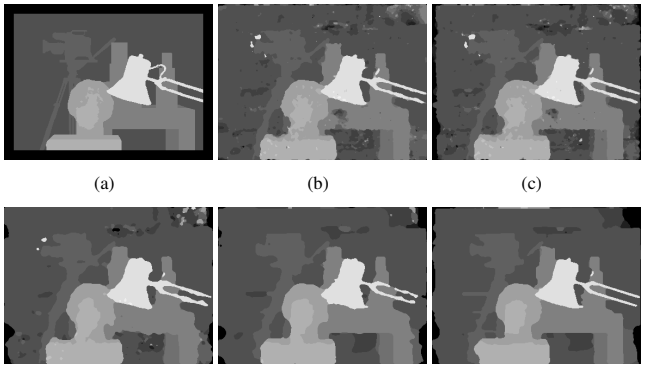
$$\begin{aligned}\mathbf{P}_s^k &= (1-\alpha)\overline{\mathbf{W}}\mathbf{P}_s^k + \alpha\mathbf{P}_0^k \\ &= \alpha(I-(1-\alpha)\overline{\mathbf{W}})^{-1}\mathbf{P}_0^k \\ &= \mathbf{R}\mathbf{P}_0^k\end{aligned} \quad (5)$$

where $\mathbf{P}_s^k$ is the cost which reaches to the steady-state. $\mathbf{R}$ can be interpreted as affinity scores between two pixels in the initial cost plane $\mathbf{P}_0^k$. With the given steady-state solution in Equation 5, disparity can be simply selected by WTA measure as following:

$$\mathbf{d}(\mathbf{x}_i) = \underset{\mathbf{d}_k}{\arg\max}\, p_s(\mathbf{x}_i,\mathbf{d}_\mathbf{k}) \quad (6)$$

where $p_s(\mathbf{x}_i,\mathbf{d}_\mathbf{k})$ is an optimized steady-state probability obtained in Equation 5, and $\mathbf{d}_k \in \{\mathbf{d}_1,\ldots,\mathbf{d}_D\}$.

The correspondence matching within the RWR framework has the following advantages: 1) A non-trivial steady-state solution is guaranteed, which means that it is not needed to specify the number of iteration. In conventional methods, it is crucial to specify the number of iteration since the performance and the computation time largely depends on this parameter [1, 2]. 2) The global relationship between points or the steady-state solution can be captured by using an adjacent neighborhood only. Accordingly, the proposed method gives high quality matching performance in a semi-global manner with low complexity.



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

Figure 2: Advantage of our method. Disparity estimation results using the smallest window in cost aggregation. (a) The ground truth. 3x3 size window are used in (b) Adaptive weight [3], (c) Costfilter [4], and 4-neighbor pixels are used in (c) Anisotropic diffusion [1], (d) Geodiff [2], (e) Proposed method.

Experimental Results - The proposed method as shown in Figure 1 was compared with other state-of-the-art cost aggregation methods: adaptive weight (AW) [3], cost filter (CF) [4], anisotropic diffusion (AD) [1], and geodesic diffusion (GD) [2]. Note that AD is not the main proposal of [1], but just the part of their method. Table 1 shows the bad matching errors evaluated by the Middlebury website [5]. The symbol '*' indicates the results of the Middlebury evaluation website. It shows that the proposed method shows competitive results with state-of-the-art methods. The comparison of the computation times of AW, CF, GD, AD, and the proposed method are 12.1, 1.18, 0.93, 2.19, 1.0, respectively, when the computation time of the proposed method is normalized to 1.0. In order to compare the performance when the window size of each algorithm is similar, we conducted another experiment by changing the window size of AW and CF to 3x3 which is the similar to that used in AD, GD and the proposed method. Figure 2b and Figure 2c show the degraded results in AW and CF, which means that the results of these methods heavily depend on the window size.

Algorithm	*Tsukuba*	*Venus*	*Teddy*	*Cones*
AW [3]	2.77	0.46	13.2	8.60
AW [3]*	1.85	1.19	13.3	9.79
CF [4]	2.14	0.46	11.5	8.01
CF [4]*	1.85	0.39	11.8	8.24
AD [1]	3.85	1.78	14.2	8.83
GD [2]	2.96	0.45	12.4	8.65
GD [2]*	2.35	0.82	11.3	8.33
Proposed method	1.97	0.38	11.5	7.92

Table 1: Object evaluation for the proposed method

[1] D. Min and K. Sohn, "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Transactions on Image Processing*, 17(8):1431-1442, August 2008.

[2] L. De-Maeztu, A. Villanueva, and R. Cabeza, "Near Real-Time Stereo Matching Using Geodesic Diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(2):410-416, February 2012.

[3] K. Yoon and I. Kweon, "Adaptive support-weight approach for correspondence search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650-656, April 2006.

[4] C. Rhemann, A. Hosni, M. Bleyer, C. Rother, and M. Gelautz, "Fast cost-volume filtering for visual correspondence and beyond," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3017-3024, 2011.

[5] `http://vision.middlebury.edu/stereo`

Corner Matching Refinement for Monocular Pose Estimation

Dinesh Gamage
dinesh.gamage@moansh.edu
Tom Drummond
tom.drummond@monash.edu
Monash University
Australia

Accurate 3D structure calculation requires reliable feature extraction and matching. To improve the accuracy of the generated hypothesis, descriptor based feature matching which works at pixel level may not be adequate and sub-pixel level information may be needed. In this paper we propose a frequency domain optimisation technique, which yields improved results and a fast convergence rate. The fast convergence is a result of the multi-resolution nature of the solution.

We make use of the affine theorem in the frequency domain [1]. Given two image patches $I_0(\bar{x})$ and $I_1(\bar{x})$ surrounding two corresponding corners, which are related by an affine coordinate transformation $I_1(\bar{x}) = I_0\left[A^{-1}\left(\bar{x}-\bar{b}\right)\right]$, where A and b are the four non translational affine parameters and two translational parameters respectively, their 2-D Fourier transforms are related by:

$$\hat{I}_1(\bar{u}) = |det\,(A)|\, e^{-j\bar{u}\cdot\bar{b}}\hat{I}_0\left(A^T\bar{u}\right) \tag{1}$$

Here we use the six parameter affine model with an additional parameter. The seventh parameter compensates for energy changes caused by different local illumination conditions. If we select $\bar{\beta} = \{\beta_1 \ldots \beta_7\}$ to be the parameter set and absorb the $|det\,(A)|$ of the equation 1 into β_7 we have:

$$\beta_7\hat{I}_1(\bar{u}) = e^{-j\bar{u}\cdot\bar{b}}\hat{I}_0\left(A^T\bar{u}\right) \qquad \text{where} \quad A = \begin{pmatrix}\beta_1 & \beta_2\\ \beta_3 & \beta_4\end{pmatrix} \quad \text{and} \tag{2}$$

$$\bar{b} = \begin{pmatrix}\beta_5\\ \beta_6\end{pmatrix} \tag{3}$$

$$\text{so} \quad \beta_7 e^{j\bar{u}\cdot\bar{b}}\hat{I}_1(\bar{u}) = \hat{I}_0\left(A^T\bar{u}\right) \tag{4}$$

Thus the error r, for a frequency $\bar{u}$ can be written as,

$$r\left(\bar{u},\bar{\beta}\right) = \beta_7 e^{j\bar{u}\cdot\bar{b}}\hat{I}_1(\bar{u}) - \hat{I}_0\left(A^T\bar{u}\right) \tag{5}$$

The above equation enables us to model the affine transformation as a phase change of $\hat{I}_1$ and a warp of $\hat{I}_0$ with respect to matrix A. The Jacobian J_i of partial derivatives of r with respect to β_i can then be computed:

$$\bar{J} = \left[-\frac{\partial I_0}{\partial u}u,\ -\frac{\partial I_0}{\partial u}v,\ -\frac{\partial I_0}{\partial v}u,\ -\frac{\partial I_0}{\partial v}v,\ -\beta_7\tilde{I}_1 u,\ -\beta_7\tilde{I}_1 v,\ \tilde{I}_1\right] \tag{6}$$

Given a set of frequencies $\{u_j\}$, the errors $r(u_j)$ and the Jacobian J_{ij} can be used to obtain the parameters $\bar{\beta}$ that minimise $E = \sum_j \|r(u_j)\|^2$ using Gauss-Newton.

FFT requires a periodic signal. So each patch has to be compensated for edge effects at the border. Secondly, the presence of a large DC component in the signal corrupts low frequency components of the signal in the frequency domain (those where $\|u\|$ is small). To remove edge effects from the image patches, we multiply it by a Gaussian weighting window ($G(x,y)$) centered at the detected landmark before taking the Fourier transform. Before doing that, the DC component of the patch which appears as a large spike at $u = 0$ in the frequency domain can be removed by subtracting the average, which gives a new patch. After alleviating both of these effects we get a new patch $I'(x,y)$ defined as:

$$I'(x,y) = G(x,y)\left(I(x,y) - \frac{\sum_{x,y}G(x,y)I(x,y)}{\sum_{x,y}G(x,y)}\right) \tag{7}$$

This gives a patch with a 0 DC coefficient. The frequency response of the Gaussian multiplied patch, $\mathscr{F}[I']$ has a direct relationship with the Gabor filter with an identical Gaussian support. We use this relationship to select the useful frequency range (in order to eliminate possible aliasing effects) for the optimisation in a multi resolution manner.

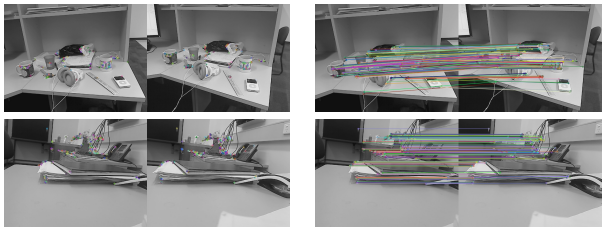


Figure 1: Pose estimation after sub-pixel refinement

Because the phase of a particular Gabor filter response changes linearly under spatial translations of the signal, it can be used to estimate spatial disparity of two instances of a the same signal with a relative shift [2]. This phase disparity is useful only if the displacement is smaller than a half a wavelength of the tuning frequency [2] of the Gabor filter, i.e the domain of convergence for the phase is $\pm\pi$. This imposes an upper frequency limit for the useful frequency range. If we assume a maximum displacement of d pixels for a 1-D signal this criteria suggests a frequency f such that $f \leq 1/2d$. In 2-D this requirement can be met by limiting the useful frequency range radially to a maximum of $1/2d$ radius. After estimating the translation (and other parameters) using small frequencies for large displacements finer refinements can be done gradually by increasing the radius, incorporating higher frequency responses to the optimisation.

Though subtracting the DC component spatially as in equation 7 can mostly reduce its effect, for better results we have to impose a lower frequency limit as well. We select the minimum frequency using the one octave bandwidth criteria which has been suggested in the literature for Gabor filter based disparity estimations. The one octave bandwidth in the frequency domain for the Gabor filter shows the spatial support to be:

$$\sigma = \frac{1}{2\pi f}\left(\frac{2^\alpha+1}{2^\alpha-1}\right) \tag{8}$$

If we select above frequency f keeping the spatial support σ a constant, in order to eliminate any DC distortion the minimum frequency should be:

$$f \geq \frac{1}{2\pi\sigma}\left(\frac{2^\alpha+1}{2^\alpha-1}\right) \tag{9}$$

Combining the minimum and the maximum criteria for frequency selection gives:

$$\frac{1}{2d} \geq f \geq \frac{1}{2\pi\sigma}\left(\frac{2^\alpha+1}{2^\alpha-1}\right) \tag{10}$$

At the end of each iteration we can expect the displacement d to reduce, increasing the useful frequency range. Higher frequencies carry finer details about the translation, which improves the final solution. This naturally enables a multi-resolution framework for refinement without any additional computations.

[1] RN Bracewell, K.Y. Chang, AK Jha, and Y.H. Wang. Affine theorem for two-dimensional fourier transform. *Electronics Letters*, 29(3): 304, 1993.

[2] D.J. Fleet, A.D. Jepson, and M.R.M. Jenkin. Phase-based disparity measurement. *CVGIP: Image understanding*, 53(2):198–210, 1991.

Unsupervised Feature Selection via Hypergraph Embedding

Zhihong Zhang[1]
zhihong@cs.york.ac.uk
Peng Ren[2]
pengren@upc.edu.cn
Edwin R. Hancock [1]
erh@cs.york.ac.uk

[1] Department of Computer Science,
The University of York,
York, UK.

[2] College of Information and Control Engineering,
China University of Petroleum,
Qingdao, China.

For the task of feature selection addressed in this paper, we introduce a hypergraph embedding view of feature selection by subspace learning. The method jointly evaluates the utility sets of features rather than individual features. There are three novel ingredients. The first is that by incorporating hypergraph representation into feature selection, we can be more effectively capture the higher order relations among samples. Secondly, inspired from the recent works on mutual information [1], we determine the weight of a hyperedge using an information measure referred to as multidimensional interaction information (MII) which precisely preserves the higher order relations captured by the hypergraph. The advantage of MII is that it is sensitive to the relations between sample combinations, and as a result can be used to seek third or even higher order dependencies among the relevant samples. Thus, the structural information latent in the data can be more effectively modeled. Finally, we describe a new feature selection strategy through hypergraph embedding, which casts the feature discriminant analysis into a regression framework that considers the correlations among features. As a result, we can evaluate joint feature combinations, rather than being confined to consider them individually.

Hypergraph Construction: we establish a novel hypergraph framework which is used for characterizing the multiple relationships within a set of samples. Based on the higher order similarity measure, we establish a hypergraph framework for characterizing a set of high dimensional samples. A hypergraph is defined as a triplet $H = (V,E,w)$. Here V denotes the vertex set, E denotes the hyperedge set in which each hyperedge $e \in E$ represents a subset of V, and w is a weight function which assigns a real value $w(e)$ to each hyperedge $e \in E$. We only consider K-uniform hypergraphs (i.e. those for which the hyperedges have identical cardinality K) in our work. Given a set of high dimensional samples $\mathbf{X} = [x_1, \cdots x_N]^T$ where $x_i \in \mathbb{R}^d$, we establish a K-uniform hypergraph, with each hypergraph vertex representing an individual sample and each hyperedge representing the Kth order relations among a K-tuple of participating samples. A K-uniform hypergraph can be represented in terms of Kth order matrix, i.e. a tensor $\mathcal{W}$ of order K, whose element $W_{i_1,\cdots,i_K}$ is the hyperedge weight associated with the K-tuple of participating vertices $\{v_{i_1}, \cdots, v_{i_K}\}$. In our work, the hyperedge weight associating with $\{x_{i_1}, x_{i_2}, \cdots, x_{i_K}\}$ is computed as follows

$$W_{i_1,\cdots,i_K} = K \frac{I(x_{i_1}, x_{i_2}, \cdots, x_{i_K})}{H(x_{i_1}) + H(x_{i_2}) + \cdots H(x_{i_K})} \ . \tag{1}$$

It is clear that $W_{i_1,\cdots,i_K}$ is a normalized version of K-th order Interaction Information. The greater the value of $W_{i_1,\cdots,i_K}$ is, the more relevant the K samples are. On the other hand, if $W_{i_1,\cdots,i_K} = 0$, the K samples are totally unrelated.

Hypergraph Representation: In our work, we consider the transformation of a K-uniform hypergraph into a graph. Accordingly, the associated hypergraph tensor $\mathcal{W}$ is transformed to a graph adjacency matrix $\boldsymbol{A}$, and the higher order information exhibited in the original hypergraph can be encoded in an embedding space spanned by the related matrix representation. In this scenario, one straightforward way for the transformation is marginalization which computes the arithmetical average over all the hyperedge weights $W_{i_1,\cdots,i_{K-2},i,j}$ associated with the edge weight $A_{i,j}$

$$\tilde{A}_{i,j} = \sum_{i_1=1}^{|V|} \cdots \sum_{i_{K-2}=1}^{|V|} W_{i_1,\cdots,i_{K-2},i,j} \tag{2}$$

The edge weight $\tilde{A}_{i,j}$ for edge ij is generated by a uniformly weighted sum of hyperedge weights $W_{i_1,\cdots,i_{K-2},i,j}$. However, the form appearing in (2) behaves as a low pass filter, and thus results in information loss through marginalization.

To make the process of marginalization more comprehensive, we use marginalization to constrain the sum of edge weights and then estimate their values through solving an over-constrained system of linear equations. Our idea is motivated by the so called *clique average* introduced in the higher order clustering literature [4]. We characterize the relationships between $\boldsymbol{A}$ and $\mathcal{W}$ as follows

$$W_{i_1,\cdots,i_K} = \sum_{\{i,j\} \subseteq \{i_1,\cdots,i_K\}} A_{i,j} \tag{3}$$

There are $\binom{|V|}{2}$ variables and $\binom{|V|}{K}$ equations in the system of equations described in (2). When $K > 2$, the linear system (2) is over-determined and cannot be solved analytically. We thus approximate the solution to (2) by minimizing the least squares error

$$\widehat{\boldsymbol{A}} = \underset{\boldsymbol{A}}{\operatorname{argmax}} \sum_{i_1,\cdots,i_K} \left(\sum_{\{i,j\} \subseteq \{i_1,\cdots,i_K\}} A_{i,j} - W_{i_1,\cdots,i_K} \right)^2 \tag{4}$$

In practical computation, we normalize the compatibility tensor $\mathcal{W}$ by using the extended Sinkhorn normalization scheme [2], and constrain the element of $\boldsymbol{A}$ to be in the interval $[0,1]$ to avoid unexpected infinities. Effective iterative numerical methods are used to compute the approximated solutions [3].

Feature Selection through Hypergraph Embedding: we formulate the procedure of feature extraction on a basis of hypergraph spectral embedding. One goal of spectral embedding is to represent the high dimensional data $\mathbf{X} \in \mathbb{R}^{N \times d}$ by a low dimensional representation $\mathbf{Y} \in \mathbb{R}^{N \times C}$ $(C \ll d)$ in the low dimensional feature space such that the structural characteristics of the high dimensional data are well preserved or are more "obvious". Here we use the representations $\mathbf{X} = [x_1, \cdots x_N]^T$ and $\mathbf{Y} = [y_1, \cdots, y_k, \cdots, y_C]$, where y_k is a N-dimensional vector and its N elements represent the N samples $x_1, \cdots x_N$ separately in the kth dimension of the low dimensional feature space.

The hypergraph embedding procedure can be viewed as feature extraction, and can be expressed as $\mathbf{Y} = \mathbf{X}\Phi$ where $\Phi \in \mathbb{R}^{d \times C}$ is a column-full-rank projection matrix. However, unlike feature extraction, feature selection attempts to select the optimal feature subset in the original feature space. Therefore, for the task of feature selection, the projection matrix $\Phi = [\Phi_1, \ldots \Phi_C]$ can be constrained to be a selection matrix which contains the combination coefficients for different features in approximating $\mathbf{Y} = [y_1, \ldots, y_C]$. That is, given the kth column of $\mathbf{Y}$, i.e y_k, we aim to find a subset of features, such that their linear span is close to y_k. This idea can be formulated as the minimization problem

$$\widehat{\Phi} = \underset{\Phi}{\operatorname{argmin}} \sum_{k=1}^{C} \|y_k - X\Phi_k\|^2 \ . \tag{5}$$

where $\Phi = [\Phi_1, \cdots, \Phi_k, \cdots, \Phi_C]$ and Φ_k is a d dimensional vector that contains the combination coefficients required to compute for different features in approximating y_k.

[1] Z. Zhang and E. R. Hancock. Hypergraph based Information-theoretic Feature Selection. *Pattern Recognition Letters*, 2012.

[2] A. Shashua, R. Zass and T. Hazan. Multi-way clustering using super-symmetric non-negative tensor factorization. *In Proc. ECCV.*, (4): 595-608, 2006.

[3] A. Björck. Numberical methods for least squares problems. *In Proc. SIAM.*, 1996.

[4] S. Agarwal, J. Lim, L. Zelnik-Manor, P. Perona, D. Kriegman, and S. Belongie. Beyond pairwise clustering. *In Proc. CVPR.*, pages 838-845, 2005.

Discriminative Hough Forests for Object Detection

Paul Wohlhart
wohlhart@icg.tugraz.at
Samuel Schulter
schulter@icg.tugraz.at
Martin Köstinger
koestinger@icg.tugraz.at
Peter M. Roth
pmroth@icg.tugraz.at
Horst Bischof
bischof@icg.tugraz.at

Institute for Computer Graphics and Vision
Graz University of Technology
Austria

Motivation and Method

Object detection models based on the Implicit Shape Model (ISM) [3] use small, local parts that vote for object centers in images. Since these parts vote completely independently from each other, this often leads to false-positive detections due to random constellations of parts. Thus, we introduce a verification step, which considers the activations of all voting elements that contribute to a detection. The levels of activation of each voting element of the ISM form a new description vector for an object hypothesis, which can be examined in order to discriminate between correct and incorrect detections.

In particular, we observe the levels of activation of the voting elements in Hough Forests [2], which can be seen as a variant of ISM. In Hough Forests, the voting elements are all the positive training patches used to train the Forest. Each patch of the input image is classified by all decision trees in the Hough Forest. Whenever an input patch falls into the same leaf node as a patch from training, a certain amount of weight is added to the detection hypothesis at the relative position of the object center, which was recorded when cropping out the training patch. The total amount of weight one voting element (offset vector) adds to a detection hypothesis (the total *activation*) can be calculated by summing over all input patches and trees in the forest. Stacking the activations of all elements gives an *activation vector* for a hypothesis.

We learn classifiers to discriminate correct and wrong part constellations based on these *activation vectors* and thus assign a better confidence to each detection. We use linear models as well as a histogram intersection kernel SVM. In the linear classifier, one weight is learned for each voting element. We additionally show how to use these weights, not only as a post processing step, but directly in the voting process. This has two advantages: First, it circumvents the explicit calculation of the activation vector for later reclassification, which is computationally more demanding. Second, the non-maxima suppression is performed on cleaner Hough maps, which allows for reducing the size of the suppression neighborhood and thus increases the recall at high levels of precision.

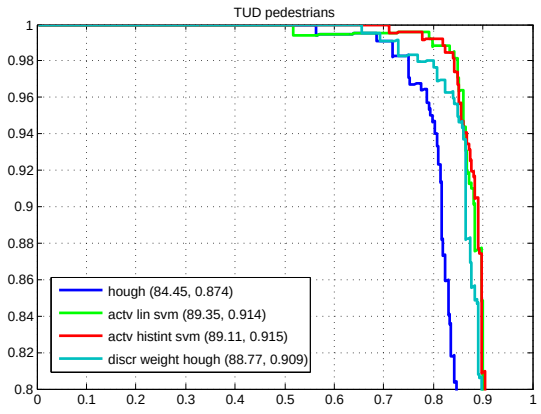


Figure 1: Precision/Recall curves on *TUD pedestrians* for standard Hough Forests [2] (*hough*), linear SVM on the activation vector (*actv lin svm*), histogram intersection kernel SVM on the activation vector (*actv histint svm*), and Hough voting with learned discriminative weights (*discr weights hough*).

Figure 2: Hough maps for an example test image from the TUD pedestrian dataset (top), with discriminative (middle) and uniform (bottom) weights

Results

The experiments on two different object classes, namely pedestrians [1] and cars [4], show significant improvements over the baseline. Visual inspection of the voting maps created with discriminatively learned voting weights (as shown for one test image in Figure 2) shows much cleaner backgrounds and clearly sharpened and pronounced peaks for correct locations. This is also reflected in the detection scores (see Figure 1 for results on *TUD pedestrians*).

This work was supported by the Austrian Science Foundation (FWF) project Advanced Learning for Tracking and Detection in Medical Workflow Analysis (I535-N23) and by the Austrian Research Promotion Agency (FFG) project SHARE in the IV2Splus program.

[1] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *Proc. CVPR*, 2008.

[2] J. Gall and V. Lempitsky. Class-specific Hough forests for object detection. In *Proc. CVPR*, 2009.

[3] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[4] B. Leibe, N. Cornelis, K. Cornelis, and L. van Gool. Dynamic 3D scene analysis from a moving vehicle. In *Proc. CVPR*, 2007.

Curvature Based Robust Descriptors

Farlin Mohideen
farlin.mohideen@anu.edu.au
Ranga Rodrigo
ranga@ent.mrt.ac.lk

Australian National University

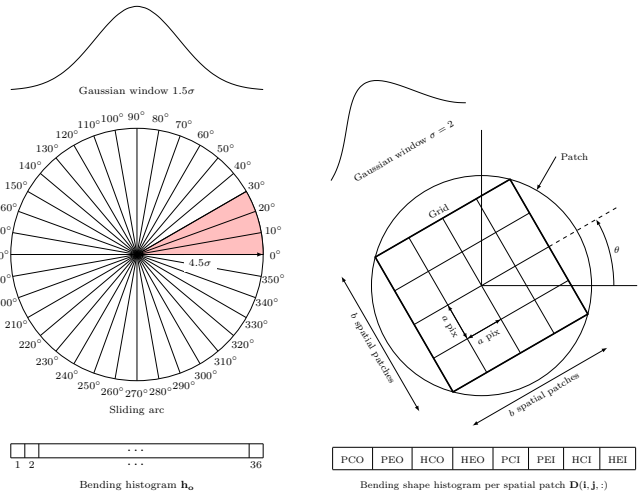
Department of Electronic and Telecommunication,
University of Moratuwa

Feature descriptors have enabled feature matching under varying imaging conditions, while mostly being backed by experimental evidence. In addition to imposing some restrictions in imaging conditions needed to ensure matching, extending the existing descriptors is not straightforward due to the lack of sound mathematical bases. In this work, by using a surface bending versus shape histogram based on the principal curvatures, we are able to produce a descriptor which is not sensitive to the errors in dominant orientation assignment. Experimental evaluations show that our descriptor outperforms existing descriptors in the areas of viewpoint, rotation, scale, zoom, lighting and compression changes, with the exception of resilience to blur. Further, we apply this descriptor for accuracy demanding applications such as homography estimation and pose estimation. The experimental results show significant improvements in estimated homography and pose in terms of residual error and Sampson distance respectively.

Eigenvalues of Hessian matrix H are the principal curvatures $\lambda_{\max}, \lambda_{\min}$ of a surface $I(x,y)$ at any given point. Let $\vec{p}$ be any point in patch S. We introduce a metric the amount of bending $m(\vec{p})$ based on rotationally invariant principle curvatures as below

$$H = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix}, \quad H\vec{v} = \lambda\vec{v}, \quad m(\vec{p}) = \sqrt{\lambda_{\max}^2 + \lambda_{\min}^2}. \tag{1}$$

We propose to represent the dominant orientation by finding where the maximum bending of the surface occurs in a sliding arc-window of 30° (Figure 1a). We compute this statistic in a circular patch and with radius proportional to scale (Figure 1a).



(a) (b)

Figure 1: An image patch divided in to 36 arc-windows. A grid superimposed on an image patch, and a grid divided into spatial patches. A Gaussian damping window is overlayed on each patch. (a) Bending histogram for dominant orientation. (b) Descriptor histogram. $a = 3\sigma$ is the width of a spatial patch. Descriptor grid width $b = 4$. Number of classification bins $c = 8$. The gaussian damping window g for descriptor is of variance $\sigma_0 = 2$. To meet these requirements and considering extra patches needed in distributing bending among adjacent bins we need to consider a patch of radius $w = \sqrt{2}a(b+1)/2 + 0.5$.

We compute the histogram h_o of bending of the surface $m(\vec{p})$ (1), vs. polar angle of the pixel $\vec{p}$ w.r.t. keypoint, multiplied by a Gaussian damping window. Thus, the bending histogram $h_o(i)$ is computed where each bin corresponds to the 10°-arc's total bending of the surface under the influence of the Gaussian. Binning is done by distributing values among adjacent bins by trilinear interpolation. Responses for the 30°-arc with 10° sliding is found by creating the final histogram $\hat{h}_o(i)$ by summing three adjacent bins of $h_o(i)$ by

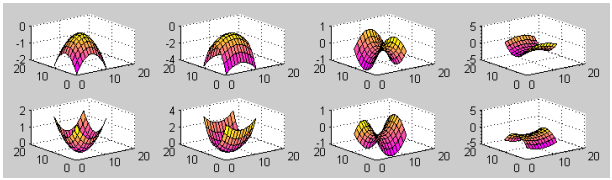


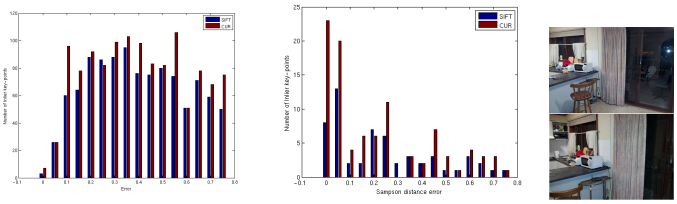
Figure 2: Shape classifications according to parabola (P), hyperbolae (H), corner (C), edge (E), outward (O) and inward (I). (a) PCO (b) PEO (c) HCO (d) HEO (e) PCI (f) PEI (g)HCI (h)HEI

$$\hat{h}_o(i) = \sum_{j=1,\ldots,3} [h_o((i+j) \bmod N)], \quad i = 1,\ldots,N. \tag{2}$$

In $\hat{h}_o(i)$ orientation values are found for the highest peak and those above 75% of the highest, followed by interpolation as in SIFT. Each dominant orientation is used to find descriptors; as in SIFT, one keypoint may have multiple descriptors. In summary, a good descriptor is preferably patch based, the grid width being proportional to the scale, resilient to misorientation due to relying on a metric like curvature which characterizes the shape of the surface at any point.

There are four steps in our feature description: (1) Computing the rotated, normalized spatial patch coordinate frame (2) Surface classification for each patch (3) Generating the descriptor vector and (4) Normalization. We use a 4×4 spatial patch grid for our descriptor, with width 12σ. For each spatial patch, we create a surface bending $m(\vec{p})$ vs shape histogram $D(i,j,:)$ (must see Figure 1b,) based on the eight classifications of the surface (must see Figure 2).

We classify the amount of surface bending according to shapes based on the ratio of principal curvatures (Figure 2). Each spatial patch produces an eight-element descriptor and all 16 spatial patches produce a 128D descriptor.



(a) Homography (b) Pose (c) Example image pair

Figure 3: Number of inlier key-points vs residual error for homography estimation and number of inlier key-points vs Sampson distance for pose estimation.

In addition to evaluation of homography estimation we evaluated pose estimation in terms of first-order geometric error (Sampson distance) by $\frac{(x'^T F x)^2}{(Fx)_1^2 + (Fx)_2^2 + (F^T x')_1^2 + (F^T x')_2^2}$. We represent the accuracy by a histogram of number of inlier key-points vs their Sampson distance as shown by Figure 3.b.

GlandVision: A Novel Polar Space Random Field Model for Glandular Biological Structure Detection

Hao Fu[1]
http://ima.ac.uk/fu
Guoping Qiu[1]
http://www.cs.nott.ac.uk/~qiu/
Muhammad Ilyas[2]
mohammad.ilyas@nottingham.ac.uk
Jie Shu[1]
http://ima.ac.uk/shu

[1] School of Computer Science
University of Nottingham
Nottingham, UK

[2] Division of Pathology
School of Molecular Medical Sciences
University of Nottingham
Nottingham, UK

Tissue diagnosis is an important part of modern day medicine. Where disease is suspected, tissue samples can be taken from the patient and viewed under the microscope by a Pathologist. In many human tissues, cells are organized into complex anatomical units called *glands*. In many disease states the glands are disrupted, often in a characteristic fashion. If automated image analysis is to be used to facilitate tissue diagnosis, then recognition of glands is essential.

A typical microscopic image of the human colon and the glands contained in it are shown in Fig.1. It can be seen there that a gland is composed of a group of cells who sit side-by-side and form the boundaries. Depending on the way the tissue has been sectioned, the shape of a gland can vary hugely and this poses significant challenge to computational algorithms for automatic gland detection.

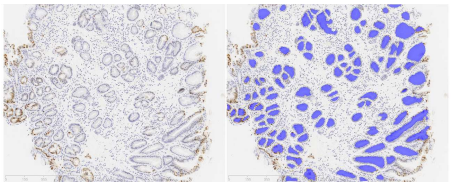


Figure 1: (a) A microscopic image of human colon tissue containing glands. (b) The glands are manually annotated in blue solid color as shown in the right image.

In this paper, we propose a novel method for detecting those glandular structures. We noticed that one of the most distinctive properties of a gland is that they usually exhibit a closed shape structure. If we place our viewpoint inside the gland, we will see a closed contour, which means if we place the co-ordinate's origin inside a gland and transform the gland to the polar space, we will see a continuous line structure along vertical direction in the polar image. Some examples are shown in Fig.2. It is seen that if the origin is inside a gland, we can see an obvious line structure in their corresponding polar image (e.g., regions circled as A, B, and C); if the origin is outside a gland there is no such line structure in its polar image (e.g., the region circled as D). Based on this observation, the problem of detecting glands can be formulated as the problem of detecting those line structures in the polar image.

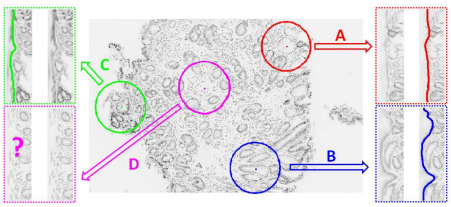


Figure 2: A, B and C correspond to cases where the polar space's origin is inside a gland while D corresponds to the case where the polar space's origin is not inside a gland. We can clearly see a continuous line structure in A, B and C, while this kind of structure can not be seen in D.

To detect the gland contours in the polar image, we developed a Conditional Random Field (CRF) model [2]. We assign each row of the polar image a label Y_i, which indicates the position of the gland contour at each row. The graphical model of our CRF contains only 360 nodes in total and is illustrated in Fig.3.

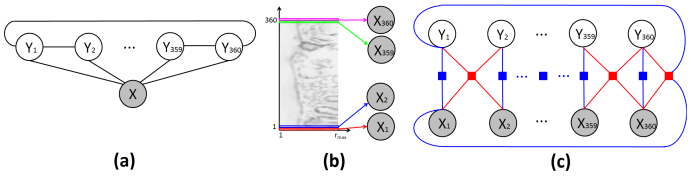


Figure 3: (a) The graphical model of our CRF model; (b) Each row of the polar image is assigned a random variable; (c) The factor graph of our CRF model.

This graph structure is a loop structure only if it contains one more edge which links Y_1 and Y_{360}, otherwise it will be a chain. To avoid the influence of this extra edge, we use two chain structures to approximate this circulate graph, thus enabling efficient inference. This is shown in Fig.4.

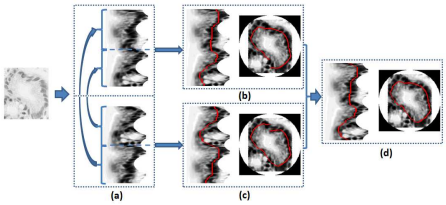


Figure 4: An Efficient Inference Strategy. (a) An image is transformed into two polar images, one's θ ranges from 0 to 2π, and the other's ranges from π to 3π. The Viterbi inference algorithm is performed separately on these two polar images, and the results are shown in (b) and (c). These two results are then combined to generate the final result shown in (d).

We treat the above random field model as a gland proposal module, and then develop another visual feature based support vector regressor (SVR) to verify if the inferred contour corresponds to a true gland. Finally, we combine [1] the outputs of the random field and the regressor to form the GlandVision algorithm for the detection of glandular structures. The flowchart of our complete GlandVision algorithm is depicted in Fig.5.

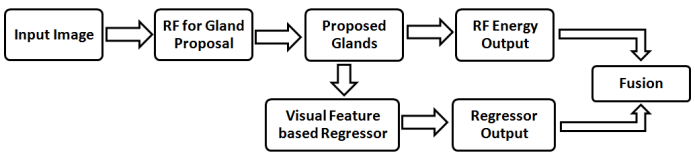


Figure 5: A Complete GlandVision Procedure

Experiments on a dataset of 20 high resolution microscopic images containing 1072 glands have shown the effectiveness of our approach.

[1] Hao Fu, Guoping Qiu, and Hangen He. Feature Combination beyond Basic Arithmetics. In *British Machine Vision Conference(BMVC)*. BMVA, 2011.

[2] John Lafferty, Andrew Mccallum, and Fernando Pereira. Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML*, 2001.

Depth Correction for Depth Cameras From Planarity

Amira Belhedi[123]
amira.belhedi@cea.fr
Adrien Bartoli[2]
http://isit.u-clermont1.fr/~ab/
Vincent Gay-Bellile[1]
vincent.gay-bellile@cea.fr
Steve Bourgeois[1]
Steve.BOURGEOIS@cea.fr
Patrick Sayd[1]
Patrick.SAYD@cea.fr
Kamel Hamrouni[3]
kamel.hamrouni@enit.rnu.tn

[1] CEA, LIST, LVIC, F-91191 Gif-sur-Yvette, France.
[2] Clermont Université, Université d'Auvergne, ISIT, BP 10448, F-63000 Clermont-Ferrand, France.
[3] Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, LR-SITI Signal Image et Technologie de l'Information, BP-37, Le Belvédère, 1002 Tunis, Tunisia.

Depth cameras open new possibilities in fields such as 3D reconstruction, Augmented Reality and video-surveillance since they provide depth information at high frame-rates. However, like any sensor, they have limitations related to their technology. One of them is depth distortion. In this article we present a new method for the correction of depth distortion.

The methods presented in literature require an accurate ground-truth for each depth-pixel. However, acquiring these reference depth is extremely difficult for several reasons. In fact, an additional system is required, *i.e.* high accuracy track line as in [3, 6] or a calibrated color camera as in [1, 7, 8] (see Figure 1). In contrast, the proposed method is more easy to use, since it does not need a large set of accurate ground truth. It is based on two steps:

Non-planarity correction (NPC): estimates a correction function: $F : \psi \rightarrow \mathbb{R}, \psi \subset \mathbb{R}^3$ such that, $F(\mathbf{Q}) = C_Z$ where ψ is a subset of $\mathbb{R}^3$ and C_Z is a scalar that represents the Z correction. NPC is based on training F: collecting a massive set of different views (different orientations and different distances) of a plan that intersect to cover all the 3D calibrated space (see Figure 2(a)), which is easy to set up. The 3D points of each view are not coplanar. This is caused by the depth distortion. The NPC principle is to train F such that the corrected points of each view tend toward coplanar points. This will constraint F up to a global 3D affine transformation A.

Affine correction (AC): estimates an affine transformation A. Any affine transformation of the corrected space will keep the planarity constraints. Estimating A (12 parameters) requires to collect a small set of ground truth measurements. AC will end up as linear least squares constraints and can be easily solved. The depth correction steps are shown in Figure 2. An iterative process is adopted to resolve the NPC step and a 3D smoothing spline, known as a 3D Thin-Plate-Spline is chosen to model F. Initially (iteration number $k = 0$), the point-to-plane distance is large (Figure 2(c)), it decreases at the next iteration (Figure 2(d)) and become very small in the last iteration (Figure 2(e)). After NPC step, the points are coplanar but not aligned with ground truth Figure 2(g). The AC step is then performed. It is shown in Figure 2(h) that after AC the obtained data are very close to the ground truth.

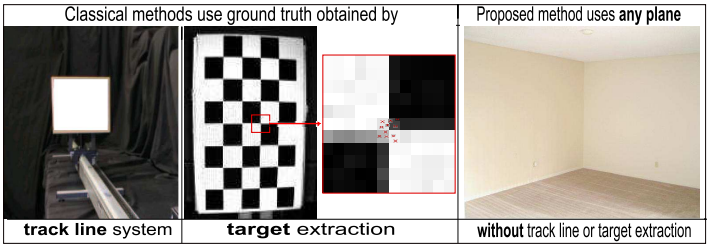
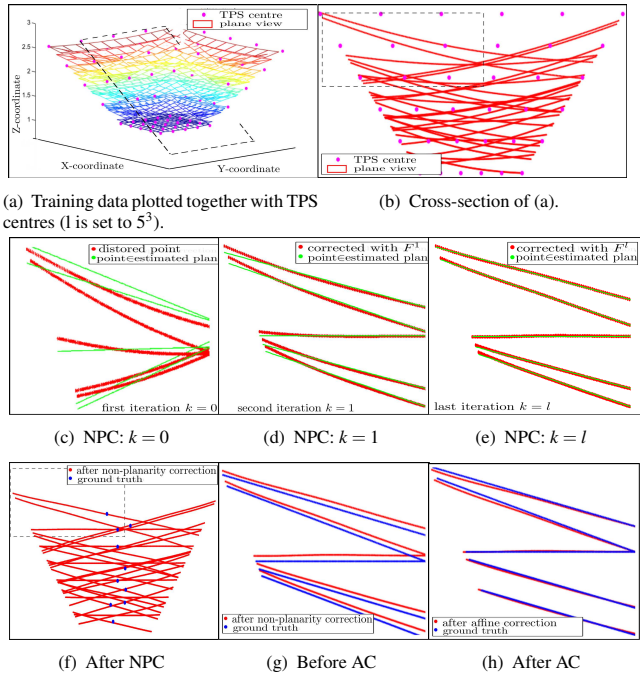


Figure 1: Classical approaches require a set of accurate ground truth that are obtained by track line system or target extraction approach. The first system is expensive. The second approach does not provide accurate ground truth: it is not feasible to extract accurate point due to the camera's low resolution (lack of accuracy at transition area): the red crosses represent the different possibilities of a corner localization. Our approach uses planar views and does not need a large number of ground truth.



(a) Training data plotted together with TPS centres (l is set to 5^3). (b) Cross-section of (a).

(c) NPC: $k = 0$ (d) NPC: $k = 1$ (e) NPC: $k = l$

(f) After NPC (g) Before AC (h) After AC

Figure 2: Simulated data results during depth correction process. (a)A part of calibrated space ranged from 1m to 2.5m. A part of (b) is considered to show obtained results at (c) first iteration, (d) second one and (e) last one of NPC. (f)A small set of reference data used to compute A plotted together with the corresponding section of training data obtained after NPC. Comparison of results (g) before and (h) after AC.

[1] A. Belhedi, S. Bourgeois, V. Gay-Bellile, P. Sayd, A. Bartoli, and K. Hamrouni. Non-parametric depth calibration of a tof camera. In *ICIP*, 2012.

[2] S. Á. Guömundsson, H. Aanæs, and R. Larsen. Environmental effects on measurement uncertainties of Time-of-Fight cameras. In *ISSCS*, 2007.

[3] T. Kahlmann, F. Remondino, and H. Ingensand. Calibration for increased accuracy of the range imaging camera SwissRanger ^{Tm} . In *IEVM*, 2006.

[4] W. Karel, P. Dorninger, and N. Pfeifer. In situ determination of range camera quality parameters by segmentation. In *Opt. 3D Meas. Tech.*, 2007.

[5] R. Lange. *3D Time-of-Flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology*. PhD thesis, University of Siegen, Germany, 2000.

[6] M. Lindner and A. Kolb. Lateral and depth calibration of PMD-distance sensors. In *ISVC*, 2006.

[7] M. Lindner and A. Kolb. Calibration of the intensity-related distance error of the PMD TOF-camera. In *IRCV*, 2007.

[8] I. Schiller, C. Beder, and R. Koch. Calibration of a PMD-camera using a planar calibration pattern together with a multi-camera setup. In *ISPRS*, 2008.

[9] C. A. Weyer, K. H. Bae, K. Lim, and D. D. Lichti. Extensive metric performance evaluation of a 3D range camera. In *ISPRS*, 2008.

Gesture-based Object Recognition using Histograms of Guiding Strokes

Amir Sadeghipour[1]
sadeghipour@uni-bielefeld.de
Louis-Philippe Morency[2]
morency@ict.usc.edu
Stefan Kopp[1]
skopp@techfak.uni-bielefeld.de

[1] Cognitive Interaction Technology - Center of Excellence,
Bielefeld University,
Bielefeld, Germany

[2] Institute for Creative Technologies,
University of Southern California,
Los Angeles, CA, USA

Humans perform iconic gestures to refer to entities through embodying their shapes. For instance, people often gesture the outline of an object (e.g. a circle for a ball) when referring to it during communication. In this paper, we present a new gesture descriptor, called Histograms of Guiding Strokes (HoGS), well-suited for automatic recognition of iconic gesture depicting 3D objects.

Iconic Gesture Dataset We created a dataset, called 3D Iconic Gesture dataset (in short, 3DIG [1]), which contains 29 subjects (20 males and 9 females) performing iconic gestures to refer to 20 different 3D objects. Using MS Kinect™, we captured 1739 gesture performances (∼87 gestures per object), each in three formats: color video, depth video and motion of the tracked skeleton (as 3D positions of 20 joints in 30 fps).

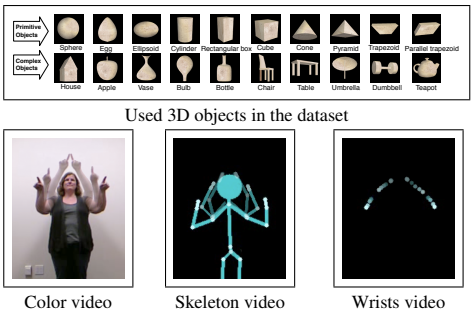


Figure 1: *top:* 3D object in the dataset, *bottom:* sample captured videos.

Analyzing the recorded gesture performances shows that people use very different techniques when performing gestures. Thus, the main challenge addressed by our HoGS descriptor is to tolerate the variations among the gesture performances, while enabling robust discriminative classification. This means that the HoGS needs to be invariant to the intra-class variabilities, while it should still represent the features which best discriminate between different classes. In the following, the mostly observed intra-class variations while analyzing the color videos of the 3DIG dataset:

- The wrists' movement direction and velocity
- The size of the used spatial space for a gesture performance
- Degree of simplification: Ignoring/considering the objects' details
- Repetition: Repeating some parts of a gestural wrist movement
- Position: The location of the used spatial space for a gesture.
- The ordering of referred components of an object.
- Handedness: The contributed hand(s).

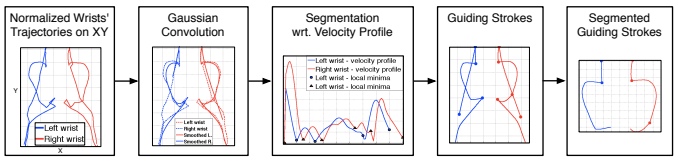


Figure 2: The processing pipeline to extract segmented guiding strokes.

Extracting Guiding Strokes To cope with the intra-class and inter-class variations, we employ the idea of decomposing the movement trajectories into *guiding strokes* [1] with respect to their spatial and kinematic features. For this aim, first we smooth the normalized 3D movement trajectories of both wrists. Then, as illustrated in Figure 2 the trajectories are split based on their velocity profiles. Finally, we segment the guiding strokes, by extracting the ones which contain the semantic content of a gesture and not the preparing and retraction movement parts. This is done automatically, through thresholding with respect to the spatiotemporal features of each guiding stroke (i.e. velocity, position and ordering). As a result, a gesture is represented as a set of guiding strokes. Each guiding stroke is a segment of the trajectory which can be parametrized. Such a parameterized representation of gestures makes it possible to extract gesture features from the attributes that reduces the intra-class variation (see Table 1).

[1]The dataset is available online at http://projects.ict.usc.edu/3dig

Computing HoGS We propose Histograms of Guiding Strokes (HoGS) as descriptors for iconic gesture recognition. To this end, we compute the histograms of the relevant attributes of the guiding strokes (see Table 1). This is done for the guiding strokes laying on each projection plane (*xy*, *xz* or *yz*) separately.

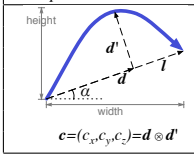
Sample GS.	*Attribue*	*Domain of values*
	Length (normalized)	$(0,1]$
	Width (normalized)	$(0,1]$
	Height (normalized)	$(0,1]$
	Curvature $= \lvert d\rvert/\lvert l\rvert$	$[0,1]$
	Curvature side $\equiv sign(c_z)$	$\{L,R\}$
	Skewness $\propto \lvert d'\rvert/\lvert l\rvert$	$[-0.5,0.5]$
	Orientation $= \alpha$	$[0,\pi)$

Table 1: The extracted attributes of guiding strokes for HoGS.

Results For the classification of iconic gestures in 3DIG dataset, we applied 1-to-1 Support Vector Machines (SVM) using HoGS as descriptor. In order to validate our approach, we designed different experiments to test and compare it with respect to three aspects. First, we compared the performance of the proposed HoGS descriptor to the descriptors applied in sketch recognition applications. Second, the performance of the SVM-based classification is evaluated against common gesture recognition approaches working with instantaneous features. Third, in an online study we asked people to recognize the captured gesture based on different video types (see Figure 1, bottom row). We applied the resulted human judgment performance as a scoring ground-truth for this task.

Classifier	*Descriptor*	*F_1 of all gestures*	*F_1 of drawing gestures*
SVM	**HoGS**	**0.61**	**0.77**
SVM	pen gesture features [2]	0.53	0.57
Bag of Trees	HoGS	0.48	0.53
NN+DTW	Instantaneous	0.44	0.40
HMM	Instantaneous	0.33	0.44
Humans	**Color video**	**0.74**	**0.76**
Humans	Skeleton video	0.38	0.44
Humans	Wrists video	0.34	0.43

Table 2: The classification results of 20 objects (i.e. chance level is 0.05)

Our approach (SVM with HoGS features) outperforms the other alternatives and compares favorably to human judgment performance. Instantaneous features, which are used in common gesture recognition approaches, do not achieve any accurate result in this dataset. The gestures performed through drawing technique are better classified in all conditions. This difference can be observed even in humans' judgment, yet very slightly when watching color videos.

[1] Stefan Kopp and Ipke Wachsmuth. Synthesizing multimodal utterances for conversational agents: Research articles. *Comput. Animat. Virtual Worlds*, 15(1):39–52, 2004.

[2] D.J.M. Willems and R. Niels. Definitions for features used in online pen gesture recognition. Technical report, NICI, Radboud University Nijmegen, 2008.

Prime Shapes in Natural Images

Qi Wu
http://www.cs.bath.ac.uk/~qw219
Peter Hall
http://www.cs.bath.ac.uk/~pmh

Media Technology Research Centre
Department of Computer Science
University of Bath,
Bath, UK

Shape has been well studied in many disciplines, yet to the best of our knowledge the question as to whether there is a set of elementary planar shapes that appear commonly in the world around us has never been asked. If such a set exists, then the elemental shapes could play a similar role in shape analysis as the primary colours do in colour analysis. This paper uses a fully unsupervised framework to find out the 'primary shapes' in image segmentations. It conclude that the most common of those found are familiar enough to be named: shapes such as triangles, squares and circles (more exactly, these shapes up to affine transformation). We propose to use qualitative shapes as features in future applications. For example, hierarchies of qualitative shape can be used for cross-modal matching [2].

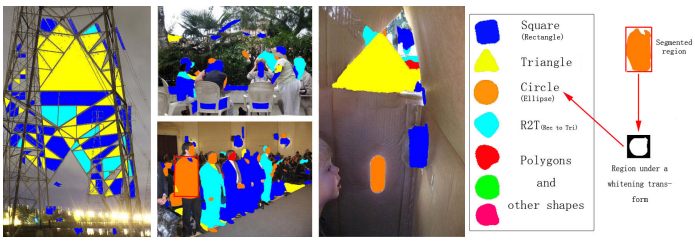


Figure 1: Segmented regions classified by prime shapes obtained from MIT database. Example: the segmented torso of a man is classified as an ellipse/circle.

Our proposition has its roots in Art, most particularly 20th century Western Art. Painters such as Picasso (*e.g.* Seated Woman with Wrist Watch), Leger (*e.g.* Card Players), and schools such as Italian Futurism, Tubism, and Orphism, depicted objects (and motions) as being composed of just a few basic geometric forms: cones, cylinders, bricks and so on. Additionally, it is very common for artists to make initial sketches using simple shapes to layout a scene, as any book on drawing instruction will testify. Empirical evidence that aligns with artistic intuition has existed since at least the 1970's, when psychologists such as Rosch [4] showed simple shapes (specifically triangles, squares, and circles) are easier for humans to recall that other shapes. This paper provides evidence that simple shapes are integral to what might be called 'the visual signal'. As Figure 1 shows, we can classify image regions into qualitative shape that have been learned without supervision.

Our experiment is designed to find out whether common simple shapes objectively exist in image segmentations. We wish to remove as much bias as we can, so supervised methods are ruled out and we have been sure to use a range of segmentation methods, shape descriptors, and databases. Importantly, *we will not define simple shapes in advance, rather they should be an emergent property based on image statistics*. Our approach is to automatically cluster regions that have been segmented from images, and compare these clusters with those created from a database of randomly created images; there is no human interaction at all.

Three main steps are included:

- Shape Production: we choose three types of segmentation methods to produce shapes, which are *Thresholding*, *MSER* [3] and *Berkeley Segmentor* [1].

- Shape Description: we opted for *Zernike Moments* [5] and *Chebyshev Moments* as the shape descriptor. Before the computing its description, we apply a whitening transfer that brings the region into the unit disc. This will map any triangle into equilateral form, any rectangle into a square, and any ellipse into a circle. However, scaling into the unit disc changes the effective sample rate. To make sure that this plays no role in moments competition, we resample the shapes into a 50^2 binary images.

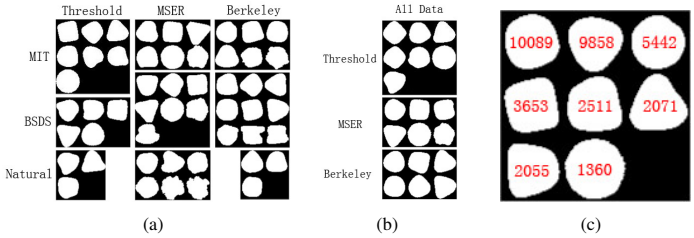


Figure 2: A matrix of final results. (a) Each entry shows the shape icons yielded by different databases, different segmentation methods. (b) Shape icons for different segmentation methods yielded by combining all three databases, different segmentation methods. (c): Final grouping result by combing all databases and segmentations. The number of each prime shape is plotted in each corresponding icon. The total number of segmented regions, classified or not, is 56992

- Shape Classification: this phase consists of two steps, mean-shift comes first and then an agglomerative clustering algorithm is applied to do the second times grouping. To locate statistically significant clusters in shapes drawn from an image database we count the total number of shapes in a cluster of a given size to get $p(m|D)$, which is the probability of observing a cluster of size m, given source $D \in \{\text{Image Database}, \text{Random}\}$. We keep only those clusters of size m for which $p(m|\text{Image Database}) > p(m|\text{Random})$.

Final shapes for each database and each segmentation method can be seen in Figure 2. The shapes tend to be be simple and nameable; shapes such as circle, square and triangle are common. There are some irregular looking shapes too, but these are not often observed compared to the regular shapes.

During the research, we also noticed that the priors on different prime shapes depends on the database used, and these contain different sorts of photograph. This suggests a scene classification application which can be found in our paper.

The main contribution of this paper is a discovery which is unique, so far as we know: regions in image segmentations naturally form classes that correspond to simple, easily recognisable shapes. In fact, more than half of the segmented regions in the datasets can be classified into prime shapes. We found this to be true no matter what segmentation algorithm we used, no matter what databased we used, an no matter how we described the shape of segmented regions. There are no arbitrary parameters in our clustering algorithm, which is fully unsupervised. In short, we have provided empirical evidence to suggest that *natural images contain simple shapes to a statistically significant degree*. And the results clearly show prime shapes emerging from segmentations: they are 'features in the signal', and as such may be of use to many applications in Computer Vision and maybe elsewhere, not just scene classification.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2294 – 2301, June 2009.

[2] A. Balikai, P. Rosin, Y.-Z. Song, and P.M. Hall. Shapes fit for purpose. In *British Machine Vision Conference*, 2008.

[3] J Matas, O Chum, M Urban, and T Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761 – 767, 2004. ISSN 0262-8856.

[4] Eleanor H. Rosch. Natural categories. *Cognitive Psychology*, 4(3): 328 – 350, 1973. ISSN 0010-0285.

[5] M.R. Teague. Image analysis via the general theory of moments. *JOSA*, 70(8):920–930, 1980.

An Evaluation of Local Shape Descriptors in Probabilistic Volumetric Scenes

Maria I. Restrepo
http://vision.lems.brown.edu/students/restrepo
Joseph L. Mundy
http://vision.lems.brown.edu/faculty/mundy

LEMS Laboratory
School of Engineering
Brown University
Providence, RI. USA

Motivation

Understanding the three-dimensional world from images is a long standing goal in computer vision, with numerous applications in the areas of robotics, autonomous navigation, city mapping and surveillance. Multi-view stereo can be thought of as the initial step towards this goal and it has been widely studied by the scientific community. However, few of the available methods can perform image-based 3-d modeling of outdoor, crowded, large scale scenes, where accurate modeling is difficult due to severe occlusion, varying illumination conditions, the presence of highly reflective surfaces and sensor errors. In this realm, probabilistic volumetric methods offer a dense representation for the solution of the multi-view stereo problem, modeling explicitly the scene's uncertainty. In particular, Pollard and Mundy [5] propose a volumetric Bayesian framework that is robust to large variations in appearance from a video sequence. This model is extended into a continuous framework by Crispell *et al*. [1] allowing for an octree subdivision of space. Finally, a recent GPU implementation [4] of this model, capable of processing one HD-resolution frame per second, guarantees scalability to large urban scenes and encourages its application to higher level 3-d computer vision tasks.

Inspired by the growing number of applications of the probabilistic volumetric model (PVM) to 3-d scene modeling and understanding, this work aims to provide the first evaluation of the performance of several local shape descriptors extracted from the PVM in terms of accuracy for object classification. Descriptors based on local histograms are of particular interest as they are the most popular and most successful for many image indexing applications. While the performance of many 3-d shape descriptors has been studied in point cloud data (from range sensors, or computer generated meshes) it is unclear that their descriptiveness and robustness to noise successfully extends to the diffuse surface probability distributions of the PVM. Surfaces can be poorly localized due to the presence of highly reflective materials, large regions of constant intensity and challenging illumination conditions in the input image data. This work takes a step towards the characterization of the PVM as a new representation for 3-d scene understanding that provides a dense continuous representation of scene geometry despite of all these sources of ambiguity.

Evaluation Framework

This papers evaluates four popular shape descriptors, namely Spin Images (SI) [3], 3-d Shape Contexts (SC) [2] , Signatures of Histograms of OrienTations (SHOT) [8] and Fast Point Feature Histogram (FPFH) [7]. For evaluation, the proposed framework starts by learning the probabilistic models for 17 urban scenes from publicly available aerial imagery collected from a helicopter flying around Providence, RI USA. After learning the PVM, surface normals are computed by convolving three-dimensional Gaussian derivative kernels with the volumetric surface probabilities. Surface normals and their locations are used to compute the local 3-d descriptors. Finally, the descriptors are used to learn *Bag of Words* models for five object categories: planes, cars, houses, buildings and parking lots. During learning and categorization, the objects are manually segmented and labeled using the bounding boxes provided in [6]. The *Bag of Words* framework follows common practices, where a common vocabulary is learned for all categories using k-means clustering. Naive Bayes, SVM and Nearest Neighbors classifiers are compared during the classification stage. To ensure robustness of the evaluation, the results are reported for various vocabulary sizes, descriptor parameters and classifiers across multiple splits of the train/test data. A special effort is made to report all parameters used through out this work, such that this evaluation can provide precise guidance to future works using the PVM. An overview of the proposed pipeline is presented in Figure 1.

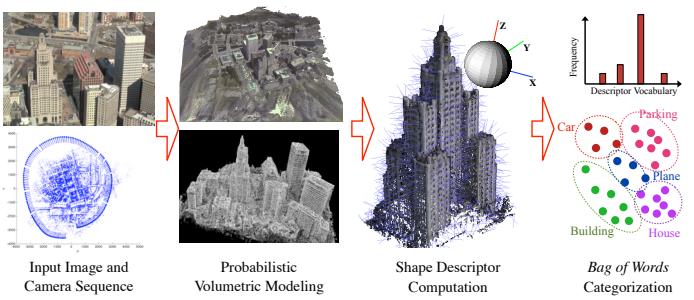


Figure 1: Framework overview. First, an input image and camera sequence are used to learn the PVM. Then, surface normals and shape descriptors are computed at highly likely surface locations. Finally, a *Bag of Words* model is used to learn and classify 5 object categories.

Results

Some of the results presented in this paper are shown in Figure 2. Under the different test scenarios, the FPFH obtained high recall while having the advantage of being compact and fast to compute. Spin Images underperformed, in particular when recognizing buildings. The SVM classifier was the more effective classifier. Shape contexts achieve adequate classification performance, but have very high storage requirements, posing run-time challenges during batch k-means. The results indicate that distribution-based descriptors effectively extract salient characteristics of the shape information in the PVM for object categorization. This work provides guidance on the selection of descriptors and parameters for characterization of the PVM, making a fundamental step on the understanding of the shape information in the PVM.

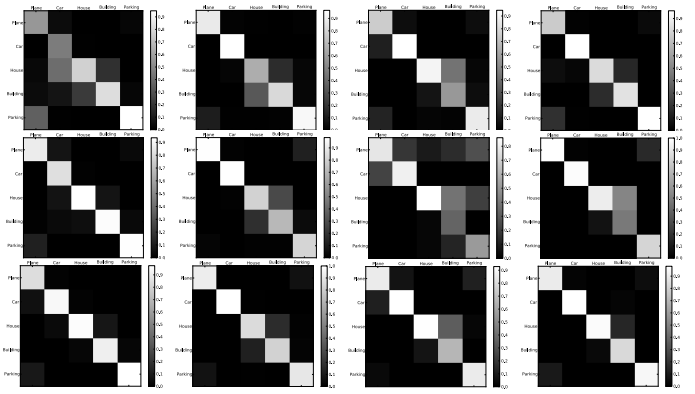


Figure 2: Average confusion matrices (across trials) . Columns (left to right): FPFH, SHOT, SI, SC. Rows (top to bottom): Nearest Neighbor, Bayes, SVM. In all cases $r_{supp} = 30$ and $K = 500$

[1] D Crispell, J Mundy, and G Taubin. A Variable-Resolution Probabilistic Three-Dimensional Model for Change Detection. *IEEE Trans. Geosci. Remote Sens.*, 2011.

[2] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing Objects in Range Data Using Regional Point Descriptors. In *ECCV*, 2004.

[3] A.E Johnson and M Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *PAMI*, 1999.

[4] Andrew Miller, Vishal Jain, and Joseph Mundy. Real-time Rendering and Dynamic Updating of 3-d Volumetric Data. In *Workshop on GPGPU*, 2011.

[5] T Pollard and J.L Mundy. Change Detection in a 3-d World. In *CVPR*, 2007.

[6] Restrepo, M.I, Mayer, B.A, Ulusoy, A.O. and Mundy, J.L. Characterization of 3-d Volumetric Scenes for Object Recognition. *IEEE J. Sel. Topics Signal Process.*, 2012.

[7] R.B Rusu, N Blodow, and M Beetz. Fast Point Feature Histograms (FPFH) for 3D Registration. In *ICRA*, 2009.

[8] Federico Tombari, Samuele Salti, and Luigi Di Stefano. Unique signatures of histograms for local surface description. In *ECCV*, 2010.

Learning geometrical transforms between multi camera views using Canonical Correlation Analysis

Christian Conrad
conrad@vsi.cs.uni-frankfurt.de
Rudolf Mester
mester@vsi.cs.uni-frankfurt.de

Visual Sensorics and Information Processing Lab
Goethe University, Frankfurt am Main, Germany
Computer Vision Laboratory
Electr. Eng. Dept. (ISY), Linköping University, Sweden

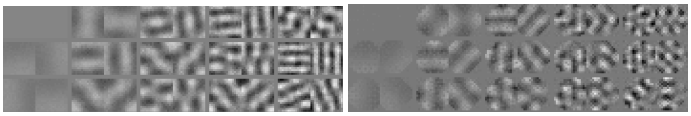


Figure 1: **CCA on natural images:** (left) Filters learnt on pairs of natural images related by a (left) 90 degree rotation and (right) 45 degree rotation.

We present an unsupervised and sampling-free approach to learn the correspondence relations between pairs of cameras in closed form employing a linear model known as Canonical Correlation Analysis (*CCA*). The only assumption we make is that the relative orientation between the cameras involved is fixed. In a two stage algorithm, we first learn the inter-image transformation based on CCA. This analysis usually has to be done in a multi-scale framework, as applying CCA directly to full resolution images may be computationally prohibitive. In the second stage we employ the learnt transformation which is given only implicitly and predict for a given pixel in a first view its corresponding region within a second view. We denote these regions as correspondence prior.

CCA has been introduced by Hotelling [2] as a method of analyzing the relations between two sets of variates and can be applied in closed form. Consider two random vectors $\mathbf{x}$ and $\mathbf{y}$ where $\mathbf{x} \in \mathbb{R}^N$ and $\mathbf{y} \in \mathbb{R}^M$. The goal of CCA is to find basis vectors for which the correlation between $\mathbf{x}$ and $\mathbf{y}$ when projected onto the basis vectors are mutually maximized [3]. In the case of a single pair of basis vectors $\mathbf{u} \in \mathbb{R}^N, \mathbf{v} \in \mathbb{R}^M$ the projections are given as $a = \mathbf{u}^T\mathbf{x}$ and $b = \mathbf{v}^T\mathbf{y}$. Assuming $\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{y}] = 0$, the correlation ρ between a and b can be written as:

$$\rho(a,b) = \frac{\mathbb{E}[ab]}{\sqrt{\mathbb{E}[aa]\mathbb{E}[bb]}} = \frac{\mathbb{E}[(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})]}{\sqrt{\mathbb{E}[(\mathbf{u}^T\mathbf{x})(\mathbf{u}^T\mathbf{x})]\mathbb{E}[(\mathbf{v}^T\mathbf{y})(\mathbf{v}^T\mathbf{y})]}}, \quad (1)$$

$$= \frac{\mathbf{u}^T\mathbb{E}[\mathbf{x}\mathbf{y}^T]\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbb{E}[\mathbf{x}\mathbf{x}^T]\mathbf{u}\mathbf{v}^T\mathbb{E}[\mathbf{y}\mathbf{y}^T]\mathbf{v}}} = \frac{\mathbf{u}^T\mathbf{C}_{xy}\,\mathbf{v}}{\sqrt{\mathbf{u}^T\mathbf{C}_{xx}\,\mathbf{u}\mathbf{v}^T\mathbf{C}_{yy}\,\mathbf{v}}}. \quad (2)$$

Note that (2) does not depend on the actual scaling of $\mathbf{u}$ or $\mathbf{v}$, therefore in the case of a single pair of basis vectors CCA can formally be defined as solving the following optimization problem:

$$\max_{\mathbf{u},\mathbf{v}} \mathbf{u}^T\mathbf{C}_{xy}\,\mathbf{v} \qquad \text{s.t.} \quad \mathbf{u}^T\mathbf{C}_{xx}\,\mathbf{u} = \mathbf{v}^T\mathbf{C}_{yy}\,\mathbf{v} = 1. \quad (3)$$

It can be shown that (3) can be cast as a generalized eigenproblem with $K = min(N,M)$ solutions, defining two sets of basis vectors $\{\mathbf{u}_k\}$ and $\{\mathbf{v}_k\}$ with $k = 1,..,K$ [1]. The projections $a_k = \mathbf{u}_k^T\mathbf{x}$ and $b_k = \mathbf{v}_k^T\mathbf{y}$ are uncorrelated which implies that $\mathbf{u}_i^T\mathbf{u}_j = 0$ and $\mathbf{v}_i^T\mathbf{v}_j = 0$ for $i \neq j$. Assuming that $N = M$, and arranging the basis vectors column wise such that $\mathbf{U} = \{\mathbf{u}_k\}$ and $\mathbf{V} = \{\mathbf{v}_k\}$ define an orthogonal basis for the random vectors $\mathbf{x}$ and $\mathbf{y}$, respectively [1]. Figure 1 shows filters learned with CCA on pairs of natural images.

Given a binocular image stream, we learn the inter-image transformation by applying CCA on data matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{T\times N}$, where each row in $\mathbf{X}$ and $\mathbf{Y}$ correspond to a subsampled version of the original image, respectively. We whiten the two data matrices before applying CCA such that $\mathbf{C}_{xx} = \mathbf{C}_{yy} = \mathbb{I}$ holds. Then the two constraints within (3) relax to $\mathbf{u}^T\mathbf{u} = \mathbf{v}^T\mathbf{v} = 1$ and the sets of basis vectors determined by CCA will form orthonormal bases which allows us to perform the prediction in closed form. We whiten the data matrices using PCA from which we obtain matrices $\mathbf{W}_\mathbf{x}$, $\mathbf{W}_\mathbf{y} \in \mathbb{R}^{N\times N}$ containing the PCA basis vectors for our given data in $\mathbf{X}$, and $\mathbf{Y}$. Given that CCA is able to perfectly learn the transformation then the correlation between a pair of corresponding patches $(\mathbf{x},\mathbf{y})$ will be maximum iff

$$\mathbf{U}_w^T(\mathbf{W}_\mathbf{x}^T\mathbf{x}) = \mathbf{V}_w^T(\mathbf{W}_\mathbf{y}^T\mathbf{y}). \quad (4)$$

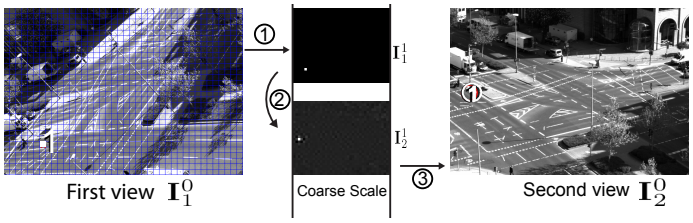


Figure 2: Visual description of how correspondence priors are generated, once the transformation between two views has been learnt via CCA.

This implies that we can predict $\mathbf{y}$ from $\mathbf{x}$ and vice versa which is the same as applying the learnt transformation to $\mathbf{x}$ or $\mathbf{y}$. Solving (4) for $\mathbf{x}$ we can predict $\mathbf{x}$ from $\mathbf{y}$ as:

$$\mathbf{x} = \mathbf{W}_\mathbf{x}(\mathbf{U}_w(\mathbf{V}_w^T(\mathbf{W}_\mathbf{y}^T\mathbf{y}))). \quad (5)$$

Correspondence priors are then generated as follows: Let (x,y) be the spatial coordinates of a pixel in $\mathbf{I}_1^0$, where the superscript 0 denotes the original resolution. Next, determine the pixels' coordinates (u,v) within the low resolution, and generate a binary image of the same size as the low resolution where the pixel at (u,v) is set to 1 (step 1 in Fig. 2). Using (5) we apply the learned transformation to the binary image and obtain the predicted image. When there is a one-to-one pixel correspondence and the transformation has perfectly been determined, the predicted image will be binary again where a single pixel is set to 1 marking the correspondence. However, due to noise one typically obtains predictions that encode regions of high probability containing the correspondence (step 2 in Fig. 2). Interpreting the predicted images as an empirical bivariate correspondence distribution over the spatial image coordinates we encode the prediction by means of a 2×2 covariance matrix $\mathbf{C}_p$. Based on an eigenvalue analysis of $\mathbf{C}_p$, we consider that a correspondence exists if both eigenvalues of $\mathbf{C}_p$ are small. Finally, the correspondence prior to the pixel at $(x,y) \in \mathbf{I}_1^0$ in the second view is given as the covariance error ellipse from $\mathbf{C}_p$ projected onto $\mathbf{I}_2^0$ (step 3 in Fig. 2). Figure 3 shows several correspondence priors and correspondence maps for different real world setups.

[1] T.W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, 1958.

[2] H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936.

[3] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*. Academic Press, 1980.

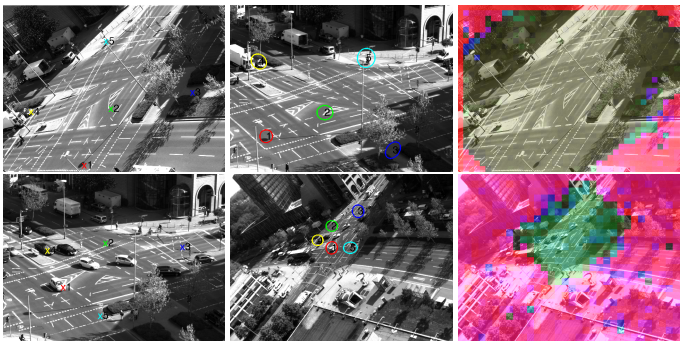


Figure 3: Correspondence priors and maps for real world setups. For selected pixels within the first view (first column), regions of high probability containing the true corresponding pixel in second view (middle column) are determined. (right column) Correspondence maps. Regions colored in shades of purple have a high probability of not being visible in the first view.

Structured Learning for Multiple Object Tracking

Wang Yan
wy109@cs.rutgers.edu
Xiaoye Han
xiaoye@cs.rutgers.edu
Vladimir Pavlovic
vladimir@cs.rutgers.edu

Department of Computer Science,
Rutgers, The State University of
New Jersey
USA

Adaptive tracking-by-detection methods use previous tracking results to generate a new training set for object appearance, and update the current model to predict the object location in subsequent frames. Such approaches are typically bootstarpped by manual or semi-automatic initialization in the first several frames. However, most adaptive tracking-by-detection methods focus on tracking of a single object or multiple unrelated objects. Although one can trivially engage several single object trackers to track multiple objects, such solution is frequently suboptimal because it does not utilize the inter-object constraints or the obejct layout information [2].

We propose in this paper an adaptive tracking-by-detection method for multiple objects, inspired by recent work in [1] and [2]. The constraints for structured Support Vector Machine (SVM) in [1] are modified to localize multiple objects simultaneously with both appearance and layout information. Moreover, additional binary constraints are introduced to detect the existences of respective objects and to prevent possible model drift. Thus the method can handle frequent occlusion in multiple object tracking, as well as objects entering or leaving the scene. Those binary constraints make the optimization problem significantly different from the original Structured SVM [3]. The inter-object constraints, embedded in a linear programming technique similar to [2] for optimal position assignment, are applied to diminish false detections.

In single object tracking case, given a set of frames $\{x_1, x_2, \ldots, x_n\}$ indexed by time, and the corresponding set of labeling, i.e. bounding box, $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, structured SVM tries to find a model $f(x, \mathbf{y})$, such that the task of predicting object location in a testing frame x could be conquered by maximizing:

$$f(x, \mathbf{y}) = \langle \mathbf{w}, \Psi(x|_{\mathbf{y}}) \rangle, \tag{1}$$

where $x|_{\mathbf{y}}$ is the patch of frame x within bounding box $\mathbf{y}$ or the features extract from it, and $\Psi(\cdot)$ is the mapping from input space to the implicit features space. The resulted optimization problem is the following,

$$\min_{\mathbf{w}, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{n} \xi_i \qquad \text{s.t.} \tag{2}$$

$$\langle \mathbf{w}, \Psi(x_i|_{\mathbf{y}_i}) - \Psi(x_i|_{\mathbf{y}}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \quad , \quad i = 1, 2, \ldots, n, \quad \mathbf{y} \neq \mathbf{y}_i \tag{3}$$

where $\xi_i \geq 0$, $\mathbf{y} \neq \mathbf{y}_i$ implies bounding box $\mathbf{y}$ in (3) could be anywhere else other than groundtruth $\mathbf{y}$, and $\Delta(\mathbf{y}_i, \mathbf{y}) = 1 - \frac{\mathbf{y}_i \cap \mathbf{y}}{\mathbf{y}_i \cup \mathbf{y}}$ is the loss of predicting $\mathbf{y}$ when groundtruth is $\mathbf{y}_i$.

The tracker could track slowly changing object due to its adaptive nature, but it is also likely to drift when the object is occluded or out of the scene. For selective adaptation and suppression of drifting, binary constraints are added. Suppose Z is the training set of the object detector, and each $\mathbf{z} \in Z$ has the label $l_{\mathbf{z}} \in \{+1, -1\}$. For each $\mathbf{z} \in Z$, the binary constraint is

$$l_{\mathbf{z}}(\langle \mathbf{w}, \Psi(\mathbf{z}) \rangle + b) \geq 1 - \eta_{\mathbf{z}} \quad , \quad \forall \mathbf{z} \in Z, \tag{4}$$

where b is the bias and $\eta_{\mathbf{z}} \geq 0$. (4) favors the sample $\mathbf{z}$ which is correctly classified by current model. The overall objective function (2) becomes

$$\min_{\mathbf{w}, b, \xi, \eta} \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{n} \xi_i + C_2 \sum_{\mathbf{z} \in Z} \eta_{\mathbf{z}}. \tag{5}$$

Objective function (5), constraints (3)(4) and slack variable constraints lead to a new optimization problem, which could be recognized as a combination of structured SVM and binary SVM.

In multiple object case, compared with the single object version, we add constraint that two or more objects can not appear in the same location in one frame, as well as the objects layout information. The training set includes a frame set $\{x_1, x_2, \ldots, x_n\}$ indexed by time, and $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n\}$ is the correspond set of structured labels, where $\mathbf{Y}_i = (\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \ldots, \mathbf{y}_i^{(K)})$ indicates the bounding boxes corresponding to K objects in frame i. If the k-th object does not appear in the i-th frame, $\mathbf{y}_i^{(k)} = null$. We design a function $f(x, \mathbf{Y})$ such that the object locations $\mathbf{Y}^*$ in frame x are given by maximizing

$$f(x, \mathbf{Y}) = \sum_{k=1}^{K} \langle \mathbf{w}^{(k)}, \Psi(x|_{\mathbf{y}^{(k)}}) \rangle + \langle \mathbf{v}, \Phi(\mathbf{Y}; \mathbf{Y}_{i-1}) \rangle, \tag{6}$$

where $\mathbf{Y}_{i-1}$ is the layout in $i-1$-th frame and $\Phi(\mathbf{Y}; \mathbf{Y}_{i-1})$ is the layout feature of size $\binom{K}{2} \times 2$, whose k-l-j-th element is

$$\Phi_{klj}(\mathbf{Y}; \mathbf{Y}_{i-1}) = \begin{cases} \left| \left(\mathbf{y}_{i-1}^{(k)}(j) - \mathbf{y}_{i-1}^{(l)}(j) \right) - \left(\mathbf{y}^{(k)}(j) - \mathbf{y}^{(l)}(j) \right) \right| & \text{if } \mathbf{y}_{i-1|i}^{(k|l)} \neq null \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

while $\mathbf{y}(1)$ and $\mathbf{y}(2)$ are the horizontal and vertical coordinates of the bounding box $\mathbf{y}$'s center, respectively. The model leads the following optimization.

$$\min_{\mathbf{w}, \mathbf{v}, \xi, \eta} \frac{1}{2}\left(\sum_{k=1}^{K} \|\mathbf{w}^{(k)}\|^2 + \|\mathbf{v}\|^2\right) + C_1 \sum_{i=2}^{n} \xi_i + C_2 \sum_{k=1}^{K} \sum_{\mathbf{z} \in Z} \eta_{\mathbf{z}} \quad \text{s.t.} \tag{8}$$

$$\sum_{k=1}^{K} \langle \mathbf{w}^{(k)}, \Psi(x_i|_{\mathbf{y}_i^{(k)}}) - \Psi(x_i|_{\mathbf{y}^{(k)}}) \rangle + \langle \mathbf{v}, \Phi(\mathbf{Y}_i; \mathbf{Y}_{i-1}) - \Phi(\mathbf{Y}; \mathbf{Y}_{i-1}) \rangle \geq \Delta^M(\mathbf{Y}_i, \mathbf{Y}) - \xi_i, \forall i, \mathbf{Y} \neq \mathbf{Y}_i \tag{9}$$

$$l_{\mathbf{z}^{(k)}}(\langle \mathbf{w}^{(k)}, \Psi(\mathbf{z}^{(k)}) \rangle + b^{(k)}) \geq 1 - \eta_{\mathbf{z}^{(k)}}, \quad \forall k, \quad \forall \mathbf{z}^{(k)} \in Z^{(k)}, \tag{10}$$

(9) is the structured constraint, where $\mathbf{Y}_i$ is the groundtruth object location set of frame i, $\mathbf{Y}$ is the set of locations other than groundtruth, and $\Delta^M(\mathbf{Y}_i, \mathbf{Y}) = \sum_{k=1}^{K} \Delta(\mathbf{y}_i^{(k)}, \mathbf{y}^{(k)})$ is a combination of losses on each objects. (10) is the binary constraint.

Figure 1: Tracking results.

Fig.1 shows the tracking results across time on 2 different video clips, 'motinas-multi-face-fast' and 'toys' respectively. We compared against 5 methods. Evaluation results show that our method works better than other adaptive single object tracker when tracking multiple objects, and outperforms non-adaptive association-based multiple object tracking methods when tracking objects without enough training samples.

[1] S. Hare, A. Saffari, and P.H.S. Torr. Struck: Structured output tracking with kernels. In *IEEE International Conference on Computer Vision*, 2011.

[2] H. Jiang, F. Fels, and J. Little. A linear programming approach for multiple object tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[3] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *International Conference on Machine Learning*, 2004.

Metric Learning from Poses for Temporal Clustering of Human Motion

Adolfo López-Méndez[1]
adolf.lopez@upc.edu
Juergen Gall[2]
juergen.gall@tue.mpg.de
Josep R. Casas[1]
josep.ramon.casas@upc.edu
Luc van Gool[3]
vangool@vision.ee.ethz.ch

[1] Technical University of Catalonia (UPC)
Barcelona, Spain
[2] MPI for Intelligent Systems
Tuebingen, Germany
[3] ETH Zurich
Switzerland

Segmenting human motion into distinct actions is a highly challenging problem. From the motion analysis perspective, segmentation is difficult due to large stylistic variations, temporal scaling, changes in physical appearance, irregularity in the periodicity of human motions and the huge number of actions and their combinations. From a semantic viewpoint, segmentation is inherently elusive and difficult because in the vast majority of cases it is not clear when a set of poses describes an action. For instance, punching with the left hand and punching with right hand can be different actions, but it might be also regarded as punching or even more general as boxing.

We propose to learn what makes a sequence of poses different from others such that it should be annotated as an action, as illustrated in Fig. 1.

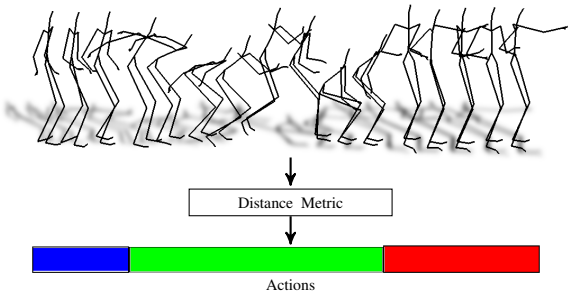


Figure 1: System Overview: Human motion sequences are clustered into different actions using a learned distance metric. We use annotations available in a mocap dataset to learn a distance metric that captures the semantic similarity between skeleton motion.

We make use of already annotated motion capture datasets and formulate action segmentation as a weakly supervised temporal clustering problem for an unknown number of clusters. Since publicly available datasets might contain different motions and action labels than the test sequences, we can not use the annotation directly for action segmentation. Instead, we use the annotations to learn a distance metric for skeleton motion using relative comparisons in the form of *samples of the same action are more similar than they are to a different action*. This is very intuitive since the sequences of a single database are usually labeled based on a semantic similarity.

We obtain a set of 14 relevant joint positions $\{\mathbf{q}_1,\dots,\mathbf{q}_{14}\}$ that can be easily obtained in different datasets [1]; see Fig. 2. We define a feature vector using these joint positions, and their velocity and acceleration:

$$\mathbf{x} = \{\mathbf{q}_1,\dots,\mathbf{q}_{14},\dot{\mathbf{q}}_1,\dots,\dot{\mathbf{q}}_{14},\ddot{\mathbf{q}}_1,\dots,\ddot{\mathbf{q}}_{14}\} \quad (1)$$

In the paper, we propose a set of relative constraints for pose features in order to capture the semantic similarity between poses given action labels of mocap datasets. We then rely on Information Theoretic Metric Learning (ITML) [3] in order to find the distance metric d_A, parameterized by a positive semi-definite matrix $\mathbf{A}$:

$$d_A(\mathbf{x}_i,\mathbf{x}_j) = (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{A}(\mathbf{x}_i-\mathbf{x}_j) \quad (2)$$

Since for each feature $\mathbf{x}_i$ we have only an action label y_i, we define the constraints based on triplets of points $(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)$ with class labels (y_i,y_j,y_k), where feature vectors with the same label should be closer to each other than to feature vectors with different labels. As an example, if $y_i = y_j \wedge y_j \neq y_k$ then the learned metric should hold $d_A(\mathbf{x}_i,\mathbf{x}_j) \leq \min(d(\mathbf{x}_i,\mathbf{x}_k),d(\mathbf{x}_j,\mathbf{x}_k))$.

The learned distance metric is then used to cluster the feature vectors in a test sequence into k motion primitives. The obtained primitives are provided to a hierarchical Dirichlet process (HDP)[5] that clusters the motion sequence into distinct behaviors (see Fig. 2). Provided that HDPs are non-parametric Bayesian models for infinite component mixtures, the number of actions (clusters) in a motion sequence is automatically estimated.

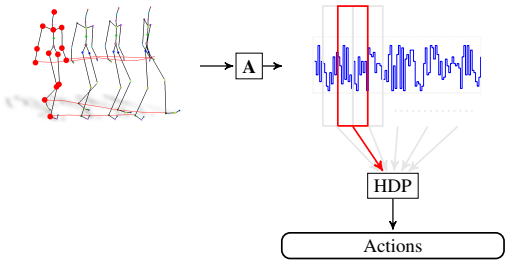


Figure 2: Detailed overview of our approach. A set of pose-based features are extracted using 14 relevant joints (marked with red spheres). These features are subsequently clustered into primitives using a metric (**A**) learned on related action sequences. In order to infer the different actions in a sequence, we first group the primitives using a sliding window. Then, we provide the resulting sets of primitives to a hierarchical Dirichlet process.

We conduct experiments on two publicly available mocap datasets: the CMU dataset [2], and the HDM05 dataset [4]. Specifically, we carry cross-dataset experiments in order to validate that the learned metric can be used for unseen actions and across datasets.

Details about the proposed constraints, implementation and evaluation methodology are given in the paper. Our conclusion, supported on the experimental results, is that the learned metrics improve the clustering results even across datasets and do not require that the actions of the test sequences are present in the training data. Furthermore, the method does not require to know the actual number of actions in a sequence. This makes our semi-supervised temporal clustering approach a compelling alternative to other unsupervised methods.

[1] J. Barbič, A. Safonova, J. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004*, GI '04, pages 185–194, School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, 2004. Canadian Human-Computer Communications Society.

[2] Carnegie Mellon University Motion Capture Database. http://mocap.cs.cmu.edu. URL `http://mocap.cs.cmu.edu`.

[3] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 209–216, 2007.

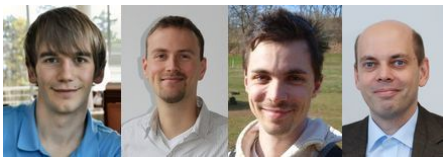
[4] HDM05 Mocap Dataset. http://www.mpi-inf.mpg.de/resources/hdm05/index.html. URL `http://www.mpi-inf.mpg.de/resources/HDM05/index.html`.

[5] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

Divergence-Based One-Class Classification Using Gaussian Processes

Paul Bodesheim
paul.bodesheim@uni-jena.de
Erik Rodner
erik.rodner@uni-jena.de
Alexander Freytag
alexander.freytag@uni-jena.de
Joachim Denzler
joachim.denzler@uni-jena.de

Computer Vision Group
Friedrich Schiller University of Jena, Germany
http://www.inf-cv.uni-jena.de



Detecting samples of unknown classes is a key task for active learning [1] and **one-class classification (OCC)** [2]. Starting from a set of only positive training samples, we want to estimate a soft membership score for every new test sample. Applying OCC methods is especially beneficial in situations where either negative data is difficult to model with given samples or where negative samples are hard to obtain.

We present an information theoretic framework for OCC, which allows for deriving several new novelty scores. With these scores, we are able to rank samples according to their novelty and to detect outliers not belonging to a learnt data distribution. **Our new framework sheds light on OCC from a completely different theoretical perspective.** The key idea of our approach is to measure how strongly a new test sample would influence the current model if it was used for training. This is carried out in a probabilistic manner using Jensen-Shannon divergence and the Gaussian process (GP) regression framework. An overview of our presented approach, which is based on mutual information (MI) and divergence measures of information theory, can be seen in Figure 2. Although our formulation is strongly related to active learning [1], we only consider OCC [2] in the paper. In the following, we assume a given training set $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ with labels $\mathbf{y} = \mathbf{1} = (1, \ldots, 1)^T$ and estimate a membership score of a test sample $\mathbf{x}^*$. We score the sample based on the resulting model change after treating it as an additional training sample.

To evaluate the change of the model, we once assume that $\mathbf{x}^*$ belongs to the target class ($y^* = 1$) and once assume the opposite ($y^* = -1$). Since we have no precise knowledge about the correct label of new samples, we model the assumed label $y^* \in \{1, -1\}$ as a random variable. For reasons explained in the paper, we approximate the change of the whole model by relying on a neighborhood of infinite small size and only taking the new sample itself into account. Therefore, we introduce a second random variable $Y^* \in \{1, -1\}$ to evaluate the model on $\mathbf{x}^*$ after the model update, *i.e.*, the variable Y^* is the reclassification result of $\mathbf{x}^*$. The dependencies between the assumed label y^* and the reclassification result Y^* can be measured using the conditional MI:

$$\mathrm{I}(Y^*, y^* \mid \mathbf{D}^*) = \mathrm{H}(Y^* \mid \mathbf{D}^*) - \mathrm{H}(Y^* \mid y^*, \mathbf{D}^*) \tag{1}$$

that depends on the available data $\mathbf{D}^* = (\mathbf{X}, \mathbf{y}, \mathbf{x}^*)$, which contains the training set as well as the new test sample $\mathbf{x}^*$. A low conditional entropy $\mathrm{H}(Y^* \mid y^*, \mathbf{D}^*)$ indicates that the reclassification result Y^* is almost certain given the assumed label y^*. Since the reclassification of $\mathbf{x}^*$ and thus

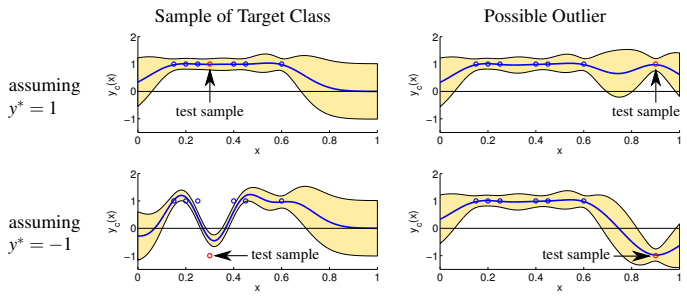


Figure 1: Visualization of our divergence approach. While both label assumptions $y^* \in \{1, -1\}$ of a possible outlier can be verified by reclassification using the model additionally trained with the test sample (blue curve), the assumption $y^* = -1$ will lead to a weak reclassification of a test sample stemming from the target class. Our approach exploits this difference. Classification uncertainty is visualised by shaded areas.

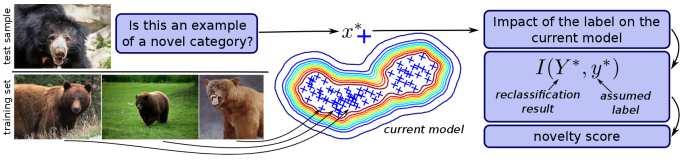


Figure 2: Outline of our approach based on mutual information.

the value of Y^* heavily depends on the training data, one achieves a low conditional entropy if the test sample $\mathbf{x}^*$ is far away from the training samples and the reclassification is mainly influenced by the choice of y^*. Summing up, a low conditional MI is induced by a strong membership to the target class and vice versa. The conditional MI of Eq. 1 turns out to be equal to the Jensen-Shannon (JS) divergence [3]:

$$\begin{aligned} \mathrm{I}(Y^*, y^* \mid \mathbf{D}^*) &= \mathrm{D}_{\mathrm{JS}}^{\pi}(\mathbf{p}_1 || \mathbf{p}_{-1}) && (2) \\ &= \pi \cdot \mathrm{D}_{\mathrm{KL}}(\mathbf{p}_1 || \mathbf{m}) + (1-\pi) \cdot \mathrm{D}_{\mathrm{KL}}(\mathbf{p}_{-1} || \mathbf{m}) \quad , && (3) \end{aligned}$$

where $\mathrm{D}_{\mathrm{KL}}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence, $\mathbf{m} = \pi \cdot \mathbf{p}_1 + (1-\pi) \cdot \mathbf{p}_{-1}$ the mixture of the two conditional probability distributions $\mathbf{p}_1 = \mathrm{p}(Y^* \mid y^* = 1, \mathbf{D}^*)$ and $\mathbf{p}_{-1} = \mathrm{p}(Y^* \mid y^* = -1, \mathbf{D}^*)$, and π the prior probability: $\pi = \mathrm{p}(y^* = 1 \mid \mathbf{D}^*)$. Therefore, we propose using the **negative JS divergence as an OCC score**. For the computation, we only need the parameter π as well as conditional probability distributions $\mathbf{p}_1$, and $\mathbf{p}_{-1}$. In this paper, we propose using posterior probabilities of a GP.

In the case of GP regression, continuous outputs y_c are assumed to be generated according to $y_c(\mathbf{x}) = f(\mathbf{x}) + \varepsilon$, where f is a latent function and ε is a noise term. Following a Bayesian framework, output values of unknown samples $\mathbf{x}^*$ are predicted in a probabilistic fashion by marginalising over latent functions f. Using assumptions mentioned in the paper, the predictive output y_c^* for a new sample $\mathbf{x}^*$ given data $\mathbf{D}^*$ is normally distributed as well with moments μ_* and σ_*^2 given in closed form. We compute probabilities $\pi = \mathrm{p}(y^* = 1 \mid \mathbf{D}^*)$ based on these moments via:

$$\pi = \mathrm{p}(y^* = 1 \mid \mathbf{D}^*) = \mathrm{p}(y_c^* > 0 \mid \mathbf{D}^*) = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left(\frac{-\mu_*}{\sqrt{2\sigma_*^2}}\right) \quad , \tag{4}$$

where $\mathrm{erf}(\cdot)$ is the error function. We also need to compute probabilities of the conditional distributions $\mathbf{p}_1$ and $\mathbf{p}_{-1}$ in a similar vein to Eq. (4). The conditional probabilities will arise from a GP model learnt with $N+1$ training samples by treating the current test sample $\mathbf{x}^*$ and its assumed label y^* as training data as well. The assumption $y^* = 1$ is still an OCC setting whereas the assumption $y^* = -1$ leads to a highly imbalanced binary classification scenario. The difference between the behaviour of samples stemming from the target class and outliers is visualised in Figure 1 using a solid line for the predictive mean and shaded areas for the predictive variance of the GP model.

Evaluations on machine learning and image categorization datasets are described in the paper. Our conclusion is that we reach state-of-the-art performance while offering a completely new access to the challenging problem of one-class classification.

[1] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision*, 88(2):169–188, 2010.

[2] Michael Kemmler, Erik Rodner, and Joachim Denzler. One-class classification with gaussian processes. In *ACCV*, pages 489–500, 2010.

[3] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Hierarchical Sparse Spectral Clustering for Image Set Classification

Arif Mahmood and Ajmal S. Main
{arifm,ajmal}@csse.uwa.edu.au

School of Computer Science and Software Engineering,
The University of Western Australia, Crawley WA 6009

We present a structural matching technique for robust classification based on image sets. In set based classification, a probe set is matched with a number of gallery sets and assigned the label of the most similar set. We represent each image set by a *sparse* dictionary and compute a similarity matrix by matching all the dictionary atoms of the gallery and probe sets. The similarity matrix comprises the sparse coding coefficients and forms a fully connected directed graph. The nodes of the graph are the dictionary atoms and the edges are the sparse coefficients. The graph is converted to an undirected graph with positive edge weights and spectral clustering is used to cut the graph into two balanced partitions using the normalized cut algorithm. This process is repeated until the graph reduces to critical and non-critical partitions. A critical partition contains atoms with the same gallery label along with one or more probe atoms whereas a non-critical partition either consists of only probe atoms or atoms with multiple gallery labels with no probe atom. Using the critical partitions, we define a novel set based similarity measure and assign the probe set the label of the gallery set with maximum similarity. The proposed algorithm is applied to image set based face recognition using two standard databases. Comparison with existing techniques shows the validity and robustness of our algorithm in the presence of outlier images.

A schematic diagram of the proposed algorithm is shown in Fig. 1. The intrinsic data dimensionality in the image sets is often less than the apparent dimensions. Therefore, we reduce the data dimensionality using PCA basis computed from the training (gallery) sets. For each reduced dimensionality gallery set, we pre-compute sparse dictionaries of varying sizes. A sparse dictionary must be able to represent all images in an image set as a sparse linear combination of its atoms. Given an image set $X_i = \{x_j\}_{j=1}^{n_i} \in \mathcal{R}^{m\times n_i}$, its dictionary $D_i \in \mathcal{R}^{m\times p_i}$ should be able to minimize a cost function $\frac{1}{n}\sum_{j=1}^{n_i} f(x_j, D_i)$. Each column of D_i represents a basis vector for the image set X_i. We use the convex ℓ_1 formulation of the Lasso as the cost function [3]

$$\min_{\alpha_i, D_i} \Big(\frac{1}{n_i} \sum_{j=1}^{n_i} \frac{1}{2} ||x_j - D_i\alpha_i||_2^2 + \lambda ||\alpha_i||_1 \Big) \ . \qquad (1)$$

Sparse dictionaries D_i of various sizes for each of the training set are learned from (1).

We start from the smallest size dictionary with p_i atoms per gallery set. A sparse dictionary with p_q atoms is learned for the query (probe) image set as well. Let D_G be the set of learned dictionaries for the gallery image sets and D_q be the dictionary for the query image set. Each dictionary atom in D_G inherits a label from its parent image set whereas a test label t is assigned to each atom in D_q. Let L_G be the labels of the gallery image sets and L_q be the labels for the query image set. We append dictionaries in an array $D_s = [D_G|D_q]$ and the labels as well $L_s = [L_G|L_q]$.

As a similarity measure, we compute the sparse coefficients required to represent a particular dictionary atom as a linear combination of the remaining atoms [1] in D_s. We take one atom d_i out of D_s, replace it by zeros, and represent d_i as a sparse linear combination of the remaining atoms. We use a fast implementation of LARS to find the sparse coding coefficients α_i of d_i computed as

$$\min_{\alpha_i} ||d_i - D_s\alpha_i||_2^2 \ \ s.t. \ \ ||\alpha_i||_1 \leq \lambda \ . \qquad (2)$$

We append all α_i as columns in a similarity matrix $S = \{\alpha_i\}_{i=1}^{P+p_q}$.

Considering each dictionary atom in D_s as a node in a fully connected graph G, the sparse linear coefficients in S are the edge weights connecting any two nodes in G. Thus the similarity matrix S forms an adjacency matrix for G, which is a directed graph. To form a positive weight undirected graph, the modified adjacency matrix is computed as $A = |S| + |S^t|$, where $|\cdot|$ stands for absolute value. In order to apply spectral clustering [4] on A, we first compute the degree matrix $D(i,j) = \sum_{i=1}^{P+p_q} A(i,j)$ if $i = j$ and $D(i,j) = 0$ if $i \neq j$. Using D and A, we compute un-normalized graph Laplacian matrix $L = D - A$ and then the normalized Laplacian [6]

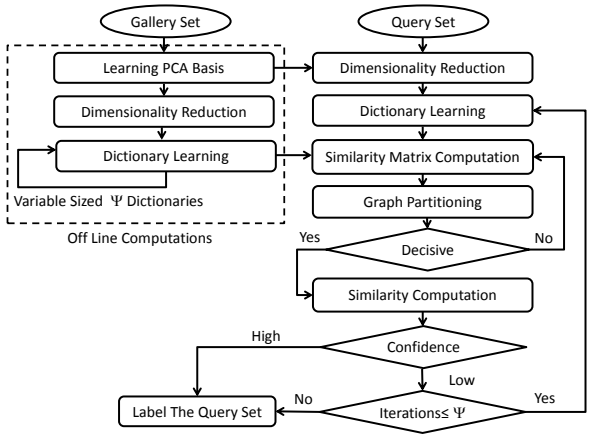


Figure 1: Block diagram of the Hierarchical Sparse Spectral Clustering (HSSC) algorithm.

$L_w = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. To cut the graph into two balanced partitions, we compute the eigen vectors of L_w. Using the sign of the elements of the second eigenvector we divide the set of all dictionary atoms into two partitions.

We recursively perform binary partitioning of the graph G until each partition is identified as a decisive cluster which may contain atoms from only one gallery set along with query atoms (critical cluster), only query atoms (non-critical cluster) or zero query atom and one or more gallery atoms (non-critical cluster). For each gallery set, we count the number of atoms in all critical clusters and the corresponding number of query atoms in those clusters as well. The product of both counts represents a similarity score of the query set with that particular gallery set. Based on the distribution of query atoms in the critical clusters, a confidence score is also defined. If the confidence is high, the algorithm stops and a label is predicted for the query set based on the maximum similarity score. If the confidence is low, the full process is repeated with an increased dictionary size. If confidence remains low for consecutive dictionary sizes, however the predicted query label remains consistent, that label will soon accumulate high confidence and the algorithm will stop. If confidence remains low over a number of dictionary sizes and the predicted label is inconsistent, the algorithm will stop after executing the maximum number of iterations and the the final label will be predicted as the label with the maximum mode over all iterations. This may occur in the case of difficult matches and the predicted label confidence will remain low.

Experiments are performed on the Honda/UCSD [2] and CMU Mobo data [5] for face recognition based on image sets. Comparison with existing techniques shows the efficacy of the proposed algorithm. We also test robustness to outliers by mixing an increasing number of imposter images in the probe set. The proposed algorithm demonstrates significant robustness by achieving 100% recognition rate on the Honda database in the presence of up to 11 imposters selected randomly from a random gallery set and mixed with the probe sets.

[1] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *CVPR*, pages 2790–2797. IEEE, 2009.

[2] K. C. Lee, J. Ho, M. H. Yang and D. Kriegman. Video-Based Face Recognition Using Probabilistic Appearance Manifolds. In *CVPR*, pages 313–320, 2003.

[3] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *In NIPS*, pages 801–808. NIPS, 2007.

[4] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, December 2007. ISSN 0960-3174.

[5] R. Gross and J. Shi. The CMU Motion of Body (MoBo) Database. Technical Report CMU-RI-TR-01-18, Robotics Institute, 2001.

[6] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, 2000.

Manifold-enhanced Segmentation through Random Walks on Linear Subspace Priors

Pierre Yves Baudin[1,2,4,5,6]
pierre-yves.baudin@ecp.fr
Noura Azzabou[5,6,7]
n.azzabou@institut-myologie.org
Pierre Carlier[5,6,7]
p.carlier@institut-myologie.org
Nikos Paragios[2,3,4]
nikos.paragios@ecp.fr

[1] SIEMENS Healthcare, FR
[2] Center for Visual Computing, FR
[3] Université Paris-Est, FR
[4] Equipe Galen, INRIA Saclay, FR
[5] Institute of Myology,Paris, FR
[6] CEA,IdM NMR Laboratory, Paris, FR
[7] UPMC University Paris 06, FR

Segmentation of skeletal muscles in 3D Magnetic Resonance Imaging, which poses some very specific issues: simultaneous multi-object segmentation, non-discriminative appearance of the muscles, partial contours between them, large inter-subject variations, spurious contours due to fat infiltrations. For these reasons, it is necessary to impose knowledge-based shape priors into segmentation methods. In [2, 3, 4], segmentation is achieved with multi-object elastical deformable models, using either a reference atlas or a hierarchical statistical prior model. A discrete optimization procedure on a graph framework is used in [7], where a higher-order pose invariant shape prior is imposed on surface landmarks nodes. A pixel-wise region-based approach was proposed in [1] where prior knowledge is enforced by embeding the model into a statistical low-dimensional non-linear manifold through PCA in the Isometric Log-Ratio space. Random Walks for image segmentation, presented in [6], is a numerical method for computing globally large discrete regions-based segmentation methods, and is notoriously robust to partial contours. Prior knowledge on intensity within the RW formulation was introduced in [5].

In this paper, we propose a segmentation method based on RW, in which shape deformation is constrained to remain close to a PCA shape space built from training examples. Using the PCA allows us to model complex non-rigid shape variations relying on a few eigen-modes. Our method also benefits from the high performances of the RW optimization process.

The RW method amounts to computing the probability x_i^s that the node v_i is assigned to the label s . It was shown ([6]) that this probabilities minimize the functional:

$$E_{\mathrm{RW}}^s(x^s) = x^{sT} L x^s \tag{1}$$

where L is the combinatorial Laplacian matrix of the graph, defined as:

$$\forall (i,j) \in \mathcal{E},\ L_{ii} = \sum_k w_{ik},\ L_{ij} = -w_{ij}. \tag{2}$$

It is common practice to use as a Gaussian weighting function for w_{ij}. Once computed x^s for each label s, the segmentation is obtained by retaining the label of maximum probability: $\hat{s}_i = \arg\max_s x_i^s$.

Since minimizing (1) is an independent process for each label s, the whole RW process can be equivalently synthesized in one equation, via concatenation of the x^s and diagonal concatenation of L:

$$E_{\mathrm{RW}}(x) \quad = \quad x^T \bar{L} x. \tag{3}$$

The principle of a shape space is to design a low dimensional affine space approximating this implicit space. Assume we possess T co-registered segmented training volumes. We perform the PCA on vectors $\left\{\hat{x}^i\right\}_{i=1\ldots T}$, which are ground truth segmentations represented as probability vectors. Retaining the n eigen-modes of greatest variance, the projection of any segmentation in the shape space is represented as:

$$\tilde{x} = \bar{x} + \Gamma\gamma.$$

Thus, any segmentation x can be expressed as:

$$x = dx + \Gamma\gamma + \bar{x} \tag{4}$$

where $dx \in \mathbb{R}^{KN\times 1}$ is the deviation of x from the shape space.

In order to obtain a segmentation which remains close to the shape space, we want to minimize the objective function (3) with respect to both dx and γ, while keeping dx small. This leads us to the following functional:

$$E_1(dx,\gamma) = (dx + \Gamma\gamma + \bar{x})^T \bar{L} (dx + \Gamma\gamma + \bar{x}) + \lambda \|dx\|^2. \tag{5}$$

The minimum of (5) verifies:

$$\left(A^T \bar{L} A + \lambda B\right) y = A^T \bar{L} \bar{x} \tag{6}$$

with:

$$y = \begin{bmatrix} dx \\ \gamma \end{bmatrix},\ A = [I_{KN}\ \Gamma],\ B = \begin{bmatrix} I_{KN} & 0 \\ 0 & 0 \end{bmatrix}. \tag{7}$$

The system of equations (6) can be solved with iterative methods like Bi-Conjugate Gradient. Computation time is approximately 15 min per segmentation on a 2.8 GHz Intel process with 4GB of RAM. We present results obtained with our method on a set of 3D MR volumes of muscles. Our data set is comprised of 30 volumes of the right thigh of healthy subjects, covering a wide range of ages and morphologies. On figure 1, we show the effect of the PCA shape prior on one example of our dataset.

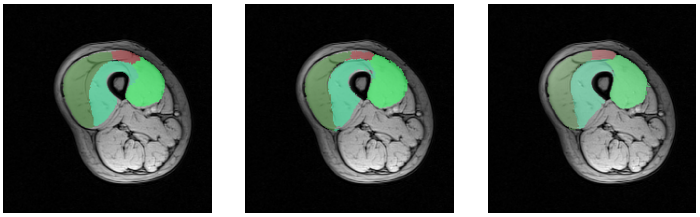


Figure 1: Effect of the PCA shape prior: (left) mean segmentation using $x = \bar{x}$, (middle) shape space segmentation using $x = \Gamma\gamma + \bar{x}$, (right) segmentation with shape prior using $x = dx + \Gamma\gamma + \bar{x}$. The shape space segmentation fits the boundaries better than the mean segmentation, but has fuzzy contours due to the approximation of projecting complex shapes into a linear subspace.

[1] Shawn Andrews, Ghassan Hamarneh, Azadeh Yazdanpanah, Bahareh HajGhanbari, and W Darlene Reid. Probabilistic multi-shape segmentation of knee extensor and flexor muscles. In *MICCAI*, volume 14, pages 651–8. 2011.

[2] Salma Essafi, Georg Langs, and Nikos Paragios. Hierarchical 3D diffusion wavelet shape priors. In *ICCV*, pages 1717–1724. IEEE, September 2009.

[3] Benjamin Gilles and Nadia Magnenat-Thalmann. Musculoskeletal MRI segmentation using multi-resolution simplex meshes with medial representations. *Medical image analysis*, 14(3):291–302, June 2010.

[4] Benjamin Gilles and Dinesh K. Pai. Fast musculoskeletal registration based on shape matching. In *MICCAI*, volume 5242 of *Lecture Notes in Computer Science*, pages 822–829. January 2008.

[5] L. Grady. Multilabel Random Walker Image Segmentation Using Prior Models. In *CVPR*, volume 1, pages 763–770, 2005.

[6] Leo Grady. Random walks for image segmentation. *Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006.

[7] Chaohui Wang, Olivier Teboul, Fabrice Michel, Salma Essafi, and Nikos Paragios. 3D knowledge-based segmentation using pose-invariant higher-order graphs. In *MICCAI*, volume 6363 of *Lecture Notes in Computer Science*, pages 189–196. 2010.

Towards Longer Long-Range Motion Trajectories

Michael Rubinstein[1]
mrub@mit.edu
Ce Liu[2]
celiu@microsoft.com
William T. Freeman[1]
billf@mit.edu

[1] MIT CSAIL
[2] Microsoft Research New England

Although dense, long-rage, motion trajectories are a prominent representation of motion in videos, there is still no good solution for constructing dense motion tracks in a truly long-rage fashion. Ideally, we would want every scene feature that appears in multiple, not necessarily contiguous, parts of the sequence to be associated with the same motion track. Despite this reasonable and clearly stated objective, there has been surprisingly little work on general-purpose algorithms that can accomplish that task. State-of-the-art dense motion trackers process the sequence incrementally in a frame-by-frame manner, and associate, by design, features that disappear and reappear in the video, with different tracks, thereby losing important information of the long-term motion signal.

In this paper, we propose a novel divide and conquer approach to long-range motion estimation. Given a long video or image sequence, we first produce high-accuracy *local* track estimates, or *tracklets*, and later propagate them into a *global* solution, while incorporating information from throughout the video. Tracklets are computed using state-of-the-art motion trackers [2, 3] that have become quite accurate for short sequences as demonstrated by standard evaluations. Our algorithm then constructs the long-range tracks by linking the short tracks in an optimal manner. This induces a combinatorial matching problem that we solve simultaneously for all tracklets in the sequence.

The main contributions of this paper are: (a) a novel divide-and-conquer style algorithm for constructing dense, long-rage motion tracks from a single monocular video, and (b) Novel criteria for evaluating long-range tracking results with and without ground-truth motion trajectory data. We evaluate our approach on a set of synthetic and natural videos, and explore the utilization of long-range tracks for action recognition.

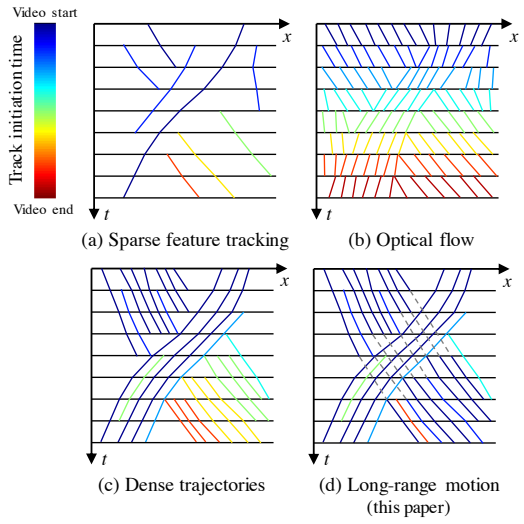


Figure 1: **Comparison of several motion representations.** Sparse feature point tracking (a) can establish long-range correspondences (*e.g.* between hundreds of frames), but only a few feature points are detected. While useful for some applications, it is a very incomplete representation of the motion in a scene. On the other hand, dense optical flow (b) reveals more about the moving objects, but the integer-grid-based flow fields cannot reliably propagate to faraway frames. A natural solution, therefore, is to combine feature point tracking and dense optical flow fields to a set of *spatially dense* and *temporally smooth* trajectories (or particles, tracks) [2], as show in (c). Despite recent advances in obtaining dense trajectories from a video sequence [1, 3], it is challenging to obtain *long-range* dense trajectories. In this paper, we propose a novel divide and conquer approach to long-range motion estimation (d), where point trajectories are assigned to the same tracks despite occlusion and deformation.

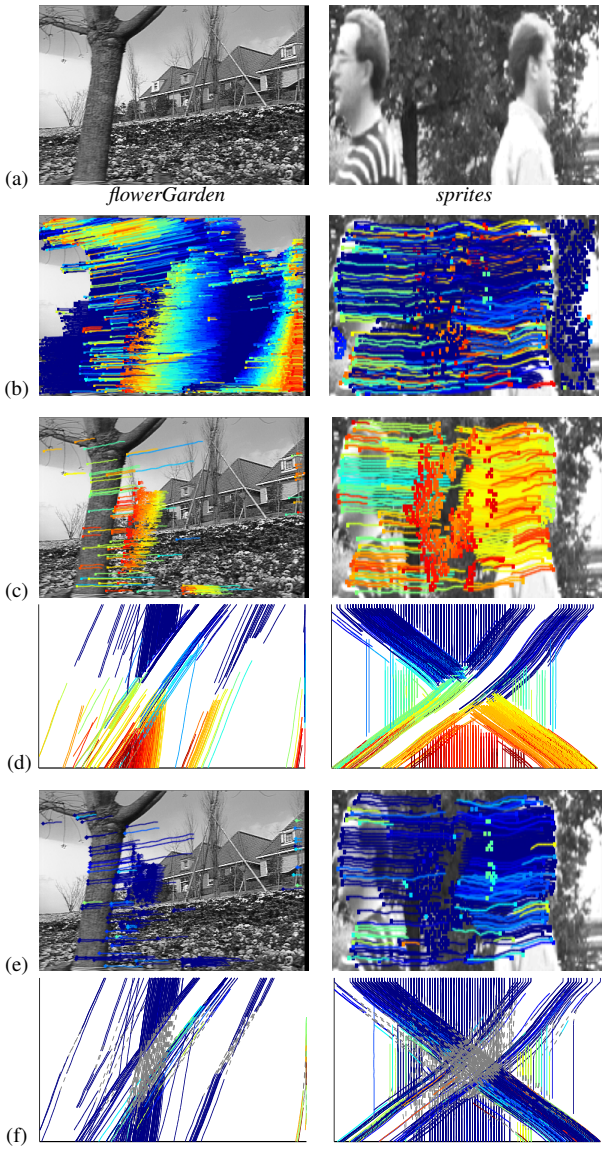


Figure 2: **Long-range motion trajectory results**. For each video (column), (a) is a representative frame from the sequence, (b) are the resulting long-range motion tracks, (c) and (e) focus on the tracks involved in the linkage (tracks which are left unchanged are not shown), before (c) and after (e) they are linked. (d) and (f) show XT views of the tracks in (c) and (e), respectively, when plotted within the 3D video volume (time advancing downwards). The tracks are colored according to their initiation time, from blue (earlier in the video), to red (later in the video). Track links are shown as dashed gray lines in the spatiotemporal plots (d) and (f). For clarity of the visualizations, random samples ($25-50\%$) of the tracks are shown.

[1] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2011.

[2] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2195–2202. IEEE, 2006.

[3] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. *Computer Vision–ECCV 2010*, pages 438–451, 2010.

Non-parametric synthesis of laminar volumetric textures from a 2D sample

Radu Urs[1, 2]
radu.urs@ims-bordeaux.fr
Jean-Pierre Da Costa[1, 2]
Jean-Marc Leyssale[3]
Gérard Vignoles[3]
Christian Germain[1, 2]

[1] University of Bordeaux, IMS, UMR 5218, F-33400 Talence, France
[2] CNRS, IMS, UMR 5218, F-33400 Talence, France
[3] CNRS, LCTS, UMR 5801 CNRS-UB1-CEA-Safran, F33600 Pessac, France

Motivation

Many tools in image processing and computer graphics involve texture analysis and synthesis. The field of texture synthesis is particularly dynamic with notable applications in image compression, inpaiting, extrapolation or texture mapping. This research field has led to the development of many 2D synthesis techniques, but their extension to the 3D environment remains unstable proving itself as a very complex and computational issue. 3D textures are mainly used for texturing volumetric objects trying to increase the realism of 3D scenarios, but they can also be observed in 3D vision when exploring for instance material structure or seismic data.

In this article, we are interested in non-parametric algorithms that achieve 3D texture synthesis from a single 2D sample. In particular, we investigate their capability to produce laminar textures, i.e. textures made of anisotropic sheets stacked along a given direction. The algorithms under study are variants of the original algorithm proposed by Wei and Levoy [1]. In spite of its versatility and its good computational properties, this algorithm is also known to produce output textures that are more regular than the examples they try to mimic. Several authors have proposed variants of these algorithms that intend to better reproduce, in the output texture, the diversity learned in the input sample. How do these algorithms perform on laminar textures with strong anisotropy? How are their properties modified when inferring 3D from 2D? That is what we intend to investigate in this work.

Algorithms under study

The first algorithm we implemented is a 2D/3D extension of Wei and Levoy's algorithm [1] that uses only a single 2D image as source of synthesis. Based on a Markov field hypothesis [4], the method relies on texture locality and texture stationarity. Starting from a random initialization, we synthesize voxel by voxel, examining the 2D neighbourhoods of the current voxel from two orthogonal views of the 3D block. This phase implies a search of the best match for each of these two neighbourhoods in the same input image. The output voxel value is updated with the average of the two found voxels and we reiterate until reaching the same neighbourhoods after two consecutive iterations.

Next we turned our attention to the approach proposed by Kopf et *al.* [2]. It aims at an optimal combination of information from the front view and the side view as suggested by Kwatra et *al.* [5] and, in the same time, at adding a colour histogram matching mechanism in the texture optimization procedure. But the results are, more or less, affected by blurring or missing textural patterns.

To improve these results, we addressed the algorithm of Chen and Wang [6]. Their optimization procedure, constrained by two new kinds of matching histograms – position and index histograms, indirectly achieves colour histogram matching while preserving the input texture by distributing uniformly the input pixels in the output block.

Whatever the method, a problematic issue remains the choice of the neighbourhood system. The causality of the neighbourhood is strongly related to the scan type used for synthesis. Alternative to the random walk – that allows the synthesis of a pixel by freeing itself from its past, or the lexicographical walk – making the synthesis of a pixel dependent on previous pixels, we propose new scan types, namely the 3D extensions of space filling curves (e.g. Hilbert Curve in Fig. 1).

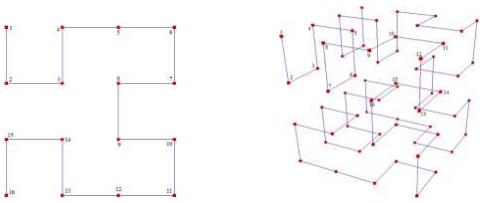
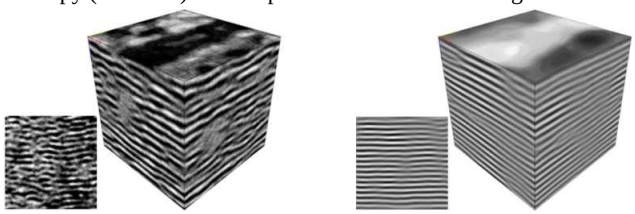


Figure 1: Illustration of a fractal scan type: on the left, the Hilbert curve in 2-dimensions and on the right, its extension to 3-dimensions.

Experimental benchmark and application

Experimental evaluation is performed on laminar textures of dense carbons observed by High Resolution Transmission Electronic Microscopy (HRTEM) and we present some results in Fig. 2.



a *b* *c* *d*

Figure 2: Volumetric results: *b* - the 3D view of the solid texture obtained from the raw HRTEM sample in *a*; and *d* is the volumetric texture obtained by synthesizing the filter HRTEM sample from *c*.

Beyond the traditional subjective evaluation of the synthesized textures as in [1-5], we propose a genuine quantitative benchmark for the analysis of the synthesized textures which consists in comparing input and output image characteristics. Both grey level statistics and pattern morphology are studied. Precisely, this original study focuses on the one hand, on grey level dynamics (1st order statistics) and spatial statistics (autocorrelations), and on the other hand on morphological properties (fringe lengths, tortuosity and orientation).

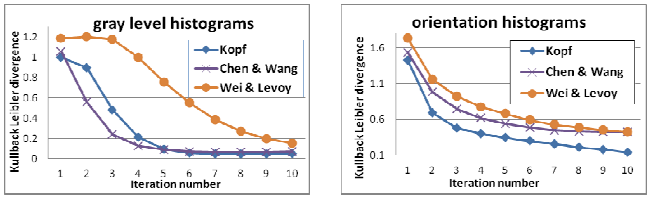


Figure 3: Indicators of the objective comparison of the 3D textures: comparing the gray level or orientation histograms of the sample and the ones obtained after a multi-2D blocks analysis.

Our objective comparison of the evolution of the 3D blocks statistics towards the ones of the 2D sample strengthens the visual assumptions relative to the improvements brought by the histogram matching, while Wei and Levoy's algorithm is not able to conserve the contrast. In terms of preserving texture structure, our study reveals the 3D textures tendency towards the same structure observed on the 2D sample. Computationally, using Kopf's approach with a fractal scan type to reduce the convergence time proves to be the most efficient method.

[1] L.-Y. Wei and M. Levoy. *Texture synthesis from multiple sources*. Proceedings of ACM SIGGRAPH 2003 Sketches & Applications.

[2] J. Kopf, C.W. Fu, D. Cohen-Or, O. Deussenn, D. Lischinski and T.T. Wong. *Solid Texture Synthesis from 2D Exemplars*. ACM SIGGRAPH, TOG, vol. 26, issue 3, 2007.

[3] J. Chen and B. Wang. *High quality solid texture synthesis using position and index histogram matching*. The Visual Computer, vol.26, pp. 253-262, 2010.

[4] L.-Y. Wei and M. Levoy. *Fast texture synthesis using tree-structured vector quantization*. Proc. of ACM SIGGRAPH, pp. 479-488, 2000.

[5] V. Kwatra, A. Schodl, I. Essa, G. Turk and A. Bobick. *Graphcut textures: image and video synthesis using graph cuts*. ACM SIGGRAPH, pp. 277-286, 2003.

[6] J. Han, K. Zhou, L.-Y. Wei, M. Gong, H. Bao, X. Zhang, B. Guo. *Fast example-based surface texture synthesis via discrete optimization.* Visual Computer, 22(9), pp. 918–925, 2006.

Motion Models That Only Work Sometimes

Cristina García Cifuentes[1]
http://visual.cs.ucl.ac.uk/pubs/MotionModelPrediction
Marc Sturzel[3]

Frédéric Jurie[2]

Gabriel J. Brostow[1]

[1] University College London, UK
[2] University of Caen, France
[3] EADS Innovation Works, France

It is too often that tracking algorithms lose track of interest points in image sequences. This persistent problem is difficult because the pixels around an interest point change in appearance or move in unpredictable ways. In this paper we explore how classifying videos into categories of camera motion improves the tracking of interest points, by selecting the right specialist motion model for each video. As a proof of concept, we enumerate a small set of simple categories of camera motion and implement their corresponding specialized motion models. We evaluate the strategy of predicting the most appropriate motion model for each test sequence. Within the framework of a standard Bayesian tracking formulation, we compare this strategy to two standard motion models. Our tests on challenging real-world sequences show a significant improvement in tracking robustness, achieved with different kinds of supervision at training time.

We implement specialized dynamic models for four types of camera motion and two standard tracking models (Brownian and constant velocity). After training an SVM, we categorize each new video sequence to its most appropriate dynamic model.

The performance of the six motion models is computed over the whole dataset. No individual model performs extraordinarily overall, but each one does well on its "kind" of videos. The experiments show how predicting the most appropriate motion model leads to significant improvement of tracking robustness. We propose that such predictions can be made using a set of training videos, with shared motion properties, and using a classifier to predict the category of a new test videos. There is good reason to expect that further very specialized motion models, that are not good in general but outstanding under known circumstances, are worth developing.

frame 0 frame 10 frame 20

Figure 1: Example sequence with overlaid box showing the output of our specialized "forward" motion model, where the velocity and scale of objects approaching the camera tend to increase. Neither Brownian nor constant-velocity motion models are as successful at tracking interest points here.

	Individual motion models						Ideal predictions			Our method
	Br	CVel	TRight	TLeft	Fwd	Bwd	best{Br, CVel}	best{all}	manual labels	
tracking robustness ($\cdot 10^{-2}$)	42.3	43.2	37.9	37.2	44.7	43.7	49.3	56.1	52.6	51.9
$\pm$ std. dev. random runs	0.4	0.4	0.7	0.5	0.2	0.1	0.6	0.4	0.2	0.1
% best choice (± 2)	21	11	12	20	20	16	32	100	52	50

Table 1: White background: average tracking robustness of each individual motion model over all videos (default parameters). Bottom row: percentage of times *that* motion model (among six) was the best choice. Dark gray: best-case results if model is selected by either a performance-based oracle, or our inspection-based labels. Right column: tracking each video using our classifier's suggested motion model, using inspection-based training data.

Depiction Invariant Object Matching

Anupriya Balikai
anupriyabalikai@gmail.com
Peter Hall
pmh@cs.bath.ac.uk

Department of Computer Science,
University of Bath
Bath,
UK

Matching objects no matter how they are depicted (photo, painting, drawing, etc.) is an important open problem. We propose that the way in which images are described is a key to matching performance. To test this we use a hierarchical descriptor with regions as node to encode structure. The nodes are labelled with photometric descriptors in one case, and with non-photometric descriptors in the other. Measuring performance across a photos-only database yields comparable results, but we see a marked improvement for the non-photometric descriptor when either an art-only or a mixed database is used. We further improve performance using an MRF based matcher of our own design.

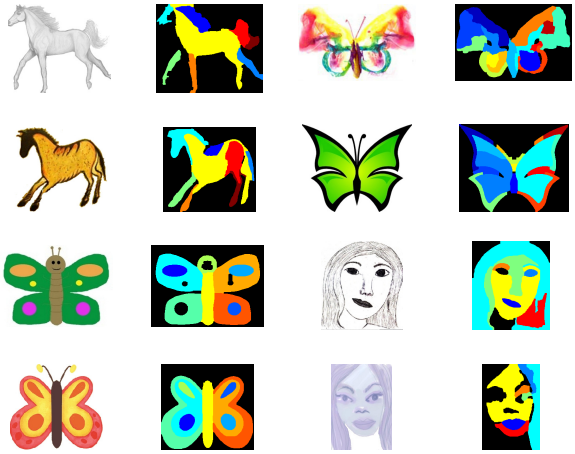


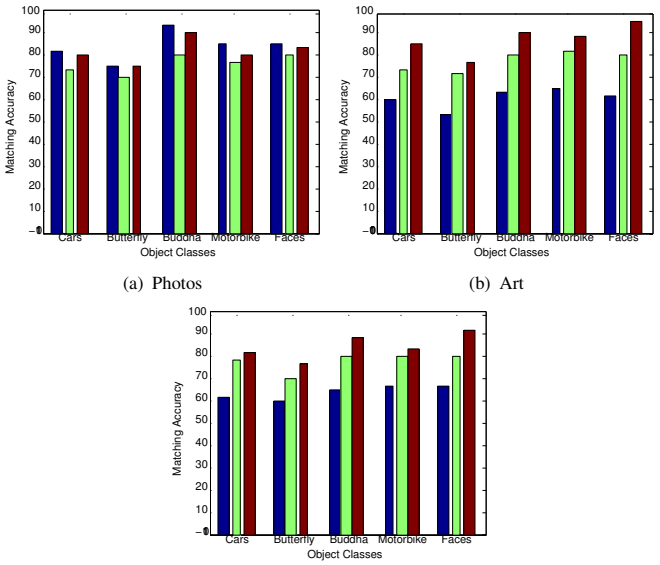
Figure 1: Matching across depictions using structure and depiction-invariant features. Each pair of images is colour-coded to show regions that have been matched using the proposed method.

We are motivated by the human ability to recognise across depictive boundaries. Object matching is an area that has received continuous and consistent attention within Computer Vision. The state of the art is now robust to challenges such as occlusions, viewpoint invariance, pose invariance and so on. However, little attention has been given to matching objects across depictive styles.

There has been little work on matching across depictive styles [2, 3]. Our approach to object matching differs to these methods by finding a generic representation for an object, without the need for any learning. The main contribution of this paper lies in the introduction of an object descriptor that combines global structure and local non-photometric features.

Our object description is based on a labelled graph. An object's structure is encapsulated by a graph constructed using the output of any hierarchical segmentor. In our experiments we find that the Berkeley segmentor [1] output the best segmentations. Since the number of levels in the hierarchy are quite large, we reduce the levels in the hierarchy by using the Laplacian Energy of the graph as explained in [4]. Every region in the segmentation tree is assigned as a vertex in the graph. Vertices at the same hierarchical level are connected by an edge if they share a common boundary. Vertices across consecutive levels are connected by an edge if they are related by containment. To remove noise from the graph we have come up with novel yet simple technique based on discarding regions that lie below a certain area threshold. A recursive algorithm is introduced to ensure that meaningful regions are not discarded. Each region in the hierarchy is described using self similarity descriptors [2] augmented by geometric terms relating to shape and orientation. We now have full description for the entire graph, and hence the object.

Two methods are presented for matching across pairs of images. The first method, based on Feature Correspondence Graph Matching [5], is used to compute a mapping between the segmentation graphs of the two objects. However, inconsistent segmentation of different instances



(a) Photos (b) Art

(c) Mixed

Figure 2: Matching accuracies obtained for graph matching with photometric features (blue), graph matching with SSD features (green) and sliding window search with SSD features (red), for 5 object categories over the three datasets.

of an object brings about failure cases with this method. To counter this problem, a second new approach to object matching is introduced based on a sliding window search that finds best matches for the segmented regions in the first image, across the second image. The overall match is then computed using max-sum on a Markov Random Field [6].

Experiments were conducted on three datasets of images. called *photos only*, *art only* and *mixed*. The first being a subset of the Caltech-101 object categories dataset, while the next two have been introduced in this paper. The accuracy of the matcher is measured against human labelled ground truth. In order to provide a comparison between matching photos to photos, and depiction- invariant matching, experiments included the use of photometric and SSD features using the proposed matchers.

Figure 2 shows that matching performance is contingent upon representation: our descriptor is comparable to the state of the art photometric methods for the dataset of photographs alone, but outperforms the state of the art for the other two datasets.

We conclude that description is important to the problem of cross-domain matching, and that learning is not necessary to achieve that task.

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[2] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[3] Abhinav Shrivastava, Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Data-driven visual similarity for cross-domain image matching. *SIGGRAPH ASIA*, 2011.

[4] Y.Z. Song, P. Arbelaez, P. Hall, C. Li, and A. Balikai. Finding semantic structures in image hierarchies using laplacian graph energy. *European Conference on Computer Vision (ECCV)*, 2010.

[5] L. Torresani, V. Kolmogorov, and C. Rother. Feature correspondence via graph matching: Models and global optimization. *European Conference on Computer Vision (ECCV)*, 2008.

[6] T. Werner. A linear programming approach to max-sum problem: A review. *Pattern Analysis and Machine Intelligence (PAMI)*, 2007.

BiCov: a novel image representation for person re-identification and face verification

Bingpeng Ma
bingpeng.ma@unicaen.fr
Yu Su
yu.su@unicaen.fr
Frédéric Jurie
frederic.jurie@unicaen.fr

GREYC, CNRS UMR 6072,
University of Caen Basse-Normandie,
Caen, France

Person re-identification and face verification tasks are both consisting in recognizing an individual through different images (e.g. images coming from cameras in a distributed network or from the same camera at different time). The key requirement of approaches addressing these tasks is their ability to measure the similarity between two person/face-centered bounding boxes, i.e. to predict if they represent to the same person, despite changes in illumination, pose, viewpoint, background, partial occlusions and low resolution.

In this paper, we propose the new BiCov image representation allowing to measure effectively the similarity between two persons/faces without requiring any pre-processing step (e.g. background subtraction or body part segmentation).

The proposed method includes two stages. In the first stage, Biologically Inspired Features (BIF) are extracted, through the use of Gabor filters (S1 layer) and MAX operator (C1 layer). In the second stage, the covariance descriptor [3] is applied to compute the similarity of BIF features at neighboring scales. While the Gabor filters and the covariance descriptors improve the robustness to the illumination variation, the MAX operator increases the tolerance to scale changes and image shifts. Furthermore, we argue that measuring the similarity of neighboring scales limits the influence of the background. By overcoming illumination, scale and background changes, the performance of person re-identification and face verification is greatly improved.

Stage 1. Considering the great success of BIF, the first step of BiCov consists in extracting such features to model image low-level properties. For an image $I(x,y)$, we compute its convolution with Gabor filters: $G(\mu,\nu) = I(x,y) * \psi_{\mu,\nu}(z)$ where μ and ν are scale and orientation parameters respectively. In BiCov, μ is quantized into 16 while the ν is quantized into 8.

In practice, we have observed that for person re-identification task, image representations $G(\mu,\nu)$ of different orientations can be averaged without significant loss of performance. Thus, we replace $\psi_{\mu,\nu}(z)$ in Eq. by $\psi_{\mu}(z) = \frac{1}{8}\sum_{\nu=1}^{8}\psi_{\mu,\nu}(z)$. This simplification makes the computations much more efficient.

In BiCov, two neighborhing scales are grouped into one band (we therefore have 8 different bands) by applying "MAX" pooling over two consecutive scales: $B_i = \max(G(2i-1), G(2i))$ "MAX" pooling operation increases the tolerance to small scale changes which often appear in person and face images since they are only roughly aligned. We call B_i $i \in [1,\cdots,8]$ as *BIF Magnitude Images*.

Stage 2. For each pixel on the BIF Magnitude Image B_i, a 7-dimensional vector is computed to capture the intensity, texture and shape statistics:

$$f_i(x,y) = [x, y, B_i(x,y), B_{i_x}(x,y), B_{i_y}(x,y), B_{i_{xx}}(x,y), B_{i_{yy}}(x,y)] \quad (1)$$

where x and y are the pixel coordinates, $B_i(x,y)$ is the raw pixel intensity at position (x,y), $B_{i_x}(x,y)$ and $B_{i_y}(x,y)$ are the derivatives of image B_i, while $B_{i_{xx}}(x,y)$ and $B_{i_{yy}}(x,y)$ are the second-order derivatives.

We divide the BIF Magnitude Image B_i into small regions with equal size and overlap. In this way, the spatial information of the images can be kept. Then, each region is represented by a covariance descriptor [3]:

$$C_{i,r} = \frac{1}{n-1}\sum_{(x,y)\in region\ r}(f_i(x,y)-\bar{f}_i)(f_i(x,y)-\bar{f}_i)^T \quad (2)$$

where $\bar{f}_i$ is the mean of $f_i(x,y)$ over region r and n is the size of region r. Covariance descriptors can capture shape, location and color information, and their performances have been shown to be better than other methods

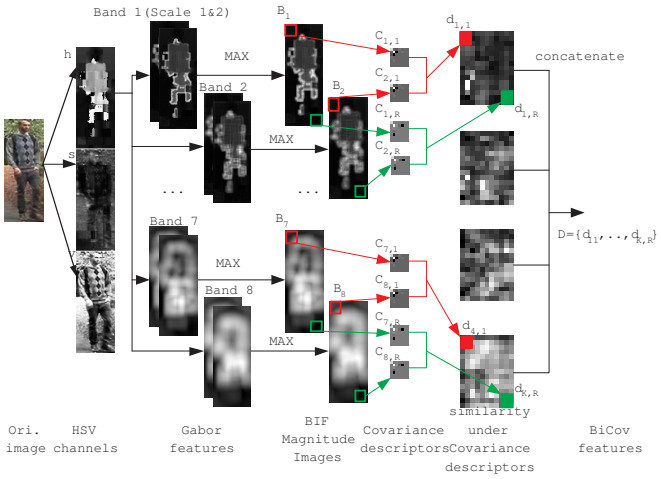


Figure 1: The flowchart of BiCov.

in many situations, as rotations and illuminations changes are absorbed by the covariance matrix [3].

In the traditional covariance-based methods, covariance matrices computed by Eq.2 are considered as image representation. Differently, in this paper, we compute for each region the difference of covariance descriptors between two consecutive bands

$$d_{i,r} = d(C_{2i-1,r}, C_{2i,r}) = \sqrt{\sum_{p=1}^{P}\ln^2\lambda_p(C_{2i-1,r}, C_{2i,r})} \quad (3)$$

where $\lambda_p(C_{2i-1,r}, C_{2i,r})$ is the pth generalized eigenvalues of $C_{2i-1,r}$ and $C_{2i,r}$, $i = 1,2,3,4$. BiCov avoid computing the difference of covariance descriptors of probe image and every gallery image which could be extremely time-consuming when the gallery is huge.

Finally, the differences are then concatenated to form the image representation: $D = (d_{1,1},\cdots,d_{1,R},\cdots,d_{K,1},\cdots,d_{K,R})$, where R is the number of regions and K is the number of band pairs (4 in our case). The distance between two images I_i and I_j is obtained by computing the Euclidian distance between these representations D_i and D_j.

Better person re-identification performance is usually obtained by combining different type of image descriptors. In this paper, we follow the same methodology and combine BiCov descriptor with other two: (a) Weighted Color Histograms (wHSV) and (b) Maximally Stable Color Regions (MSCR), as defined in [1].

To show the effectiveness of BiCov, this paper conducts experiments on two person re-identification tasks (VIPeR and ETHZ) and one face verification task (LFW), on which it improves the current state-of-the-art performance.

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[2] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, 1999.

[3] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.

Fast Pedestrian Detection by Cascaded Random Forest with Dominant Orientation Templates

Danhang Tang
http://www.iis.ee.ic.ac.uk/~dtang
Yang Liu
http://www.iis.ee.ic.ac.uk/~yliu
Tae-Kyun Kim
http://www.iis.ee.ic.ac.uk/~tkkim

Department of Electrical Engineering,
Imperial College,
London, UK

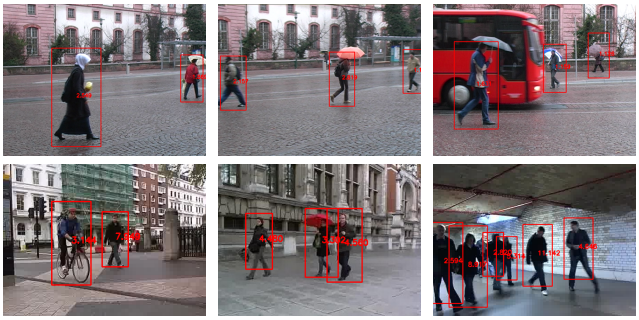


Figure 1: 1^{st} row: results of TUD pedestrian dataset; 2^{nd} row: results of our own video sequences. Numbers within bounding boxes indicate the voting scores.

In this paper, we present a new pedestrian detection method combining Random Forest and Dominant Orientation Templates(DOT) to achieve state-of-the-art accuracy and, more importantly, to accelerate run-time speed. Our method consists of a 2-level cascade: At first the scale space is divided into G overlapped groups, and a holistic detector(1^{st} level), designed with RF and our novel split function based on DOT, is applied on the first layer of each group and thus areas of interests are identified. After that a patch-based detector(2^{nd}), improved from HF [1] using the novel split function and clustering votes, is applied within these areas to achieve final bounding boxes.

Main contribution The prerequisite of this method is to adopt DOT as a descriptor, for its binary form is apt for bitwise operation. Besides orientations, we also encode the hue channel of colour information into binary format as another feature. Utilising DOT allows down-sampling images, which is the key reason for speeding up. However, since a significant amount of information(magnitude) is discarded, it loses some discriminative information. To compensate, we propose a novel similarity measurement to incorporate more dimensions. This measurement develops into a non-linear split function which better split the feature space whist maintaining the complexity of an axis-align split. This novel split function drastically improves both the detection *accuracy* of RF with 2-pixel tests on *DOT*, and the detection *speed* of RF with 2-pixel tests on *HOG* .

We define a template $\mathcal{T}$ as a n-dimension DOT sample selected from a positive training set $\mathbf{S}^P$. The similarity between $\mathcal{T}$ and any sample $\mathcal{S}$ can be measured with:

$$F(\mathcal{S},\mathcal{T}) = \sum_{\substack{P_d^{\mathcal{S}} \in \mathcal{S} \\ P_d^{\mathcal{T}} \in \mathcal{T}}} \delta(P_d^{\mathcal{S}} \otimes P_d^{\mathcal{T}} \neq 0), d = 1,...,n \qquad (1)$$

where $P_d^{\mathcal{S}}$ and $P_d^{\mathcal{T}}$ refer to the d^{th} dimension of $\mathcal{S}$ and $\mathcal{T}$ respectively. $\otimes$ is the bitwise AND operation. δ is an impulse function that is zero except when any bit in $P_{\mathcal{S}} \otimes P_{\mathcal{T}}$ is 1. To accelerate this matching process, SSE optimization is employed similar to [2]. Therefore although this function measures the distance between samples in a feature space as a non-linear split, it is efficient since only binary bitwise operations and addition are involved.

With the measurement above, we can then define the split function h_i as:

$$h_i(\mathcal{S}) = \begin{cases} 0, & F(\mathcal{S},\mathcal{T}_i) \leq \tau_i \\ 1, & F(\mathcal{S},\mathcal{T}_i) > \tau_i \end{cases}, \qquad (2)$$

where $\mathcal{T}_i$ means a chosen template and τ_i is a threshold of the i^{th} node.

During training, a set of positive $\mathbf{S}^P$ and negative $\mathbf{S}^N$ samples are used to construct a set of randomised decision trees. At the i-th non-leaf node,

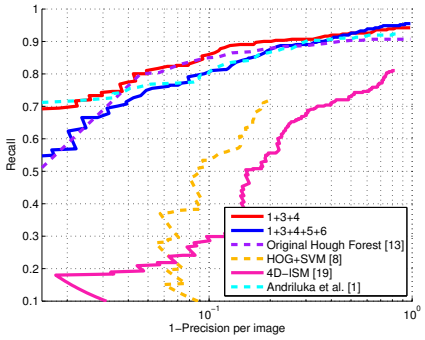


Figure 2: Comparison of accuracy with different components: 1. Dominant Orientations; 2. 2-Pixel Test; 3. Template Matching; 4. Dominant Colours; 5. Cascade; 6. Clustered Votes. (b) Comparison between our detector and State-of-the-art.

Method	Time(s)
Original Hough Forest	4.15
Our method(patch-based+DOT)	0.62
Our method(cascade)	0.33
Our method(cascade+clustering)	0.20

Table 1: Comparison of efficiency with 24 scales(0.17∼0.87) on 640x480 images.

we have these parameters:

$$\theta_i = \{\mathcal{T}_i, \tau_i\}, \quad \mathcal{T}_i \in \mathbf{S}_i^P, \qquad (3)$$

where $\mathcal{T}_i$ is a template chosen from positive samples, and τ_i is a threshold. We randomly generate a set of θ_i, and select the optimal one in terms of information gain.

Although adopting DOT allows down-sampling in scanning-windows, we still need to perform dense classification and voting to obtain satisfying results. Thus it is a natural option to construct a cascade to filter out unlikely regions before performing the patch-based detector. The design of this cascade and training tricks are described in detail in the full version of our paper.

Result In the experiment section, we first revisit different methods of training RF and come up with an optimal combination. The rest tests are performed accordingly. We compare our method against the original HF and other state-of-the-art methods. Figure 2 shows that adopting DOT with orientation and colour has better accuracy than the original HF and achieve 85% correct rate at 10^{-1} 1-precision. Cascaded version achieves more than 20 times of speed improvement whilst keeping a comparable accuracy. (Table 1) Note that our speed optimisation is mainly done about features rather than scales, therefore it can be combined with those works optimising multi-scale detection and obtain further speed-up. Also, we emphasise the inherent benefits of our RF framework for scalability, quick training, multi-class or multi-part detection.

[1] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *CVPR*, 2009.

[2] S. Hinterstoisser, V. Lepetit, S. Ilic, P. Fua, and N. Navab. Dominant orientation templates for real-time detection of texture-less objects. In *CVPR*, 2010.

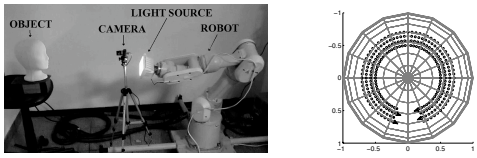
A method for improving consistency in photometric databases

Felipe Hernández-Rodríguez
felipe.hernandez@cinvestav.edu.mx
Mario Castelán
mario.castelan@cinvestav.edu.mx

Robotics and Advanced Manufacturing
Centro de Investigación y de Estudios Avanzados del I.P.N.
Ramos Arizpe, Coah., México.

Building photometric databases usually requires the gathering of images of a still object under different light source directions. During this process, unexpected artifacts such as noise, shadows, inter-reflections and other unwanted effects introduced by the sensibility of the camera may appear along the database, diminishing its consistency as a whole and therefore its suitability for the purposes of photometric analysis. This paper describes a method for improving photometric consistency in image databases acquired under photometric rigs. The main idea of our approach is to build and analyze a luminance matrix storing the reflectance behavior of each pixel under the different light source directions. To this end, we propose to fit sinusoidal functions to the singular vectors of this luminance matrix in order to improve its agreement with Lambertian reflectance. Experiments demonstrate that our method improves the photometric consistency of the database, providing stability for the purposes of photometric analysis of the database and surface shape recovery.

The data acquisition process for the photometric sampling database developed in our lab is described in Figure 1. A database of $k = k_1 \times k_2$ images of the observed object under k different light source directions was constructed, with k_1 azimuth angles and k_2 zenith angles of the light source direction vector.



(a) Experimental setting (b) Illumination patterns

Figure 1: **Data acquisition process scheme for the photometric sampling database.** In (a), the elements of the photometric sampling database acquisition scheme are shown, while (b) presents a projection of the light source positions, for four zenith and sixty-five azimuth angles. Separation between samples was 5°.

The photometric correction is based on the photometric sampling concept introduced by [1] that is related to the image irradiance equation for a Lambertian surface, that establishes the relationship of the surface normals $\mathbf{n}(u,v) \in \mathbb{R}^3$ and the light source direction $\mathbf{l}(u,v) \in \mathbb{R}^3$ to calculate the luminance for each pixel in the image: $i(u,v) =< \mathbf{n}, \mathbf{l} >$.

In accordance with the gathered images during the data acquisition process, the luminances of a pixel will draw a sinusoidal function if the illumination variations imposed around the object are circular, i.e., the different light source direction vectors are circles lying on the surface of a virtual sphere. The sine curve can be decomposed in the three parameters: amplitude ($\mathbf{A}$), phase ($\mathbf{B}$) and shift ($\mathbf{C}$) as $I(\theta) = \mathbf{A}sin(\theta + \mathbf{B}) + \mathbf{C}$, where $I(\theta)$ is the pixel luminance at each θ variation in azimuth.

The photometric correction commences by generating, for each pixel, a matrix $\mathbf{M}_{k_1 \times k_2}$ storing the k pixel intensity values recorded at each (azimuth, zenith) configuration pair. This matrix, which we refer to as the *luminance matrix*, contains the pixel reflectance history along the two main variations of the light source trajectory. For every pixel, the observed reflectance may be decomposed by the principal axis of the luminance matrix. The study of these axis allows identifying regions which best fit a sinusoidal behavior, i.e., close to a Lambertian behavior. The signal is finally corrected once the sine curve parameters have been calculated and the signal replaced with a sine function. We use SVD to decompose the luminance matrix,

$$\mathbf{M} = \mathbf{U}_{k_1 \times r} \Sigma_{r \times r} \mathbf{V}^T_{r \times k_2}, \quad (1)$$

where $r = rank(\mathbf{M})$. The column and row spaces of $\mathbf{M}$ are decomposed into the orthogonal basis $\mathbf{U}$ (left singular vectors) and $\mathbf{V}$ (right singular vectors), respectively. The singular values, contained in the diagonal elements of Σ explain the degree of retained variability in both right and left singular vectors. In our context $\mathbf{U}$ is related to the azimuth variations while $\mathbf{V}$ refers to variations in zenith.

Figure 2 shows a visual sketch of the light source trajectory variations over a surface normal of the mannequin and the human face. Since real images may include noisy variations in reflectance (i.e., bottom row in Figure 2), using least squares for calculating the three sine parameters (amplitude, phase and shift) may lead to poor estimations in the sought parameters, as behaviors departing from Lambert's law may occupy a large region of the singular vector.

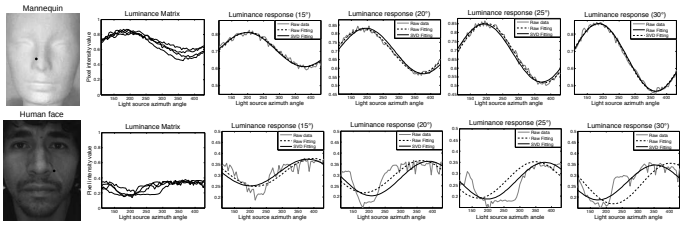


Figure 2: **Photometric correction for two different reflectance examples.** A mannequin and a human face are respectively shown in the top and bottom rows of the figure. For each case, the values in the luminance matrix for a single pixel (marked with a dark dot) are plotted in the second column. The rest of the columns depict sine fitting results on each vector of the luminance matrix.

To overcome this problem, we performed fitting using smaller fixed-sized signal periods starting from each point along the singular vector, then chose the phase, amplitude and shift parameters appearing in the majority of the cases. The fitting procedure is roughly illustrated by Figure 3. Once the sine parameters are estimated over each of the singular vectors, a new luminance matrix $\mathbf{M}' = \mathbf{U}'\Sigma\mathbf{V}'^T$ is generated to improve photometric consistency on the database. The new fitted columns of $\mathbf{U}'$ and $\mathbf{V}'$ contain the sine-fitted singular vectors from the original matrices $\mathbf{U}$ and $\mathbf{V}$ in Eq. 1.

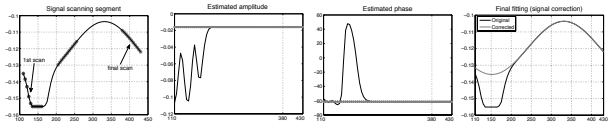


Figure 3: **The sinusoidal fitting procedure.** The figure illustrates the proposed sine fitting procedure, based on scanning sine segments through the singular vectors of the luminance matrix.

To determine the efficiency of the proposed method two different photometric databases was corrected to induce consistency, the photometric sampling database and the Yale B photometric stereo database.

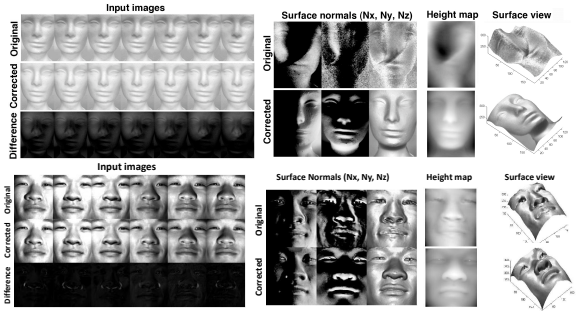


Figure 4: **Experimental results .** The figure presents original and corrected results for the photometric sampling database and the Yale B database.

[1] S.K. Nayar, K. Ikeuchi, and T. Kanade. Shape and reflectance from an image sequence generated using extended source. *Proceedings of IEEE ICRA*, 1:28–35, 1989.

Object Instance Sharing by Enhanced Bounding Box Correspondence

Santosh K. Divvala
santosh@cs.cmu.edu
Alexei A. Efros
efros@cs.cmu.edu
Martial Hebert
hebert@ri.cmu.edu

The Robotics Institute,
Carnegie Mellon University
Pittsburgh, PA. USA.

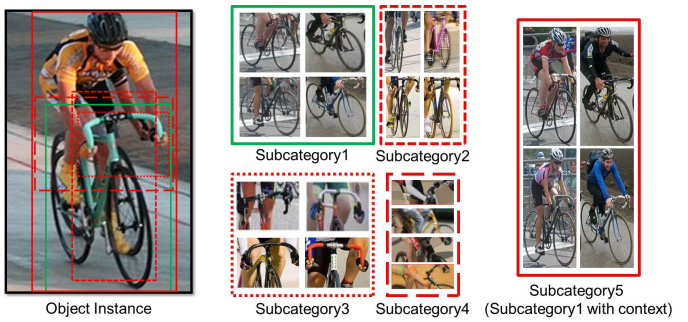


Figure 1: (left) A bicycle instance with its ground-truth bounding box shown in solid green. (center) Four (of the 25) subcategories discovered by our approach (few sample instances within each subcategory are shown). We allow the bicycle instance to be used multiple times with different bounding box representation for training the subcategory models. The different bounding box extents used per subcategory model are color coded accordingly e.g., subcategory3's match is shown using red dotted box, subcategory4's match shown in red dashed box, etc. (right) Subcategory1 shown after adaptively enlarging the bounding box to include local contextual cues around it.

Figure 2: Subcategories composed of only a few instances, specifically in case of truncation, can gather more data from other training examples. Each row displays (left) a sample training instance from a subcategory, (right) new samples generated from existing training instances. Red box is the new sample, green box is the human annotation.

Figure 3: Human-annotated bounding boxes (green box) are automatically enlarged (red box) to leverage local contextual cues (adapted to the subcategory). There is a wide variation in the types of context captured per subcategory e.g., rail tracks under a 'train', people seated at 'dining table', etc.

Consider the task of building a sliding-window object detector. The standard learning-based approach is to first turn each human-labeled bounding box into a feature vector using some feature descriptor, e.g. HOG, and then train a classifier, e.g. SVM, on a stack of these feature vectors to discriminate them from the rest of the visual world. This is a reasonable strategy for older datasets, such as "INRIA person", where object instances are largely in correspondence, i.e. aligned such that each feature vector dimension has the same visual meaning for all object instances. However, modern datasets, such as PASCAL VOC, are much less restricted and do not guarantee good correspondence, with often huge variations between annotated bounding box instances.

The way modern approaches usually tackle this problem is by using mixture models [1, 2]. The idea is to somehow segregate instances within a category into disjoint groups (subcategories) and then train separate classifiers for each such subcategory. Each subcategory has reduced appearance diversity (via improved alignment), leading to a simpler learning problem. The recent success of the discriminatively-trained mixture model framework of Felzenszwalb et al., [1] has led to a wide popularity of such models for object detection. While reasonable, this assumes that a lot of training data is available for each subcategory. But this is often not the case, especially for occluded/truncated instances.

Consider the image shown in Figure 1(left). The human-labeled "bicycle" bounding box is indicated by the solid green box. Given this ground-truth framing for the object instance, it is most similar to instances in the "45°-view bicycle" subcategory, so, in a standard mixture-model detector, it would be assigned to subcategory1. However, by relaxing the bounding box framing for this instance, subregions of it can also match to the other subcategory models (subcategory2, subcategory3, subcategory4) as shown using the red bounding boxes. Furthermore, looking *outside* the bounding box might also allow us to capture consistencies in the local context surrounding the object, discovering new subcategories such as 'person riding a bicycle' (subcategory5).

What we propose in this paper is the idea of *training data reuse*. Conceptually, we would like to allow different object subcategories to be able to share (subregions of) each others training instances by providing *extra* correspondences between instances that were not part of the original human-supplied bounding box annotations. We operationalize this by two complementary operations: bounding box shrinking, which aims to find subregions of an instance that could be shared (Figure 2); and bounding box enlarging, which aims to create new subcategories by enlarging instances to include their local context (Figure 3).

Approach Overview Our approach builds upon the latent bounding box fitting method introduced in [1, 3], where the human-annotated bounding box is treated as being partially *latent* i.e., the bounding box is allowed to move within a local neighborhood (down to 70% overlap). Intuitively, this can be understood as locally "wiggling" the bounding box representation such that it best *aligns* with the rest of the object instances within a category (or subcategory). In this paper, we apply a very similar mechanism, but rather than just making local adjustments, we use it to *search* for bounding box representations that capture new correspondences between instances in the training data. The main difference is that the latent bounding box fitting assumes that each object instance is represented by a *single* bounding box belonging to a single subcategory, whereas our aim is to find *many different* bounding boxes for the same instance, so that it can be shared across multiple subcategories.

[1] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, September 2010.

[2] Robert Jacobs, Michael Jordan, Steven Nowlan, and Geoffrey Hinton. Adaptive mixture of local experts. In *Neural Computation*, 1991.

[3] P. A. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2005.

MaxFlow Revisited: An Empirical Comparison of Maxflow Algorithms for Dense Vision Problems

Tanmay Verma
tanmay08054@iiitd.ac.in
IIIT-Delhi
Delhi, India
Dhruv Batra
dbatra@ttic.edu
TTI-Chicago
Chicago, USA

Motivation. Over the past two decades, algorithms for finding the maximum amount of flow possible in a network (or max-flow) have become the workhorses of modern computer vision and machine learning – from optimal (or provably-approximate) inference in sophisticated discrete models [2, 12] to enabling real-time image processing [16]. Perhaps the most prominent role of max-flow is due to the work of Hammer [10] and Kolmogorov and Zabih [12], who showed that a fairly large class of energy functions – sum of submodular functions on pairs of boolean variables – can be efficiently and *optimally* minimized via a reduction to max-flow. Max-flow also plays a crucial role in approximate minimization of energy functions with multi-label variables, triplet or higher order terms, global terms, and terms encoding label costs.

Given the wide applicability, it is important to ask *which* max-flow algorithm should be used. There are numerous algorithms for max-flow with different asymptotic complexities and practical run-time behaviour. Broadly speaking, there are three main families of max-flow algorithms:

1. Augmenting-Path (AP) variants: algorithms [1, 5, 6, 7, 9] that maintain a valid flow during the algorithm, *i.e*. always satisfying the capacity and flow-conservation constraints.
2. Push-Relabel (PRL) variants: algorithms [4, 8] that maintain a *preflow*, *i.e*. satisfy the capacity constraints but may violate the conservation constraints to have flow excess at nodes (but never a deficit).
3. Pseudoflow (HPF) variants: algorithms [3, 11] that maintain a *pseudoflow*, *i.e*. satisfy the capacity constrains but may violate the conservation constrains to allow flow excess and deficit at nodes.[1]

Boykov and Kolmogorov [1] compared AP and PRL algorithms on a number of computer vision problems, and found that their own algorithm (BK) was the fastest algorithm in practice, even though they could only prove a very loose asymptotic complexity bound of $O(n^2mC)$, where n is the number of nodes, m is the number of edges and C is the max-flow value.

Goal. The central thesis of this work is that since this comparison a decade ago, the models used in computer vision and the *kinds* of inference problems we solve have changed significantly. Specifically, while [1] only considered 4-connected grid MRFs, the models today involve high-order terms, long-range connections, hierarchical MRFs and even global terms. The effect of all these modifications is to make the underlying max-flow graph significantly denser, thus causing the complexity of the algorithm of [1] to become a concern. It is time to revisit this comparison.

Contribution. The goal of this paper is to compare the runtimes of different max-flow algorithms, to investigate if the conclusions of Boykov and Kolmogorov [1] are still valid for current-day *dense* problems, and find out which algorithm is most suited for modern vision problems. One key contribution of our study is that it includes recently proposed algorithms – Pseudoflow [3, 11] and Incremental Breadth-First-Search (IBFS) [9] – which were not developed at the time [1] was written, and thus were absent from their comparison.

Problems. We tested a number of max-flow algorithms on the following:

1. **Synthetic Instances.** We created synthetic max-flow graphs with a basic grid structure and randomly added long-range edges depending on a density parameter.
2. **ALE Graphs.** We used the max-flow graphs created during alpha-expansion by the Automatic Labeling Environment (ALE) of Ladicky *et al*. [13] on PASCAL VOC 2010 segmentation images.
3. **Deconvolution.** We used the QPBO max-flow graph on the binary image deconvolution CRF instance from Rother *et al*. [15]. This is an extremely dense graph and the problem is not submodular.
4. **Super Resolution.** We used the QPBO max-flow graph on the super-resolution CRF instances of [15].
5. **Texture Restoration.** We used the QPBO max-flow graph on the Brodatz texture D103 model from [15].
6. **DTF Graphs.** Decision Tree Field (DTF) [14] is a recently introduced model that combines random forests and conditional random fields. We used the 100 instances provided by Nowozin *et al*. [14] and saved the QPBO-graphs to file.
7. **3D Segmentation.** Finally, we also evaluated all algorithms on the standard benchmark for such studies, the binary 3D (medical) segmentation instances from the University of Western Ontario `http://vision.csd.uwo.ca/maxflow-data/`.

We note that all of the previous studies were restricted to 3D segmentation, and problems 2-6 have never been used to evaluate max-flow algorithm, yet they are in some sense more representative of modern problems.

Findings: Our paper has the following findings:

1. **Choice of Algorithm *Matters*.** In all applications, the fastest algorithms is *orders of magnitude* faster than the slowest algorithm.
2. **BK Scales Poorly with Density.** Our results show that the motivating hypothesis of this study is correct. In a number of cases (Synthetic, Deconv, 3Dseg) BK starts out fairly competitive at low densities but very quickly becomes the slowest algorithm.
3. **New Kids in Town: IBFS and HPF.** In a number of applications we considered, both IBFS and HPF significantly outperform BK, IBFS more consistently so than HPF.
4. **Clever Data-structures Matter.** We found the data-structures used by BK to be particularly efficient. In a number of applications (see *e.g*. SuperRes, Texture-Restoration), BK maxflow time is longer than IBFS but the maxflow+initialization time is shorter.

We hope that the results of our study guide practitioners in picking the correct implementation for their problems.

[1] Interestingly, the key difference between Push-Relabel and Pseudoflow algorithms is not the concept of pseudoflow rather the admissibility of certain push schemes.

[1] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 26(9):1124–1137, 2004.
[2] Yuri Boykov, Olga Veksler, and Ramin Zabih. Efficient approximate energy minimization via graph cuts. *PAMI*, 20(12):1222–1239, 2001.
[3] Bala G. Chandran and Dorit S. Hochbaum. A computational study of the pseudoflow and push-relabel algorithms for the maximum flow problem. *Oper. Res.*, 57:358–376, March 2009. ISSN 0030-364X. doi: 10.1287/opre.1080.0572.
[4] Boris V. Cherkassky and Andrew V. Goldberg. On implementing the push-relabel method for the maximum flow problem. *Algorithmica*, 19(4):390–410, 1997.
[5] E. A. Dinic. Algorithm for Solution of a Problem of Maximum Flow in a Network with Power Estimation. *Soviet Math Doklady*, 11:1277–1280, 1970.
[6] Jack Edmonds and Richard M. Karp. Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, 19:248–264, April 1972. ISSN 0004-5411.
[7] L. R. Ford and D. R. Fulkerson. Maximal Flow through a Network. *Canadian Journal of Mathematics*, 8:399–404, 1956.
[8] Andrew V. Goldberg and Robert E. Tarjan. A new approach to the maximum-flow problem. *J. ACM*, 35:921–940, October 1988. ISSN 0004-5411. doi: http://doi.acm.org/10.1145/48014.61051. URL `http://doi.acm.org/10.1145/48014.61051`.
[9] Andrew V. Goldberg, Sagi Hed, Haim Kaplan, Robert E. Tarjan, and Renato F. Werneck. Maximum flows by incremental breadth-first search. In *Proceedings of the 19th European conference on Algorithms*, ESA'11, pages 457–468, 2011. ISBN 978-3-642-23718-8.
[10] P.L. Hammer. Some network flow problems solved with pseudo-boolean programming. *Operations Research*, 13:388–399, 1965.
[11] Dorit S. Hochbaum. The pseudoflow algorithm: A new algorithm for the Maximum-Flow problem. *Operations Research*, 56(4):992–1009, July 2008.
[12] Vladimir Kolmogorov and Ramin Zabih. What energy functions can be minimized via graph cuts? *PAMI*, 26(2):147–159, 2004.
[13] L. Ladickỳ, C. Russell, P. Kohli, and P. H. S. Torr. Associative hierarchical CRFs for object class image segmentation. *ICCV*, 2009.
[14] Sebastian Nowozin, Carsten Rother, Shai Bagon, Toby Sharp, Bangpeng Yao, and Pushmeet Kholi. Decision tree fields. In *ICCV*, 2011.
[15] C. Rother, V. Kolmogorov, V. Lempitsky, and M. Szummer. Optimizing binary MRFs via extended roof duality. In *CVPR*, June 2007. doi: 10.1109/CVPR.2007.383203.
[16] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "Grabcut": interactive foreground extraction using iterated graph cuts. *SIGGRAPH*, 2004.

Scalable Cascade Inference for Semantic Image Segmentation

Paul Sturgess[1]
paul.sturgess@brookes.ac.uk
Ľubor Ladický[2]
lubor@robots.ox.ac.uk
Nigel Crook[1]
ncrook@brookes.ac.uk
Philip H. S. Torr[1]
philiptorr@brookes.ac.uk

[1] Department of Computing
Oxford Brookes University
Oxford,UK

[2] Department of Engineering Science
University of Oxford
Oxford, UK

Semantic image segmentation (SIS) is a problem of simultaneous segmentation and recognition of an input image into regions and their associated categorical labels, such as person, car or cow. A popular way to achieve this goal is to assign a label to every pixel in the input image and impose simple structural constraints on the output label space. Such approaches have been successfully formulated as pairwise conditional random fields (CRF) and higher order CRFs [3]. These approaches are now practically solvable for some problems due to advances in inference techniques. Currently the α-expansion [1] algorithm has proved to be perhaps the most efficient approximation algorithm for the SIS problem and is amongst the state-of-the art for quantitative performance. Empirically the algorithm's runtime is linear in the number of labels, making it practical only when working in a specific domain that has few classes-of-interest ($10-20$ for example). However when working in a more general setting where the number of classes could easily reach tens of thousands, sub-linear complexity is required. In this paper we propose to meet this requirement by dividing the large label set into smaller more manageable ones, and then only solving for some of these subsets. Since the SIS problem is concerned with categorical labels a natural way to subdivide the label set is by building a hierarchy, or taxonomy. Given a hierarchy we propose a cascade architecture that can reject whole portions of the label space at the early stages of the optimisation. We also dynamically subdivide the image into smaller and smaller regions during inference to gain further efficiency. The use of a cascade is motivated by the observation that even with a large label domain, a single image will usually only contain a small subset of classes.

We demonstrate the effectiveness of the approach with quantitative evaluation of performance on the SUN09 database [2] that has 107 labels.

1 Cascaded Inference

In order to obtain scalable SIS we propose a to perform cascade style inference. In this section we specify the details of our approach. First we define two general functions:

$$\begin{aligned} &\textit{variable selection} && T_\delta : V \to V', \\ &\textit{variable assignment} && T_V : \delta \to \delta', \end{aligned}$$

that can be applied respectively to the variables (vertices) and the label domain of the cost function; T_δ transforms the current set of variables, given a domain; T_V modifies the current domain, given some variables. We can specify these transformation functions in different ways such that their evaluation performs a move for many move making algorithms. Here we are interested in specifying them in order to perform cascaded inference over a tree structured label space, or taxonomy τ. We define such a space with reference to an unstructured domain Δ as recursive subdivision into disjoint subsets δ such that the root node contains all the elements of Δ and leaf nodes contain the elementary labels $l \in \Delta$. Now, let δ denote a group of siblings, that is a set of children that share the same direct parent in the tree and thus forms a subdomain of Δ. Also let $\pi(\delta)$ signify the domain that the shared parent belongs to, i.e. If the domain $\Delta = \{cat, dog, car, van\}$, then we could have the following groupings that form our tree; The head node would be $everything = \{cat, dog, car, van\}$ and it may have two children, such as $animal = \{cat, dog\}$ and $vehicle = \{car, van\}$. In turn these would then have two leaf nodes as children. Then $\pi(vehicle)$ points to the domain $everything$ and $\pi(dog)$ points to the label domain $\{cat, dog\}$. Thus a tree defines a set of domains $\{\delta_1, \ldots \delta_{n+1}\}$, where n is the number of sibling groups. For convenience we also maintain an index δ_i^j to the j^{th} elementary label contained within the i^{th} domain, i.e. $vehicle^1 = car$, as does $everything^3$. Given these notations variable selection and assignment based on a tree is then defined as:

$$T_v(\delta) = \begin{cases} \delta_i & \text{if } \delta_i^j \in f^*_{\pi(\delta)} \text{ and } \delta \neq \emptyset \\ \emptyset & \text{otherwise,} \end{cases} \tag{1}$$

$$T_\delta(v) = v \in \{\mathbf{I}(f(T_\delta(v)) \neq \inf\} \tag{2}$$

where $\emptyset$ is the empty set, $\mathbf{I}$ is an indicator function, $f^*_{\pi(\delta)}$ is a given solution for the a labelling problem defined on the domain $\pi(\delta)$ and variables V' and

$$f(T_\delta(v)) = \begin{cases} c^\tau(v, f(v)) & \text{if } f(v) \in \delta \\ \infty & \text{otherwise} \end{cases}, \tag{3}$$

$$c^\tau(v, f(v)) = \arg \min_{f(v) \in \delta_i} c(v, f(v))). \tag{4}$$

For the first layer of the tree $f^*_{\pi(\delta)}$ is trivial since $\pi(\delta)$ is the single label domain of the head node, i.e. $f : V \to [1]$. This means that we have to solve a k label problem at the start of our cascade, where k is the number of children of the head node. In our running example this would be the $\{animal, vehicle\}$ domain on all variables V of the original graph. However when we visit all the nodes in the tree in the following fashion:-

for all i minimize:

$$Q(f) = \sum_{v \in T_{\delta_i}(v)} c^\tau(v, f(v)) + \sum_{(u,v) \in \mathcal{E}'} w(u,v) \cdot d(f(u), f(v))$$

subject to:

$$\begin{aligned} & f : v \to \alpha && \forall v \in T_{\delta_i}, \exists \alpha \in T_v \\ & d(\alpha, \alpha) = 0 && \forall \alpha \in T_v \\ & d(\alpha, \beta) = d(\beta, \alpha) \geq 0 && \forall \alpha, \beta \in T_v \\ & d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\gamma, \beta) && \forall \alpha, \beta, \gamma \in T_v \\ & w(u,v) \geq 0 && \forall u, v \in T_{\delta_i}, \end{aligned} \tag{5}$$

many sub-problems will be trivial such as:- no labels, $|\delta| = \emptyset$; a single label $|\delta| = 1$; no finite cost variables $\forall v \in T_v : c(v, f_\delta(v)) = \infty$. In these cases, we need not evaluate the function at all, saving computation time. In the cases where the cost is non-trivial with binary $\delta = \{\alpha, \beta\}$, or a multi-class domain with $|\delta| > 2$ and $\exists v \in T_v : c(v, f_\delta(v)) \neq \infty$. The cost function remains metric since we only modify the data term $c(.,.)$, thus we can approximately solve it using α-expansion or other suitable methods. We show that our cascaded approach achieves a good approximation, $Q(\bigcup_{i \in leafs} Q(f^*_{\delta_i})) \approx Q(f^*_\Delta)$.

[1] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23:2001, 2001.

[2] Myung Jin Choi, Joseph J. Lim, Antonio Torralba, and Alan S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010.

[3] Lubor Ladicky, Chris Russell, Pushmeet Kohli, and P. H. S. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform

Ali Mosleh[1]
mos_ali@encs.concordia.ca
Nizar Bouguila[2]
bouguila@ciise.concordia.ca
A. Ben Hamza[2]
hamza@ciise.concordia.ca

[1] Department of Electrical and Computer Engineering
[2] Concordia Institute for Information Systems Engineering
Concordia University
Montréal, QC, Canada

Figure 1: Several steps of the text detection technique. (a) Original image. (b) Edge map using bandlet transform. (c) SWT of the image. (d) Text and non-text CCs classification. (e) Merging text CCs to produce the final result.

A slew of semantic image content analysis techniques are specialized in extracting text embedded in images since it is a vital source of semantic information. A robust text detection step is the basic requirement for a scheme designed to extract text information from images. Text detection is still a challenging issue due to unconstrained color, sizes, alignments of characters, lighting and also various shapes of fonts, even though various methods have been proposed in the past years [2]. Existing text detectors can be broadly classified into two main groups: texture (also called region) based and connected component (CC) based methods.

The general scheme of our proposed method consists in producing (Fig. 1) the image edge map and then finding CCs based on stroke width transform (SWT) [1] guided by the generated edge map. Next, precise feature vectors are formed using the properties of CCs from SWT and pixel domain. An unsupervised clustering is performed on the image CCs to detect the candidate text CCs. Finally, text candidate components are linked to form text-words. The method is considered as a CC-based technique and the contribution is twofold: 1) Since accurate edge maps drastically enhance SWT results, a precise edge detection approach adaptive to text-regions is proposed by employing the bandlet transform. 2) A feature vector based on text properties and stroke width values is employed in k-means clustering in order to detect text CCs.

Bandlet transform [3] effectively represents the geometry of an image. The image coefficients are dyadically segmented in squares S for polynomial flow approximation of the geometry before the bandletization process. Since the image coefficients are all warped along local dominant flows in the bandlet transform, the final bandlet coefficients generated for each segmentation square S have the form of approximation, and high-pass filtering values appear in the wavelet transform of a 1D signal. We benefit from the bandlet-based resulting 1D high-pass frequency coefficients that are adapted to the directionality of the edge that exists in each segmentation square S in order to find a binary map of the edge positions in the image. Since the approximation part of the bandlet transform resulting coefficients consists of coarse information of the original signal, we discard it and only process the high-pass coefficients. The first-order derivatives of the fine-detail bandlet coefficients are computed. Then, local maxima of the resulting gradient signal are found using a contextual filter as follows:

$$M_i = \begin{cases} 1 & \text{if } g_i > T \ \wedge \ g_i > g_j, \ \forall j \in [i-L, i-1] \ \wedge \\ & g_i > g_j, \ \forall j \in [i+1, i+L] \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where each point in the gradient signal is denoted by x_i. g_i represents the gradient value for x_i and g_j indicates gradient value of neighboring pixels of x_i that exist is a window with size $2L+1$ centered at x_i. T is a threshold value and M is a map of local maxima of the gradient signal. The corresponding locations of 0's of M in the bandlet fine (high-pass) coefficients are set to 0, for all the bandlet squares S. Then, the inverse bandlet transform is performed in order to have the final edge locations of the original image. The quality of the edge map depends on the value of the threshold T. Therefore, the edge detection process is performed with two different values of T. Once, a low value is chosen to assure that no reasonable edge information is lost. Then, a higher value is assigned to T to only capture the significant edges. Filially, the two results are combined to produce the finial edge map (Fig. 1(b)).

Element	Description
V_{SWT}	variance of stroke width values
μ_{SWT}	mean of stroke width values
M_{SWT}	median of stroke width values
R_s	ratio of the component diameter and M_{SWT}
V_G	variance of gradient directions of all the CC's edge pixels
SK_G	skewness estimation for the gradient directions of all the CC's edge pixels
C_L	contrast value of the CC and the background in the CC's bounding box
R_{asp}	aspect-ratio of the bounding box of the CC

Table 1: Elements of the feature vector generated or each CC.

Our text detection approach obtains features for CCs produced by SWT, then decides which CC is a text candidate using k-means clustering. In the first step, we find the edges of the input image by means of the proposed bandlet-based edge detection method. A ray shooting process is performed from each edge pixel along its gradient direction. The number of pixels which lie on the ray between two edge pixels with opposite gradient directions is considered as the stroke width for those pixels. Using a 4-neighboring pixels search, adjacent pixels are grouped if the ratio of their stroke width values is higher than 0.3 and lower than 3 (Fig. 1(c)). Then, features of the produced CCs are extracted and used to find text candidates. Table 1 summarizes the elements of the feature vector generated for each CC which has the following form:

$$\overrightarrow{F} = \{V_{SWT}, \ \mu_{SWT}, \ M_{SWT}, \ R_s, \ V_G, \ SK_G, \ C_L, \ R_{asp}\} \quad (2)$$

The produced $\overrightarrow{F}$ of all the CCs of the image are fed to a k-means scheme and consequently clustered into two groups, non-text and text components (Fig. 1(d)). In order to identify which cluster is associated to the texts and which is not, at the beginning of the process we append a sample text to the end of each input image. Hence, the resulting cluster that contains the sample text components is considered as the group of text components and the rest of the components are discarded. In the last step, the remaining text components which are horizontally aligned and have reasonable distance to each other, for example as far as a character width, are grouped together and form the word components (Fig. 1(e)).

[1] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2963–2970, June 2010.

[2] K. Jung, K. I. Kim, and A. K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, 37(5):977 – 997, 2004.

[3] S. Mallat and G. Peyre. A review of bandlet methods for geometrical image representation. *Numerical Algorithms*, 44:205–234, Mar. 2007.

Higher-order Co-occurrence Features based on Discriminative Co-clusters for Image Classification

Takumi Kobayashi
takumi.kobayashi@aist.go.jp

National Institute of Advanced Industrial Science and Technology
1-1-1, Umezono, Tsukuba, Japan

Motivation

Co-occurrence based image features have attracted keen attentions due to the promising performances for image classification tasks [1, 2, 3, 6, 7]. For extracting the co-occurrences, it is common to transform the *quantitative* data into *qualitative* data (symbols) by means of quantization (clustering) at first; *e.g.*, continuous gradient orientation is coded into orientation bins [3], RGB colors are indexed [2] and local features are categorized into visual words [7]. Such point-wise clustering, however, is not necessarily suitable to characterize the pair-wise co-occurrences. And the higher-order co-occurrences beyond pair-wise has been rarely considered due to the exponential increase of the dimensionality by using those point-wise symbols. In this paper, we propose a method to extract image features based on effective higher-order co-occurrences. The proposed method constructs the co-clusters to discriminatively quantize joint primitive quantitative data, such as pair-wise pixel intensities, unlike the standard co-occurrence methods that utilize simple clusters trained in an unsupervised manner for quantizing point-wise data. The discriminative co-clusters effectively exploit the co-occurrence characteristics even by a fewer number of cluster components, resulting in low-dimensional co-occurrence features, which enables us to develop the higher-order co-occurrence features of feasible dimensionality.

Proposed method

Let $\mathcal{R}$ be the quantitative data space, $x_{\boldsymbol{p}} \in \mathcal{R}$ be the quantitative data at pixel position $\boldsymbol{p}$ in an image plane. We consider the general form for extracting co-occurrences:

$$\boldsymbol{M} = \left\{ \sum_{\{\boldsymbol{p},\boldsymbol{q}\}\in\mathbb{N}} \omega(\boldsymbol{p},\boldsymbol{q}) g_k(x_{\boldsymbol{p}},x_{\boldsymbol{q}}) \right\}_{k=1,\cdots,D} \in \Re^D,$$

where $\mathbb{N}$ indicates the set of local neighbor pairs and ω is the weighting function on those pairs. We introduce the function $g_k(x_{\boldsymbol{p}},x_{\boldsymbol{q}}) : \mathcal{R}\times\mathcal{R} \to \Re_+$ to assign the pair $(x_{\boldsymbol{p}},x_{\boldsymbol{q}})$ with the k-th cluster ($k = 1,\cdots,D$) in the joint space $\mathcal{R}\times\mathcal{R}$, called *co-cluster*. The co-clusters g_k directly measure the co-occurrences and we construct them in a discriminative manner.

Discriminative co-cluster (Fig. 1). Suppose a two-class problem of images I_n with the class label $y_n \in \{+1,-1\}$. From the image I_n, we first extract primitive co-occurrence features $\tilde{\boldsymbol{M}}_n$ on $\mathcal{R}\times\mathcal{R}$ as in GLCM [1]; in practice, the space $\mathcal{R}$ which is usually continuous is finely partitioned into (large number of) L bins, resulting in $\tilde{\boldsymbol{M}}_n \in \Re^{L\times L}$. Then, the *linear* SVM is applied to those $(\tilde{\boldsymbol{M}}_n, y_n)$ in order to produce the classifier weight $\boldsymbol{W}$ on $\mathcal{R}\times\mathcal{R}$, actually $\boldsymbol{W} \in \Re^{L\times L}$. The classifier weight exploits the discriminative information: the positive weights in $\boldsymbol{W}$ contribute to '+1' class, while the negative ones to '-1' class. Finally, we perform clustering on the weight matrix $\boldsymbol{W}$ to produce the co-cluster assignment function g_k which is determined as the membership function to the k-th co-cluster on $\mathcal{R}\times\mathcal{R}$. We separately treat the weight $\boldsymbol{W}$ in terms of its sign (positive/negative) as the positive weight $\boldsymbol{W}^+ = \max(\boldsymbol{W},0)$ and the negative $\boldsymbol{W}^- = \max(-\boldsymbol{W},0)$, $\boldsymbol{W} = \boldsymbol{W}^+ - \boldsymbol{W}^-$, and apply the EM clustering method [5] to those respective weights; the cluster component is represented by Gaussian function $\mathcal{N}_k$ with the prior weights α_k. The function g_k is finally determined by

$$g_k(x_1,x_2) = \frac{\alpha_k \mathcal{N}_k(x_1,x_2)}{\sum_k \alpha_k \mathcal{N}_k(x_1,x_2)}, \ \forall x_1,x_2 \in \mathcal{R},$$

which is the posterior probability at (x_1,x_2), resulting in the normalized g_k: $\sum_k g_k(x_1,x_2) = 1$.

Higher-order co-occurrence (Fig. 2). By using the co-clusters g_k, the proposed higher-order co-occurrence features are defined by

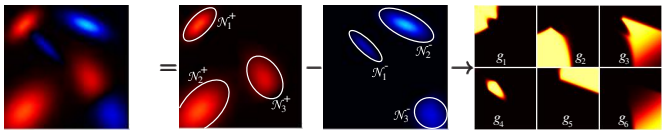
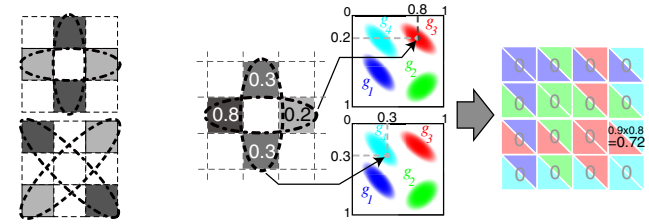


Figure 1: Construction of discriminative co-clusters g_k.



(b) *cross* quadruplet (c) higher-order co-occurrence

Figure 2: Higher-order co-occurrence.

$$\boldsymbol{H} = \left\{ \sum_{\{\boldsymbol{p},\boldsymbol{q},\boldsymbol{r},\boldsymbol{s}\}\in\mathbb{Q}} \omega(\boldsymbol{p},\boldsymbol{q},\boldsymbol{r},\boldsymbol{s}) g_k(x_{\boldsymbol{p}},x_{\boldsymbol{q}}) g_l(x_{\boldsymbol{r}},x_{\boldsymbol{s}}) \right\}_{k,l=1,\cdots,D} \in \Re^{D\times D},$$

where $\mathbb{Q}$ indicates the quadruplets. In this higher-order co-occurrence, it is important how to determine the quadruplets $\mathbb{Q}$, forms of which could be combinatorially increased. Co-occurrences are based on pairs which are oriented in various directions, and we configure the quadruplets, the pairs of pairs, in the form of *cross* as shown in Fig. 2a in order to extract diverse characteristics in image textures; the pairs in the cross are maximally (orthogonally) separated.

Results

We applied the proposed method to cancer detection using biopsy images [6] and pedestrian detection using the Daimler Chrysler dataset [4], both of which result in two class classifications. The primitive quantitative data $x_{\boldsymbol{p}}$ are pixel intensities for biopsy images and gradient orientations for pedestrian images. Even for smaller number of the co-clusters, D, the proposed method produces superior performances to the other methods, including the standard co-occurrence method; in equal error rate, 94.29% for cancer detection and 94.32% for pedestrian detection.

[1] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transaction on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

[2] J. Huang, S. R. Kumar, M. Mitra, W-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 762–768, 1997.

[3] T. Kobayashi and N. Otsu. Image feature extraction using gradient local auto-correlations. In *European Conference on Computer Vision*, pages 346–358, 2008.

[4] S. Munder and D. M. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.

[5] Y. Xiao and G. Xuan. Fast em algorithm of multi-dimensional histogram in medical images. In *International Conference on Diagnostic Imaging and Analysis*, pages 328–333, 2002.

[6] A. Yaguchi, T. Kobayashi, K. Watanabe, K. Iwata, T. Hosaka, and N. Otsu. Cancer detection from biopsy images using probabilistic and discriminative features. In *International Conference on Image Processing*, pages 1641–1644, 2011.

[7] Y. Yang and S. Newsam. Spatial pyramid co-occurrence for image classification. In *International Conference on Computer Vision*, pages 1465–1472, 2011.

An Assessment of Visual Discomfort Caused by Motion-in-Depth in Stereoscopic 3D Video

Sang-Hyun Cho
cshgreat@catholic.ac.kr
Dept. of Computer Engineering,
The Catholic University of Korea

Hang-Bong Kang
hbkang@catholic.ac.kr
Dept. of Digital Media,
The Catholic University of Korea

Stereoscopic image viewing comfort is one of the main problems that should be solved before the mass market proliferation of stereoscopic 3D content services. Recent research suggests that motion-in-depth could play a more important role in generating visual discomfort than lateral motion on vertical and horizontal axes in stereoscopic 3D displays [1,2,3]. However, previous studies did not consider other factors like viewing time and display size in evaluating visual discomfort. The main contribution of this paper is two-fold: (1) We analyze the effects of motion-in- depth, viewing time and display size in measuring visual discomfort. (2) The evaluation method for visual discomfort is proposed by integrating a subjective test such as a questionnaire, and an objective test such as eye blink rate detection.

The design of the experimental environment was in line of the recommendations of ITU-R BT.500-13 [4]. The experimental setup is shown in Figure 2 with the following specifications:

- Size: 55inch (passive type), 27inch (passive type)
- Aspect ratio: 16 : 9
- Spatial Resolution: 1920 * 1080
- Environmental luminance on the screen: 200 lux
- Participants: 20 subjects (14 males and 6 females, ages 20~35: medical condition checked)

Lighting conditions were held constant for all participants during all sessions. Any external illumination was completely blocked out by thick curtains. The temperature and humidity were maintained constantly and there were no vibrations or strong odors.

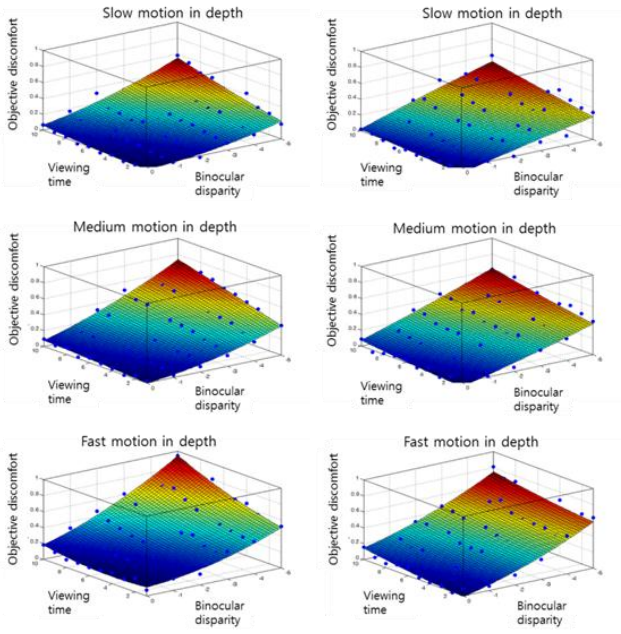
(a) (b)

Figure 1: Experimental setup: (a) 27 inch 3D TV ; (b) 55 inch 3D TV.

The subject closed their eyes and rested for 5 minutes. This stage was intended to eliminate eyestrain resulting from the subject's previous activities, and to achieve a normalized baseline for the experiment's diverse subjects. Then, the following eight questions were answered in a period of 2 minutes to check the subject's pre-stimulus subjective eyestrain. Next, the participant watched the 3-, 5- and 10-minute stereoscopic 3D video clips as shown in Table 2. While the subject was wearing polarized glasses equipped with an eye tracking device, we detected her eye blinking using eye blinking method as in [5] and measured her eyestrain response at one minute intervals with a hand held slider similar to [6]. The position of the slider could be adjusted along a graphical scale and including at regular intervals the adjective terms, [extremely uncomfortable]-[uncomfortable]-[middle]-[comfortable]-[very comfortable], in accordance with the ITU recommendation [4]. After watching stereoscopic 3D video, the subject re-answered the previously mentioned eight questions in a span of 2 minutes to measure the post-stimulus subjective eyestrain. The survey scores, representing the amount of subjective discomfort, were normalized between 0 and 1 after subtracting the pre-stimulus score from the corresponding post-stimulus score.

Visual discomfort was measured in respect to three kinds of motion-in-depth (slow, medium and fast motion), viewing time and display size. To begin with, we present our subjective assessment results based on the participants' questionnaires. An analysis of variance (ANOVA) was performed on the individual ratings of discomfort obtained through the questionnaires for each size of display. Note that we only use the median 50% of rating scores for analysis, while the upper and lower 25% of rating scores were removed as statistical outliers. As a result, only 6 questions were statistically significant across both sizes of display with a 95% significance level. Thus, we set discomfort value as an average of Q1, Q3, Q4, Q5, Q6, and Q7 rating scores. To model the subjective visual discomfort, we fit a two-dimensional function of binocular disparity and viewing time to the data for each motion in depth. To convert eye blinking rates into objective visual discomfort, we correlate the eye blinking rate with viewers' visual discomfort responses. The relationship between eye blinking rates and visual discomfort is modeled using the polynomial function.

By composite eye blinking rates and viewers' discomfort responses, we construct an objective visual discomfort model. For each size of display, observed eye blink corresponds to visual discomfort value by function h. From calculated visual discomfory values, we use the polynomial function to obtain objective visual discomfort model. Regression analysis was performed to find the optimal value of the coefficients. This is shown in Figure 2. Visual discomfort increases rapidly as the degree of binocular disparity increases in case of fast motion-in-depth. In addition, the smaller-sized display results in more eyestrain than the larger-sized display.



(a) (b)

Figure 2: Objective visual discomfort : (a) 27 inch; (b) 55 inch

[1] S. Yano, S. Ide, T. Mitsuhashi, and H. Thwaites, "A study of visual fatigue and visual comfort for 3D HDTV/HDTV images," *Displays*, vol. 23, pp. 191–201, 2002.

[2] F. Speranza, W. J. Tam, R. Renaud, and N. Hur, "Effect of disparity and motion on visual comfort of stereoscopic images," in Proc. of SPIE, vol 6055, pp. 94-103, 2006.

[3] Y.J. Jung, S. Lee, H. Sohn, H. W. Park and Y. M. Ro, "Visual comfort assessment metric based on salient object motion information in stereoscopic video," *Journal of Electron Imaging*, vol. 21, Issue 1, Feb, 2012.

[4] ITU, Methodology for the subjective assessment of the quality of television pictures, Recommendation BT.500-13 2010.

[5] S.-H, Cho and H.-B. Kang, "The measurement of eyestrain caused from diverse binocular disparities, viewing time and display sizes in watching stereoscopic 3D content," In *Proc CVPRW*, pp.23-28, June 2012.

[6] M. Lambooij, W.A. IJsselsteijn, I. Heynderickx, "Visual discomfort of 3D TV: Assessment methods and modeling," *Displays*, vol. 32, Issue 4, Oct. 2011.

MCMC Supervision for People Reidentification in Nonoverlapping Cameras

Boris Meden[1]
boris.meden@cea.fr
Frédéric Lerasle[2]
lerasle@laas.fr
Patrick Sayd[1]
patrick.sayd@cea.fr

[1] CEA, LIST,
Laboratoire Vision et Ingénierie des Contenus,
BP 94, F-91191 Gif-sur-Yvette, France

[2] CNRS ; LAAS ;
Université de Toulouse ; UPS, LAAS ;
F-31077 Toulouse Cedex 4, France

We present a pedestrian tracking system that uses re-identification to monitor non-overlapping cameras. As tracking, re-identification is an assignment problem, the difficulties being to generate an accurate representation and to prune unlikely pairings. The assignments are realised in two stages.

First, a Markovian multi-target tracking-by-detection framework which includes identification in the search space is run in the cameras, following the approach of [1]. Inspired by [2], the appearance model used is composed of horizontal stripes of HSV histograms weighted by their distances to the symmetry axis. The use of topology allows to instantiate new identities from the feeding areas in an identity database, which we compare with to perform re-identification. The mixed-state formalism [3] uses that database and samples in this identity space. That way the tracker produce a tracklet and re-identification probabilities in the database representing the belief of the tracker. The resulting tracklets are sent to the supervisor along with their probabilities of identity, their time of existence and their areas.

At the network level, the supervisor resorts to deferred logic to optimize the assignment between the received tracklets using re-identification distributions and network topology information. The combinatorial space is efficiently explored through MCMC sampling. Tracks output by the supervisor are optimized to represent the activity of the same person, as shown in figure (1).

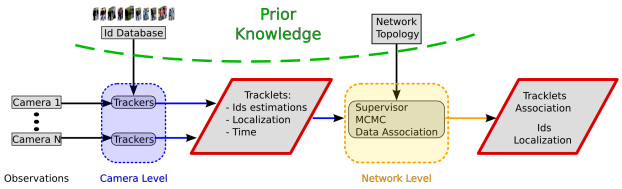


Figure 1: Synoptic diagram of the combination between markovian tracking-by-reidentification and MCMC data association at the network level.

Camera Level: Mixed-state Tracking by Reidentification The weight $w_{tr}^{(p)}$ associated with the p-th particle of tracker tr is computed integrating the distance to the associated detection d^*, the colorimetric similarity to the appearance model $w_{App}(.)$ and the colorimetric similarity to the identity of the particle $w_{Id}(.)$. $Id(p)$ represents the identity taken by particle p. This is the discrete parameter of p.

$$w_{tr}^{(p)} = \underbrace{\alpha \cdot \mathcal{I}(tr) \cdot p_{\mathcal{N}}(d^* - p)}_{\text{distance to the detection}} + \underbrace{\beta \cdot w_{App}(d,tr)}_{\text{appearance model}} + \underbrace{\gamma \cdot w_{Id}(d, id(p))}_{\text{identity}} \quad (1)$$

where α, β and γ are weighting coefficients empirically set, and $\mathcal{I}(tr)$ is a boolean signifying the existence or not of an associated detection to the tracker.

The state estimation is a two-stage process. First we compute the Maximum A Posteriori over the discrete parameter relatively to the current observation $\mathbf{Z}_t$ with equation (2), *i.e.* the most likely identity at time step t.

$$\hat{id}_t = \arg\max_j P(id_t = j|\mathbf{Z}_t) = \arg\max_j \sum_{p \in \Upsilon_j} w_{tr}^{(p)}(t), \quad (2)$$

$$\text{where } \Upsilon_j = \left\{ p | \mathbf{X}_t^{(p)} = (\mathbf{x}_t^{(p)}, j) \right\}$$

Then, the continuous components are estimated over the subset of particles $\hat{\Upsilon}$ which have that most likely identity, following equation (3).

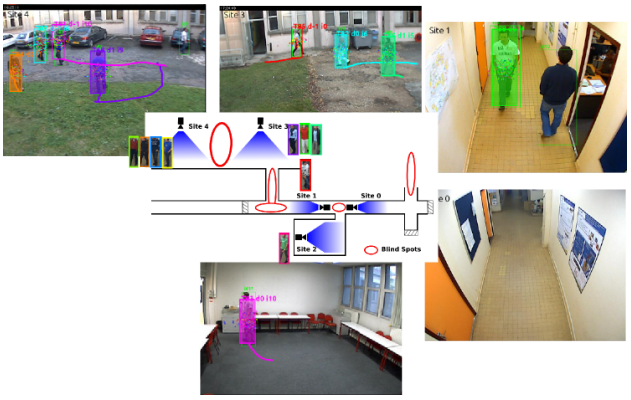


Figure 2: Overview of the monitored network.

$$\hat{\mathbf{x}}_t = \sum_{p \in \hat{\Upsilon}} w_{tr}^{(p)}(t) \cdot \mathbf{x}_t^{(p)} / \sum_{p \in \hat{\Upsilon}} w_{tr}^{(p)}(t), \quad (3)$$

$$\text{where } \hat{\Upsilon} = \{ p | \mathbf{X}_t^{(p)} = (\mathbf{x}_t^{(p)}, \hat{y}_t)^\mathsf{T} \}$$

That way, on top of target image position estimation, each filter provides a discrete identity distribution for its target.

Network Level: Topologic and Appearance Driven MCMC Optimisation The likelihood a collection of tracklets τ_i have to be associated to an identity id mixes topology and identity distributions: $p(Y|H) = \mathcal{P}_{Topo}(Y|H) \cdot \mathcal{P}_{MSR}(Y|H)$, with:

$$\mathcal{P}_{Topo}(Y|H) = \prod_{i=1}^{|\tau_n|-1} p_{\mathcal{N}}(d_{topo}(a_{i-1}^{out}, a_i^{in})), \quad (4)$$

where $d_{topo}(.)$ is the distance between two nodes of the topological graph, $a_i^{in/out}$ are the area of beginning (*resp.* ending) of the i-th tracklet, $p_{\mathcal{N}}(.)$ is a gaussian kernel to transform the distance into a similarity between 0 and 1 and $|\tau_n|$ is the cardinal of the tracklet set τ_n.

As comparing directly descriptors taken from different cameras yields an homogeneity problem, we use instead the mixed-state trackers belief on the tracklet identity, resulting from online comparison with the database:

$$\mathcal{P}_{MSR}(Y|H) = \prod_{i=1}^{|\tau_n|-1} ids_i(id), \quad (5)$$

where ids_i is the discrete probability distribution over the identity database for the i-th tracklet. That way $ids_i(id)$ represents the probability that tracklet i has the identity id.

We use MCMC to optimise the tracklet-to-identities assignment $H = \{\tau_i\}_{i=1...N}$ The tracking results obtained on a large ground-truthed dataset demonstrate the effectiveness of the approach. Figure 2 provides an example of the method running on 5 cameras.

[1] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. *PAMI*, 2010.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

[3] B. Meden, P. Sayd, and F. Lerasle. Mixed-State Particle Filtering for Simultaneous Tracking and Re-Identification in Non-Overlapping Camera Networks. In *SCIA*, 2011.

Efficient Exemplar Word Spotting

Jon Almazán
www.cvc.uab.es/~almazan
Albert Gordo
agordo@cvc.uab.es
Alicia Fornés
afornes@cvc.uab.es
Ernest Valveny
ernest@cvc.uab.es

Computer Vision Center
Departament de Ciències de la Computació
Universitat Autònoma de Barcelona
Barcelona, Spain

This work addresses the problem of word spotting: given a query word image, the goal is to retrieve locations in a set of document images where this word may be present. Traditionally, word spotting systems have followed a well defined flow. First, an initial layout analysis is performed to segment word candidates. Then, the extracted candidates are represented as sequences of features, and, by using a similarity measure – commonly a Dynamic Time Warping (DTW) or a Hidden Markov Model (HMM)-based similarity –, the query word is compared and candidates are ranked. An example of this framework is the work of Rath and Manmatha [4].

One of the main drawbacks of these systems is that they need to perform a costly and error prone segmentation step to select candidate windows. Any error introduced in this step will negatively affect the following stages, and so it is desirable to avoid segmenting the image whenever possible. Unfortunately, since the comparison of window regions, represented by sequences, is based on costly approaches such as a DTW or a HMM, it is not feasible to perform this comparison exhaustively with a sliding window approach over the whole document image. Another important drawback is that best performance techniques – such as HMMs or Neural Networks – need a large amount of annotated training images. However, this is not a common case in real scenarios.

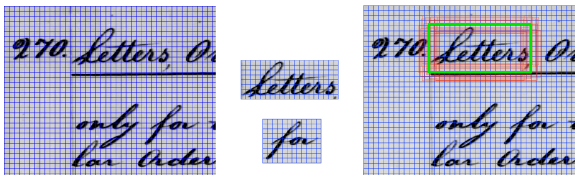
The recent [5] addresses the segmentation problem by representing regions with a fixed-length descriptor based on the bag of visual words framework. To further improve the system, unlabeled training data is used to learn a LSI space, where the distance between words is more meaningful than in the original space. We follow [5] and address the word spotting problem in an unsupervised, segmentation-free setting, and argue that the current methods can be improved in several ways.

First, they can be improved in the choice of low level features. We address these issues by using HOG descriptors, which have been shown to obtain excellent results when dealing with large variations in the difficult tasks of object detection and image retrieval. In our baseline system, the document images are divided in equal-sized cells (see Fig 1a) and represented with HOG histograms. Queries are represented analogously using cells of the same size in pixels (Fig 1(b)). The score of a document region is computed using the cosine similarity, *i.e.*, calculating the dot-product between the L2 normalized descriptors. Following this approach, we can compute the similarity of all the regions in the document image with respect to the query using a sliding window and rank the results.

Second, spotting methods can be improved in the learning of a more semantic space. We propose to perform this unsupervised learning once the query has been observed, and adapt the learning to the query. For this task, we propose to use a similar approach to the Exemplar SVM framework of [3]. We need first a set $\mathcal{P}$ of relevant regions to the query. This set is constructed by slightly shifting the window around the query word to produce many almost identical, shifted positive samples (see Fig 1(c)). Then we need a set of non-relevant regions. To produce this negative set $\mathcal{N}$, we sample random regions over all the documents. Given these sets, we can solve the following optimization problem

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{\mathbf{y}_p \in \mathcal{P}} L(\mathbf{w}^T \mathbf{y}_p) + C \sum_{\mathbf{y}_n \in \mathcal{N}} L(-\mathbf{w}^T \mathbf{y}_n), \qquad (1)$$

where $L(\mathbf{x}) = \max(0, 1-\mathbf{x})$ is the hinge loss and C is the cost parameter. Solving this optimization produces a weight vector $\mathbf{w}$, which can be seen as a new representation of the query. This new representation has been directly optimized to give a high positive score to relevant regions, and a high negative score to non-relevant regions when using the dot-product with L2 normalized regions.



(a) HOG grid. (b) (c)

Figure 1: a) Grid of HOG cells. b) Two random queries. The windows adjust around the HOG cells. c) A query (in green) and some positive samples (in red) used to learn the Exemplar SVM.

Finally, these methods can be improved in the cost of storing the descriptors of all the possible windows of all the dataset items. Assuming HOG descriptors of 31 dimensions represented with single-precision floats of 4 bytes each, and 50,000 cells per image, storing as few as 1,000 precomputed dataset images would require 5.8GB of RAM. Since documents will not fit in RAM memory when dealing with large collections, it would produce a huge performance drop in the speed at query time. To address this problem, we propose to encode the HOG descriptors using Product Quantization (PQ) [2]. Encoding the descriptors with PQ would allow us to preserve a much larger amount of images in RAM at the same time. As a side effect, computing the scores of the sliding window also becomes significantly faster. In our case, we can reduce the size of the HOG descriptor to one byte with minimal loss, and achieve a 10-fold improvement in computational time.

We evaluate our approach on two public datasets: The George Washington (GW) and the Lord Byron (LB). Figure 2 shows the mean Average Precision of our vanilla method and its improvements – EWS and EWS+PQ – on the GW dataset as a function of the HOG cell size and the sliding window step size.

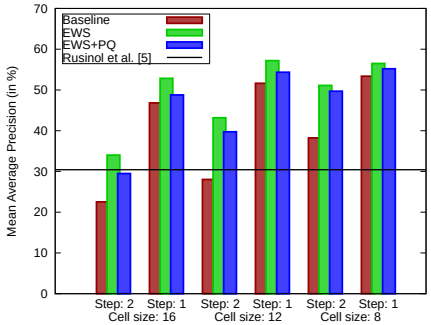


Figure 2: Mean Average Precision of our method for different setups.

We publish the MATLAB code implementation for training and testing the Exemplar Word Spotting on the web page [1].

[1] J. Almazán, A. Gordo, A. Fornés, and E. Valveny. Exemplar Word Spotting library. URL `http://almazan.github.com/ews/`.

[2] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 2011.

[3] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of Exemplar-SVMs for object detection and beyond. In *ICCV*, 2011.

[4] T. Rath and R. Manmatha. Word spotting for historical documents. *IJDAR*, 2007.

[5] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós. Browsing heterogeneous document collections by a segmentation-free word spotting method. In *ICDAR*, 2011.

Transductive Kernel Map Learning and its Application to Image Annotation

Dinh-Phong Vo
vo@telecom-paristech.fr
Hichem Sahbi
sahbi@telecom-paristech.fr

LTCI CNRS Telecom ParisTech
46 rue Barrault, 75013, Paris, France

We introduce in this paper a novel image annotation approach based on maximum margin classification and a new class of kernels. The method goes beyond the naive use of existing kernels and their restricted combinations in order to design "model-free" transductive kernels applicable to interconnected image databases.

Let $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_l, \ldots, \mathbf{x}_m\}$ denote an image database described in an n-dimensional input space. We assume that only the first l ($l \ll m$) vectors of $\mathcal{S}$ are labeled (a.k.a annotated), i.e., $\{\mathbf{y}_1, \ldots, \mathbf{y}_l\}$ are known; here $\mathbf{y}_i \in \{-1, +1\}^r$ and r is the number of possible labels used for annotation.

Our approach considers image annotation as a multi-label classification problem in which a sample $\mathbf{x}_i$ may have more than one label, i.e., $r > 1$, with $\mathbf{y}_{ik} = +1$ iff $\mathbf{x}_i$ has the k^{th} label and $\mathbf{y}_{ik} = -1$ otherwise. Our objective is to build an optimal *kernel map* and a decision criterion f in order to infer the unknown label vectors $\{\mathbf{y}_{l+1}, \ldots, \mathbf{y}_m\}$.

We adopt the max-margin classification [4] approach in order to learn a classifier $f(\mathbf{x}_i) = \mathbf{W}'\phi(\mathbf{x}_i)$ that balances training error and model complexity. This classifier corresponds to

$$\underset{f}{\operatorname{argmin}}\ \mathcal{R}(f) + \gamma_c \sum_{i=1}^{l} \ell\left(f(\mathbf{x}_i), \mathbf{y}_i\right), \tag{1}$$

where $\mathcal{R}$ is a regularizer, $\ell(f(\mathbf{x}_i), \mathbf{y}_i)$ is the loss associated with a prediction $f(\mathbf{x}_i)$ when the true output is $\mathbf{y}_i$ and $\gamma_c > 0$ balances these two terms. For nonlinear classification, ϕ maps the input data (in $\mathcal{S}$) into a high dimensional space $\mathcal{H}$ such that $\mathbf{W}$ can separate labeled data $\{\mathbf{x}_i\}_{i=1}^{l}$.

Following the kernel trick [3], the function f may also be expressed as a linear combination of symmetric, continuous and positive (semi) definite kernel functions. A kernel (denoted κ) is defined on two samples $\mathbf{x}_i, \mathbf{x}_j$ as $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$. The closed form of $\kappa(\cdot, \cdot)$ may also be defined among a collection of existing kernels including linear, polynomial and histogram intersection; but the underlying mapping $\phi(.) \in \mathcal{H}$ is usually *implicit*, i.e., it does exist but it is not necessarily known and may be infinite dimensional.

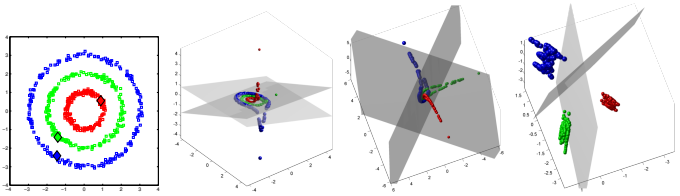
Our proposed method, in contrast to usual kernel methods, finds an *explicit* and finite dimensional kernel map. According to Vapnik's VC-theory [4], a finite dimensional kernel map, with a bounded related VC-dimension, avoids loose generalization bounds and may guarantee better performance.

Our goal is to find hyperplane parameters $\mathbf{W}$ as well as a Gram (kernel) matrix $\mathbf{K} = \Phi'\Phi$ where each column Φ_i corresponds to an explicit mapping of $\mathbf{x}_i$ into a high dimensional space (i.e., $\phi(\mathbf{x}_i) = \Phi_i$). The learned mapping Φ must i) guarantee linear separability of data in $\mathcal{S}$, ii) ensure good generalization performance by maximizing the margin, iii) approximate the input data, and also iv) ensure positive definiteness of $\mathbf{K}$ by construction, i.e., without adding further constraints. Considering $\mathcal{R}(f) = \|\mathbf{W}\|_F^2$ and $\ell(f(\mathbf{x}_i), \mathbf{y}_i) = \|f(\mathbf{x}_i) - \mathbf{y}_i\|_2^2$, the map Φ and the classifier parameters $\mathbf{W}$ are found by solving

$$\begin{array}{ll} \underset{\mathbf{B},\Phi,\mathbf{W}}{\min} & \frac{\mu}{2}\|\Phi\|_F^2 + \frac{1}{2}\|\mathbf{W}\|_F^2 + \frac{\gamma_c}{2}\left\| \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{B} & \mathbf{0}_{n\times p} \\ \mathbf{0}_{r\times p} & \mathbf{W}' \end{bmatrix} \begin{bmatrix} \Phi \\ \Phi\mathbf{C} \end{bmatrix} \right\|_F^2 \\ \text{s.t} & \|\mathbf{B}_i\|_2^2 = 1, \forall i = 1, \ldots, p \end{array} \tag{2}$$

here $\mathbf{C} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with $\mathbf{C}_{ii} = 1_{\{1\le i\le l\}}$, $\mathbf{0}_{n\times p}$ and $\mathbf{0}_{r\times p}$ are $n \times p$ and $r \times p$ zeros matrices respectively, $\mathbf{X} \approx \mathbf{B}\Phi$ is factorized using an overcomplete basis $\mathbf{B} \in \mathbb{R}^{n\times p}$ (i.e., $p > n$) and a new kernel map $\Phi \in \mathbb{R}^{p\times m}$.

According to [4], the VC-dimension (related to a family of classifiers) depends also on the dimension of the learned kernel map and this may affect generalization, especially if this dimension is very high. Since the actual (intrinsic) dimension of the learned kernel map Φ is unknown, we choose the number of basis p to be sufficiently large such that the factorization term (in right-hand side of Eq. 2) tends to zero for an infinite number of solutions. Then, the actual (intrinsic) dimension is found



(a) input data (b) $\gamma_s = 10, \gamma_c = 1$ (c) $\gamma_s = 10^2, \gamma_c = 1$ (d) $\gamma_s = 10^3, \gamma_c = 10^2$

Figure 1: (a) This figure shows the input data where different colors stand for different classes; red-colored data are annotated with $(1\ \ -1)'$, blue-colored data with $(-1\ \ 1)'$ and green-colored data with $(1\ \ 1)'$. Note that just one training sample per class (diamond-shaped) is labeled while others are unlabeled. Figures (b,c,d) are the learned kernel maps (shown in 3d) and the obtained decision hyperplanes for different setting of the parameters γ_c and γ_s.

by regularizing Eq. 2 using the Frobenius norm $\|\Phi\|_F^2$ which has similar effect as the nuclear norm where $\mu \ge 0$ controls the rank of $\mathbf{K}$.

For a better conditioning of Eq. 2, we adopt transductive inference [1, 5] which assumes that close data in a high-density area of the input space should have similar labels [2]. This assumption, therefore, enables label diffusion from training to the test data (see toy example in Fig. 1).

By representing $\mathbf{x}_i$'s as vertices $\{v_i\}$ and their pairwise similarities as edges $\{e_{ij}\}$, the smoothness assumption between v_i and v_j is modeled by the differences between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$, i.e.,

$$\frac{1}{4}\sum_{i=1,j=1}^{m} \|\mathbf{W}'\Phi_i - \mathbf{W}'\Phi_j\|^2 \mathbf{A}_{ij} \Longleftrightarrow \frac{1}{2}\mathrm{tr}\left(\mathbf{W}'\Phi\mathbf{L}\Phi'\mathbf{W}\right), \tag{3}$$

where the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{A}$ is defined by the affinity matrix $\mathbf{A}$ whose elements $\mathbf{A}_{ij} = 1_{\{v_j \in \mathcal{N}_k(v_i)\}} \cdot s(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{D} = \mathrm{diag}(\mathbf{A1})$ with $\mathbf{1}$ being the all-one vector of length m. Here $s(\cdot, \cdot)$ is a visual similarity and $\mathcal{N}_k(v_i)$ is the set of the k-nearest neighbors of v_i.

Now, we obtain the complete form of our transductive learning problem as

$$\begin{array}{ll} \underset{\mathbf{B},\Phi,\mathbf{W}}{\min} & \frac{\mu}{2}\|\Phi\|_F^2 + \frac{1}{2}\mathrm{tr}\left(\mathbf{W}'\left(\mathbf{I}_p + \gamma_s\Phi\mathbf{L}\Phi'\right)\mathbf{W}\right) + \\ & \quad + \frac{\gamma_c}{2}\left\| \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{B} & \mathbf{0}_{n\times p} \\ \mathbf{0}_{r\times p} & \mathbf{W}' \end{bmatrix} \begin{bmatrix} \Phi \\ \Phi\mathbf{C} \end{bmatrix} \right\|_F^2, \\ \text{s.t} & \|\mathbf{B}_i\|_2^2 = 1, \forall i = 1, \ldots, p \end{array} \tag{4}$$

with $\mathbf{I}_p$ the $p \times p$ identity matrix and again $\mathbf{C}$ is the diagonal $m \times m$ matrix for which the i^{th} diagonal element is fixed to 1 for a labeled sample, and 0 for an unlabeled one.

Solving this minimization problem makes it possible to learn both a decision criterion and a kernel map that guarantee linear separability in a high dimensional space and good generalization performance (see Fig. 1). Experiments conducted on image annotation, show that, indeed, our obtained kernel achieves at least comparable results with related state of the art methods on the MSRC and the Corel5k databases.

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.*, 7:2399–2434, December 2006.

[2] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[3] B. Schölkopf and AJ. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, December 2001.

[4] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[5] V. Vapnik and A. Sterin. On structural risk minimization or overall risk in a problem of pattern recognition. *Automation and Remote Control*, 10(3):1495–1503, 1977.

Multi-camera Pedestrian Detection with a Multi-view Bayesian Network Model

Peixi Peng
pxpeng@jdl.ac.cn
Engineering Laboratory for Video Technology, Peking University

Yonghong Tian
yhtian@pku.edu.cn
Engineering Laboratory for Video Technology, Peking University

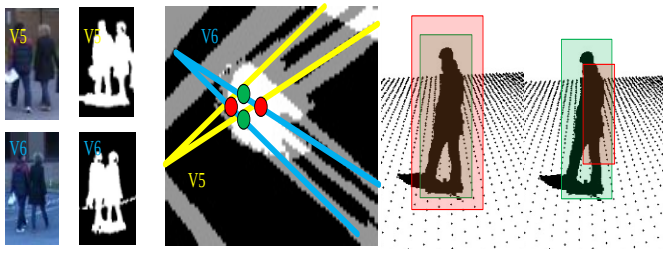
Yaowei Wang
ywwang@jdl.ac.cn
Department of Electronic Engineering, Beijing Institute of Technology

Tiejun Huang
tjhuang@pku.edu.cn
Engineering Laboratory for Video Technology, Peking University

In recent years, more and more cameras are widely deployed for video surveillance in a cooperative manner. In such scenarios, multiple-pedestrian detection has become an essential technology for many applications such as crowd behaviour analysis. Often, occlusions among pedestrians will complicate the detection process and make it difficult for the system to accurately detect the pedestrians after heavy occlusion. In this sense, the availability of multi-view information will make pedestrian detection easier and more accurate..

In this paper, we estimate the occupancy possibility of each location that can then be used to predict the occurrence of a pedestrian in this location. We integrate occupancy possibility of all views together as the final occupancy possibility on the ground plane. The general method [1], intersecting the view lines from multi-cameras on the ground plane, yield satisfying results when people are well-separated in multi-views. When occlusion becomes more frequently, these approaches will cause many "phantom" phenomena. Phantoms are the intersections of viewing rays at locations that are not occupied by any pedestrians (as shown in Figure 1(a)), which has also been reported in previous work [2].

To address this problem, we first classify the phantoms in a single view into two categories. The first-class phantoms are those who occlude some pedestrians (e.g., the right one in Figure 1(b)). Often, the phantoms of this kind are generated due to the projection of inaccurate foreground extraction results on the ground plane. In this case, if these phantoms are directly treated as detection results, the matching degree with the foreground masks should be much less than the pedestrians which are occluded by them. In order to reduce the first-class phantoms in the multi-view projection, the key point is to make the detection results best match the foreground masks using the occlusion relationship among phantoms and pedestrians. On the other hand, the second-class phantoms denote those that are occluded by pedestrians, despite they can also match the foreground masks well (e.g., the left one in Figure 1(b)). The reason for generating the phantoms of this kind is usually due to the non-invertible mapping from 3D world coordinates to 2D image coordinates. These phantoms always be occluded by pedestrians mostly. Thus to reduce the second-class phantoms, we need to estimate the non-occluded parts for each phantom. Hence, our method proposed in this paper try to reduce phantoms by analysing the occlusion relationship among potential pedestrians at different locations in all views.



. Figure 1: (a) An example of phantom: locations (red circles) are occupied by phantoms and locations (green circles) are occupied by pedestrians; (b) The first-class (left) and second-class (right) phantoms, where the red ones denote phantoms

By summarizing the two cases above, we can conclude that the key problem to reduce the possible phantoms in the multi-view projection is to effectively model and utilize the occlusion relationship among potential pedestrians at different locations in all views. It is notable that a 1st phantom in one view may be the 2nd phantom in another view and vice versa. Considering it, a multi-view Bayesian network (MBN) is proposed in this paper. In general, a MBN is constructed with the locations on the ground plane and several single Bayesian networks (SBNs), where each SBN is used to characterize the potential occlusion relationship of all locations in a single view, while the locations on the ground plane is used to establish the correspondence among all SBNs through the geometric constraints among cameras (See Figure 2). Moreover, we also model the "subjective supposing" node states (SSNS) as a set of Boolean parameters of MBN, which are then used to denote whether a pedestrian occurs at the locations. In fact, SSNS can distinguish the 1st phantoms with the 2nd pedestrians. During MBN inference, we can estimate the part of the pedestrian at this location which are not occluded by other pedestrians based on SSNS. A learning algorithm is then proposed to estimate the SSNS parameters of the MBN, by finding such a configuration that the final occupancy possibility can best explain the image observations (i.e., foreground masks) from different views. The overall framework of our method is shown in Figure 2.

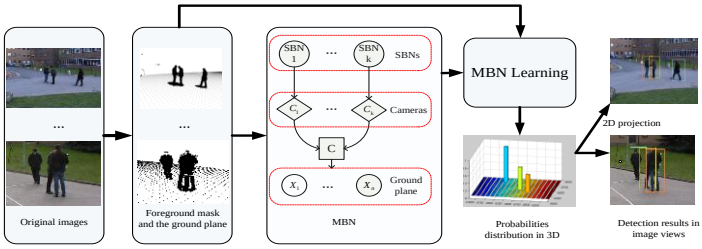


Figure 2:The framework of our approach

Implementation details of MBN model and SSNS learning is described in the paper. The experimental results on PETS2009S2L1 and APIDIS benchmark datasets demonstrate the effectiveness of our method compared with other state-of-the-art methods [2] [3][4].

References

[1] A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa. Object detection, tracking and recognition for multiple smart cameras. *Proc. of the IEEE*, 96(10):1606–1624, 2008.

[2] W. Ge and R. T. Collins. Crowd detection with a multiview sampler. In *Proc. of ECCV*, pages 324–337, 2010

[3] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):267–282, 2008.

[4] Alahi A., Jacques L., Boursier Y., and Vandergheynst P., "Sparsity Driven People Localization with a Heterogeneous Network of Cameras,"*Journal of Mathematical Imaging and Vision*, vol.41,pp.39-58, 2009.

Online Feedback for Structure-from-Motion Image Acquisition

Christof Hoppe[1]
hoppe@icg.tugraz.at
Manfred Klopschitz, Markus Rumpler,
Andreas Wendel, Horst Bischof, Gerhard Reitmayr [1]
{klopschitz,rumpler,wendel,bischof,reitmayr}@icg.tugraz.at
Stefan Kluckner[2]
stefan.kluckner@siemens.com

[1] Institute for Computer Vision and Graphics
Graz University of Technology
Graz, Austria
[2] Research Group Video Analytics
Corporate Technology
Siemens AG Austria, Graz

The quality and completeness of 3D models obtained by Structure-from-Motion (SfM) heavily depend on the image acquisition process. If the user gets feedback about the reconstruction quality already during the acquisition, he can optimize this process. The goal of this paper is to support a user during image acquisition by giving online feedback of the current reconstruction quality. We propose an online SfM method that integrates wide-baseline still-images in an online fashion into a consistent reconstruction and we derive a surface model given the SfM point cloud. To guide the user to scene parts that are captured not very well, we colour the mesh according to redundancy and resolution information. In the experiments, we show that our approach makes the final SfM result predictable already during image acquisition. The method is suited for large-scale reconstructions as obtained by flying micro aerial vehicles as well as on small indoor environments.

We propose a method that supports a user in the acquisition process in two ways: (a) sparse online SfM with accuracy close to offline methods and (b) surface extraction and quality visualization. The workflow of our method is shown in Figure 1.

1 Online SfM for Wide-Baseline Still-Images

To speedup the SfM process to work in realtime on wide-baseline images, we weaken the assumption of most batch-based SfM pipelines that images are captured in random order. We assume that a freshly acquired input image I has an overlap to an already reconstructed scene part. This allows us to split the SfM problem in two tasks that are easier to solve: A localization and a structure expansion part. Hence, we can first localize I within the reconstructed scene according to the method proposed by Irschara et. al. [1]. We compute visual similarity scores to all reconstructed images and perform feature matching between I and the top n scored images. This results in 2D-3D correspondences between I and the reconstructed point cloud. We then localize I by solving the 3-point pose problem in a RANSAC loop. Finally, we expand the map by triangulating new 3D points. To avoid scene drift, we optimize the reconstructed scene by bundle adjustment in a parallel thread. Online SfM allows us to provide feedback within less than 2 seconds if the new acquired image has been integrated into the reconstruction (see Figure 2), which reduces the number of acquired images that are not suited for SfM. Furthermore, we show in the experiments that the accuracy is close to that obtained by offline methods like Bundler [2].

2 Surface Extraction and Quality Visualization

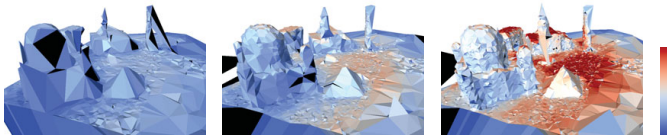
Because it is difficult to estimate the reconstruction quality on a point cloud, we derive a surface model given the sparse points obtained by SfM. The idea is to generate a 3D triangulation of all sparse points that embeds the real surface. To extract the subset of triangles which are on the object's surface, an energy functional is defined and minimized by graph cuts. This method extracts a surface even from a very low number of 3D points as shown in Figure 3. Since the number of 3D points is relatively low, this can be computed within seconds.

The quality of a final (dense) reconstruction mainly depends on two parameters: (a) Redundancy and (b) Ground Sampling Distance (GSD). Since the camera positions and the surface mesh are available, we compute both values and visualize them on the surface model by colouring. The user interactively selects which data is visualized to decide for a new camera position. This supports the user to obtain an equally distributed scene sampling. An example is shown in Figure 4.

Our interactive method makes the image acquisition for SfM more efficient and allows the user to inspect the final reconstruction already on site. It opens SfM for applications where a certain completeness and accuracy of the reconstruction has to be guaranteed.

(a)

Figure 2: Online feedback of image integration. The result of the image integration is provided within less than 2 seconds to the user. Red bordered images could not be aligned into the reconstruction. This allows the user to adopt the image acquisition.



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

Figure 3: (a) - (c) The resulting mesh of the City-of-Sights after 10, 20 and 50 reconstructed images. (d) Colormap. Blue indicates that a low number of cameras observe a triangle. Red indicates that a triangle is seen more than 30 times.

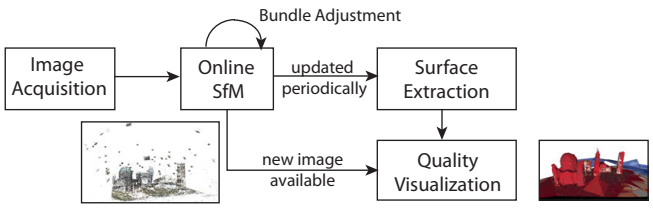
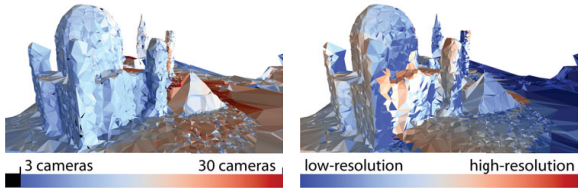


Figure 1: Workflow. Still-images are acquired by the user and integrated into the reconstruction. Periodically, we extract a surface mesh and visualize quality information.



(a) (b)

Figure 4: (a) Surface mesh extracted from a sparse point cloud with overlaid redundancy information. (b) Coloring according to maximum GSD. Best viewed in color.

[1] A. Irschara, C. Zach, J. M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009.

[2] N. Snavely, S. M. Seitz, and R. S. Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, November 2008.

Stixmentation - Probabilistic Stixel based Traffic Scene Labeling

Friedrich Erbs
friedrich.erbs@daimler.com
Beate Schwarz
beate.schwarz@daimler.com
Uwe Franke
uwe.franke@daimler.com

Image Understanding
Daimler AG
Boeblingen, Germany

The detection and segmentation of moving objects like vehicles, pedestrians or bicycles from a mobile platform is one of the most challenging and most important tasks for driver assistance and safety systems. For this purpose, we present a multi-class traffic scene segmentation approach based on the Dynamic Stixel World, an efficient super-pixel object representation for traffic scenes.

Related works often performs pixel-wise segmentation and is mainly focused on static scenes. Using the Stixel World creates dramatic advantages in comparison with such traditional approaches. The relevant information in the scene is represented with a few hundreds Stixels instead of hundreds of thousands of individual dense stereo depth and optical flow measurements. This compression of the input data volume also reduces the computational burden for a subsequent segmentation step by at least three orders of magnitude, thus enabling real-time capability. Besides that, the Stixel World turns out to be extremely stable with respect to outliers due to a global optimization used in its calculation. Subsequent algorithms profit strongly from this high reliability of data. Taking into account motion information creates the possibility to discriminate between different objects which cannot be separated based on depth or appearance information alone.

This work presents a probabilistic conditional random field framework for segmenting moving objects into different motion classes. The main steps of our segmentation process are summarized in Figure 1. It starts from dense stereo depth maps obtained by the Semi-Global Matching (SGM) stereo algorithm [1] as shown in Figure 1(a). Then, the multi-layered Stixel World [3] and the Dynamic Stixel World which is extended by motion information [2] (Figure 1(b)) are computed. The final segmentation result depicted in Figure 1(c) separates the image into different motion classes. These include oncoming, forward-moving, right-moving and static background (shown in yellow, magenta, cyan and black respectively).

For the segmentation, we seek the most probable labeling minimizes the following log-likelihood energy E

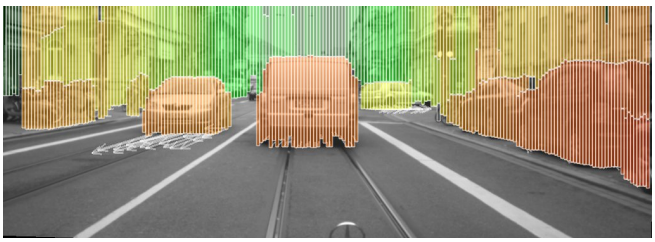
$$\begin{aligned} E &= -\log p\left(L^t \mid \mathcal{Z}^t, L^{t-1}\right) \\ &\sim \sum_{i=1}^{N} \psi\left(l_i^t \mid \mathcal{Z}^t, L^{t-1}\right) + \lambda \cdot \sum_{(i,j)\in\mathcal{N}_2} \phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right). \end{aligned} \quad (1)$$

In this context, $L^t = \{l_1^t, \ldots, l_N^t\}^{\mathrm{T}}$ denotes a labeling for a given input image I^t containing N dynamic Stixels and the observations for all Stixels are combined in a measurement array $\mathcal{Z}^t = \{\vec{z}_1^t, \ldots, \vec{z}_N^t\}$. $\mathcal{N}_2$ denotes the set of all neighboring Stixels and the term λ is a scaling factor for the binary term $\phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right)$. The unary terms are modeled

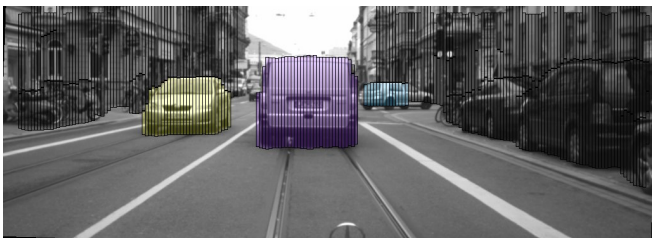
$$\begin{aligned} \psi\left(l_i^t \mid \mathcal{Z}^t, L^{t-1}\right) &= -\log p\left(l_i^t \mid \mathcal{Z}^t, l_i^{t-1}\right), \text{ where} \\ p\left(l_i^t \mid \mathcal{Z}^t, l_i^{t-1}\right) &= p\left(l_i^t \mid \vec{z}_i^t, l_i^{t-1}\right) \\ &\propto p\left(\vec{z}_i^t, l_i^{t-1} \mid l_i^t\right) \cdot p\left(l_i^t\right) \\ &\approx \underbrace{p\left(\vec{z}_i^t \mid l_i^t\right)}_{\text{Data Term}} \cdot \underbrace{p\left(l_i^{t-1} \mid l_i^t\right)}_{\text{Temporal Expectation}} \cdot \underbrace{p\left(l_i^t\right)}_{\text{Prior Term}}. \end{aligned} \quad (2)$$

The smoothness term $\phi\left(l_i^t, l_j^t \mid \mathcal{Z}^t, L^{t-1}\right)$ is modeled as a Potts model, this way favoring neighboring Stixels to belong to the same class. The unary potential terms are defined to be the negative log-likelihoods of statistical probability distributions. These distributions for the different object classes in typical urban traffic scenes were set up from a large training ground truth database containing manually labeled Stixels as training examples. This database contains about 38,000 images and about ten million Stixels. All parameters of the energy function defined in 1 including the weight parameter λ were learned from this dataset.

(a) Dense SGM Stereo reconstruction [1]. The color represents the distance to the obstacle with red being close and green far away.



(b) Dynamic Stixel World [3]. The arrows point to the predicted Stixel position within the next half second.



(c) Segmentation result with three moving objects shown in yellow, magenta and cyan. The static background is shown in black.

Figure 1: Example results for the different steps of our segmentation process chain.

In order to evaluate the performance of the presented approach, the segmentation results were compared with another challenging data set, containing about 8000 images recorded from our experimental vehicle. All experiments have been performed with a single parameter set, and thus without any manual parameter tuning. The experimental results yield highly accurate segmentation of urban traffic scenarios, the average labeling accuracy is 98.06%. Additionally distinct features have been omitted in order to test their influence on the final segmentation result. Using the Stixel World allows to compute the alpha-expansion graph cut inference in real time in 1 ms on a single CPU core. The key conclusion from the experiments is that learning statistical relations from sufficient training data sets yields a powerful and robust segmentation apparatus with no need for any manual parameter tuning. As shown by the results, the approach generalizes unseen new traffic scenes well.

[1] H. Hirschmueller. Accurate and efficient stereo processing by semiglobal matching and mutual information. *CVPR*, 2005.

[2] D. Pfeiffer and U. Franke. Efficient representation of traffic scenes by means of dynamic stixels. *IV*, 2010.

[3] D. Pfeiffer and U. Franke. Towards a global optimal multi-layer stixel representation of dense 3d data. *BMVC*, 2011.

MoT - Mixture of Trees Probabilistic Graphical Model for Video Segmentation

Ignas Budvytis
ib255@cam.ac.uk
Vijay Badrinarayanan
vb292@cam.ac.uk
Roberto Cipolla
cipolla@eng.cam.ac.uk

Department of Engineering,
University of Cambridge,
Cambridge, UK

We present a novel mixture of trees (**MoT**) graphical model for video segmentation. Each component in this mixture represents a tree structured temporal linkage between super-pixels from the first to the last frame of a video sequence. Our time-series model explicitly captures the uncertainty in temporal linkage between adjacent frames which improves segmentation accuracy. We provide a variational inference scheme for this model to estimate super-pixel labels and their confidences in nearly realtime. The efficacy of our approach is demonstrated via quantitative comparisons on the challenging SegTrack joint segmentation and tracking dataset [6].

Motivation. It is a common practice in computer vision problems to establish mappings between frames via optic flow algorithms [4] or long term point trajectories. However for tasks requiring semantic label propagation in video sequences, satisfactory results are not achieved: [1]. Poor performance can be attributed to a lack of robust occlusion handling, label drift caused by round-off errors, high cost of multi-label MAP inference or sparsity of robust mappings. These issues have led to the use of label inference over short overlapping time windows ([6]) as opposed to a full length video volume. To address these issues, we have developed a novel super-pixel based mixture of trees (**MoT**) video model, motivated by the work of Budvytis et. al [3]. Our model alleviates the need to use short time window processing and can deal with occlusions effectively. It requires no external optic flow computation, and instead, infers the temporal correlation from the video data automatically. We also provide an efficient structured variational inference scheme for our model, which estimates super-pixel labels and their confidences. The uncertainties in the temporal correlations are also inferred, unlike the joint label and motion optimisation method of [6] where only a MAP estimate is obtained.

Model. Let $S_{i,j}$ denote super-pixel j at frame i, and $Z_{i,j}$ denote the corresponding missing label. We associate the temporal mapping variable $T_{i,j}$ to super-pixel $S_{i,j}$. $T_{i,j}$ can link to super-pixels in frame $i-1$ which have their centers lying within a window $W_{i,j}$, placed around the center of $S_{i,j}$. Let $S_i = \{S_{i,j}\}_{j=1}^{\Omega(i)}$, $Z_i = \{Z_{i,j}\}_{j=1}^{\Omega(i)}$ and $T_i = \{T_{i,j}\}_{j=1}^{\Omega(i)}$ denote the set of super-pixels, their labels and the correponding temporal mapping variables respectively at frame i. $\Omega(i)$ denotes the number of super-pixels in frame i. Our proposed mixture of trees (**MoT**) probabilistic model for the video sequence factorises as follows:

$$p(S_{0:n}, Z_{0:n}, T_{1:n}|\mu) = \frac{1}{\mathcal{Z}(\mu)} \prod_{i=1:n} \prod_{j=1:\Omega(i)} \Psi_a\left(S_{i,j}, S_{i-1,T_{i,j}}\right) \times \Psi_l\left(Z_{i,j}, Z_{i-1,T_{i,j}}|\mu\right) \Psi_u\left(Z_{i,j}\right) \Psi_u\left(Z_{0,j}\right) \Psi_t\left(T_{i,j}\right), \quad (1)$$

where $S_{i-1,T_{i,j}}$ indexes the super-pixel mapped to by $T_{i,j}$ in frame $i-1$ and similarly for $Z_{i-1,T_{i,j}}$. To define the *appearance factor* $\Psi_a(.)$ of the MRF on the R.H.S of Eqn. 1, we first find the best match pixel in frame $i-1$ for a pixel in frame j by performing patch cross-correlation within a pre-fixed window. The appearance factor is then defined using the number of pixels in super-pixel $S_{i,j}$ which have their best matches in $S_{i-1,T_{i,j}}$ as follows:

$$\Psi_a\left(S_{i,j}, S_{i-1,T_{i,j}}\right) \triangleq \#\text{shared pixel matches}. \quad (2)$$

Note that more sophisticated super-pixel match scores can also be substituted here as in [4]. The *label factor* $\Psi_l(.)$ is defined between the multinomial super-pixel label random variables as follows:

$$\Psi_l\left(Z_{i,j} = l, Z_{i-1,T_{i,j}} = m|\mu\right) \triangleq \mu\,(\text{if } l = m) \text{ or } 1-\mu\,(\text{if } l \neq m), \quad (3)$$

where l, m take values in the label set $\mathcal{L}$. μ is a parameter which controls label affinity. We set it to a value of 0.95 in our experiments. The single node potential for the temporal mapping variables $\Psi_t(.)$ is similar to a box prior and is defined as follows:

$$\Psi_t\left(T_{i,j}\right) \triangleq 1.0\left(\text{if } T_{i,j} \in W_{i,j}\right) \text{ or } 0.0\,(\text{if outside}). \quad (4)$$

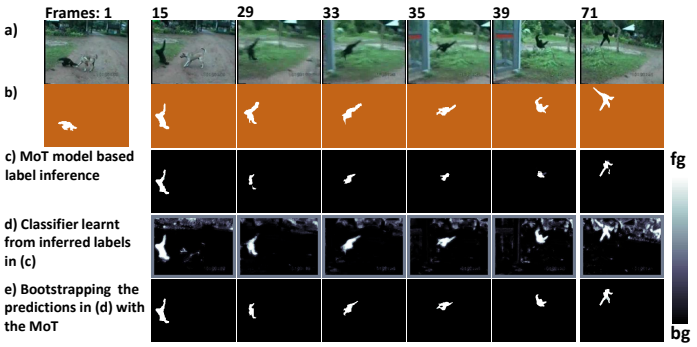


Figure 1: The first two rows show the image sequence Monkey-dog from the SegTrack dataset [6] and the corresponding ground truth. The segmentation algorithm in this sequence has to cope with fast shape changes, motion blur and overlap between foreground and background appearances. Row (c) is the inferred labels using the **MoT** time-series and with flat unaries. Row (d) are the Random Forest predictions when trained using the posteriors in row (c). Fusing these predictions with the MoT time-series results in an improved segmentation in row (e). Bright white and dark black correspond to confident foreground and background respectively.

The super-pixel label unary factors $\Psi_u\left(Z_{i,j}\right)$ are defined as output of Random Decision Forest Classifier (see Fig. 1 above and Sec. 4 in the paper). From Eqn.1 we note that the temporal mapping variable is present both in the appearance and label factor. Thus these variables are *jointly* influenced by both object appearance and semantic labels, a property which is desirable for interactive video segmentation systems.

Inference. We use structured variational inference scheme [5] where we assume the following form for the *approximate variational posterior* of the latent variables.

$$Q(Z_{0:n}, T_{1:n}) \triangleq Q(Z_{0:n}) \prod_{i=1:n} \prod_{j=1:\Omega(i)} Q\left(T_{i,j}\right). \quad (5)$$

The temporal mappings are assumed independent in the approximate posterior, however, the super-pixel latent labels do not factorise into independent terms, thereby maintaining *structure* in the posterior. The observed data log likelihood $\log(S_{0:n}|\mu)$ is lower bounded using the approximate posterior in Eqn. 5. To maximise the above lower bound we employ calculus of variations [2]. Finally, to compute the approximate super-pixel label and required pair-wise marginals we use variational message passing [2].

Evaluation. We evaluated the performance of our approach in a tracking and segmentation setting using the challenging SegTrack [6] dataset. Fig. 1 illustrates qualitative results of different stages of our algorithm on a Monkey-dog sequence from SegTrack. A detailed qualitavive and quantitative comparisons with some of the recent state of the art approaches are provided in the paper.

[1] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. In *CVPR*, 2010.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] I. Budvytis, V. Badrinarayanan, and R. Cipolla. Semi-supervised video segmentation using tree structured graphical models. In *CVPR*, 2011.

[4] A. Fathi, M. Balcan, X. Ren, and J. M. Rehg. Combining self training and active learning for video segmentation. In *BMVC*, 2011.

[5] L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In *NIPS*, 1996.

[6] D. Tsai, M. Flagg, and J. M. Rehg. Motion coherent tracking with multi-label mrf optimization. In *BMVC*, 2010.

Improved Initialisation and Gaussian Mixture Pairwise Terms for Dense Rnadom Fields with Mean-field Inference

Vibhav Vineet
vibhav.vineet-2010@brookes.ac.uk
Jonathan Warrell
jwarrell@brookes.ac.uk
Paul Sturgess
paul.sturgess@brookes.ac.uk
Philip H.S. Torr
philiptorr@brookes.ac.uk

Department of Computing
Oxford Brookes University
Oxford, UK
cms.brookes.ac.uk/research/visiongroup/

Many labelling problems in computer vision are often modelled as discrete optimisation problems such as object class segmentation, stereo correspondence, image de-noising etc. Generally, these problems are solved in a markov random field (MRF) or conditional random field (CRF) framework, where the basic model includes pairwise terms defined over a grid with 4 or 8 neighbours. A more expressive model is to allow dense connectivity which captures long range interactions between variables. However the complexity increases with more interactions.

Recently, Krahenbuhl and Koltun [1] proposed an efficient bilateral filtering based method for inference in dense pairwise CRFs, where pairwise weights take the form of a weighted combination of Gaussian kernels. They formulate multilabelling problems as performing approximate maximum posterior marginal (MPM) inference in a mean-field approximation to the CRF. Empirically they achieve a significant speed-up compared to graph-cuts based methods, and observe improvements in the accuracy on object-class segmentation problems. However, in its current form, their method is associated with two limitations.

The first issue is related to the fact that the mean-field approximation assumes complete factorisation over the individual variables. Though this simplified model leads to efficient and tractable models for learning and inference, the mean-field inference methods are generally sensitive to initialisation. In this work, we propose a hierarchical mean-field approach to improve the quality of the solutions by providing good initial conditions. We perform mean-field inference at the coarser level, and transfer the labels from the coarser level to the finer level for better initialisation. At the coarser level, we use a SIFT-flow [3] based label transfer method for better initialisation. SIFT-flow provides an algorithm for finding the correspondence between two images, where images are taken from different view points but share similar high level scene characteristics. Suppose we have a large training set of annotated ground truth images with per pixel class labels. Now, given a test image, we first find the K-nearest neighbour images using GIST features. In general, we restrict our set to 30 nearest neighbours. We then compute a dense correspondence using the SIFT-flow method from the test image to each of 30 nearest neighbours. We re-rank those nearest neighbours based on the flow values, and pick the best nearest neighbour. Once we have recovered our best candidate, we warp the corresponding ground truth of the candidate image to the current test image. We use these warped labels to initialize the mean-field inference method. These transferred labels provide good initial conditions, and act as a soft constraint on our solutions.

The second issue relates to the form of the pairwise weights in [1] which are a linear combination of Gaussian kernels. Although they learn the standard deviation and weighting co-efficient of each component, they allow each Gaussian component to take only zero mean. Further, they use the same combination of Gaussian kernels for each label pair. In this paper, we propose an approach that extends this model. Specifically, we allow our model to take a more general Gaussian mixture model for every pair of labels, where we learn the mean, the co-variance matrix and the mixing co-efficient of each Gaussian component. Assuming there are M Gaussian mixture components, our energy function takes the following form:

$$E(\mathrm{X}|\mathrm{I}) = \sum_i \psi(x_i) + \sum_{i<j} \kappa(x_i,x_j) \sum_{v=1}^{V} w^{(v)} k^{(v)}(\mathbf{f}_i,\mathbf{f}_j) - \lambda \sum_{i<j} \sum_{m=1}^{M} \alpha_m^{(x_i,x_j)} \mathcal{G}_m(\mathrm{I},\mu_m,\Sigma_m) \quad (1)$$

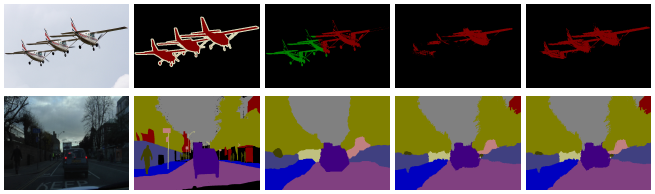


Figure 1: Qualitative results on PascalVOC-10 (first row) and CamVid (last row) datasets. From left to right: input image, ground truth, output from [2] (U+P), output from [1] (dense CRF), output from our dense CRF with better initialisation and Gaussian mixture.

where $\kappa(x_i,x_j)$ is the label compatibility function between pair of labels, λ is a weighting function, μ_m, and Σ_m are the mean, and co-variance matrix of mth mixture component, α_m^{ij} is the m^{th} mixing coefficient between the i^{th} and j^{th} labels, and I is an image derived feature. Given the number of mixture components, we learn the mixing co-efficient $\alpha^{(\ldots)}$, the mean μ_m, and the co-variance matrix Σ_m. We propose a piecewise learning framework where given a set of data points, we fit Gaussian mixture model to those points. We use a variation of the *Expectation Maximization (EM)* algorithm for estimating the parameters determined by maximum likelihood.

1 Experiments

We demonstrate the accuracy and efficiency offered by our approach on object-class segmentation problems on CamVid and PascalVOC-10 segmentation dataset. We assess the average union/intersection measure per class (defined in terms of the true/false positives/negatives for a given class as TP/(TP+FP+FN)). On CamVid dataset, we observe an overall improvement of 6.5% compared to graph-cuts based α-expansion and 5.5% compared to dense CRF method [1]. Here we use only unary and pairwise connections. Further, our model with unary and pairwise connections perform better than [4] who use unary, pairwise and higher order terms by almost 0.2%. On PascalVOC-10 segmentation dataset, with our final model which includes the mixture components and the better initialisation strategy, we observe an improvement of 3.5% and 3% over the α-expansion method. Further, in our model which includes unary and pairwise terms, we observe 0.5% improvement in U/I score and almost 6% improvement in the average recall scores over the work of [2] who include higher order terms, detector potentials, and object co-occurrence terms along with unary and pairwise potentials. We also observe a qualitative improvement in the results on both CamVid and PascalVOC dataset, some of the images are shown in the Fig. 1.

[1] Philipp Krahenbuhl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[2] L. Ladicky, C. Russell, P. Kohli, and P. Torr. Graph cut based inference with co-occurrence statistics. *ECCV*, 2010.

[3] Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T. Freeman. Sift flow: Dense correspondence across different scenes. In *ECCV (3)*, pages 28–42, 2008.

[4] Paul Sturgess, Karteek Alahari, Lubor Ladicky, and Philip H. S. Torr. Combining appearance and structure from motion features for road scene understanding. In *BMVC*, 2009.

Image Segmentation using Dual Distribution Matching

Tatsunori Taniai[1]
taniai@nae-lab.org
Viet-Quoc Pham[2]
quocviet.pham@toshiba.co.jp
Keita Takahashi[3]
keita.takahashi@ieee.org
Takeshi Naemura[1]
naemura@nae-lab.org

[1] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
[2] Toshiba Corporate Research and Development Center, Kanagawa, Japan
[3] Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan

This paper addresses the problem of foreground-background image segmentation where only the approximate color distributions of the foreground and background regions are given as the input. Our aim is to derive a fundamental algorithm with this primitive setup that can find foreground and background regions that are consistent with the given input distributions. The essential question here is how to measure consistencies between the given distributions and the segmentation.

Local measures are widely adopted [2] by virtue of their simplicity. Each pixel is *individually* evaluated to determine how likely it is to belong to the foreground or background based on its color. However, local-measure-based methods are subject to the shrinking bias, which often results in shortcutting across thin structures.

Recent studies (*e.g.* BMGC [1]) have shown that methods based on *global measures* outperform conventional local-measure-based methods. The global consistency is measured by the similarity between a given distribution and the resulting distribution from the extracted region. We introduce a new distribution matching method named dual distribution matching (DDM) in order to increase the robustness of global measures. In this method, *the consistencies between two input distributions (the foreground and background distributions) and the resulting segmentation are enforced simultaneously*. Our method makes it possible to achieve robust and accurate segmentations even with not-so-accurate input distributions.

Dual Distribution Matching Binary segmentation is formulated as a problem that involves finding a label $\boldsymbol{L}$ for the set of pixels P, as $\boldsymbol{L} = \{L_p | L_p \in \{F,B\}, \forall p \in P\}$, where p denotes a pixel, and F/B denotes the foreground/background label. The foreground/background region is the set of all pixels with F/B and is denoted as $\mathbf{R}_l^{\boldsymbol{L}} = \{p \in P | L_p = l\}$ $(l = F,B)$. The probability distribution of colors (or intensities) within region $\mathbf{R}_l^{\boldsymbol{L}}$ is written as $\mathcal{P}_l^{\boldsymbol{L}}$ $(l = F,B)$.

Let us assume that only the approximate distributions for both the foreground and background are given as $\mathcal{H}_F \simeq \mathcal{P}_F^{\boldsymbol{L}^*}$ and $\mathcal{H}_B \simeq \mathcal{P}_B^{\boldsymbol{L}^*}$, where $\boldsymbol{L}^*$ is the ground truth of $\boldsymbol{L}$. Here, $\boldsymbol{L}^*$ is inferred as the label that minimizes the following energy function $\mathcal{E}(\boldsymbol{L})$:

$$\mathcal{E}(\boldsymbol{L}) = \underbrace{\lambda_F \mathcal{M}_F(\boldsymbol{L})}_{\text{Foreground Matching}} + \underbrace{\lambda_B \mathcal{M}_B(\boldsymbol{L})}_{\text{Background Matching}} + \underbrace{\lambda_S \mathcal{S}(\boldsymbol{L})}_{\text{Smoothness}}, \quad (1)$$

where $\mathcal{M}_l(\boldsymbol{L})$ is the negative of the distribution similarity measure $\mathcal{B}(,)$:

$$\mathcal{M}_l(\boldsymbol{L}) = -\mathcal{B}\left(\mathcal{P}_l^{\boldsymbol{L}}, \mathcal{H}_l\right) \quad (l = F,B). \quad (2)$$

The $\mathcal{S}(\boldsymbol{L})$ is a smoothness function composed of pairwise discontinuity penalties. This is called *dual distribution matching* or DDM, because both the foreground and background distributions are matched simultaneously. The term $\mathcal{B}(,)$ is the Bhattacharyya coefficient that measures the amount of overlap between two distributions f and g, which takes 1 as the maximum when $f = g$:

$$\mathcal{B}(f,g) = \sum_{z \in Z} \sqrt{f(z)g(z)} \leq 1 \quad (3)$$

With the definitions above, $\mathcal{E}(\boldsymbol{L})$ with $\lambda_B = 0$ or $\lambda_F = 0$, which we define as $\mathcal{E}_F(\boldsymbol{L})$ or $\mathcal{E}_B(\boldsymbol{L})$ respectively, is equivalent to the single distribution matching of the BMGC method [1]. We refer to the BMGC method with $\mathcal{E}_F(\boldsymbol{L})$ or $\mathcal{E}_B(\boldsymbol{L})$ as F-BMGC or B-BMGC. As illustrated in Fig.1, those methods cannot capture the true solution $\boldsymbol{L}^*$ if the input distribution $\mathcal{H}_F$ or $\mathcal{H}_B$ is inaccurate. In contrast, *our method is more likely to capture the true solution by using both constraints simultaneously*. We show in this paper that $\mathcal{M}_F(\boldsymbol{L})$ *and* $\mathcal{M}_B(\boldsymbol{L})$ *should be weighted in proportion to the size of the foreground and background areas* so that the minimum solution of the energy function $\mathcal{E}(\boldsymbol{L})$ captures the true solution $\boldsymbol{L}^*$.

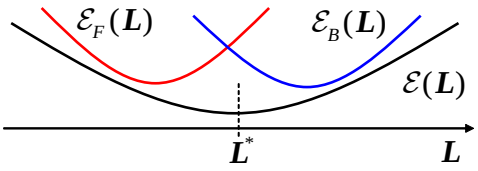


Figure 1: Dual matching energy function.

Experimental Results Figure 2 shows segmentation results of our method DDM, DDM with fixed weighting parameters, single distribution matching methods (F-BMGC and B-BMGC) [1], and local-measure-based method (interactive graph cuts) [2], where approximate input distributions are produced from foreground and background regions of lasso-trimap. Our method achieved the best accuracy in this experiment.

Figure 2: Segmentation results with approximate input distributions.

Also, we compared local and global consistency measures while varying the accuracy of the input distributions. Figure 3 shows that *the proposed method outperforms the others at high and medium accuracies, whereas interactive graph cuts performed the best at very low accuracies.*

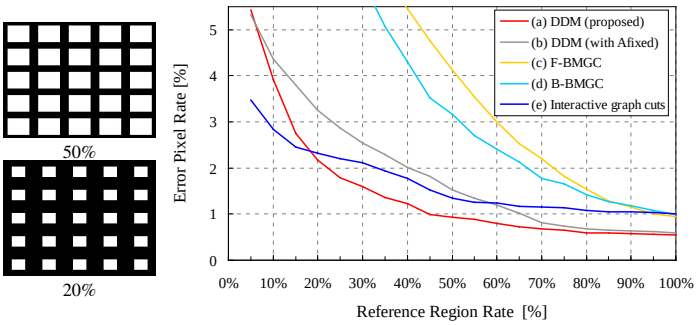


Figure 3: Comparison between local and global models according to input distribution accuracy. The input distributions were purposely made inaccurate by limiting the reference region using masks (left).

[1] Ismail Ben Ayed, Hua-Mei Chen, Kumaradevan Punithakumar, Ian G. Ross, and Shuo Li. Graph cut segmentation with a global constraint: Recovering region distribution via a bound of the bhattacharyya measure. In *Proc. CVPR*, pages 3288–3295, 2010.

[2] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proc. ICCV*, pages 105–112, 2001.

Keynote

Visual Tracking in the 21st Century

Jiří Matas

Czech Technical University (CTU), Prague

Visual tracking is an old area that has recently seen a surge in activity. The interest has been fueled by progress in related fields like detection, segmentation and optic flow as well as by application-driven demand and the increase in the available computing power.

The published tracking methods differ in many aspects such as the speed, the complexity of the model of the tracked entity, the (geometric) transformations assumed, the mode of operation (casual and non-causal), the ability to adapt and learn, the robustness to occlusion and assumptions about the observer. I will review the dataset used in recent publications and show that the ”tracker space” is still wide open with large areas to be explored.

I will then present three trackers developed by me and my collaborators that operate at very different points in the speed-robustness-flexibility space that are close to the ”convex hull” of published methods: the TLD tracker, the Flock-of-Trackers and the Zero-Shift-Point tracker. I will focus on a common aspect shared by the trackers: mechanisms for prediction and handling of tracking errors. Such mechanisms contribute to tracker robustness, which will be demonstrated live.

Jiří Matas received his MSc degree in cybernetics (with honours) from the Czech Technical University, Prague, Czech Republic, in 1987 and his PhD from the University of Surrey, UK, in 1995. From 1991 to 1997, he was a research fellow at the Centre for Vision, Speech and Signal Processing at the University of Surrey. In 1997, he joined the Center for Machine Perception at the Czech Technical University in Prague. Since 1997, he has held various positions at these two institutions.

He has published more than a hundred papers in refereed journals and conferences. His publications have more than 1100 citations in the Science Citation Index. He received the best paper prize at the British Machine Vision Conferences in 2002 and 2005 and at the Asian Conference on Computer Vision in 2007. Jiří Matas has served in various roles at major international conferences (e.g. ICCV, CVPR, ICPR, NIPS), co-chairing ECCV 2004 and CVPR 2007 and is on the editorial board of IEEE Transactions on Pattern Analysis and Machine Intelligence and the International Journal of Computer Vision.

His research interests include object recognition, sequential pattern recognition, ensemble methods, invariant feature detection, and Hough Transform and RANSAC-type optimization.

Image Retrieval for Image-Based Localization Revisited

Torsten Sattler[1]
tsattler@cs.rwth-aachen.de
Tobias Weyand[2]
weyand@vision.rwth-aachen.de
Bastian Leibe[2]
leibe@vision.rwth-aachen.de
Leif Kobbelt[1]
kobbelt@cs.rwth-aachen.de

[1] Computer Graphics Group
RWTH Aachen University, Germany
[2] Computer Vision Group
RWTH Aachen University, Germany

Image-based localization is the task of determining the exact location from which a query photo was taken. In this paper, we consider image-based localization relative to a 3D point cloud of a scene, obtained from Structure-from-Motion, which allows an accurate estimate of the full camera pose from correspondences between 2D features and 3D points. To quickly establish the required 2D-to-3D matches, Irschara *et al*. use image retrieval methods [5] to find database images (used for the reconstruction) similar to the query image [1]. Since the relation between 2D features and 3D points is known for the database images, the correspondences for the query image can be computed by feature matching between images. Recent work has demonstrated that directly matching the features against the points outperforms retrieval-based methods in terms of the number of images that can be localized successfully [4]. Yet, direct matching is inherently less scalable than retrieval-based approaches since it needs to keep SIFT descriptors [3] in memory at all times.

In this paper, we therefore analyze the algorithmic factors that cause the gap in registration performance. We show that using *selective voting* schemes enable retrieval methods to outperform state-of-the-art direct matching methods and explore how both selective voting and correspondence search can be accelerated by using a Hamming embedding [2].

Selective Voting

The main cause for the performance gap are the *incorrect votes* that are cast by image retrieval-based approaches such as [1]. Fig. 1 illustrates this problem. Although the query feature (pink) corresponds to only a single 3D point (red), inverted file scoring also casts a vote for every image that has a feature (black) matched to the same visual word. Dealing with these incorrect votes is challenging even for advanced re-ranking schemes such as *tf∗idf* weighting [5] or *probabilistic ranking* [1]. Since pose estimation is only attempted for the top-k images, failure to rank any of the relevant images among the top-k negatively impacts localization performance.

Two *selective voting* schemes can be used to avoid incorrect votes. *Correspondence voting* finds the two nearest neighbors among all descriptors of 3D points having the same visual word and votes for the image that contains the nearest neighbor if the SIFT ratio test [3] is passed. This scheme essentially uses the correspondences found by the direct matching approach from Sattler *et al*. [4] to vote for database images. The camera pose is then estimated from correspondences found with pairwise image matching. Since *correspondence voting* requires that SIFT descriptors are kept in memory at all times, a *selective voting* scheme using Hamming embedding [2] can be used to the reduce memory requirements. The resulting *Hamming voting* only casts a vote for an image containing a point if the Hamming distance between the binary embeddings of the query feature and the point is below a certain threshold (*cf*. Fig. 1(right)). Using 64-bit for the embedding requires only little memory overhead to store the embeddings in the words, while 10^6 Hamming distance computations can be done in about 2ms on a modern CPU. Thus, *Hamming voting* preserves the scalability of retrieval-based methods.

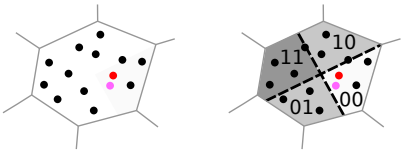


Figure 1: (Left) The query feature (pink) corresponds to a single 3D point (red), yet unrelated inverted file entries (black) cause *false positive votes*. (Right) By thresholding Hamming distances of a Hamming embedding, *Hamming voting* can avoid casting many of the incorrect votes.

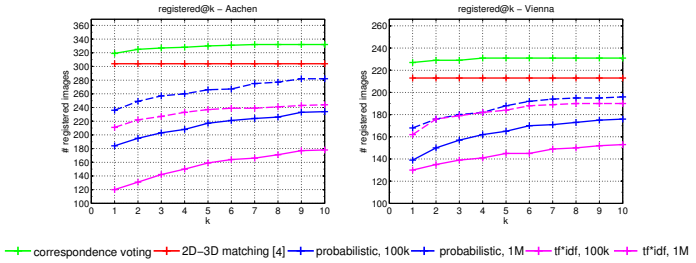


Figure 2: The *correspondence voting* scheme is able to achieve significantly better results than standard ranking schemes due to its ability to discard incorrect votes. It also outperforms the direct matching approach.

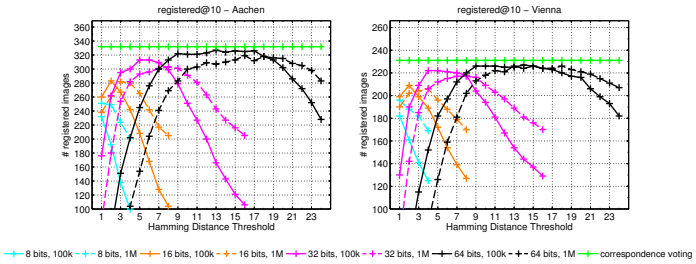


Figure 3: *Hamming voting* using 32- and 64-bit vectors achieves nearly the same performance as *correspondence voting* with SIFT descriptors.

Results

We compare *selective voting*-based localization to classical image retrieval-based methods and the state-of-the-art direct matching approach from [4]. We measure the performance of the methods in the number of images for which a pose can be estimated successfully. Two large-scale datasets are used for the evaluation, including our novel Aachen dataset consisting of 1.5M 3D points and 369 query images [1].

Fig. 2 compares *correspondence voting* to retrieval methods using different ranking schemes with different visual vocabulary sizes and the method from [4]. Using this *selective voting* scheme significantly improves image retrieval-based localization and enables us to outperform the state-of-the-art direct matching method [4].

The evaluation of *Hamming voting* with different sizes for the resulting binary descriptors and different vocabulary sizes in Fig. 3 shows that a performance similar to *correspondence voting* can be achieved using nearly one order of magnitude less memory. Thereby, using a coarser vocabulary yields better results due to less quantization errors.

Further results on accelerating the matching between images required for correspondence search can be found in the paper.

[1] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From Structure-from-Motion Point Clouds to Fast Location Recognition. In *CVPR*, 2009.

[2] H. Jégou, M. Douze, and C. Schmid. Packing bag-of-features. In *ICCV*, 2009.

[3] D. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2), 2004.

[4] T. Sattler, B. Leibe, and L. Kobbelt. Fast Image-Based Localization using Direct 2D-to-3D Matching. In *ICCV*, 2011.

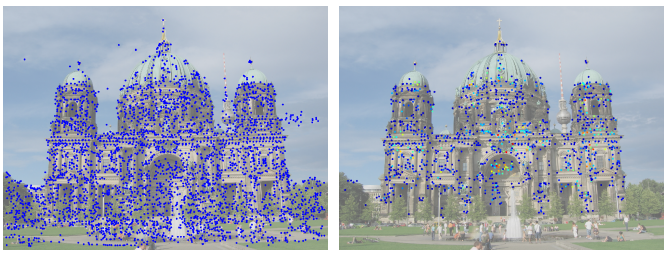
[5] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, 2003.

[1] Available at `http://www.graphics.rwth-aachen.de/localization`.

Improved Geometric Verification for Large Scale Landmark Image Collections

Rahul Raguram
rraguram@cs.unc.edu
Joseph Tighe
jtighe@cs.unc.edu
Jan-Michael Frahm
jmf@cs.unc.edu

Department of Computer Science
University of North Carolina
Chapel Hill, NC, USA.



(a) (b)

Figure 1: (a) All detected features for a single image (b) Features filtered based on the results of geometric verification – only visual words that were inliers in at least 10 previous image pairs are shown. The features in (b) are heatmap colour coded based on the inlier counts.

Figure 2: The top and bottom ten verified images, according to classifier scores. Our approach uses the output of geometric verification to generate a training set in an online manner, without manual labeling. The resulting classifier is able to reliably distinguish between landmark and non-landmark images.

In this work, we address the issue of geometric verification, with a focus on modeling large-scale landmark image collections gathered from the internet. In particular, we show that we can compute and learn descriptive statistics pertaining to the image collection by leveraging information that arises as a by-product of the matching and verification stages.

In designing a 3D reconstruction system for internet photo collections, one of the key considerations is robustness to "clutter" – when operating on datasets downloaded using keyword searches on community photo sharing websites (such as Flickr), it has been observed that invariably, a large fraction of images in the collection are unsuitable for the purposes of 3D reconstruction [3, 4]. Thus, one of the fundamental steps in a 3D reconstruction system is *geometric verification*: the process of determining which images in an internet photo collection are geometrically related to each other. This is a computationally expensive process, and much work in recent years has focused on developing efficient ways to perform this step. For example, Agarwal et al. [1] use image retrieval techniques to determine, for every image in the dataset, a small set of candidate images to match against. An alternate approach, adopted by Frahm et al. [2], is to first cluster the images based on global image descriptors and to then perform the verification within each cluster. While these approaches are extremely promising, there are still some limitations. For instance, even the carefully optimized approach described in [2] spends approximately 50% of the processing time simply verifying image pairs against each other. In addition, the approach in [2] suffers from "incompleteness"; due to the coarse clustering, a large fraction of images are discarded following the clustering and verification steps. In this work, we aim to overcome these limitations.

Thus far, the typical way to perform geometric verification has been to estimate the geometric relationship between pairs *independently*, which does not fully exploit the specific characteristics of the dataset being processed. Our main idea in this work is simple: as the geometric verification progresses, we learn information about the image collection, and subsequently use this learned information to improve efficiency and completeness. More specifically, since images of the same geometric structures are being repeatedly verified against each other, this process of repeated matching reveals useful information about two things:

(a) the stability and validity of low-level image features

(b) the global appearance of the various landmarks in the dataset

As a motivating example, consider Figure 1(a), which shows all detected SIFT features for a single image. Note that a large number of features lie in areas of the image that are very unlikely to pass any geometric consistency check (for e.g., features on vegetation, people, and in the sky). Now, if we have previously verified *other* images of the same scene, we can weight each visual word in the current image by the number of times that the word has previously passed the geometric consistency check in other image pairs (see Figure 1(b)). Note, in particular, that this weighting emphasizes visual words that are stable, reliable, and more likely to be geometrically consistent, while also suppressing spurious visual words. Our first contribution in this work is to integrate this learned information into a RANSAC procedure, which results in an appreciable improvement in efficiency compared to current techniques.

As a second contribution, we show that it is possible to learn additional useful information capturing higher-level information about the dataset. For instance, once we have obtained a sufficiently large set of successfully verified image pairs, we hypothesize that this set captures useful information about the *global* appearance of various landmarks present in the dataset. We observe that this information can then be used to train a classifier that distinguishes between landmark and non-landmark images (see Figure 2). We then employ this classifier during the image registration step. In this context, having a trained model of *landmark appearance* is very useful, since this allows us to only verify those images that are likely to be landmark images and discard the rest.

In summary, this work presents techniques for taking advantage of the information generated during geometric verification, to improve the overall efficiency of the process. Our approach thus integrates online knowledge extraction seamlessly into structure-from-motion systems, and is particularly relevant for large-scale image collections. Our results demonstrate both improved efficiency, as well as higher image registration performance, potentially yielding more complete 3D models for these large-scale datasets.

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building Rome in a Day. In *International Conference on Computer Vision*, 2009.

[2] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building Rome on a Cloudless Day. In *European Conference on Computer Vision*, volume 6314, pages 368–381, 2010.

[3] L. Kennedy, S.-F. Chang, and I. Kozintsev. To Search or To Label?: Predicting the Performance of Search-based Automatic Image Classifiers. In *ACM Multimedia Information Retrieval Workshop (MIR 2006)*, 2006.

[4] Rahul Raguram, Changchang Wu, Jan-Michael Frahm, and Svetlana Lazebnik. Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs. *Int. J. Comput. Vision*, 95(3):213–239, 2011.

Transfer Learning by Ranking for Weakly Supervised Object Annotation

Zhiyuan Shi
zhiyuan.shi@eecs.qmul.ac.uk
Parthipan Siva
psiva@eecs.qmul.ac.uk
Tao Xiang
txiang@eecs.qmul.ac.uk

School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London E1 4NS, UK

Object detectors [5] locate objects of interest in images and have many applications including image tagging, consumer photography, and surveillance. Most existing object detectors take a fully supervised learning (FSL) approach, where all the training images are manually annotated with the object location. However, manual annotation of hundreds of object categories is time-consuming, laborious, and subjective to human bias. To reduce the amount of manual annotation, a weakly supervised learning (WSL) [3, 6] approach is desired. In WSL, the training set is only annotated with a binary label indicating the presence or absence of the object of interest, not the location or extent of the object (Fig. 1(a)).

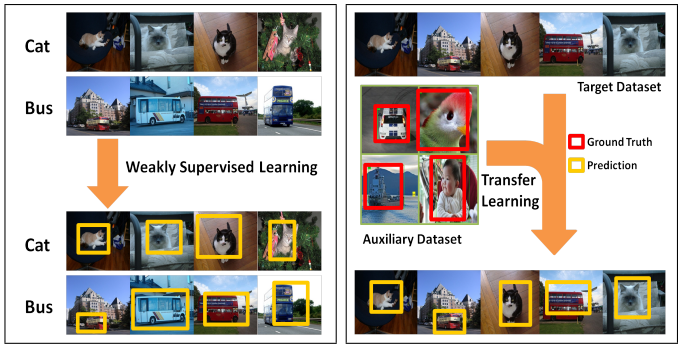


Figure 1: (a) Weakly supervised learning approach for automatic annotation of objects [3, 6]. (b) Our transfer learning approach for automatic annotation of objects.

Typically three information cues, saliency, inter-class, and intra-class, are used to locate or annotate the object of interest in images known to contain the object of interest (positive images). Saliency information ensures that the annotated region is a foreground region. Inter-class information ensures that the annotated regions look dissimilar to all images without the object of interest (negative images). Intra-class information ensures that the annotated regions in all positive images look similar to each other. Methods that use saliency alone [1] select salient regions in each positive image independently. Methods that use inter-class and intra-class information [3, 6] typically use saliency to limit the search space of each image by only looking at the most salient regions; then they select one of these salient regions by maximising the inter-class and intra-class information.

In this paper we utilise a fourth information cue (Fig. 1(b)) which is typically neglected by other approaches: an auxiliary fully annotated dataset. While we want to reduce manual annotation when learning new object categories, we cannot ignore the fact that there exist many datasets which already have manual annotation of object locations [4]. However, these auxiliary datasets seem unhelpful since they often contain object categories that are unrelated to the target object category we wish to annotate. For example, an auxiliary dataset might contain annotations of cars, birds, boats and person but a target object category might be cats and buses (Fig. 1(b)). So what information can we actually transfer? When adopting the strategy of selecting the optimal object location from a set of candidate salient regions [3, 6], the performance of the selection can obviously be measured by examining the degree of overlap between the selected region and the ground truth region (Fig. 2(a)). One can safely assumes that the more a salient region overlaps the ground truth region, the more similar the two's appearances are. In other words, there exists a mapping relationship between the degree of overlap (hence the accuracy of annotation) and the appearance similarity. This relationship should hold true regardless of the object category and is what we propose to learn and transfer to the target data. To quantify this mapping relationship, one must take into consideration the high dimensionality typical for representing object appearance and the inevitable noise. To this end, we formulate a ranking based transfer learning model which, once learned, takes appearance similarity as input and predicts the ranking order among all the candidate salient regions according to their degree of overlap with the (unknown) true object location.

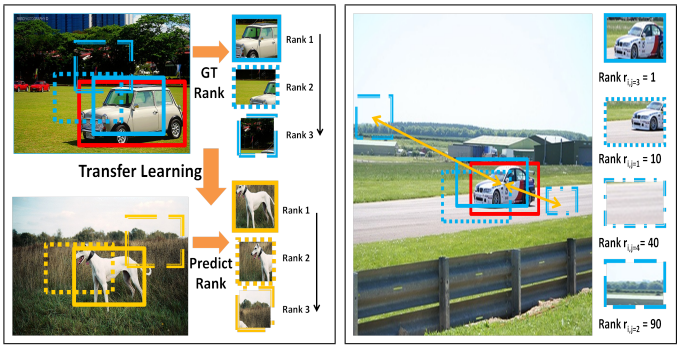


Figure 2: (a) The mapping relationship between the degree of overlap and the appearance similarity between salient regions and the ground truth location is transferred from the auxiliary data to the target data. (b)A higher rank is given to salient regions (blue) with higher overlap to the ground truth (red). In this image, a road region ($j = 4$) which contains more relevant context to the car, is ranked higher than a sky region ($j = 2$) according to their distances to the ground truth region (red).

More specifically, for each image $i \in \mathcal{A}$, we represent each of the N salient regions with an unnormalised BOW histogram $x_{i,j}$, where $j = 1 \ldots N$. To compute a feature from $x_{i,j}$ that is independent of the object category we define a difference vector $d_{i,j}$ as the feature of interest:

$$d_{i,j} = \left| \frac{x_{i,j}}{\|x_{i,j}\|_1} - \frac{g_i}{\|g_i\|_1} \right|, \tag{1}$$

All N salient regions are sorted by its overlap with the ground truth bounding box (Fig. 2(b)), where overlap is defined by [4] as the intersect area divided by union area. They will form enormous preference pairs for a moderate number of images and salient regions in each image (N). For efficient learning, we use the primal-based pairwise RankSVM algorithm proposed in [2] to minimise the objective function:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{(k,l)\in\mathcal{P}} \ell(\mathbf{w}^T \hat{d}_{i,k} - \mathbf{w}^T \hat{d}_{i,l}), \tag{2}$$

We show that our novel transfer learning model outperforms the state-of-the-art WSL approaches on the challenging PASCAL VOC 2007 dataset.

[1] B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *CVPR'10*, 2010.

[2] O. Chapelle and S.S. Keerthi. Efficient algorithms for ranking with svms. *Inf. Retr.*, 13(3):201–215, June 2010.

[3] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. ECCV'10, 2010.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.

[6] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV'11*, 2011.

Enhancing Exemplar SVMs using Part Level Transfer Regularization

Yusuf Aytar
yusuf@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Department of Engineering Science
University of Oxford
Parks Road
Oxford, OX1 3PJ, UK

Content based image retrieval (CBIR), the problem of searching digital images in large databases according to their visual content, is a well established research area in computer vision. In this work we are particularly interested in retrieving subwindows of images which are similar to the given query image, i.e. the goal is detection rather than image level classification. The notion of *similarity* is defined as being the same object class but also having similar viewpoint (e.g. frontal, left-facing, rear etc.). A query image can be a part of an object (e.g. head of a side facing horse), a complete object (e.g. frontal car image), or a composition of objects (visual phrases, e.g. person riding a horse). For instance, given a query of a horse facing left, the aim is to retrieve any left facing horse (intra-*class* variation) which might be walking or running with different feet formations (exemplar deformation).

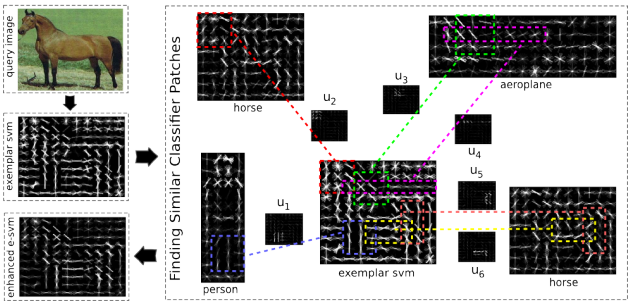


Figure 1: **Overview of the EE-SVM learning procedure.** The box on the right shows mining classifier patches from existing classifiers by matching subparts of E-SVM trained from the given query image. Comparing E-SVM and EE-SVM, better suppression of the background can be seen from the visualized classifiers. Note, here and in the rest of the paper we only visualize the positive components of the HOG classifier.

Recently exemplar SVMs (E-SVM) [1], where an SVM is trained with only a single positive sample, have found applications in the areas of CBIR [2] and object detection [1]. Since the E-SVM is trained from a single positive sample (together with many negatives), it is specialized to that given sample. This means that it can be strict (on viewpoint for example), and the negatives give some background suppression. However, the single positive is also a limitation: only so much can be learnt about the foreground of the query, and more significantly it can lead to lack of generalization. In our context, *generalization* refers to intra-class variation and deformation whilst maintaining the viewpoint. Learning such generalization from a single positive is challenging given the lack of examples of allowable deformations and intra-class variation.

In this work we propose a transfer learning approach for boosting the performance of E-SVMs using part-like patches of previously learned classifiers. The formulation softly constrains the learned template to be constructed from classifiers that have been fully trained (i.e. using many positives). For instance, the neck of a horse can be transferred from the tail of an aeroplane (see figure 1), or a jumping bike can borrow part of wheel patches from regular side facing bike or motorbike classifiers (see figure 2). The intuitive reason behind borrowing patches from other well

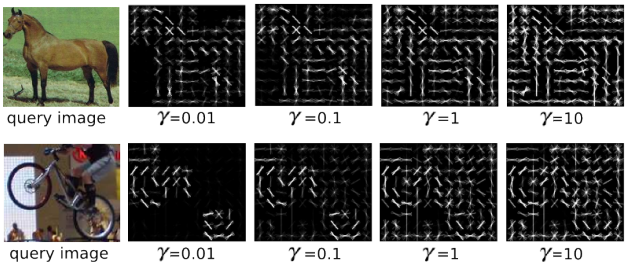


Figure 2: **Two limits of EE-SVM from reconstruction($\gamma = 0.01$) to E-SVM($\gamma = 10$).** Learned EE-SVM templates with varying γ values are displayed. λ is fixed to 1.

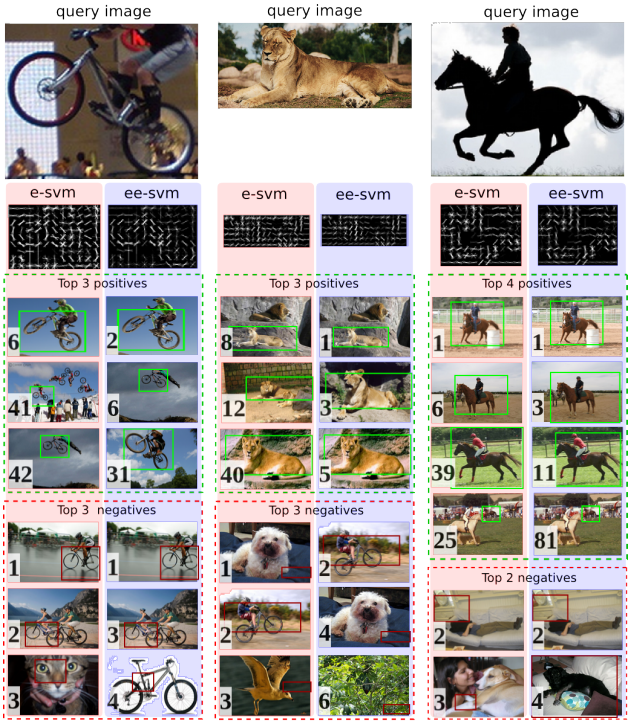


Figure 3: Retrieval of unusual poses on ImageNet. A visual phrase retrieval is also shown on the rightmost column.

trained classifiers is that these classifier patches bring with them a better sense of discriminative features and background suppression. The classifier patches also bring some generalization properties which an E-SVM may lack because it is only trained on a single positive sample. The result of the transfer learning is an enhancement of background suppression and tolerance to intra-class variation. However, these enhancements incurs no (significant) additional cost in learning and testing. We term the boosted E-SVM, Enhanced Exemplar SVM (EE-SVM).

We make the following contributions: (a) introduce the EE-SVM objective function; (b) demonstrate the improvement in performance of EE-SVM over E-SVM for CBIR; and, (c) show that there is an equivalence between transfer regularization and feature augmentation for this problem and others, with the consequence that the new objective function can be optimized using standard libraries.

Enhanced E-SVM incorporates the part based transfer regularization using the objective:

$$min_{w,b,\alpha} \quad \lambda||w - \sum_i^M \alpha_i u_i||^2 + \gamma \sum_i^M \alpha_i^2 + \sum_i^N \max\left(0, 1 - y_i(w^\mathsf{T} x_i + b)\right) \quad (1)$$

where λ and γ controls the balance between the two regularization terms as well as the tradeoff between error term and regularization terms. u_i's are the classifier patches cropped from source classifiers and relocated on a w sized template padded with zeros other than the classifier patch (see Figure 1), and α_i's are transfer weights. Note that given a fixed set of u_i's the formulation is convex.

EE-SVM is evaluated both quantitatively and qualitatively on the PASCAL VOC 2007 and ImageNet datasets for pose specific object retrieval. It achieves a significant performance improvement over E-SVMs, with greater suppression of negative detections and increased recall, whilst maintaining the same ease of training and testing.

[1] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *Proc. ICCV*, 2011.

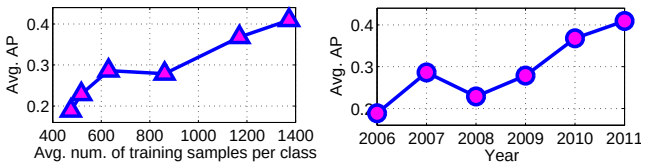
[2] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Trans. Graph.*, 30(6), 2011.

Do We Need More Training Data or Better Models for Object Detection?

Xiangxin Zhu[1]
xzhu@ics.uci.edu
Carl Vondrick[2]
vondrick@mit.edu
Deva Ramanan[1]
dramanan@ics.uci.edu
Charless C. Fowlkes[1]
fowlkes@ics.uci.edu

[1] Computer Science Department
University of California
Irvine, CA, USA

[2] CSAIL
Massachusetts Institute of Technology
Cambridge, MA, USA
(Work performed while at UC Irvine)



Much of the impressive progress in object detection is built on the methodologies of statistical machine learning, which makes use of large training datasets. Consider the benchmark results of the well-known PASCAL VOC object challenge over the past 5 years (above). We see a clear trend in increased performance over the years as methods have gotten better and training datasets have become larger. In this work, we ask a meta-level question about the field: will continued progress be driven faster by increasing amounts of training data or the development of better object detection models?

To answer this question, we collected a massive training set that is an order of magnitude larger than existing collections such as PASCAL [4]. We follow the dominant paradigm of scanning-window templates trained with linear SVMs on HOG features [1, 2, 5, 6], and evaluate detection performance as a function of the amount of positive training data (N) and the model complexity (K), where K is measured by the amount of mixture components capturing variations in object sub-categories, 3D viewpoint, etc.

We found there is a surprising amount of subtlety in scaling up training data sets in current systems. For a given model, one would expect performance to generally increase with the amount of data, but eventually saturate. Empirically, we found the bizarre result that off-the-shelf implementations often decrease in performance with additional data! One would also expect that to take advantage of additional training data N, it is necessary grow the model complexity K. However, we often found scenarios in which performance was relatively static even as model complexity and training data grew (Fig 2).

In this paper, we offer explanations and solutions for these phenomena. First, we found it crucial to set model regularization as a function of the amount of training data N using cross-validation, a standard technique not typically deployed in current object detection systems. Second, existing strategies for discovering subcategory structure, such as clustering aspect ratios [5] and appearance features [3] may not suffice. We found this was related to the inability of classifiers to deal with "polluted" data when mixture labels were improperly assigned (Fig. 3). Increasing model complexity K is thus only useful when mixture components capture the "right" sub-category structure (Fig. 4). Finally, we found that is was easier to capture the "right" structure with compositional representations; we show that one can implicitly encode an exponentially-large K by composing parts together, yielding substantial performance gains over explicit mixture models (Fig.5). We conclude that there is currently little benefit to simply increasing training dataset sizes. But there may be significant room to improve current representations and learning algorithms, even when restricted to existing feature descriptors.

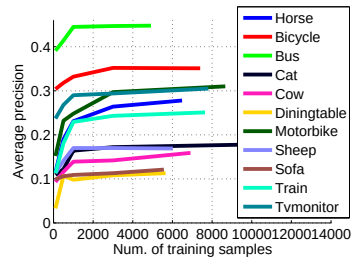


Figure 2: We plot the best performing mixture-models at varying amount of training data for 11 PASCAL categories. All the curves saturate with a relatively small amount of training data. In this work, we analyze how these apparant limits on performance can be broken.

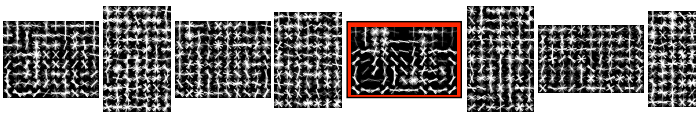


Figure 3: A single clean `bicycle` template (marked with red) alone achieves ap=29.4%, which is almost equivalent to the performance of using all 8 mixtures (ap=29.7%). This suggests that scaling up model complexity by simply adding additional mixture components may not suffice. Both models strongly outperform a single-mixture model trained on the full training set. This suggests that SVMs are sensitive to noisy examples, and one should train with "clean" data that does not pollute a template.

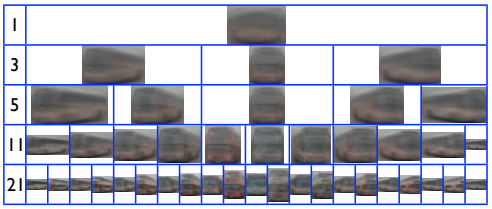


Figure 4: We describe supervised methods for hierarchically structuring data. In this case, we learn separate mixture components corresponding to `bus` viewpoints and object type (single vs double-decker).

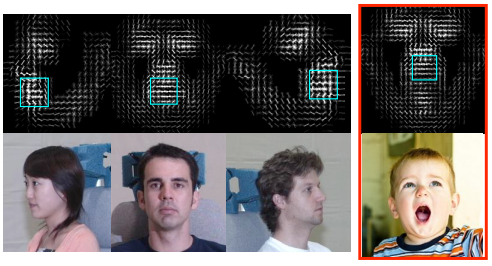


Figure 5: We describe two methods for increasing the performance of mixture models. First, we share spatially-localized regions (the blue "part") between mixture components, shown on the [left]. Second, we allow parts to be composed in novel spatial arrangements not seen in the training data [right]. These modifications define a spectrum of representations between classic mixture models and deformable part models.

[1] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR 2005*.

[3] S. Divvala, A. Efros, and M. Hebert. How important are deformable parts in the deformable parts model? *CoRR*, abs/1206.3714, 2012.

[4] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 2010.

[6] T. Malisiewicz, A. Gupta, and A.A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

An Object Co-occurrence Assisted Hierarchical Model for Scene Understanding

Xin Li
xinli@temple.edu
Yuhong Guo
yuhong@temple.edu

Computer and Information Science
Temple University
Philadelphia
PA 19122, USA

Historically, there is a controversy between cognitive psychology and computer vision on the task of scene recognition, the main source of which is about achieving scene recognition using low-level features to directly capture the gist of a scene versus using intermediate semantic representations [1]. Following this controversy, two main directions have been explored on this task. One attempts to use supervised classifiers that directly operate on low-level image features such as color, texture, and shape [4]. The other direction ventures to bridge the gap between low-level image properties and the semantic content of a scene using intermediate semantic representations that can be obtained by processes such as segmentation and object recognition [2, 3]. In this work, we propose a novel three-level (superpixel level, object level and scene level) generative hierarchical model for scene understanding, which does not require tedious object annotations over the training data.

The proposed hierarchical probabilistic graphical model, shown in Figure 1, integrates both low-level representations and intermediate semantic modeling to explain an image from three different levels: the superpixel level, the object level and the scene level. It captures the high-level contextual information expressed in form of object co-occurrences using a probabilistic chain structure over the object class assignment variables in each image. The rationale behind this design is to capture the possible object category correlation information without inducing more complicated inference problems.

In the model setting, the total number of different objects for the whole image set is assumed to be known. But as an unsupervised model at the object level, it does not require the object annotations to be provided. Instead, object annotation will be accomplished implicitly as an intermediate result in our approach. In particular, object classes are not pre-associated with fixed human defined concepts (*e.g.* desk, computer, and sky), but are simply represented using consecutive index integers from 1 to the number of object classes, which is 30 in our experiments. Unsupervised learning at the object-level is expected to automatically capture useful concepts for each object class.

Following the proposed model, the resulting joint distribution of a given scene with class C, the appearances of object classes, O_1, O_2, …, O_n, the objects $\mathbf{tO}$, the region features $\mathbf{tR}$, and the image patch features $\mathbf{tX}$ can be expressed as:

$$\begin{aligned} &P(C, O_1, O_2, \ldots, O_n, \mathbf{tO}, \mathbf{tR}, \mathbf{tX} | \alpha, \beta, \gamma, \eta) = P(C) \cdot P(O_1 | C, \eta_1) \\ &\times \prod_{i=2}^{N_o} P(O_i | O_{i-1}, C, \eta_i) \cdot \prod_{l=1}^{Nr} (P(tO_l | O_1, O_2, \ldots, O_n, \gamma) \\ &\times \prod_{k=1}^{N_F} p(tR_{lk} | tO_l, \alpha_k) \cdot \prod_{m=1}^{N_p} P(tX_{lm} | tO_l, \beta)) \end{aligned} \tag{1}$$

To learn the model parameters automatically, we derive a collapsed Gibbs sampling algorithm. Details are discussed in the paper.

With the trained model, we predict the most likely scene class for an image from the new test image set. We use the visual components of the proposed model to compute the posteriori probability of each scene class by integrating out the latent object variable Os and tOs:

$$\begin{aligned} &P(C = c | \mathbf{tR}, \mathbf{tX}) = \frac{P(C = c, \mathbf{tR}, \mathbf{tX})}{P(\mathbf{tR}, \mathbf{tX})} \\ &\propto \prod_{n=1}^{N_r} \sum_{\{O_1, O_2, \ldots, O_{N_o}\}} \big(P(O_1, O_2, \ldots, O_{N_o} | C = c) \cdot \\ &\sum_{o} P(\mathbf{tR}_n | tO_n = o) \cdot P(\mathbf{tX}_n | tO_n = o) \cdot P(tO_n = o | O_1, \ldots, O_{N_o})\big) \end{aligned} \tag{2}$$

The most likely scene class can then be determined as:

$$c^* = \arg\max_{c \in \mathcal{C}} P(C = c | \mathbf{tR}, \mathbf{tX}) \tag{3}$$

Moreover, we exploit an ensemble prediction strategy by training multiple models and take the majority vote of the multiple models as the final scene label of the test image.

The proposed model is evaluated on the LabelMe dataset, comparing to a golden standard method OB+SVM, which used supervised object detectors, and two other baselines, the models without either the ensemble strategy or the chain structure. The test accuracy results are presented in Table 1, which show the proposed approach is almost as good as the OB+SVM, and produces consistent superior performances over the other two baseline methods.

Method	Proposed	w/o Ensemble	w/o Chain	OB+SVM
bathroom	0.765	0.604	0.565	0.938
bedroom	0.704	0.673	0.573	0.568
airport	0.676	0.638	0.584	0.459
coast	0.920	0.875	0.607	0.534
corridor	0.786	0.757	0.550	0.964
livingroom	0.471	0.464	0.447	0.765
office	0.938	0.822	0.675	0.938
park	0.660	0.749	0.630	0.849
speech	0.769	0.592	0.438	0.846
street	0.718	0.688	0.425	0.875
Average	0.741	0.686	0.549	0.774

Table 1: Comparison results of scene classification.

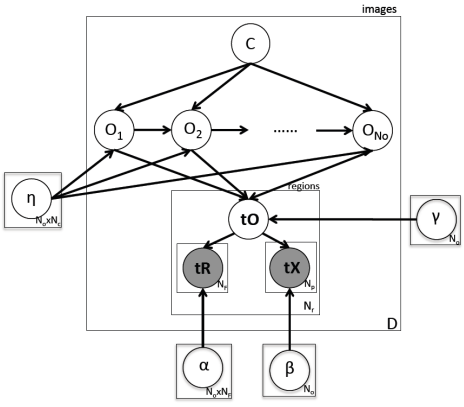


Figure 1: The proposed model. Nodes denote random variables and edges indicate dependencies. The variables at the right lower corner of each box denote the numbers of replications. The box indexed by D represents a single image in the image set of size D. The box indexed by N_r denotes the visual information of the image. N_c, N_o, N_F, and N_p denote the number of different scenes, objects, region features, and patches respectively. $\alpha, \beta, \gamma, \eta$ are the parameters of the distributions associated with the variables. We omitted the distribution hyperparameters for clarity's sake.

[1] P. Espinace, T. Kollar, A. Soto, and N. Roy. Indoor scene recognition through object detection. In *Proceedings of ICRA*, 2010.

[2] E. Xing L. Li, H. Su and F. Li. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Proceedings of NIPS*, 2010.

[3] L. Li, R. Socher, and F. Li. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Proceedings of CVPR*, 2009.

[4] I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *Proceedings of ICRA*, 2000.

Efficient Kernels Couple Visual Words Through Categorical Opponency

Ioannis Alexiou

Anil Anthony Bharath
http://www.bg.ic.ac.uk/research/a.bharath/

Biologically Inspired Computer Vision
Department of Bioengineering
Imperial College London, UK

Vision systems, designed for tasks such as object recognition and categorization, are based on a series of well-defined processing stages. Typically, these stages will include keypoint detection schemes, which sample a dense, transform-domain representation of an image. Commonly used keypoints detectors are Harris, Hessian, MSER and DoG, which are able to produce a sparsified representation of an image, although recent work has shown that dense sampling can drastically increase the recognition performance. A "selection" mechanism may also be employed to sparsify the dense descriptor representation where combined sparse coding and max-pooling 'distil' the features over defined image regions.

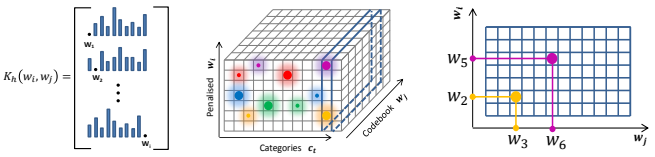
There are numerous methods that can bind those mid-level features into a compact representation, whilst ensuring that spatial information is ecoded: spatial pyramids, neural networks and other, high-order of spatial features. Spatial pyramids incorporate coarse spatial relationships producing good recognition rates. Yet another efficient approach binds descriptors into multiplets of features yielding much improved recognition performance compared to a bag of features approach. The latter approach attempts to encode spatial relationships using relative distances (*Correspondence transform*). Co-occurences of these high-order features are mapped onto an offset space where occurence counts are assigned to the features which satisfy predefined distance criteria.

Inspired by the observation that binding descriptors may lead to quasi-contour construction, we may assume that parts of contours can be a flexible tool to improve recognition rates and pose invariance. Specifically, constellation methods have shown that fixed numbers of parts, which come from standard descriptors, is a powerful representation tool. Typical disadvantages of such approaches involve a predetermined, fixed number of parts where computationally expensive search assigns parts to locations and set a hypothesised center for the model. In this research we focus on whether a selection mechanism can reduce the dictionary size of high-order features. Specifically, we focus on the transition from visual words to 2_{nd}, 3_{rd}, $\ldots$, k_{th} visual word order, which increases the computational complexity by a factor of N^k where N is the size of the dictionary. This work proposes efficient kernels to create small dictionaries of paired words utilizing *categorical opponency* to unveil such pairs. Secondly, we examine how proximity of such pairs can be scalable, and how the proximity of pairing affects performance. Thirdly, coupling kernels are applied per image using decision functions to detect co-occurrences using an approach that is compatible with fast indexing.

Dictionaries of higher-order (than simple visual words) can raise the size of the dictionary exponentially. Even paired words can yield numerous combinations depending on the size of the initial codebook. Among these numerous pairs a subset can exist that makes the problem tractable. In addition it is assumed that specific pairs are very unique to each class and others non-informative. Considering the aforementioned, a data-driven method is described to mine those pairs from the train set given the ground truth of the object classes.

$$K_h(w_i, w_j) = P(w_j|w_i) \cdot G_i \tag{1}$$

The kernelized expression (1) provides a statistical medium to monitor specific visual words down-weighted by the "G_i" term. We look at the probability that a keyword occurs conditional on another specified word, and estimate a function similar to a joint probability of occurrence. Specifically, a histogram of visual words is constructed by removing the candidate w_t. By removing the w_t we monitor the co-occurrence of other(non-identical) words in the image. Often the word w_t can be assigned to several keypoints then the votes are accumulated to the correspondent row $P(w_j|w_t)$ in the kernel $K_h(w_t, w_j)$. This procedure is done per object then the kernels are L1 normalized and arranged according to object classes. The creation of these kernels is not much slower than computing histograms of visual words. The words are assigned once for every image, then the look up is achieved by indexed retrieval

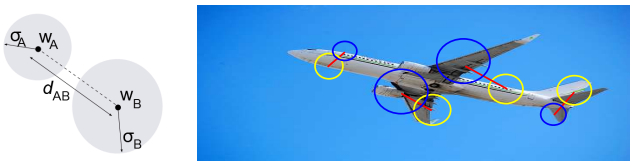


(a) (b) (c)

Figure 1: Starting from left to right: (a) illustrates how a kernel may be constructed using conditional histograms. (b) This figure shows the concept of first and second maxima. Each colour represents a different pair, where the size of the coloured/shaded blobs illustrates the 1^{st} and 2^{nd} maxima. (c) Once a candidate category is selected, we focus on the maximum votes that belong to this category. Finally, these pairs are selected in order from highest to lowest ratios.

$$K_{Op}(w_i, w_j) = \frac{\max\limits_{c_t} K_h(w_i, w_j, c_t)}{\max\limits_{c_t' \neq c_t} \left[K_h(w_i, w_j, c_t') \leqslant \max\limits_{c_t} K_h(w_i, w_j, c_t) \right]} \tag{2}$$

The parameter c_t characterises an object label assigned to a normalized kernel. Presumably if a histogram of words can provide a rough discrimination among categories then pair might must exist enhancing this behaviour. Expression (2) is a analytic approach to detect which pairs have high dominance over a specific class. Searching along the categories $c_t \in \{1, \ldots, Total\,Categories\}$ the maximum of kernel entry is divided by the second maximum found in another category. This forms a ratio where the higher the ratio the higher the opponency of this class against the others. This means that for a given combination of visual words this tends to be unique to the examined class.



(a) (b)

Figure 2: (a) illustrates a word pair comprised of two words w_A and w_B. The associated keypoint information captures the effective descriptor (spatial) radius (σ_A and σ_B), and relative Euclidean distance (d_{AB}), which is used to derive the paired words. (b) illustrates the single word pair occurring 4 times; multiple occurrences such as this are used in Equation (3).

$$\mathcal{B}^{(p)} = \sum_{m|n=1}^{Q} \max_{n|m=1}^{Q} K_{N\times M}^{(p)}(\phi_{w_j}^{(n)} = w_j, \phi_{w_i}^{(m)} = w_i) \tag{3}$$

where Q is taken to be the smaller dimension of the $K_{N\times M}$ matrix, i.e. either M or N.

Experiments in Caltech 101 database using 30 training examples per class are summarised in the next table to show the amount of improvement using pairs of visual words.

	Pyramids (10500)	Pairs (5000)
Keypoint-Based	$52.30_{\pm 1.28}$	$71.00_{\pm 2.00}$
Grid-Based	$60.16_{\pm 1.56}$	$73.88_{\pm 2.10}$

Table 1: Average accuracies per method for Caltech 101.

In conclusion, pairing up visual words can increase the classification rates comparing to spatial pyramids with hard word assignments. Future aims are to explore alternative pairing kernels, merge spatial pyramids and sparse coding approaches with pairs of visual words.

Fast Line Description for Line-based SLAM

Keisuke Hirose
hikeisuke@hvrl.ics.keio.ac.jp
Hideo Saito
http://hvrl.ics.keio.ac.jp/saito/

Graduate School of Science and Technology,
Keio University

Simultaneous localization and mapping (SLAM) is a technique to simultaneously perform mapping of environments and localization of a camera in real-time. Vision based SLAM is used for real applications such as augmented reality. Most existing monocular vision based SLAM techniques employ point features as landmarks[3].

Our approach uses line segments rather than points as landmarks, since there are some advantages in using line segments. Images of artificial environments with little texture contain many line segments, whereas few point features can be localized in such a scene. Moreover, line segment matching is more robust than point matching with respect to partial occlusion and view-point changes. With regard to the SLAM based on line segments, existing line based SLAM systems[1, 4] don't have any descriptions of line segments. In order to establish 2D and 3D correspondences, they simply use the distance between line segments in the image space. Therefore, wrong correspondences often occur in complicated scenes that include many line segments.

We propose here a real-time SLAM system that uses line segments as landmarks, and a fast line descriptor (LEHF:Line-based Eight-directional Histogram Feature) in order to establish correct 2D and 3D correspondences.

In localization, line segments are detected by the line segment detector (LSD) method[2] at every frame. LEHF that is our new line descriptor is computed for each detected line segment. We based the development of LEHF on the mean standard deviation line descriptor (MSLD)[5], which uses a SIFT-like strategy. Fig.1 shows how LEHF is computed.

For each point that is uniformly taken around the line segment, the gradient vector is computed based on differential values (dx, dy). As shown in the figure, we obtain 14 eight-directional gradient histograms by summing 45 gradient vectors along the line segment. 14 computed eight-directional gradient histograms are merged to obtain a line descriptor referred to as LEHF. However, if we simply merge all the histograms, computed LEHFs are not matched between the images that one image is rotated 180 degrees since the directions of the eight-directional gradient histogram are not matched. Therefore 14 histograms are merged symmetrically.

In order to estimate a camera pose, 2D and 3D line correspondences are established. First 3D line segments are projected into the image space by a previous camera pose. In existing methods, the projected 3D line segment simply corresponds to a detected 2D line segment that is nearest from the projected line segment in the image space. This often results in wrong correspondences being detected. We compute LEHF distances between the projected 3D line segment and some detected 2D line segments. Then the detected 2D line segment that has the minimum Euclidean distance between LEHFs is chosen.

The RANSAC algorithm and the line-based orthogonal iteration algorithm (LBOI)[6] estimate a camera pose from 2D and 3D correspondences. All 3D line segments are re-projected by the estimated camera pose to compute re-projection errors. Each 3D line segment is determined inlier or outlier based on the computed re-projection error.

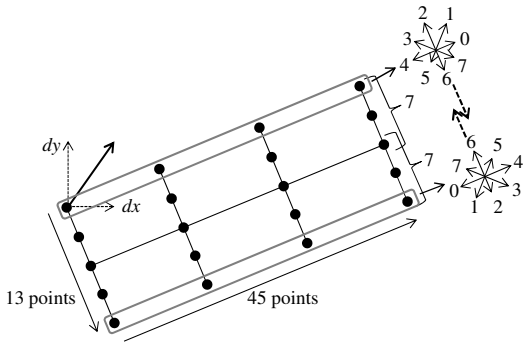


Figure 1: Overview of LEHF

In the mapping, line segments that are not used for localization are tracked between frames as new line segments. In order to track line segments correctly, computed LEHF is used. Tracked line segments in a number of frames are reconstructed for mapping 3D line segments.

We conducted experiments with our SLAM system. The experimental results of demonstrating our SLAM system in a desktop environment are shown in Fig. 2.

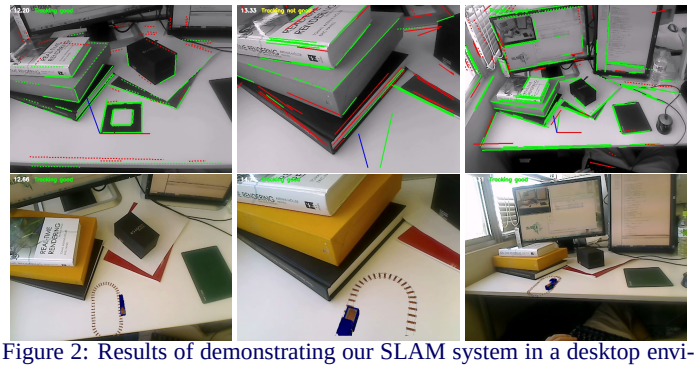


Figure 2: Results of demonstrating our SLAM system in a desktop environment.

The green line segments are inlier line segments and red line segments are outlier line segments. The mean processing time of 1087 frames was 87.78ms. Moreover, we evaluated our SLAM system by using synthetic data. In the synthetic data experiment, we compared our SLAM system with the existing approach using the nearest neighbor search. We show that the use of LEHF provides better accuracy of estimated camera poses.

We have presented a real-time SLAM system based on a line feature called a LEHF. By using LEHF, 2D and 3D correspondences are established correctly and the camera poses are robustly estimated. Our SLAM was demonstrated for augmented reality in a desktop environment and evaluated by using synthetic data.

[1] P.G. Andrew and W. Mayol-Cuevas. Real-time model-based slam using line segments. In *2nd International Symposium on Visual Computing*, 2006.

[2] R. Grompone von Gioi, J. Jakubowicz, J.M. Morel, and G. Randall. Lsd: A fast line segment detector with a false detection control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4): 722–732, 2010.

[3] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225–234. IEEE, 2007.

[4] P. Smith, I. Reid, and A. Davison. Real-time monocular slam with straight lines. In *British Machine Vision Conference*, volume 1, pages 17–26, 2006.

[5] Z. Wang, F. Wu, and Z. Hu. Msld: A robust descriptor for line matching. *Pattern Recognition*, 42(5):941–953, 2009.

[6] X. Zhang, K. Wang, Z. Zhang, and Q. Yu. A new line-based orthogonal iteration pose estimation algorithm. In *Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on*, pages 1–4. IEEE, 2009.

A local Rayleigh model with spatial scale selection for ultrasound image segmentation

Djamal Boukerroui
http://www.hds.utc.fr/~dboukerr

Université de Technologie de Compiègne
Heudiasyc UMR CNRS 7253
BP 20529 - 60205 Compiègne Cedex, France.

Ultrasound data are very noisy, with poor contrast, and often presents missing boundaries of the object of interest due to problems of specular reflection, shadows, signal dropout and attenuation. As a consequence, conventional intensity gradient-based methods have had limited success on typical clinical images [5]. Note also that segmentation methods based on global statistical models, regardless of the used framework, fail on this type of data, mainly because of the attenuation problem. Adaptive solutions robust to attenuation exist in the literature [1, 2, 5]. Local image statistics were used for the estimation of the segmentation model's parameters. Recently, there has been a reinvestigation of the use of local statistics by the image segmentation community, but in a variational framework [3, 4, 6]. These recent studies show a better behavior of these local models on images with strong intensity inhomogeneities. This contribution falls under this context.

First, we propose the adaptation of the model proposed by Sarti et al. [7]. The latter assumes a global Rayleigh model envelope image statistics. Let $I:\Omega\rightarrow\mathbb{R}^+$ denote a given observed image and $\mathcal{C}$ be a closed contour represented as the zero level set of a signed distance function ϕ. The interior Ω_i and the exterior Ω_e of $\mathcal{C}$ are defined by a smooth approximation of the Heaviside function respectively by: $H_i(\phi)=H(\phi)$ and $H_o(\phi)=1-H(\phi)$. We seek the partition of Ω that maximizes the likelihood function of the observed data. Given the independence assumption, this leads to the minimization of the following energy function [7]

$$E(\phi)=-\sum_{r\in\{i,o\}}\int_\Omega H_r(\phi)\log p(I(\mathbf{x}))d\mathbf{x}+\lambda\int_\Omega\delta(\phi)|\nabla\phi|d\mathbf{x}, \quad (1)$$

where the first two terms are the data terms and the last one is a length regularisation with a weight penalty λ. We will further assume that the random intensity $I(\mathbf{x})$ follows a Rayleigh pdf with a parameter σ^2:

$$p(I(\mathbf{x}))=\frac{I(\mathbf{x})}{\sigma^2}\exp\left(-\frac{I(\mathbf{x})^2}{2\sigma^2}\right) \text{ and } \widehat{\sigma^2}_{\mathrm{ML}}=\frac{\int_{\Omega_r}I(\mathbf{x})^2d\mathbf{x}}{2\int_{\Omega_r}d\mathbf{x}},$$

where $\widehat{\sigma^2}_{\mathrm{ML}}$ is a Maximum Likelihood estimates under the assumption that all the observed pixels in the domain Ω_r are identically distributed. In Sarti et al. [7], only two global domains were used, Ω_i for the inside and Ω_e for the outside pixels. Therefore the hypothesis of identically distributed observations is generally false for ultrasound images because of the presence of strong intensity inhomogeneities. However, the assumption remains true if the estimate is made locally in a region centered around each pixel of the domain Ω. Thus the energy corresponding to the inside term of (1) is given by:

$$E_i(\phi)=\int_\Omega H(\phi)\left[\frac{I(\mathbf{x})^2}{2\sigma_i^2(\mathbf{x})}+\log(\sigma_i^2(\mathbf{x}))\right]d\mathbf{x} \quad (2)$$

$$\text{and}\quad \sigma_i^2(\mathbf{x})=\frac{\int_\Omega H(\phi)K(\mathbf{x}-\xi)I(\xi)^2d\xi}{2\int_\Omega H(\phi)K(\mathbf{x}-\xi)d\xi}, \quad (3)$$

where $K(\cdot)$ is any given kernel defining the spatial locality around the position $\mathbf{x}$. Here, a Gaussian kernel with a standard deviation σ_K is used.

Local region-based segmentation models are surely a better alternative to global ones in the presence of intensity inhomogeneities. Such models however may be more sensitive to initialisation if the chosen local spatial scale is not appropriate. A decrease of robustness to noise is also observed when small scales are used. To our knowledge, two pixel dependent scale selection methods have been introduced recently [6, 8]. The second contribution of this paper is the proposition of a novel Intersection of Confidence Intervals (ICI) rule for the spatial scale selection. Our approach is based on the idea of choosing the largest scale that gives the best estimate of the segmentation model parameters. Specifically, in the presence of intensity inhomogeneities, the hypothesis of identically distributed data in the local window will become less and less valid as the scale of the kernel K grows and will lead to an increasingly biased estimations. This means that there exists a bias-variance balance that gives the ideal scale. We can make use of the ICI algorithm to search for the largest local window (minimising variance) that gives us the best estimate of σ^2 (minimising bias).

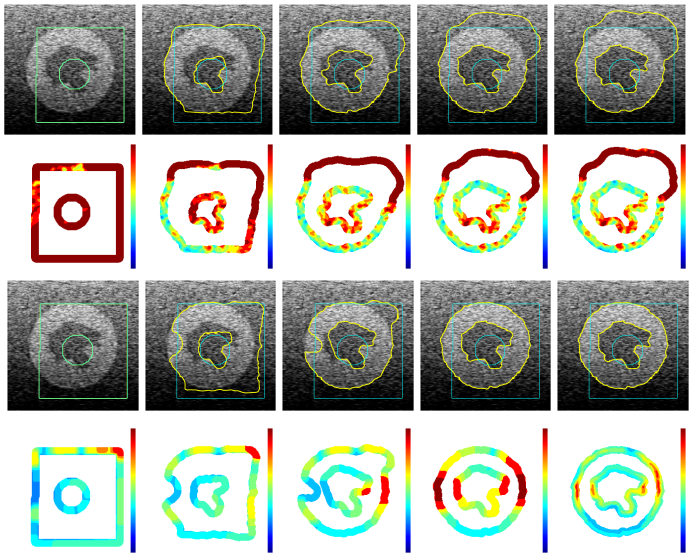


Figure 1: 1st & 3rd lines: Example of contour evolution of the local Rayleigh segmentation model for iterations 1, 10, 25, 75 and 120 respectively when using the scales estimated with [6] (2nd line) and with the proposed ICI rule (4th line). The colormap Blue-Red corresponds to scales from 5 to 80. Image size 256×256.

In order to demonstrate the usefulness of the proposed approach and quantify its performances, we chose to test it on realistic US simulations. To this end, we have used the simulation program Field-II, to synthesize phantom data with known ground truth. Two phantoms with two scatters amplitudes and three levels of tissue attenuations were simulated. We also used several dB ranges for the envelope logarithmic compression to simulate different image contrasts. A quantitative evaluation is then conducted on 240 images and statistics of the Dice similarity measure and the Mean Absolute Distance are shown. The results show the robustness and the superiority of the proposed segmentation approach in comparison to [3, 7]. The efficiency and the genericity of the proposed scale selection strategy is also demonstrated.

[1] E. A. Ashton and K. J. Parker. Multiple resolution bayesian segmentation of ultrasound images. *Ultrasonic Imag.*, 17(4):291–304, October 1995.

[2] D. Boukerroui, A. Baskurt, J.A. Noble, and O. Basset. Segmentation of ultrasound images–multiresolution 2D and 3D algorithm based on global and local statistics. *Pattern Recognit. Lett.*, 24:779–790, 2003.

[3] T. Brox and D. Cremers. On local region models and a statistical interpretation of the piecewise smooth Mumford-Shah functional. *Int. J. Comput. Vis.*, 84(2):184–193, 2009.

[4] S. Lankton and A. Tannenbaum. Localizing region-based active contours. *IEEE Trans. Image Process.*, 17(11):2029–2039, 2008.

[5] J. A. Noble and D. Boukerroui. Ultrasound image segmentation: A survey. *IEEE Trans. Med. Imag.*, 25(8):987–1010, 2006.

[6] J. Piovano and T. Papadopoulo. Local statistic based region segmentation with automatic scale selection. In *ECCV*, pages 486–499. Springer, 2008.

[7] A. Sarti, E. Mazzini, C. Corsi, and C. Lamberti. Maximum likelihood segmentation of ultrasound images with rayleigh distribution. *IEEE Trans. Ultra. Fer. Freq. Control*, 52(6):974–960, June 2005.

[8] Q. Yang and D. Boukerroui. Optimal spatial adaptation for local region-based active contours: An intersection of confidence intervals approach. In *IMAGAPP*, pages 87–93, Algarve, Portugal, March 5-7 2011.

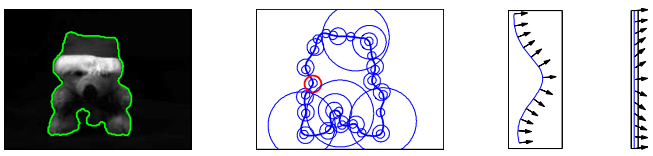
Object Matching Using Boundary Descriptors

Ognjen Arandjelović
ognjen.arandjelovic@gmail.com
Swansea University, UK

The problem of recognizing 3D objects from images has been one of the most active areas of computer vision research in the last decade. This is a consequence not only of the high practical potential of automatic object recognition systems but also significant breakthroughs which have facilitated the development of fast and reliable solutions. These mainly centre around the detection of robust and salient image loci (keypoints) or regions, and the characterization of their appearance (local descriptors). While highly successful in the recognition of textured objects even in the presence of significant viewpoint and scale changes, these methods fail when applied on texturally smooth (i.e. nearly textureless) objects. Unlike textured objects, smooth objects inherently do not exhibit appearance from which well localized keypoints can be extracted.

Since their texture is not informative, characteristic discriminative information of smooth objects must be extracted from shape instead. Considering that it is not possible to formulate a meaningful prior which would allow for the reconstruction of an accurate depth map for the general class of smooth 3D objects, the problem becomes that of matching apparent shape as observed in images. This is a most challenging task because apparent shape is greatly affected by out of plane rotation of the object. What is more, the extracted shape is likely to contain errors when the object is automatically segmented out from realistic, cluttered images. The bag of boundaries (BoB) method of Arandjelović and Zisserman was the first to address this problem explicitly [1].

Boundary keypoint detection The problem of detecting characteristic image loci is well researched and a number of effective methods have been described in the literature. When dealing with keypoints in images, the meaning of saliency naturally emerges as a property of appearance (pixel intensity) which is directly measured. This is not the case when dealing with curves for which saliency has to be defined by means of higher order variability which is computed rather than directly measured. In this paper we detect characteristic boundary loci as points of local curvature maxima, computed at different scales. Starting from the finest scale after localizing the corresponding keypoints, Gaussian smoothing is applied to the boundary which is then downsampled for the processing at a coarser scale. Having experimented with a range of factors for scale-space steps, we found that little benefit was gained by decreasing the step size from 2 (i.e. by downsampling finer than one octave at a time). We estimate the curvature at a vertex as the curvature of the circular arc fitted to three consecutive boundary vertices: the vertex of interest, its predecessor, and its successor. An example of a boundary contour and the corresponding interest point loci are shown respectively in Figures 1(a) and 1(b). The method used to perform Gaussian smoothing of the boundary is explained next.



(a) Contour (b) Keypoints (c) Descriptor

Figure 1: (a) Original image of an object overlaid with the object boundary (green line), (b) the corresponding boundary keypoints detected using the proposed method and (c) an illustration of a local boundary descriptor based on the profile of boundary normals' directions (the interest point is shown in red in (b)).

Boundary curve smoothing. The most straightforward approach to smoothing a curve such as the object boundary is to replace each of its vertices by a Gaussian-weighted sum of vectors corresponding to its neighbours. However, this method introduces an undesirable artefact which is demonstrated as a gradual shrinkage of the boundary. In the limit, repeated smoothing results in the collapse to a point – the centre of gravity of the initial curve. We solve this problem by applying two smoothing operations, with the second update to the boundary vertices being applied in the "negative" direction and weighted by a constant such that in the limit repeated smoothing does not change the circumference of the boundary.

Local boundary descriptor Following the detection of boundary keypoints, our goal is to describe the local shape of the boundary. After experimenting with a variety of descriptors based on local curvatures, angles and normals, using histogram and order preserving representations, we found that the best results are achieved using a local profile of boundary normals' directions.

To extract a descriptor, we sample the boundary around a keypoint's neighbourhood (at the characteristic scale of the keypoint) at n_s equidistant points and estimate the boundary normals' directions at the sampling loci. This is illustrated in Figure 1(c). For each sampling point, a circular arc is fitted to the closest boundary vertex and its two neighbours, after which the desired normal is approximated by the corresponding normal of the arc, computed analytically. The normals are scaled to unit length and concatenated into the final descriptor with $2n_s$ dimensions. After experimenting with different numbers of samples, from as few as 4 up to 36, we found that our method exhibited little sensitivity to the exact value of this parameter. For the experiments in this paper we used $n_s = 13$.

We apply this descriptor in the same way as Arandjelović and Zisserman did theirs, or indeed a number of authors before them using local texture descriptors. The set of training descriptors is first clustered, the centre of each cluster defining the corresponding descriptor word. An object is then represented by a histogram of its descriptor words. Since we too do not encode any explicit geometric information between individual descriptors we refer to our representation as a bag of normals (BoN).

Evaluation. To evaluate the effectiveness of the proposed method we used the publicly available *Amsterdam Library of Object Images* and performed three experiments:

- We compared the BoB and BoN representations in terms of their robustness to viewpoint change. The representations of all 1000 objects learnt from a single view were matched against the representations extracted from viewpoints at 5–85° yaw difference. Each object image was used as a query in turn.
- We compared the BoB and BoN representations in terms of their robustness to segmentation errors. The representations of all 1000 objects learnt from a single view were matched against the representations extracted from the same view but using distorted segmentation masks. In this experiment we distorted the segmentation mask by morphological erosion using a 3×3 'matrix of ones' structuring element. As before, each object image was used as a query in turn.
- We compared the BoB and BoN representations in terms of their robustness to segmentation errors. This time we distorted the segmentation mask by morphological dilation using a 3×3 'matrix of ones' structuring element. As before, each object image was used as a query in turn.

Overall, the performance of the BoB and BoN representations was found to be similar. Some advantage of the BoN was observed in rank-1 matching accuracy: each 5° change in yaw can be estimated to decrease the BoB performance by approximately 12% and the BoN performance by approximately 10%. In the second and third experiments the superiority of the proposed BoN representation was more significant. For example, the distortion of the segmentation mask by two erosions reduces the rank-1 matching rate of the BoB by 30% and that of the BoN by half that i.e. 15%. The negative effects of dilation of the mask were less significant for both representations but qualitatively similar: repeated twice, dilation reduces the rank-1 matching rate of the BoB by 25% and that of the BoN by only 10%.

[1] R. Arandjelović and A. Zisserman. Smooth object retrieval using a bag of boundaries. *In Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 375–382, November 2011.

Hash-Based Support Vector Machines Approximation for Large Scale Prediction

Saloua Litayem
http://www-rocq.inria.fr/~litayem
INRIA Paris-Rocquencourt, France

Alexis Joly
www-sop.inria.fr/members/Alexis.Joly
INRIA Sophia-Antipolis, France

Nozha Boujemaa
http://pages.saclay.inria.fr/nozha.boujemaa
INRIA Saclay, France

How-to train effective classifiers on huge amount of multimedia data is clearly a major challenge that is attracting more and more research works across several communities. Less efforts however are spent on the counterpart scalability issue: how to apply big trained models efficiently on huge non annotated media collections ? In this paper, we address the problem of speeding-up the prediction phase of linear Support Vector Machines via Locality Sensitive Hashing. We propose building efficient hash-based classifiers that are applied in a first stage in order to approximate the exact results and filter the hypothesis space. Experiments performed with millions of one-against-one classifiers show that the proposed hash-based classifier can be more than two orders of magnitude faster than the exact classifier with minor losses in quality.

Let $h(\mathbf{x})$ be a trained linear SVM classifier defined as

$$h(\mathbf{x}) = sgn\left(\omega.\mathbf{x} + b\right) \tag{1}$$

We suppose that all features $\mathbf{x} \in \mathbb{R}^d$ are L_2-normalized, so that $\|\mathbf{x}\| = 1$. In addition, let us denote as $\mathcal{F}$, a family of binary hash functions $f : \mathbb{R}^d \rightarrow \{-1, 1\}$ such that:

$$f(\mathbf{x}) = sgn\left(\mathbf{w}.\mathbf{x}\right)$$

if $\mathbf{w} \in \mathbb{R}^d$ is a random variable distributed according to $p_w = \mathcal{N}(0, \mathbf{I})$, we get the popular LSH function family sensitive to the inner product. In this case, for any two points $\mathbf{q}, \mathbf{v} \in \mathbb{R}^d$ we have:

$$Pr\left[f(\mathbf{q}) = f(\mathbf{v})\right] = 1 - \frac{1}{\pi} cos^{-1}\left(\frac{\mathbf{q}.\mathbf{v}}{\|\mathbf{q}\|\,\|\mathbf{v}\|}\right) \tag{2}$$

The basic idea of our Hash based SVM classifier is that the inner product between $\mathbf{x}$ and ω can actually be estimated by a Hamming distance between their respective hash codes $\mathbf{F}_D(\mathbf{x})$ and $\mathbf{F}_D(\omega)$, each being composed of D binary hash functions in $\mathcal{F}$.

Definition - Hash based SVM classifier
For any linear SVM classifier $h(\mathbf{x}) = sgn\left(\omega.\boldsymbol{x} + b\right)$, and a hash function family $\mathcal{F}$, we define a Hash-based SVM classifier as :

$$\begin{cases} \hat{h}(\mathbf{x}) = sgn\left(r_{\omega,b} - d_H\left(\mathbf{F}_D(\mathbf{x}), \mathbf{F}_D(\omega)\right)\right) \\ r_{\omega,b} = \frac{D}{\pi} cos^{-1}\left(\frac{-b}{\|\omega\|}\right) \end{cases} \tag{3}$$

with $d_H\left(\mathbf{F}_D(\mathbf{x}), \mathbf{F}_D(\omega)\right)$ being the Hamming distance between $\mathbf{x}$ and ω.

Applying a Hash-based SVM classifier $\hat{h}(\mathbf{x})$ with a brute-force scan instead of the exact classifier $h(\mathbf{x})$ does not change the prediction complexity, which is still $O(N)$ in the number of images to classify. Performance gains are more related to memory usage and the overall speed when a very large number of classifiers have to be applied simultaneously (which is often the case when dealing with a large number of classes). The second main advantage is to speed up the computation of the classification function. A Hamming distance on typically $D = 256$ bits can be much faster than an inner product on high-dimensional data with a double precision (particularly when benefiting from *pop-count* assembler instructions).

We propose a filter-and-refine strategy for approximating a one-against-one linear multi-class SVM with a large number of categories and a large dataset to be classified. We consider a dataset $\mathcal{X}$ of N feature vectors $\mathbf{x}$ in $\mathbb{R}^d$ that needs to be classified efficiently across a set $\mathbf{C}$ of K classes c_k. We then consider a one-against-one linear multi-class SVM $H(\mathbf{x})$ that is assumed to have been trained to solve this classification problem. $H(\mathbf{x})$ is defined as

$$H(\mathbf{x}) = \underset{c_k \in \mathbf{C}}{\arg\max} \#\left\{h_{k,j} \mid h_{k,j}(\mathbf{x}) = 1\right\} \tag{4}$$

where $h_{k,j}$ represents the $K(K-1)/2$ one-against-one classifiers (c_k vs c_j) defined by:

$$h_{k,j}(\mathbf{x}) = sgn\left(\omega_{k,j}.\mathbf{x} + b_{k,j}\right) \tag{5}$$

Thanks to our hash-based SVM approximation method, each $h_{k,j}$ can be approximated by an efficient hash-based binary classifier $\hat{h}_{k,j}(\mathbf{x})$ such that:

$$\hat{h}_{k,j}(\mathbf{x}) = sgn\left(\hat{\kappa}(\mathbf{x}, \omega_{\mathbf{k,j}})\,\|\omega_{\mathbf{k,j}}\| + b\right) \tag{6}$$

And finally, a hash-based multi-class SVM (HBMS) can be computed as

$$\hat{H}(\mathbf{x}) = \underset{c_k \in \mathbf{C}}{\arg\max} \#\left\{\hat{h}_{k,j} \mid \hat{h}_{k,j}(\mathbf{x}) = 1\right\} \tag{7}$$

Figure 1 illustrates the mean accuracy of HBMS vs that of the exact multi-class classifier for a varying number of k kept classes in the filtering step and an increasing number of bits. The accuracy improvement might be a result of a better generalization ability owing to refinement step with the best top-k classes obtained with our SVM approximation classifier in the filtering step. According to results, small hash codes and a small number of classes can be used to approximate the exact classifier well.

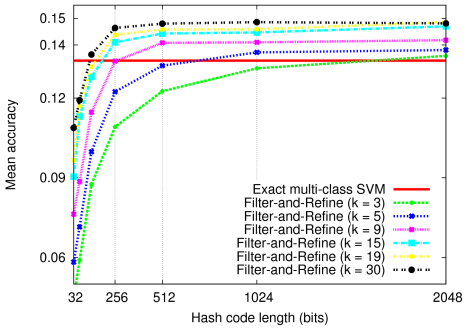


Figure 1: Exact Multi-class SVM vs Filter-And-Refine method

Figure 2 shows the average processing time per image for the filter-and-refine method for varying rates of passed classes to the refinement step and hash code lengths. If moderate losses in quality are tolerated with typically $k = K/100$ and $D = 512$ bits, then the cost of the whole filter-and-refine strategy is roughly equal to the cost of the filtering step and the whole filter-and-refine strategy is 110 times faster.

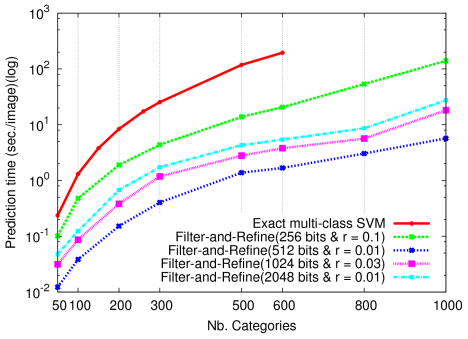


Figure 2: Filter-and-refine prediction time vs exact multi-class SVM prediction time

Leveraging over prior knowledge for online learning of visual categories

Tatiana Tommasi[12]
ttommasi@idiap.ch
Francesco Orabona[3]
francesco@orabona.com
Mohsen Kaboli[1]
mkaboli@idiap.ch
Barbara Caputo[1]
bcaputo@idiap.ch

[1] Idiap Research Institute, Martigny, CH
[2] École Polytechnique Fédérale EPFL Lausanne, CH
[3] Toyota Technological Institute Chicago, USA

The underlying main goal of all research in visual recognition is to enable vision-based artificial systems to operate autonomously in the real world. However, even the best system we can currently engineer is bound to fail whenever the setting is not heavily constrained. This is because the real world is generally too complicated and too unpredictable to be summarized within a limited set of specifications. This calls for algorithms able to support open ended learning of visual classes which can process continuously new data guided by past experience. The main issues of open ended learning has been typically addressed in a fragmented fashion in the literature. A first component is that of transfer learning, i.e. the ability to leverage over prior knowledge when learning a new class, especially in presence of few training data [3]. A second component is that of updating the learned visual class, as new samples arrive sequentially. The dominant approach in the literature here is that of online learning [1]: predictions are made on the fly and the model is progressively updated at each step, on the basis of the given true label. In this paper we propose to merge together these two components, using prior knowledge sources for initializing the online learning process on a new target task through transfer learning.

We consider binary object-vs-background problems where each image is represented by a vector $\boldsymbol{x} \in \mathbb{R}^d$ associated to a unique label $y \in \{-1,1\}$ and the prediction mechanism is based on a hyperplane which divides the instance space into two parts. This hyperplane is defined by its orthogonal vector $\boldsymbol{w} \in \mathbb{R}^d$ and the predicted label is given by $\text{sign}(\boldsymbol{w}\cdot\boldsymbol{x})$. We assume without loss of generality that $\|\boldsymbol{x}_t\| \leq 1$ and we define the hinge loss with margin 1 of a classifier $\boldsymbol{w}$ over an instance / label pair $(\boldsymbol{x},y)$ as $\ell^H(\boldsymbol{w}\cdot\boldsymbol{x},y) = \max\{0, 1-y\boldsymbol{w}\cdot\boldsymbol{x}\}$.

We adopt the Passive Aggressive (PA) algorithm [2] as our basic online learning method. A sequence of instances are presented to the learner $\boldsymbol{x}_t$, $t = 1,\ldots,T$ which generates the corresponding prediction and then receives the true label y_t which is used to update its hypothesis for future trials. Starting from an arbitrary hypothesis, $\boldsymbol{w}_1$, at the $t-$th round PA is updated solving the following optimization problem

$$\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}}{\text{argmin}} \frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2 + C\xi \quad \text{s.t.} \quad \ell^H(\boldsymbol{w}\cdot\boldsymbol{x}_t, y_t) \leq \xi \quad \text{and} \quad \xi \geq 0\,, \tag{1}$$

where C is the aggressiveness parameter that trades off the two quantities in (1).

Among the existing transfer learning approaches we consider the Multi-KT algorithm [4]. We suppose to have k binary source tasks and a discriminative model learned for each of them in terms of a linear function $h_j(\boldsymbol{x}) = \hat{\boldsymbol{w}}_j \cdot \boldsymbol{x}$ for $j = 1,\ldots,k$. For a novel target task with T available training samples $(\boldsymbol{x}_t, y_t)$ $t = 1,\ldots,T$, Multi-KT solves the following optimization problem [4]:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\left\|\boldsymbol{w} - \sum_{j=1}^{k}\beta_j\hat{\boldsymbol{w}}_j\right\|^2 + \frac{C}{2}\sum_{t=1}^{T}(y_i - \boldsymbol{w}\cdot\boldsymbol{x}_t - b)^2\,. \tag{2}$$

Here the weights β_j assigned to each prior knowledge are found by minimizing $\sum_{t=1}^{T}\ell^H(\tilde{y}_t, y_t)$ subject to $\|\boldsymbol{\beta}\|_2 \leq 1$, where $\tilde{y}_t$ is the leave-one-out prediction for the $t-$th sample, and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_k)$.

Thus we define a learning algorithm based on two phases: at the beginning n target training samples are given as input to Multi-KT which outputs the corresponding target model, and as second step, this model is used to initialize the online learning process. This has several advantages: by using a principled transfer learning process we can study the relation between the old sources and the new target. Within this framework, few samples might be sufficient to indicate in which part of the original space the correct solution (the best in term of generalization capacity) should be sought. At the same time, by using the transfer process only at the beginning we limit its computational burden. Then PA guarantees that the updated solution is at each step close to the previous one: this helps keeping the positive effect produced by Multi-KT together with the proper introduction of new information when necessary. We show theoretically that a good initialization for the online learning process produces a tighter mistake bound compared to previous work (OTL [5]), while empirically improving the recognition performance on an unseen test set. We name this algorithm TROL: TRansfer initializes Online Learning and we also consider the possibility to reweight at each step prior and new knowledge defining the variant TROL+.

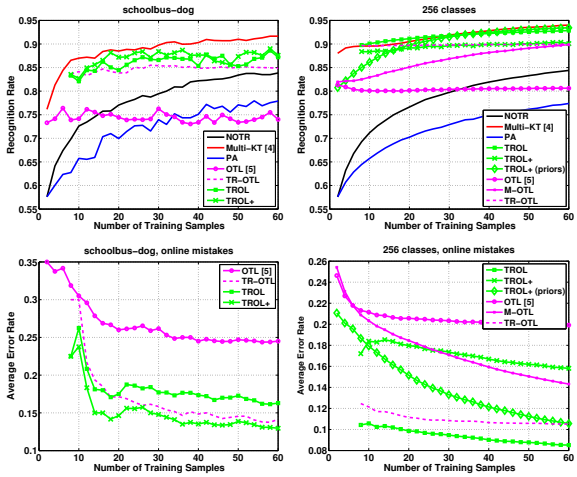


Figure 1: Left column: single source experiments. Right column: 255 object classes are considered as sources, while the remaining one defines the target problem. Top line: recognition rate results on the test set as a function of the number of training samples. Bottom line: corresponding rate of mistakes for the online learning methods.

We ran experiments on the Caltech 256 database selecting related / unrelated object classes and one or multiple prior knowledge sources, beside considering the full dataset (see Figure 1). Over all the experiments TROL and TROL+ present better results than PA, never showing negative transfer, and they are able to match the batch performance of Multi-KT on the test set. In terms of online mistakes, TROL and TROL+ outperform all the other considered baselines.

[1] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[2] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585, 2006.

[3] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

[4] T. Tommasi, F. Orabona, and B. Caputo. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[5] P. Zhao and S. C. H. Hoi. OTL: A Framework of Online Transfer Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

Unsupervised Texture Segmentation using Active Contours and Local Distributions of Gaussian Markov Random Field Parameters

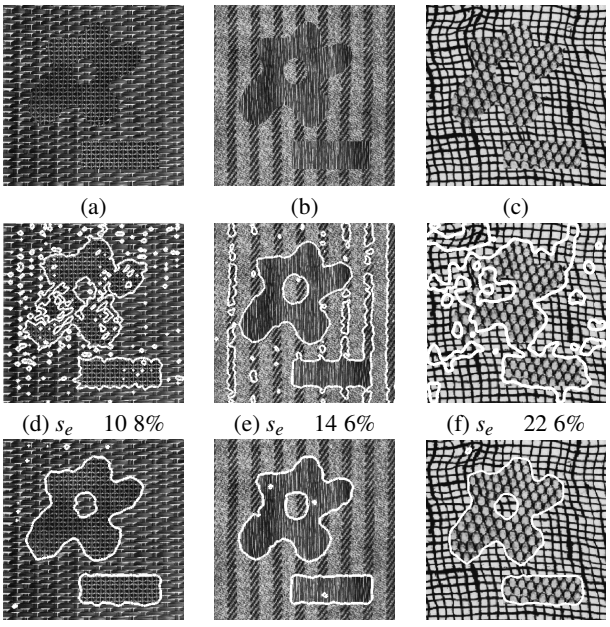
Chathurika Dharmagunawardhana[1]
cd6g10@ecs.soton.ac.uk
Sasan Mahmoodi[1]
sm3@ecs.soton.ac.uk
Michael Bennett[2]
michael.bennett@soton.ac.uk
Mahesan Niranjan[1]
mn@ecs.soton.ac.uk

[1] School of Electronics and Computer Science, University of Southampton, Southampton, UK

[2] National Institute for Health Research, Southampton Respiratory Biomedical Research Unit, University Hospital Southampton NHS Foundation Trust, Tremona Road, Southampton, UK

Gaussian Markov Random Fields (GMRF) have been exploited for modeling textures and extracting effective texture features [1, 2]. Model parameter estimates of low order GMRF have been widely used as the conventional texture feature for texture image segmentation [6]. The drawbacks of these features are firstly, their discriminative ability highly depends on model selection, yet is restricted to low order models due to computational concerns [2, 5]. Secondly, sufficiently large estimation windows should be selected to well characterize the given texture yet compromising accurate boundary localization [6]. Also the fact that the estimated model parameters obey a certain probability distribution for a given texture [3], has never been exploited when obtaining these features.

In this paper, instead of using model parameters as texture features, we exploit the variations in low order GMRF parameter estimates, obtained through model fitting in local region around the given pixel. A spatially localized estimation process is carried out by using a moderately small estimation window and modeling partial texture characteristics belonging to the local region through maximum likelihood method. Hence the estimated values inherit significant fluctuations, spatially, which can be related to texture pattern complexity. The variations occur in estimates are quantified by normalized local histograms, maintaining simplicity and efficiency and are named as PL histograms here. Selection of an accurate window size for histogram calculation is very important for a better segmentation. Since the variations occurred in estimates have a correlation with the texture pattern, the correct window size is assumed to be nearly equal to the average texture pattern size of the image. It is found via a method based on the entropy of the texture image.



(a) (b) (c)
(d) s_e 10 8% (e) s_e 14 6% (f) s_e 22 6%
(g) s_e 1 68% (h) s_e 1 09% (i) s_e 2 54%

Figure 1: Segmentation of large and medium size texture patters with low order GMRF. (a)-(c) original images, (d)-(f), conventional GMRF features [1] and (g)-(i) PL histogram. s_e segmentation error.

The novel features capture the variations that occur in model parameters which provide useful information for texture segmentation. In CGMRF features these important features are smoothed out by the estimation process. Formulation of PL histogram involves using small neighborhood sizes and estimation window sizes, hence giving lower computational cost and better boundary localization. They extend the possibility of using low order GMRF for segmenting fine to very large texture patterns (figure 1) and also improving the segmentation of textures with close characteristics (figure 2).

(a) (b) (c) s_e 17 9% (h) s_e 7 03%

Figure 2: Segmentation of images with close component textures. (a) original images, (b) segmentation target, (c) Gabor features and (d) PL histograms.

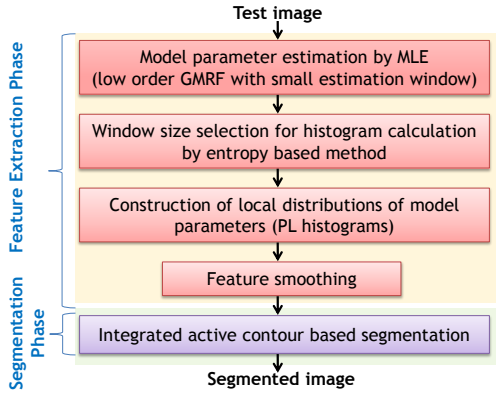


Figure 3: Proposed texture feature extraction and segmentation algorithm.

Extracted features are smoothed using diffusion via Beltrami flow and directed to the integrated active contour model [4] for unsupervised texture segmentation. The proposed method is illustrated in figure 3. Experimental results on statistical and structural component textures show improved discriminative ability of the features compared to some recent algorithms in the literature.

[1] S. Mahmoodi and S. Gunn. Snake based unsupervised texture segmentation using GMRF models. In *In Proc. ICIP*, pages 1–4, 2011.

[2] B. S. Manjunath and R. Chellappa. Unsupervised texture segmentation using Markov random field models. *IEEE PAMI*, 13(5):478–482, 1991.

[3] M. Petrou and P. G. Sevilla. *Image Processing, Dealing with Texture*. John Wiley Sons Ltd, ISBN: 0470026286, 2006.

[4] C. Sagiv, N. A. Sochen, and Y. Zeevi. Integrated active contours for texture segmentation. *IEEE IP*, 15(6):1633–46, 2006.

[5] S. Stan, G. Palubinskas, and M. Datcu. Bayesian selection of the neighbourhood order for Gauss-Markov texture models. *Pattern recognition letters*, 23:1229–1238, 2002.

[6] Y. Zhao, L. Zhang, P. Li, and B. Huang. Classification of high spatial resolution imagery using improved GMRF-based texture features. *IEEE GRS*, 45(5):1458–1468, 2007.

Spatial orientations of visual word pairs to improve Bag-of-Visual-Words model

Rahat Khan
rahat.khan@univ-st-etienne.fr
Cecile Barat
cecile.barat@univ-st-etienne.fr
Damien Muselet
damien.muselet@univ-st-etienne.fr
Christophe Ducottet
ducottet@univ-st-etienne.fr

Université de Lyon, F-42023, Saint-Etienne, France,
CNRS, UMR5516, Laboratoire Hubert Curien, F-42000, Saint-Etienne, France,
Université de Saint-Etienne, Jean Monnet, F-42000, Saint-Etienne, France.

This paper presents a novel approach to incorporate spatial information in the bag-of-visual-words (BoVW) model [1, 3] for category level and scene classification. In the traditional BoVW model, feature vectors are histograms of visual words. This representation is appearance based and does not contain any information regarding the arrangement of the visual words in the 2D image space. In this framework, we present a simple and efficient way to infuse spatial information. Particularly, we are interested in explicit global relationships among the spatial positions of visual words. For that we first introduce the notion of Pair of Identical visual Words (PIW) defined as the set of all the pairs of visual words of the same type. Then a spatial distribution of words is represented as a histogram of orientations of the segments formed by PIW. Figure 1 shows an example which gives an intuition to better understand our approach.

Our method eliminates a number of drawbacks from the previous approaches [2, 3] by i) proposing a simpler word selection technique that supports fast exhaustive spatial information extraction, ii) enabling infusion of global spatial information, iii) being robust to geometric transformations like translation and scaling.

In the conventional BoVW model, each image is represented by a set of local descriptors $\{d_1 \dots d_n\}$ extracted from n patches around interest points or regular grids. A visual vocabulary $W = \{w_1, w_2, w_3, w_4 \dots w_N\}$ is obtained by clustering a set of descriptors from all the training images. Here, N is a predefined number and the size of the vocabulary. Each patch of the image is then mapped to the nearest visual word according to the following equation:

$$w(d_k) = \arg\min_{w \in W} \mathrm{Dist}(w, d_k) \tag{1}$$

Here, $w(d_k)$ denotes the visual word assigned to the k^{th} descriptor d_k and $Dist(w, d_k)$ is the distance between the visual word w and the descriptor d_k. In the conventional BoVW method, the final representation of the image is a histogram of visual words. The number of bins in the histogram is equal to the number of visual words in the dictionary (i.e. N). If each bin b_i represents occurrences of a visual word w_i in W, b_i is defined as:

$$b_i = \mathrm{Card}(\mathcal{D}_i) \quad \text{where} \quad \mathcal{D}_i = \{d_k, k \in \{1, \dots, n\} \mid w(d_k) = w_i\} \tag{2}$$

$\mathcal{D}_i$ is the set of all the descriptors corresponding to a particular visual word w_i in the given image. $\mathrm{Card}(\mathcal{D}_i)$ is the cardinality of the set $\mathcal{D}_i$. In this final step, the spatial information of interest points is not retained. To model this information and to infuse it to the BoVW model, we propose the angle histogram of PIW. For each visual word w_i the method is as follows: first, from the set $\mathcal{D}_i$ of descriptors assigned to w_i (Equation 2), we consider all pairs of those descriptors and we build the set PIW_i constituted by the corresponding position pairs.

$$PIW_i = \{(P_k, P_l) \mid (d_k, d_l) \in \mathcal{D}_i^2, d_k \neq d_l\} \tag{3}$$

where P_k and P_l correspond to the spatial positions in the image from which the descriptors d_k and d_l have been extracted. The spatial position of a descriptor is given by the coordinates of the top-left pixel of the corresponding patch. These coordinates vary in the range of the image spatial domain. The cardinality of the set PIW_i is $\binom{b_i}{2}$, i.e. the number of possible subsets of two distinct elements among b_i elements. Second, for each pair of points of the set PIW_i, we compute the angle θ formed with the horizontal axis using the cosine law:

Figure 1: Discriminative power of spatial distribution of intra type visual words. Four images from Caltech101 dataset are shown. The black squares refer to identical visual words across all the images. For the two motorbikes in the left, the global distribution of the identical visual words is more similar than the ones in Helicopter or Bugle image. Our proposal 'PIW Angle Histogram' can capture information about these distributions.

$$\theta = \begin{cases} \arccos\left(\dfrac{\overrightarrow{P_kP_l} \cdot \vec{i}}{\left\|\overrightarrow{P_kP_l}\right\|}\right) & \text{if } \overrightarrow{P_kP_l} \cdot \vec{j} > 0 \\ \pi - \arccos\left(\dfrac{\overrightarrow{P_kP_l} \cdot \vec{i}}{\left\|\overrightarrow{P_kP_l}\right\|}\right) & \text{otherwise} \end{cases} \tag{4}$$

where $\overrightarrow{P_kP_l}$ is the vector formed by two points P_k and P_l and i and j are orthogonal unit vectors defining the image plane. Third, the histogram of all θ angles is calculated. The bins of this histogram are equally distributed between 0° and 180°. The optimal number of bins is chosen empirically. We call this histogram the PIW angle histogram for word w_i and denote it as $PIWAH_i$.

To have a global representation, we replace each bin of the BoVW frequency histogram with the $PIWAH_i$ histogram associated to w_i. The sum of all the bins of $PIWAH_i$ is normalized to the bin-size b_i of the respective bin of the BoVW frequency histogram. By this way, we keep the frequency information intact and add the spatial information. Equation 5 formalizes our global representation of an image, denoted as $PIWAH$.

$$PIWAH = (\alpha_1 PIWAH_1, \alpha_2 PIWAH_2, \alpha_3 PIWAH_3, \dots, \alpha_N PIWAH_N) \quad \text{where} \quad \alpha_i = \frac{b_i}{\|PIWAH_i\|_1} \tag{5}$$

Here, N is the vocabulary size and α_i is the normalization term. If the number of bins in each of $PIWAH_i$ is M, the size of the $PIWAH$ representation becomes MN.

For this work, we use MSRC-v2, Caltech101, 15 Scene and Graz-01 datasets for experiments. Our method imroves classification accuracy for all the datasets. The improvement is 12% for Caltech101 and 4% for 15Scene over BoVW representation. Our method also improves accuracy for Graz-01 dataset where global information is extremly difficult to model. We show that our method is complementary to method like Spatial Pyramid [1]. We also show on the MSRC-v2 dataset that our method performs as good as the existing ones [2, 3] with the advantage of being the fastest and the simplest among all.

[1] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition*, pages 2169–2178, 2006.

[2] David Liu, Gang Hua, Paul A. Viola, and Tsuhan Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.

[3] Silvio Savarese, John Winn, and Antonio Criminisi. Discriminative object class models of appearance and shape by correlatons. In *Computer Vision and Pattern Recognition*, pages 2033–2040, 2006.

Binocular Projection of a Random Scene

Miles Hansard
www.eecs.qmul.ac.uk/~milesh/

School of Elec. Eng. & Comp. Sci.
Queen Mary, University of London,
Mile End Road, London E1 4NS.

Current approaches to large-scale visual reconstruction would benefit from a statistical model of multiple-view projection. In particular, global constraints based on scene-clutter and occlusion are required. This work presents a new statistical model, starting from the simplest example of a random scene, viewed by two cameras. The most interesting aspects are revealed by working in a single epipolar plane, and supposing that the scene consists of identical discs of radius ε, as in fig. 1.

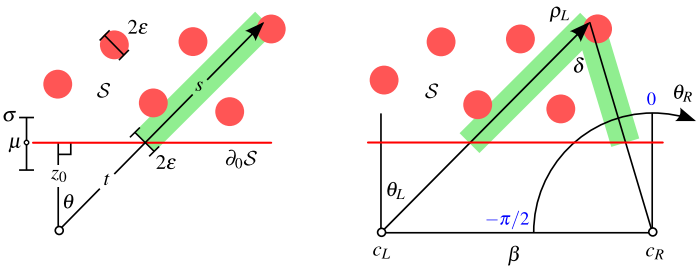


Figure 1: **Viewing geometry**. *Monocular case, left:* The scene $\mathcal{S}$ comprises discs of radius ε, with near-boundary $\partial_0\mathcal{S}$. The depth z_0 of the latter is normally distributed around μ. A ray in direction θ extends to distance t through empty space, followed by distance s through the scene, before striking an object. Equivalently, there are no disc-centres in the $2\varepsilon \times s$ green rectangle, and so s is exponentially distributed. *Binocular case, right:* both rectangles must be empty for the disc to be binocularly visible. Note that equal increments of θ_R would span increasingly long segments of $\rho_L = s + t$ as the difference-angle δ decreases.

It has previously been shown [3, 4] that, if the discs are distributed according to a Poisson process [2] of intensity λ, then the distance to a visible object follows an exponential distribution F of intensity $2\varepsilon\lambda$;

$$\mathrm{pr}(s\,|\,\lambda,\varepsilon) = F\big(s,\,2\varepsilon\lambda\big). \tag{1}$$

This is because if an object at distance s is visible, then the $s \times 2\varepsilon$ rectangle around the corresponding ray must be empty (cf. fig. 1, left). But the exponential model is not qualitatively realistic, because the mode of the distribution is at zero, implying that the optical centre is fully amid the clutter. Real range-data, in contrast, follows a two-tailed distribution along each ray, as seen in fig. 2.

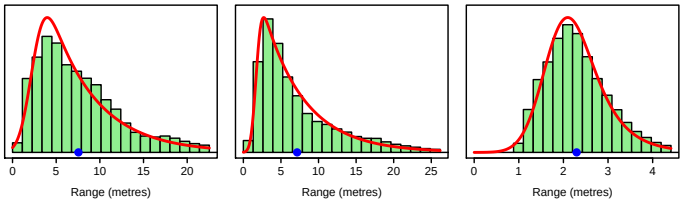


Figure 2: **Range densities**. *Left:* maximum likelihood ex-Gaussian fit to the upper 'canopy' regions of 14 range-scans of a forest scene. *Middle:* fit to the central $120° \times 40°$ 'trunks' regions, which is most representative of heterogeneous clutter. *Right:* fit to the lower 'ground' regions, which is more Gaussian. The blue dot is the mean of each data-set.

A more realistic visibility density is proposed here, in which the optical centre is displaced from the scene $\mathcal{S}$ by a shift t. If the near-boundary $\partial_0\mathcal{S}$ of the scene is locally perpendicular to the straight-ahead direction, at a normally distributed distance $G(z_0,\mu,\sigma)$, then a ray at angle θ passes a distance t through empty space, where

$$\mathrm{pr}(t\,|\,\theta,\mu,\sigma) = G\bigg(t,\,\frac{\mu}{\cos\theta},\,\frac{\sigma}{\cos\theta}\bigg). \tag{2}$$

The total distance ρ to a visible object is then $\rho = s + t$ where s has an exponential density (1) and t has a Gaussian density (2). It follows that the density H of ρ is obtained by convolution, $H = F \star G$. This can be expressed as a standard *ex-Gaussian* distribution [1] where

$$\mathrm{pr}(\rho\,|\,\theta) = H\bigg(\rho,\,2\varepsilon\lambda,\,\frac{\mu}{\cos\theta},\,\frac{\sigma}{\cos\theta}\bigg). \tag{3}$$

Figure 2 shows that this model is a good fit to real range data, acquired with an outdoor laser scanner.

It is straightforward to extend (3) to the binocular case, in which an object is jointly visible if both left and right rays are unobstructed, hence

$$\mathrm{pr}(\rho_L,\rho_R) = \mathrm{pr}(\rho_L|\theta_L) \times \mathrm{pr}(\rho_R|\theta_R). \tag{4}$$

This model is expressed in terms of scene-distances ρ_L and ρ_R, which cannot be directly observed in practice. However, given that the two rays must intersect, ρ_L and ρ_R are functions of the left and right visual directions θ_L and θ_R from optical centres from $\boldsymbol{c}_L$ and $\boldsymbol{c}_R$ respectively (cf. fig. 1, right). Hence it is possible to reparameterize (4), in order to obtain the *conditional probability of observing a point in direction θ_R from $\boldsymbol{c}_R$, given that it is observed in direction θ_L from $\boldsymbol{c}_L$*. This density, which is supported on an epipolar line, involves the Jacobian

$$J_R(\theta) = \rho_R\sqrt{\csc^2\delta + \cot^2\delta}, \quad \text{where} \quad \delta = \theta - \theta_L \tag{5}$$

such that $J_R(\theta_R)$ accounts for the variation of the combined distances ρ_L and ρ_R, with respect to θ_R. The complete conditional density is then

$$\mathrm{pr}(\theta_R\,|\,\theta_L) = \mathrm{pr}(\rho_L,\rho_R) \times J_R(\theta_R)\big/S_R(\theta_L) \tag{6}$$

where $S_R(\theta_L)$ is the normalizing constant, which can be obtained (if required) by numerical integration. The density (6) is interesting, because it balances two opposite tendencies. On one hand, the tails of the ex-Gaussian parts (3) are exponentially decreasing, which means that distant objects are *less* likely to be seen. On the other hand, the Jacobian term (5) expresses the fact that, as the two visual directions become parallel, small changes of θ_R cause big changes in ρ_L and ρ_R, making it *more* likely that an object will be seen as θ_R approaches the vanishing point. Some examples and simulations of $\mathrm{pr}(\theta_R\,|\,\theta_L)$ are shown in fig. 3.

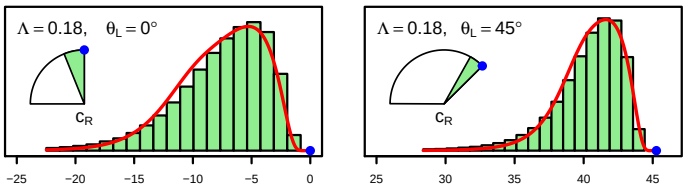


Figure 3: **Correspondence densities**. Histograms show the angular distributions $\mathrm{pr}(\theta_R|\theta_L)$ obtained by Monte Carlo simulation, as functions of θ_R, for two different directions θ_L. Red lines are *predicted* (not fitted) distributions defined by (6). The angular sectors extend from the epipole ($-90°$, cf. fig. 1, right) to θ_L, which is the vanishing-point (blue dot) of the ray through $\boldsymbol{c}_L$, as seen from $\boldsymbol{c}_R$.

In summary, the conditional probability of observing a point in a Poisson scene, given that it has already been observed in another view, has been derived. In particular, it has been shown how this probability, as a function of θ_R, is determined by the given direction θ_L and the clutter intensity λ. This provides a theoretical basis for new Bayesian priors in binocular image-matching.

[1] S.L. Burbeck and R.D. Luce. Evidence from auditory simple reaction-times for both change and level detectors. *Perception and Psychophysics*, 32:117–133, 1982.

[2] D.R. Cox and V. Isham. *Point Processes*. Chapman & Hall, 1980.

[3] J. Huang, A.B. Lee, and D. Mumford. Statistics of range images. In *Proc. CVPR*, pages 324–331, 2000.

[4] M.S. Langer. Surface visibility probabilities in 3D cluttered scenes. In *Proc. ECCV*, pages 401–412, 2008.

Visual words assignment on a graph via minimal mutual information loss

Yue Deng
http://media.au.tsinghua.edu.cn/dengyue.html
Yanjun Qian

Yipeng Li

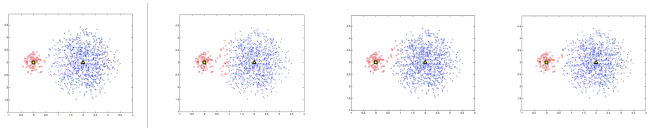
Qionghai Dai
http://media.au.tsinghua.edu.cn/qhdai.html
Guihua Er

Department of Automation
Tsinghua University
Beijing, China

Visual codewords assignment plays an important role in many Bag of Features (BoF) models for image understanding and visual recognition. It allocates image descriptors to the most similar codewords in the preconfigured visual dictionary to generate descriptive histogram for the consequent categorization. Nevertheless, existing assignment approaches, *e.g.* nearest neighbors strategy and Gaussian similarity, suffer from two problems:1) too strong Euclidean assumption and 2) neglecting the label information of the local features. Accordingly, in this paper, we propose an assignment method to simultaneously consider the above two issues in a unified model via graph learning and information theoretic criterions.

Our contributions are two-folds: 1) We propose a new local feature assignment method from the new perspective of graph learning that enables the usage of Non-Euclidean graph metric, e.g. geodesic distance and commute time for feature assignment and 2) we introduce the information theoretic penalty to reveal both the relationship of local features and their category labels. Our model exhibits both the advantages of graph learning and information theoretic learning and thus it is named Graph Assignment with minimal Mutual Information Loss (GAMIL). First of all, we describe how to assign image features via a graph.

We define the local image features set as $S=\{(f_1,l_1),(f_2,l_2)...(f_n,l_n)\}$, where $f_i \in \mathbb{R}^p$ is the image feature, *e.g.* dense sift, extracted on the original image and $l_i \in \{1,2,...C\}$ is the category label of the image that f_i is extracted from. C is the number of image categories. Therefore, in this paper, we propose to use a graph to model the samples in S. Using the manifold structure, it is possible to model the linearity among data by the locality similarity; and the global nonlinearity can be evaluated by some graph metric on the manifold, e.g. geodesic distance and commute time [3]. But the above-motioned method on a graph is only suitable to the in-sample features. For practical usage of codeword assignment, it is desirable to extend the assignment ability to the out-of-sample data. Inspired by [1], we propose to embed the graph into an Euclidean space with a linear projection matrix. In the embedded space, the original graph metric is well preserved by the Euclidean distance. Besides, it worths noting that each feature in the set S also contains the label information. During training, we know where the image feature comes from. Accordingly, the problem changes to be how to evaluate the relationship between the features F and their labels L. Fortunately, owing to the previous work [2], we know that the relationship of feature and label is always judged by the mutual information, i.e. $I(F;L)$. In probability theory and information theory, the mutual information [2] of two random variables is a quantity that measures the mutual dependence of the two random variables. It measures how much knowing one of these variables reduces uncertainty about the other. Informally, in our case, $I(F;L)$ can be interpreted as how much the uncertainty is reduced about the label L if we know the feature F. Therefore, for discriminative learning, a large mutual information score is desired. The proposed graph assignment method projects the high dimensional feature in a low dimensional space ($q<p$). Ideally we hope that the mutual information on the original graph should be kept the same in the embedding space, i.e. $I(F;L)=I(\Omega^T F,L)$. Unfortunately, reducing the dimensionality of data from high to low of course causes information loss. Therefore, instead of mutual information preservation, we propose to use the minimal mutual information loss criterion, i.e. to minimize $I(F;L)-I(\Omega^T F;L)$. Therefore, by considering both the information loss and graph similarity, the optimization for our GAMIL model is given,



(a) ground truth (b) assignment via ED. (c) assignment via GD. (d) assignment via CT.

Figure 1: A toy assignment via different methods, i.e. Euclidean Distance(ED), Geodesic Distance(GD) and Commute Time(CT).

$$\min_{\Omega} \underbrace{tr(\Omega^T F(D-W)F^T\Omega)}_{\text{graph assignment}} + \underbrace{\alpha H(L|Y=\Omega^T F)}_{\text{mutual information loss}} \quad s.t. \Omega^T FDF^T\Omega = I, \tag{1}$$

where α is a user specified parameter which trades off the graph assignment and mutual information loss. where $F=[f_1,...,f_n]\in\mathbb{R}^{p\times n}$ is the feature matrix; $\Omega\in\mathbb{R}^{p\times q}, q<p$ is the linear projection matrix. $W=[w_{ij}]$ is the weight matrix obtained on the graph which records the similarity between any two nodes i,j on the graph and $D=diag(\sum_i W_{ij})$. For an out-of-sample feature, we first project it to the subspace and then assign it to each codeword via Euclidean similarity. It is because the Euclidean distance in the embedding space represents the original nonlinear graph similarity on the manifold. For learning, the proposed model can be efficiently solved in a closed-form with the reasonable graph topology invariant approximation.

In the experiment, we randomly pick a number of images per class for training, and the left are for testing. In order to get reliable results, each experiment is repeated for 10 times (otherwise notice).

Table 1: The comparisons of GAMIL model with other state-of-the-arts on two benchmarks

Algorithms	Scene-15	Caltech-101
Hard	76.3	56.4
Soft	78.2	59.5
GAMIL	**80.7**	64.3
Info-loss[2]	74.7	-
Sparse coding[4]	80.3	**67.0**

To describe an image, we use a grid-based method to extract the dense sift features and the codebook is generated in the embedding space by K-means algorithm. For classification, we use the SVM with a histogram intersection kernel. We evaluate the proposed algorithm on two benchmarks, i.e. scene-15 and caltech-101 and our own dataset on Multiview Human Bodies (MHB). Experimental results show that our algorithm achieves state-of-the-art performances on benchmarks.

[1] Yue Deng, Qionghai Dai, Ruiping Wang, and Zengke Zhang. Commute time guided transformation for feature extraction. In *CVIU*, 2012.

[2] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *TPAMI*, 2009.

[3] Huaijun Qiu and E.R. Hancock. Clustering and embedding using commute times. In *TPAMI*, 2007.

[4] Jianchao Yang, Kai Yu, Yihong Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

Multiple queries for large scale specific object retrieval

Relja Arandjelović
relja@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Department of Engineering Science
University of Oxford
Parks Road
Oxford, OX1 3PJ, UK

The aim of large scale specific-object image retrieval systems is to instantaneously find images that contain the query object in the image database. Current systems, for example Google Goggles, concentrate on querying using a single view of an object, e.g. a photo a user takes with his mobile phone, in order to answer the question "what is this?". Here we consider the somewhat converse problem of finding *all* images of an object given that the user knows what he is looking for; so the input modality is text, not an image. This problem is useful in a number of settings, e.g. media production teams are interested in searching internal databases for images or video footage to accompany news reports and newspaper articles.

Given a textual query (e.g. "Fontana di Trevi"), our approach is to fetch images of the queried object using textual Google image search. These images are used to visually query the database to discover images containing the object of interest. We compare a number of methods for combining the multiple query images, including discriminative learning. We show that issuing multiple queries significantly improves recall and enables the system to find quite challenging occurrences of the queried object. Fig. 1 shows an example of this process, which proceeds in a matter of seconds from typing the query to receiving the ranked results.

Using multiple queries overcomes a number of the shortcomings of existing large scale specific object retrieval methods. It is important to first consider why images containing the target object are missed using a single query. Addressing this problem has been one of the main research themes in specific object retrieval research with developments in feature encoding to alleviate vector quantization (VQ) losses [4, 5, 6], and in augmentation of the bag of visual word (BoW) representation to alleviate detector and descriptor drop out (as well as, again, VQ losses) [1, 2, 3].

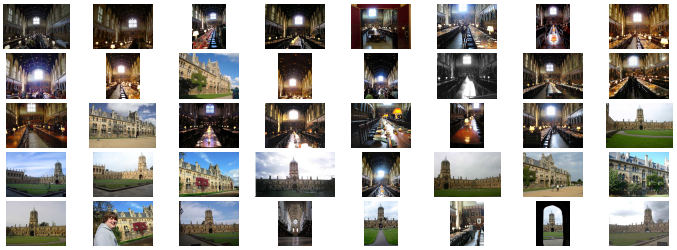
The limitation of current augmentation approaches, which are based on query expansion (QE) within the data set, is that they rely on the query to yield a sufficient number of high precision results in the first place. In more detail, in QE an initial query is issued, using only the query image, and confident matches, obtained by spatial verification, are used to re-query. There are three problems with this approach: (i) It is impossible to gain from QE if the initial query fails. (ii) If the dataset does not contain many images of the queried object QE cannot boost performance. (iii) It is not possible to obtain images from different views of the object as these are never retrieved using the initial query, for example querying using an image of a building façade will never yield results of its interior.

Table 1 shows the retrieval performance on the Oxford 105k dataset. It can be seen that all the multiple query methods are superior to the "single query" baseline, improving the performance by 29% and 52% for the Oxford queries and Google queries (with spatial reranking), respectively. It is clear that using multiple queries is indeed very beneficial as the best performance using Oxford queries (0.937) is better than the best reported result using a single query (0.891 achieved by [1]); it is even better than the state-of-the-art on a much easier Oxford 5k dataset ([1]: 0.929). All the multiple query methods also beat the "best single query" method which uses ground truth to determine which one of the images from the query set is best to be used to issue a single-query.

	Google queries		Oxford queries	
	W/o SR	With SR	W/o SR	With SR
Single query	0.464	0.575	0.622	0.725
Best single query ("cheating")	0.720	0.792	0.791	0.864
Joint-Avg	0.834	0.873	0.886	0.933
Joint-SVM	0.839	0.875	0.886	0.926
MQ-Max	0.746	0.850	0.826	0.929
MQ-Avg	0.834	0.868	0.888	0.937
MQ-ESVM	N/A	0.846	N/A	0.922

Table 1: **Retrieval performance (mAP) of the proposed methods on the Oxford 105k dataset.** SR stands for spatial reranking. The "Oxford queries" (OQ) and "Google queries" (GQ) columns indicate the source of query images, the former being the 5 predefined query images and the latter being top 8 Google images which contain the queried object. All proposed methods significantly outperform the "single query" baseline, as well as the artificially boosted "best single query" baseline.

(a) Top 8 Google Image results for the textual query "Christ Church, Oxford"



(b) Top 40 retrieved results from the Oxford 5k dataset by searching with the Google images

Figure 1: **Multiple query retrieval.** Images downloaded from Google using the "Christ Church, Oxford" textual query (a) are used to retrieve images of Christ Church college in the Oxford Buildings dataset (b). All the top 40 results of (b) do show various images of Christ Church (the dining hall, tourist entrance, cathedral and Tom tower). This illustrates the benefit of issuing multiple queries in order to retrieve all images of the queried object.

Figure 2: **Retrieved images from the TrecVid 2011 dataset.** The textual queries used to download images from Google and use them to retrieve images from the TrecVid 2011 KIS dataset are: (a) Presidential seal, (b) EA sports logo, (c) Leonardo da Vinci, (d) Fontana di Trevi, (e) I want you, (f) Comedy central logo.

[1] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, 2012.

[2] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Proc. ICCV*, 2007.

[3] O. Chum, A. Mikulik, M. Perd'och, and J. Matas. Total recall II: Query expansion revisited. In *Proc. CVPR*, 2011.

[4] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, 2008.

[5] A. Mikulik, M. Perd'och, O. Chum, and J. Matas. Learning a fine vocabulary. In *ECCV*, 2010.

[6] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. CVPR*, 2008.

A Training-free Classification Framework for Textures, Writers, and Materials

Radu Timofte[1] and Luc Van Gool[1,2]
http://homes.esat.kuleuven.be/~rtimofte

[1]ESAT-VISICS /IBBT, Catholic University of Leuven, Belgium
[2]D-ITET, ETH Zurich, Switzerland

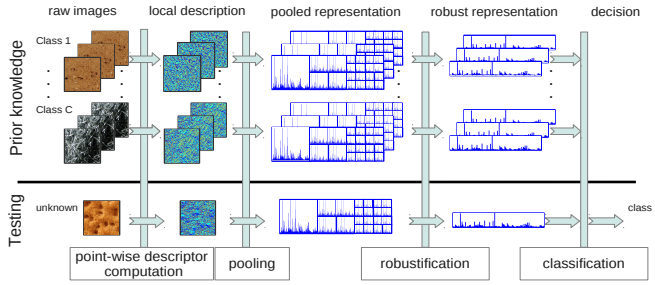


Figure 1: The scheme of our texture classification framework.

We propose a training-free texture classification scheme, outperforming methods that use training. This we demonstrate not only for traditional texture benchmarks, but also for the identification of materials and writers of musical scores. State-of-the-art methods operate using local descriptors, their intermediate representation over *trained* dictionaries, and classifiers. For the first two steps, we work with pooled local Gaussian derivative filters and a small dictionary *not* obtained through training, resp. Moreover, we build a multi-level representation similar to a spatial pyramid which captures region-level information. An extra step robustifies the final representation by means of comparative reasoning. As to the classification step, we achieve robust results using nearest neighbor classification, and s-o-a results with a collaborative strategy. Also these classifiers need no training.

Standard texture classification systems aim at i) constructing a rich representation of the image and ii) providing a classification strategy. The representation typically entails local (texture) descriptors, similarity measures, aggregating strategies, and intermediate and global (image level) descriptors. The classification strategy usually adapts its metric to the representation and aims at fixing its flaws. Class models are then built using state-of-the-art classifiers. A literature review is in the paper.

We **propose** a training-free multi-level texture classification framework (Fig. 1). It combines the robustness and simplicity of local descriptors such as BIFs [1], spatial information embedding into the global image representation in a layered fashion similar to SPM or through regions as in [3], the power of comparative reasoning [8], and s-o-a training-free classifiers [6]. Fortes of the framework are:

1) ***No need for training, and thus data independence.*** There is no need for learning a dictionary for the local descriptors (such as BIFs [1]). The system performs robustly with a fixed set of parameters on different texture, material and handwritten score datasets.

2) ***Robustness to intra-class variations.*** Robustness is provided by the local descriptors, the layered robustified representation, and the classifiers.

3) ***Layered representation embedding spatial information.*** Spatial information proved critical for object classification, and so it is for our tasks.

4) ***Robustified representations by means of comparative reasoning.*** The power of comparative reasoning (WTA-hash [8]) enhances and robustifies the representations by adding resilience to numeric perturbations.

5) ***Fast sparse and/or collaborative classification.*** Lately, sparse and collaborative representation based classifiers performed best at various tasks such as face recognition or traffic sign recognition [6].

Local Texture Descriptor (BIF) Basic Image Features (BIF) [1, 5] are defined by a partition of the filter-response space (jet space) of a set of 6 Gaussian derivative 2D filters up to 2nd order at some scale σ. The Jet space is further partitioned into 7 regions, or BIFs, corresponding to distinct types of local image symmetry. BIFs are rotation invariant. However, we can discretize orientations for BIF codes as in [5], thus obtaining Oriented Basic Image Features (oBIF). To create a more discriminative descriptor, [1] combines the descriptors at different scales on a pixelwise basis and ignores flat regions. BIF with p scales will generate 6^p distinct dictionary entries (for $p = 4$, 1296), while oBIF will generate 22^p distinct dictionary entries (for $p = 2$, 484).

Multi-Level Pooled Representation (SPM, BoR) The spatial pyramid matching (SPM) scheme pools regions at 3 or 4 pyramid levels. We continue as long as the cell/region size allows for meaningful histograms.

Table 1: Summary of texture, material, and score datasets.

Dataset	Dataset Notation	Dataset Type	Image Rotation	Controlled Illumination	Scale Variation	Significant Viewpoint	Number Classes	Sample Size	Samples per Class	Samples in Total
CUReT	D^{C}	texture	✓	✓			61	200×200	92	5612
UIUC	D^{UIUC}	texture	✓		✓	✓	25	640×480	40	1000
UMD	D^{UMD}	texture	✓		✓	✓	25	640×480	40	1000
Brodatz	D^{B}	texture					111	213×213	9	999
KTHTIPS	D^{KT}	texture		✓	✓		10	200×200	81	810
KTHTIPS2b	D^{KT2b}	texture		✓	✓	✓	11	200×200	4(×9×12)	4752
FMD	D^{FMD}	material	✓		✓	✓	10	512×384	100	1000
CVCMUSCIMA	D^{CM}	handwritten scores					50	∼2000×2000	20	1000

Table 2: Comparison of our results [%] with those achieved by state-of-the-art methods. In the brackets is the number of training samples.

	D^{C}(46)	D^{B}(3)	D^{KT}(41)	D^{UIUC}(20)	D^{UMD}(20)	D^{KT2b}(1)	D^{CM}(10)	D^{FMD}(50)
1. Our Results	**99.42**	**97.26**	**99.35**	**99.01**	**99.54**	**66.26**	**99.80**	**55.78**
3. VZ-Joint [7]	98.03	92.90(*)	92.40(*)	97.83		53.30(**)		
5. Lezebnik *et al.*	72.50(*)	88.15	91.30(*)	96.03				
7. J.Zhang *et al.*	95.30	95.90	96.10	98.70				
9. Crosier and Griffin [1]	98.60		98.50	98.80				
12. Xu *et al.*-WMFS				98.60	98.68			
14. L.Liu *et al.*-SRP [4]	99.37	97.16	99.29	98.56	99.30			48.2
15. L.Liu *et al.*-ELBP	97.29					58.10		
16. Kong and Wang [3]		96.61	99.32		99.32			
17. PRIP02 [2]							77.00	
23. Hu *et al.*								54.00

Another proposed approach [3] uses multi-levels, similar to SPM, for creating orderless region parts, allowing for overlap. The images are represented by sets of regions, called Bag-of-Regions (BoRs). BoRs cover a much larger variance in scale, translation, rotation, viewpoint, illumination by enlarging the training pool. For the test image represented as BoR, the classification score is computed for each class and region. At image level (or BoR level) the label is taken as the class with the best cumulative score over the BoR.

Robustified Representation - (WTA-hash) The power of comparative reasoning was exploited and a Winner Take All (WTA) hash technique proposed in [8]. WTA-hash transforms the input feature space into binary codes and in the resulting space the Hamming distance closely correlates with the rank similarity measures. The rank correlation measures are resilient to perturbations in numeric values and WTA-hash brings perturbation robustness to the original feature space representation.

Sparse and Collaborative Classification - (SRC, CRC, INNC) For classification we use robust classifiers in the sense of not requiring parameter tuning for different datasets: Nearest Neighbor Classifier (NNC), Sparse Representation Classifier (SRC), Collaborative Representation Classifier (CRC), and Iterative Nearest Neighbors Classifier (INNC) [6].

We have shown that training-free pipelines can outperform several s-o-a texture classification methods. We are conservative in our **experiments**, in that further fine-tuning would be possible, i.e. we only went up to the point where the methods would outperform or get on par with the s-o-a, training-based methods. We only report results using the basic image features (BIFs) and its variants as local descriptors [1, 5], SPM with one level and a simple Bag-of-Regions model [3] as intermediate representations, and classifiers such as NNC, SRC, CRC, or INNC [6].

We were somewhat surprised by the strong performance of these methods, regardless of the dataset and/or task. We believe that adding training at any level in our framework can improve the performance further.

The proposed approach is computationally simple. To a large extent, it also is training-free and data-independent. The system is validated for texture, material, and writer classification on several benchmarks. We obtain results that are at least on-par, but sometimes substantially better than state-of-the-art performance (Tables 1,2).

Details about the framework and the benchmarks are in the paper.

[1] M. Crosier and L.D. Griffin. Using basic image features for texture classification. *IJCV*, 88(3):447–460, 2010.

[2] A. Fornés, A. Dutta, A. Gordo, and J. Lladós. The ICDAR 2011 music scores competition: Staff removal and writer identification. In *ICDAR*, 2011.

[3] S. Kong and D. Wang. Multi-level feature descriptor for robust texture classification via locality-constrained collaborative strategy. *CoRR*, 2012.

[4] L. Liu, P.W. Fieguth, D.A. Clausi, and G. Kuang. Sorted random projections for robust rotation-invariant texture classification. *Pattern Recognition*, 2012.

[5] A.J. Newell and L.D. Griffin. Multiscale histogram of oriented gradient descriptors for robust character recognition. In *ICDAR*, 2011.

[6] R. Timofte and L. Van Gool. Iterative nearest neighbors for classification and dimensionality reduction. In *CVPR*, 2012.

[7] M. Varma and A. Zisserman. A statistical approach to material classification using image patch exemplars. *PAMI*, 31(11):2032–2047, 2009.

[8] J. Yagnik, D. Strelow, D.A. Ross, and R.-S. Lin. The power of comparative reasoning. In *ICCV*, 2011.

Comparing Visual Feature Coding for Learning Disjoint Camera Dependencies

Xiatian Zhu[1]
xiatian.zhu@eecs.qmul.ac.uk
Shaogang Gong[1]
sgg@eecs.qmul.ac.uk
Chen Change Loy[2]
ccloy@visionsemantics.com

[1] School of Electronic Engineering and Computer Science,
Queen Mary, University of London,
London E1 4NS, UK

[2] Vision Semantics,
London E1 4NS, UK

Problem: This work systematically investigates the effectiveness of various visual feature coding schemes for facilitating the learning of time-delayed dependencies among disjoint multi-camera views.

Related work: Quite a few studies [3, 4, 6] have been proposed to model inter-camera dependency across non-overlapping camera views. Learning time-delayed correlations among disjoint cameras in crowded public scenarios is a non-trivial task: (1) *the time gaps between camera views are unknown* therefore activities in two related views may occur at arbitrary time delays with high uncertainty; (2) *the features are inevitably noisy, ambiguous, and may vary drastically* across views owning to illumination condition, camera angles, and changes in object pose. Most state-of-the-art methods typically hand pick a few features tailored to the target environment, with the hope that those chosen features contain robust and sufficient statistics for correlating the time-delayed activity patterns across disjoint views. These manual approaches to hard selection of features are neither principled nor generalisable to different scene context.

Our solution: In this study, we wish to examine the concept that visual features should be coded and selected automatically for robust and accurate time-delayed dependency learning. The contributions of this study are two-fold: (1) We present a systematic study and evaluation to investigate the effectiveness of supervised and unsupervised feature coding methods to facilitate the learning of inter-camera activity pattern dependencies. (2) We systematically evaluate the sensitivity of inter-camera time delayed dependency learning given different training video sizes and region decomposition qualities. These factors are critical for accurate dependency learning but have been largely ignored by the published existing work in the literature.

Approach overview: We employ the Random Forest [2] as the supervised feature coding approach. In particular, given a set of localised features extracted from a region, together with people count training label over time, we first train a regression forest to learn the non-linear mapping between the crowd density and the corresponding low-level features. Given unseen data, we then construct a time series based on the predicted crowd density $\hat{y}$ obtained from the regression forest (*RF pred*), the tree-structured code (*tree code*) [5], or the combination of the two.

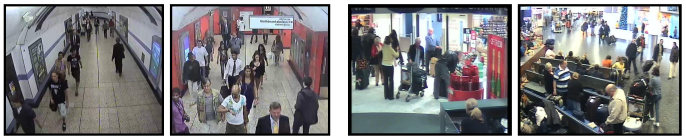
As for unsupervised coding scheme, we use the Latent Dirichlet Allocation (LDA) [1] to map the low-level features into codewords that capture the topic distribution, whereby an image region patch (document) d is treated as a collection of $j = 1 \ldots N_i$ features (words). To form the unsupervised feature codes, given a sequence of localised feature vectors detected from a region, we first perform quantisation on each feature to generate a bag-of-word representation for all image patches. Similar to text documents, these bag-of-word represented image patches are fed into the LDA, which gives us a topic-based representation. Once having the topic-based code (*topic code*), we perform k-means quantisation on them, producing the final compact topic-based code, and concatenate them over time to form a time series.

To solve the problem of using the feature codes for learning inter-camera dependencies, we adopt the Time Delayed Mutual Information (TDMI) proposed in [3] due to its reported effectiveness and simplicity. The input to TDMI are time series generated from either the supervised or the unsupervised coding scheme.

In addition to measuring deviation error in transition time, we propose a new metrics to evaluate the effectiveness of different coding methods, called Mutual Information Margin (MIM):

$$\triangle\mathcal{I} = \frac{\delta(\mathcal{I}_{\text{con}}) - \delta(\mathcal{I}_{\text{uncon}})}{\delta(\mathcal{I}_{\text{con}})}, \delta(\mathcal{I}) = \max(\mathcal{I}) - \min(\mathcal{I}), \quad (1)$$

where $\mathcal{I}_{\text{con}}$ and $\mathcal{I}_{\text{uncon}}$ denote the TDMI function yielded by the connected pairs and unconnected pairs of regions, respectively.



(a) The US dataset (b) The i-LIDS dataset

Figure 1: The example views of the US and the i-LIDS dataset.

Feature Codings	MI-MIM (US)	MI-MIM (i-LIDS)
RF pred	5.1530	7.8577
tree code	-1.7979	-1.7847
RF pred + tree code	-2.3839	-1.0335
topic code	**9.9057**	**16.6349**

Table 1: Sensitivity to the length of the training sequence: the average improvement in MIM of different feature coding methods over the k-means vector quantisation based representation. Mean improved MIM (MI-MIM) was computed by averaging individual percentage of improvement over the testing range.

Feature Codings	MI-MIM (US)	MI-MIM (i-LIDS)
RF pred	10.7670	**13.1541**
tree code	7.8714	2.0040
RF pred + tree code	7.6564	3.5522
topic code	**14.3076**	4.1265

Table 2: Sensitivity to region decomposition: Mean Improved MIM was computed following the same steps as explained in Table 1.

Experiments: We conducted extensive evaluations using two challenging multi-camera datasets: (1) an Underground Station (US) dataset, (2) the i-LIDS Multiple Camera Tracking Scenario (i-LIDS) dataset. See Fig. 1 for example.

The objective of first experiment is to compare the sensitivity of different coding schemes given different lengths of video sequence for time delayed dependency learning (see Table. 1). In the second experiment we evaluated the sensitivity of different coding schemes to the quality of region decomposition. The results are given in Table. 2.

Extensive experiments with both supervised and unsupervised feature coding methods on crowded public scene videos have demonstrated the superiority of the proposed feature coding methods to the conventional k-means vector quantisation, in terms of accuracy in time delayed dependency learning, and robustness to small training sequence size and poor region decomposition quality.

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[3] C. C. Loy, T. Xiang, and S. Gong. Incremental activity modelling in multiple disjoint cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[4] D. Makris, T. Ellis, and J. Black. Bridging the gaps between cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 205–210, 2004.

[5] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems*, 19, 2006.

[6] K. Tieu, G. Dalley, and W. E. L. Grimson. Inference of non-overlapping camera network topology by measuring statistical dependence. In *IEEE International Conference on Computer Vision*, pages 1842–1849, 2005.

Fixing the Locally Optimized RANSAC

Karel Lebeda
karel@lebeda.sk
Jiri Matas
matas@cmp.felk.cvut.cz
Ondrej Chum
chum@cmp.felk.cvut.cz

Center for Machine Perception,
Czech Technical University,
Faculty of Electrical Engineering,
Department of Cybernetics,
Karlovo namesti 13,
121 35 Prague, Czech Republic

One of the attractive properties of RANSAC [2], at least with the top-hat (inlier 1, outlier 0) cost function, is that it returns an optimal solution with a predefined, user-controllable probability. The theoretical guarantee is based on the assumption that all all-inlier (minimal) samples lead to the optimal solution. It has been observed [1, 9] that the assumption is not valid in practice and that often a significant data-dependent fraction of all-inlier samples does not lead to an acceptable solution.

To address the "not all all-inlier samples are good" problem, Chum *et al.* [1] introduced the LO-RANSAC which applies a local optimization (LO) step to promising hypotheses generated from random minimal samples. Experiments in [1] show that LO-RANSAC is superior to plain RANSAC in terms of accuracy and its probability of obtaining a correct solution is close to the theoretical value derived from the stopping criterion. The LO-RANSAC method is popular, highly cited and has been used in a number of applications.

Chum *et al.* [1] stated that the improvements of the accuracy and the probability of obtaining a correct solution may even speed the algorithm up since the increased number of found inliers triggers the stopping criterion earlier. The LO is run only rarely, the number of runs being close to the logarithm of the number of samples.

As the first contribution of the paper we show that the "no extra time" statement is true only for estimation problems with low inlier ratios. For image pairs a high fraction on inliers where a small number of random samples is sufficient for finding the solution, the original LO procedure significantly effects the running time, sometimes becoming a dominating factor that may increase the running time by an order of magnitude. To alleviate the problem and reduce the overhead we modify the iterative least squares by introducing a limit on the number of inliers used for the least squares computation. Nevertheless, the modified LO$^+$-RANSAC is slower than plain RANSAC, fortunately mainly for easy datasets where the procedure is very fast anyway (see Figure 1 for an illustration of the dependence). Essentially the result shows that the local optimization is not always a free lunch and that there is a trade-off between estimation quality (accuracy and repeatability) and the computational time.

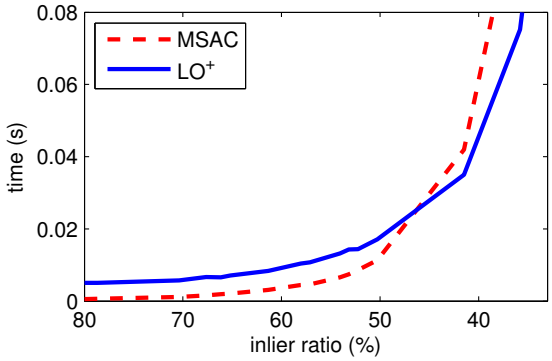


Figure 1: Dependence of the time complexity on the inlier ratio for a selected pair of images.

As a second contribution, we introduce a fast version – LO' that has execution time close to the standard RANSAC and perform close to LO-RANSAC in almost all cases. Instead of estimating models from non-minimal samples followed by iterative least squares, only a single iterative least squares are applied on each *so-far-the-best* model.

The LO procedure is relatively complex, with a high number of parameters. As a third contribution of the paper, we are making public an ultimate description of the method: a C/C++ implementation of the improved LO$^+$. The implementation has been extensively experimentally tested and performed well on dozens of geometry estimation problems with the same parameter settings. The proposed method is very stable - for many tested geometric problems it returned the identical set of inliers in 10000 out of 10000 test runs. We also show that the proposed algorithm is insensitive to the choice of the *error scale* which defines the inlier-outlier separation. In this context we confirm the slight advantage of the MSAC-like truncated quadratic [10] over the the top-hat, 0-1 loss function. The precision of the LO procedure for both methods is almost identical, but the MSAC-like kernel increases tolerance to the choice of the inlier threshold. Therefore, the proposed LO$^+$ differs from the standard LO by using the inlier limit and the truncated quadratic cost function.

The accuracy of the proposed LO method is tested within a standard Bundle adjustment method [5]. Perhaps surprisingly the bundler is rather sensitive to initialization. The LO initialized non-linear optimization is always superior in terms of residual errors to the Gold Standard method advocated by Hartley and Zissermann [3].

In our experiments, tentative correspondences were obtained by matching SIFT descriptors [6] of MSER's [7]. In the supplementary material [4], also experiments using Hessian Affine detector [8] are presented. The results on Hessian Affine features are even more favourable for the LO methods because of lower inlier ratios. Basically, they show our conclusions are independent of the selection of detectors.

The experimental evaluation shows that: (1) the LO$^+$-RANSAC with MSAC cost function offers a stable robust estimation despite its randomized nature, (2) limiting the number of inliers included in the (iterative) least squares significantly reduces execution time and often even improves the precision, (3) the speed of the minimalistic version LO' is comparable to plain RANSAC even for easy problems with very high inlier ratios, and that (4) LO-RANSAC offers significantly better starting point for bundle adjustment than the Gold Standard [3].

[1] O. Chum, J. Matas, and J. Kittler. Locally optimized RANSAC. In *DAGM-Symposium*, pages 236–243, 2003.

[2] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. of ACM*, 24(6):381–395, 1981.

[3] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[4] K. Lebeda, J. Matas, and O. Chum. Fixing the locally optimized RANSAC. Research Report CTU–CMP–2012–17, Center for Machine Perception, Czech Technical University, 2012.

[5] M. I. A. Lourakis and A. A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Software*, 36(1):1–30, 2009.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *Proc. of BMVC*, pages 384–396, 2002.

[8] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *Int. Journal of Computer Vision*, 60(1):63–86, 2004.

[9] B. Tordoff and D. W. Murray. Guided sampling and consensus for motion estimation. In *Proc. of ECCV*, pages 82–98, 2002.

[10] P. H. S. Torr and A. Zisserman. Robust computation and parametrization of multiple view relations. In *Proc. of ICCV*, pages 727 –732, 1998.

One-sided Radial Fundamental Matrix Estimation

José Henrique Brito[12]
josehbrito@gmail.com
Christopher Zach[4]
chzach@microsoft.com
Kevin Köser[3]
kevin.koeser@inf.ethz.ch
Manuel João Ferreira[2]
mjf@dei.uminho.pt
Marc Pollefeys[3]
marc.pollefeys@inf.ethz.ch

[1] Instituto Politécnico do Cávado e do Ave
Barcelos, Portugal
[2] Universidade do Minho
Guimarães, Portugal
[3] Computer Vision and Geometry Group
ETH Zurich, Switzerland
[4] Machine Learning and Perception
Microsoft Research Cambridge, UK

For modern consumer cameras, often approximate calibration data is available, making applications such as 3D reconstruction or photo registration easier as compared to the pure uncalibrated setting. In this paper we address the setting with calibrated-uncalibrated image pairs: for one image intrinsic parameters are assumed to be known, whereas the second view has unknown distortion and calibration. This situation arises e.g. when one would like to register archive imagery to recent photos. Very few existing solutions apply to the calibrated-uncalibrated setting. We propose a simple and numerically stable two-step scheme to first estimate radial distortion parameters and subsequently the focal length using novel solvers.

By using the distortion model proposed in [1], $p_u \propto \left(x_d y_d 1 + \lambda r_d^2\right)^T$ is the undistorted version of an observed image point $p_d = (x_d, y_d, 1)^T$ in an image with unknown radial distortion, $r_d^2 = (x_d-u)^2 + (y_d-v)^2$ for a known distortion center $(u,v)^T$, assumed to be at the image center, and λ is an unknown distortion parameter. The epipolar constraint becomes

$$q^T F p_u = q^T F \begin{pmatrix} x_d \\ y_d \\ 1+\lambda r_d^2 \end{pmatrix} = q^T \underbrace{\left[F \mid \lambda F_3\right]}_{=:\hat{F}} \begin{pmatrix} x_d \\ y_d \\ 1 \\ r_d^2 \end{pmatrix}, \tag{1}$$

where we introduced the 3×4-matrix $\hat{F}$. F_3 denotes the 3rd column of F. By using 9 correspondences, the nullspace of $\hat{F}$ is three-dimensional, i.e.

$$\hat{F} = x\hat{X} + y\hat{Y} + z\hat{Z}. \tag{2}$$

We can fix z to 1 due to the scale ambiguity of $\hat{F}$. The constraints $\hat{F}_4 \propto \hat{F}_3$, i.e. $\lambda\hat{F}_4 = \hat{F}_3$, now read as

$$x\hat{X}_{i4} + y\hat{Y}_{i4} + \hat{Z}_{i4} = \lambda\left(x\hat{X}_{i3} + y\hat{Y}_{i3} + \hat{Z}_{i3}\right) \tag{3}$$

for $i = 1,2,3$. First, we can eliminate λ by taking ratios, leading to 3 polynomial equations in x and y only,

$$\begin{aligned} p_{ij}(x,y) \stackrel{\text{def}}{=} &\left(x\hat{X}_{i4} + y\hat{Y}_{i4} + \hat{Z}_{i4}\right)\left(x\hat{X}_{j3} + y\hat{Y}_{j3} + \hat{Z}_{j3}\right) \\ &- \left(x\hat{X}_{j4} + y\hat{Y}_{j4} + \hat{Z}_{j4}\right)\left(x\hat{X}_{i3} + y\hat{Y}_{i3} + \hat{Z}_{i3}\right) \stackrel{!}{=} 0 \end{aligned} \tag{4}$$

for $(i,j) \in \{(1,2),(1,3),(2,3)\}$. We then compute two resultants (e.g. combining p_{12} with p_{13}, and p_{12} with p_{23}, respectively) leading to two degree 4 polynomials in x,

$$q_1(x) \stackrel{\text{def}}{=} a_1x^4 + b_1x^3 + c_1x^2 + d_1x + e_1 \stackrel{!}{=} 0 \tag{5}$$

$$q_2(x) \stackrel{\text{def}}{=} a_2x^4 + b_2x^3 + c_2x^2 + d_2x + e_2 \stackrel{!}{=} 0 \tag{6}$$

The leading monomial x^4 can now be eliminated by one step of Gaussian elimination leading to a final cubic polynomial,

$$r(x) \stackrel{\text{def}}{=} a_2q_1(x) - a_1q_2(x) \stackrel{!}{=} 0. \tag{7}$$

This can be solved in closed form leading to one or three real solutions. For each possible value of x, a corresponding y can be extracted by a similar procedure. Two of the p_{ij} polynomials (which are quadratic) yield a linear equation in y after one Gaussian elimination step. The extended fundamental matrix is given by $\hat{F} = x\hat{X} + y\hat{Y} + \hat{Z}$, and λ can be obtained as the ratio $\lambda = \hat{F}_{14}/\hat{F}_{13} = \hat{F}_{24}/\hat{F}_{23} = \hat{F}_{34}/\hat{F}_{33}$. By construction all those ratios are equal. Since we dropped the rank constraint, the estimated fundamental matrix (i.e. the 3×3 submatrix $\hat{F}[1:3,1:3]$) will generally be of full rank. We enforce rank-2 using the SVD as in the 8-point algorithm.

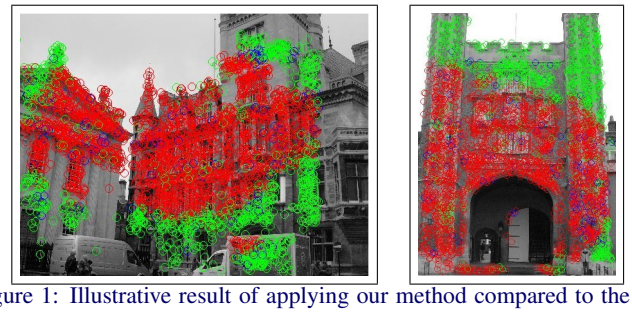


Figure 1: Illustrative result of applying our method compared to the results of the standard 8-point algorithm; red are the inliers found by both methods; green are the extra inliers found by our method; blue are inliers found by the standard 8-point not found by our method.

The focal length can be extracted in partially calibrated settings and we propose a different approach than [2]. Let F be a fundamental matrix, and K and K' camera intrinsics such that $E = (K')^T F K$ is an essential matrix. K is assumed to be known and K' is of the shape $\mathrm{diag}(f,f,1)$ for an unknown focal length f, hence we can incorporate K into F, yielding $E = \mathrm{diag}(f,f,1)F$. Plugging this expression into the trace constraint for essential matrices, $2EE^TE - \mathrm{tr}(EE^T)E = 0$, leads to a corresponding matrix constraint in terms of f,

$$\begin{aligned} G(f) \stackrel{\text{def}}{=}\ & 2\,\mathrm{diag}(f,f,1)FF^T\,\mathrm{diag}(f^2,f^2,1)F \\ & - \mathrm{tr}\left(\mathrm{diag}(f,f,1)FF^T\,\mathrm{diag}(f,f,1)\right)\mathrm{diag}(f,f,1)F \stackrel{!}{=} 0. \end{aligned} \tag{8}$$

We determine f by minimizing the algebraic error, $\|G(f)\|_F^2$. First order optimality conditions, $d\|G(f)\|_F^2/df = 0$, yields a polynomial in f^5, f^3 and f. Since we can exclude the degenerate solution $f = 0$, a double quadratic polynomial in f^4 and f^2 can be obtained, which is trivial to solve after substituting $w = f^2$. Since f has to be strictly positive, up to two possible values for f need to be checked for optimality. Our algorithm has a complexity comparable to that of the standard 8-point algorithm for classical fundamental matrix, since the main step is finding the null space created by the nine correspondences (rather than the eight correspondences in the 8-point algorithm).

To test our the method on real images, we matched a set of uncalibrated/distorted images to an image with known intrinsics using different datasets. We ran our and the standard 8-point methods in a RANSAC framework.. Results show that our method uses a higher number of inliers with an equal or lower average epipolar error. In Fig. 1 we can see two typical situations were the standard 8-point method would use only the correspondences not heavily affected by radial distortion, whereas our method would use correspondences where radial distortion is severe.

[1] A.W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 125–132, 2001.

[2] Magdalena Urbanek, Radu Horaud, and Peter Sturm. Combining off- and online calibration of a digital camera. In *International Conference on 3D Digital Imaging and Modeling (3DIM)*, pages 99–106, 2001.

Exemplar-Based Colour Constancy

Hamid Reza Vaezi Joze
hrv1@sfu.ca
Mark S. Drew
mark@sfu.ca
Simon Fraser University
Vancouver, BC, Canada

Exemplar-based learning or, equally, nearest neighbour methods have recently gained interest from researchers in a variety of computer science domains because of the prevalence of large amounts of accessible data and storage capacity. In computer vision, these types of technique have been successful in several problems such as scene recognition, shape matching, image parsing, character recognition and object detection. Applying the concept of exemplar-based learning to the well-known problem of illumination estimation seems odd at first glance since, in the first place, similar nearest neighbour images are not usually affected by precisely similar illuminants and, in the second place, gathering a dataset consisting of all possible real-world images, including indoor and outdoor scenes and for all possible illuminant colours and intensities, is indeed impossible. In this paper we instead focus on *surfaces* in the image and address the colour constancy problem by unsupervised learning of an appropriate model for each training surface in training images. We find nearest neighbour models for each surface in a test image and estimate its illumination based on comparing the statistics of pixels belonging to nearest neighbour surfaces and the target surface. The final illumination estimation results from combining these estimated illuminants over surfaces to generate a unique estimate.

The main distinctions between this work and other learning bases colour constancy methods that use spatial information by local feature descriptors such as [3, 4] is that they use this information to determined the best or combination of best possible illumination estimation algorithms while we use selected instances for illumination estimation.

We find surfaces for both training and test images by mean-shift segmentation. Since the pixels in the margin of segmented areas affect texture information, we remove margin pixels of segments by dilating segment edges as well as small segments.

In order to define a model for each surface we use both texture features and colour features. For the purpose of texture features, the MR8 filter bank [5] on three channels is selected for use because of its good performance in texture classification applications. We use the normalized histogram of frequency of appearance in that particular surface for each colour channel as our colour features. In order to make our model weakly invariant to variation in illuminant colour, We apply Max-RGB method for each surface.

Given a test surface model and its nearest neighbour surface model based on chi squared distance from training models, we can transfer the test surface's colour to its corresponding training surface's colours linearly by a 3×3 diagonal matrix.

$$e_{test} = De_{train} = \mathcal{M}_{test}^{-1} D_H \mathcal{M}_{train} e_{train} \quad (1)$$

where $\mathcal{M}$ is the weakly colour constant diagonal transformation of surface colour from the Max-RGB method and D_H is the transformation of test surface's histograms to training surface's histograms.

Algorithm 1 Illumination Estimation by Exemplar-Based method

1: surfaces ← mean-shift segmentat of the test image
2: **for all** S in surfaces **do**
3: features ← convolve S with MR8 filter
4: *label* ← *NN(features,textons)*
5: *texture hist* ← *normalized histogram of labels*
6: S_{cc} ← *MaxRGB*(S)
7: *colour hists* ← *normalized histogram of each colour channel in* S_{cc} *(10 bins)*
8: $model_S$ ← *texture hist. , colour hists.*
9: ***for all*** *i in KNN(*$model_S$*,*$models_{train}$*)* ***do***
10: $estimates_{Si}$ ← *eq. (1)*
11: ***end for***
12: **end for**
13: **return** *median*(*estimates*)

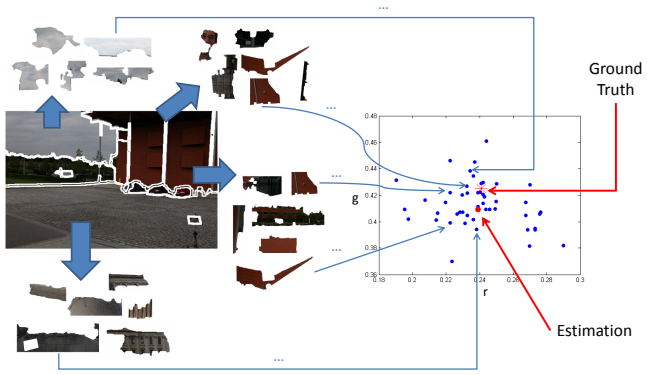


Figure 1: The procedure of estimating illuminant for a test image using exemplar-based color constancy. A test image and its nearest neighbour surface models from training images on left and estimated illuminants according to each model in *rg* chromaticity space on right.

Dataset	Color Checker		GrayBall	
Methods	Median	Mean	Median	Mean
White-Patch	5.7°	7.4°	5.3°	6.8°
Grey-World	6.3°	6.4°	7.0°	7.9°
Grey-Edge	4.5°	5.3°	4.7°	5.9°
Gamut Mapping pixel	2.5°	4.1°	5.8°	7.1°
Bottom-up+Top-down [4]	2.5°	3.5°	-	-
Natural Image Statistics [3]	2.5°	4.1°	3.9°	5.2°
Exemplar-Based	2.3°	3.1°	3.3°	4.4°

Table 1: Angular errors for two well-known color constancy datasets in term of mean and median for several algorithms.

Given a test image, we will have n large enough surfaces and M nearest neighbour surfaces from training data, or equally M illumination estimates by eq. (1) corresponding to each. The final estimate can be the median or the mean on the three channels separately after removing outliers of all of these nM estimates in *rg* chromaticity space.

We applied our proposed method to two standard colour constancy datasets of real images of indoor and outdoor scenes: the re-processed version of the Gehler colour constancy dataset [2], denoted the Color Checker dataset which include 568 images and the GreyBall dataset of Ciurea and Funt [1] which contains 11346 images. Table 1 indicates the accuracy of the proposed methods for these datasets, in terms of the mean and median of angular errors, for several colour constancy algorithms applied to this dataset.

To our knowledge, for these two standard datasets, widely used for testing colour constancy, Exemplar-Based Colour Constancy does best in terms of both mean and median angular error compared to any reported colour constancy methods, even those using a combination of algorithms such as Natural Image Statistics [3].

[1] F. Ciurea and B. V. Funt. A large image database for color constancy research. In *IS&T/SID Color Imaging Conference*, pages 160–164, 2003.

[2] P. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp. Bayesian color constancy revisited. In *In Proc. Comp. Vis. and Patt. Rec. (CVPR)*, 2008.

[3] A. Gijsenij and T. Gevers. Color constancy using natural image statistics and scene semantics. *IEEE Trans. Patt. Anal. and Mach. Intell.*, 33(4):687–698, 2011.

[4] J. van de Weijer, C. Schmid, and J. Verbeek. Using high-level visual information for color constancy. In *Proc. Int. Conf. on Comp. (ICCV)*, 2007.

[5] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *Proceedings of the 7th European Conference on Computer Vision-Part III*, ECCV '02, pages 255–271, 2002.

Indoor Scene Recognition using Task and Saliency-driven Feature Pooling

Marco Fornoni[12]
http://www.idiap.ch/~mfornoni
Barbara Caputo[2]
http://www.idiap.ch/~bcaputo

[1] Ecole Polytechnique Fédérale, Lausanne (EPFL)
Lausanne, CH
[2] Idiap Research Institute
Martigny, CH

Indoor scene recognition is as of today one of the most challenging open problems in visual place categorization. Since the seminal works of Oliva and Torralba [5] and Lazebnik et al. [3], the mainstream approach to scene recognition has been based on global, appearance-based image representations, enriched with spatial information. This approach, in various forms, has given good results for the outdoor place recognition problem, but proved to be inadequate when dealing with indoor scenes [6]. In indoor environments, indeed, the location of meaningful regions and objects varies drastically within each category. Also, the close-up distance between the camera and the subject makes the variations due to viewpoint changes even more severe. In this scenario, it becomes crucial how low-level features are spatially pooled to get the final image description, especially for the robustness of the representation. In this work we investigate this issue and propose to combine a simple spatial encoding, with a saliency-driven perceptual pooling designed to capture structural properties of the scenes, independently from their position in the image.

Saliency-driven perceptual pooling The traditional spatial encodings are designed to capture the spatial regularities in the scenes. We would instead like to let visual-structures emerge from the data, regardless of their exact position in the imaged scenes. Specifically, we are aiming to obtain a segmentation (R_1, R_2) such that R_2 captures the area of the image with a richer informative content (i.e., a high number of visual word responses), leaving to R_1 the task to collect the statistics of the remaining part. This is obtained by first computing a saliency map for each image, and subsequently using the median saliency value $\bar{s}$ of the image, to segment it in two regions: the most and least salient 50%.
To compute the saliency map, we test two approaches:

- **Itti**. The classic approach described in [2]
- **SIFT Saliency**. A novel saliency operator, directly and solely using the precomputed SIFT features to estimate the saliency map

For the latter saliency function (SIFT), we build on the work of Bruce and Tsotsos [1]. In their proposal, the probability of each pixel is locally estimated by non-parametrically fitting a distribution over the ICA projection of the RGB values of the image. Similar to [1], after computing the ICA projection $\overline{\mathbf{X}} = [\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_N]^T$ of the SIFT description $\mathbf{X}$ of an image, we estimate the probability of the j-th dimension of a descriptor i as:

$$p(\bar{\mathbf{x}}_{i,j}) = \frac{1}{N} \sum_{k=1}^{N} K(\bar{\mathbf{x}}_{i,j} - \bar{\mathbf{x}}_{k,j}), \tag{1}$$

where $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ is a one-dimensional standard Gaussian kernel. The saliency of the local descriptor $\bar{\mathbf{x}}_i$ is then computed as:

$$s(\bar{\mathbf{x}}_i) = -\sum_{j=1}^{D} \log \bar{\mathbf{x}}_{i,j} \tag{2}$$

and the final saliency map is obtained by computing the responses for all the SIFT descriptors of the image, followed by a smoothing operation.

Task-driven spatial pooling Indoor scenes are designed to support human actions and humans have a limited range of spatial mobility. For example, humans cannot easily move from the floor to the ceiling, or access facilities if they are disposed too low, or too high in the room. This reduces the spatial variability of indoor scenes to lie mostly on the horizontal axis. Given this prior, we expect that by pooling features in horizontal bands we will be able to capture the most consistent spatial patterns in indoor scenes. We instead expect less robust results by pooling descriptors in vertical bands.
To verify this intuition we thus compare the following pooling schemes:

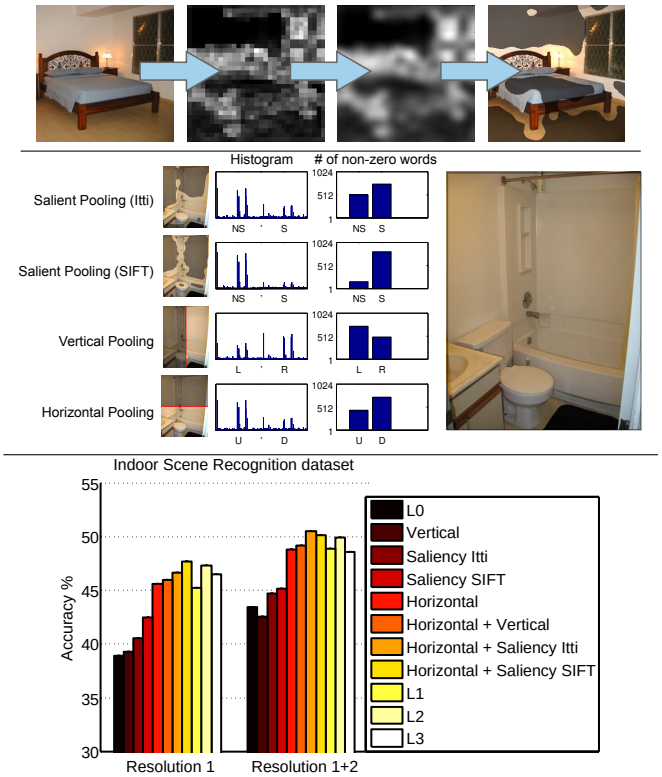


Figure 1: Top: Computation of a SIFT saliency map and resulting segmentation. Middle: Histograms obtained with different pooling techniques and number of non-zero visual words in each of the two halves of the histograms: non-salient (NS) and salient (S), left (L) and right (R), up (U) and down (D). Bottom: Performances of the different pooling strategies on the Indoor Scene Recognition [6] dataset.

- Horizontal-bands pooling. In this settings R_1 consists of the upper 50% of the image, while R_2 is its complement
- Vertical-bands pooling. In this case R_1 consists of the left-side 50% of the descriptors, and R_2 is again its complement

We performed experiments on three widely used scene recognition datasets: the Indoor Scene Recognition (*ISR*) [6], the 15-Scenes [3] and the 8-Sports [4] datasets. A visualization of the pooling strategies, together with the resulting histograms and a performance evaluation on the ISR dataset are shown in Fig. 1. We see that the salient pooling strategies perform better than the vertical one (+8.1% relative improvement). Moreover, when combined with the horizontal pooling, they always outperform the Horizontal + Vertical and the *L*1 spatial pyramid baselines, being also competitive with much higher dimensional representations, like *L*2 / *L*3.

[1] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, 2006.
[2] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 2002.
[3] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.
[4] L.J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.
[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.
[6] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

Face Recognition using Local Quantized Patterns

Sibt ul Hussain
Sibt.ul.Hussain@gmail.com
Thibault Napoléon
Thibault.Napoleon@unicaen.fr
Fréderic Jurie
Frederic.Jurie@unicaen.fr

GREYC — CNRS UMR 6072,
University of Caen Basse-Normandie,
Caen, France

We propose a novel face representation based on Local Quantized Patterns (LQP) [3]. Our this new flexible representation not only outperforms any other representation on challenging face datasets but performs equally well in the intensity space and orientation space (obtained by applying gradient or Gabor Filters) and hence is intrinsically robust to illumination variations. Extensive experiments on two challenging face recognition datasets (FERET [4] and LFW [2]) show that this representation gives state-of-the-art performance (improving the earlier state-of-the-art by around 3%) without requiring neither a metric learning stage nor a costly labelled training dataset, having the comparison of two faces being made by simply computing the Cosine similarity between their LQP representations in a projected space.

Contributions: We introduce a complete framework (*c.f*. Figure 1) for face recognition that combines *(i)* a well designed local pattern descriptor with *(ii)* a simple PCA-based similarity metric to achieve state-of-the-art accuracy rates. Our presented method is not only very simple and efficient, but also has very good generalization capability: it outperforms any existing unsupervised method and many supervised methods on all the tested datasets.

Local Quantized Patterns: LQP [3] is a generalized form of local patterns (Local Binary Patterns (LBP) [1], Local Ternary Patterns (LTP) [5], *etc*.) that uses large local neighbourhoods and/or deeper quantization with domain-adaptive vector quantization to obtain highly discriminant representation. We tailor and use these LQP features for face representation. Precisely we use Disk LQP layout (*c.f*. Figure 2) to sample pixels from the local neighbourhood and use a tolerance value (τ) to generate a pair of binary codes (as in LTP) and quantize each one using a separately learned codebook. We propose two different types of Disk LQP features for face representations: (i) *I-LQP:* LQP features are computed on simple raw intensity images; *(ii) G-LQP:* LQP features are computed from Gabor filtered images obtained by convolving the image with multi-scale multi-orientation Gabor kernels – we use 40 different Gabor kernels that span 5 different scales and 8 different orientations over the range 0 to 2π. Moreover in *G-LQP*, we concatenate the LQP computed codes from the neighbouring scales and orientations at the local pattern level. This helps to capture the patterns co-occurrence statistics over neighbouring scales and orientations and leads to a highly discriminant face descriptor.

Matching Faces via Cosine Similarity Metric: For comparing face images we use Cosine similarity in a reduced feature space. Precisely, we first use Principal Component Analysis (PCA) to project high dimensional LQP features to a low-dimensional uncorrelated space. Next, to reduce the influence of leading principal components and to have the projected features with same variance, we perform data sphering and divide all the principal components by square-roots of their corresponding eigenvalues. Finally, unlike conventional approaches (*e.g*. [1]) that use a distance-based similarity metric such as Euclidean, we use angle-based Cosine Similarity (CS) (*i.e*. $CS(d_1,d_2) = (d_1^T \cdot d_2)/(||d_1||\ ||d_2||)$) metric to compare faces in the normalized projected space. Although we also tested other metrics such as Pearson Correlation Coefficient which gave similar results, but Cosine similarity metric was preferred due to its fast computation time.

Experiments: We have experimentally validated our approach on two different face recognition tasks: i) *Face verification*, for this task we used the popular Labeled Faces in the Wild (LFW) dataset [2]; ii) *Face identification*, for this task we use the Face Recognition Technology (FERET) dataset [4]. Table 1 compares the performance of our methods with several competing supervised and unsupervised methods on FERET dataset. Table 2 reports comparative results[1] on LFW test set (View 2).

[1] Majority of the results are reproduced from the LFW dataset webpage.

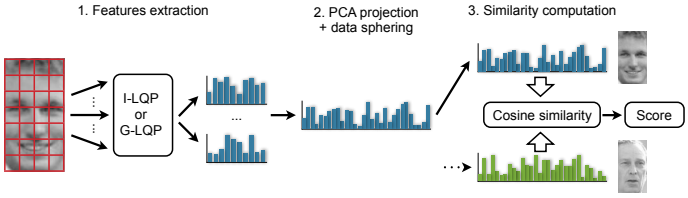


Figure 1: Overview of our face recognition framework.

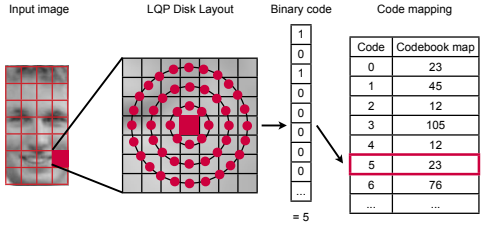


Figure 2: Overview of LQP feature computation. LQP samples pixels from a Disk layout around a central pixel and generates a binary/ternary vector and then map the resulting code to nearest codebook word via a pre-built lookup/hash table.

	Methods	Fb	Fc	Dup-I	Dup-II	Mean	*Comments*
1	HOG	90.0	74.0	54.0	46.6	66.2	
2	LBP	93.0	51.0	61.0	50.0	63.8	
3	LGBPHS	94.0	97.0	68.0	53.0	78.0	*Gabor+LBP*
4	LGBPWP*	98.1	98.9	83.8	81.6	90.6	*Gabor+LBP+WPCA*
5	POEM*	99.6	99.5	88.8	85.0	93.2	*Gradient+LBP+WPCA+R.Filtering*
6	Tan&Triggs	98.0	98.0	90.0	85.0	92.8	*Gabor+LBP+DoG Filtering+supervised*
7	**I-LQP**	99.2	69.6	65.8	48.3	70.7	Computed on intensity images
8	**I-LQP***	99.8	94.3	85.5	78.6	89.6	
9	**G-LQP**	99.5	99.5	81.2	79.9	90.0	Computed on Gabor-filtered images
10	**G-LQP***	**99.9**	**100.0**	**93.2**	**91.0**	**96.0**	

Table 1: Comparative results on FERET dataset. Superscript '*' is used to differentiate methods using PCA-projected features with Cosine similarity metric from the ones using raw features with Chi-squared distance metric.

	Methods	Accuracy (%)±S_E
UnSupervised	SD-MATCHES	64.1±0.62
	GJD-BC-100	68.5±0.65
	H-XS-40	69.5±0.48
	LARK	72.2±0.49
	POEM	75.2±0.73
	POEM*	82.7±0.59
	G-LQP	**75.3±0.26**
	G-LQP*	**82.1±0.26**
	I-LQP	**75.3±0.80**
	I-LQP*	**86.2±0.46**

	Methods	Accuracy (%)±S_E	*Comments*
Supervised	S.LE+holistic	81.2± 0.53	
	DML-eig SIFT	81.3± 0.23	
	Hybrid	84.0± 0.35	
	POEM*	84.9± 0.45	
	LARK	85.1± 0.59	
	LBP + CSML	85.3± 0.52	
	DML-eig comb	85.7± 0.56	
	Combined b/g	86.8± 0.34	*10 features+Metric learning+SVM*
	CSML + SVM	88.0± 0.37	*6 features+WPCA+CSML+SVM*
	HTBIF	88.4± 0.58	*1000+ Gabor filters+4 distances+SVM*

Table 2: Comparative results of our methods with (left) unsupervised and (right) supervised methods on aligned LFW View 2 dataset.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE TPAMI*, 2006.

[2] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07–49, UoM, 2007.

[3] S. Hussain and B. Triggs. Visual recognition using local quantized patterns. In *ECCV*, 2012.

[4] P. Phillips, H. Wechsler, J. Huang, and P. Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *TIVC*, 1998.

[5] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE TIP*, 2010.

Context-Aware Keypoint Extraction for Robust Image Representation

Pedro Martins[1]
pjmm@dei.uc.pt
Paulo Carvalho[1]
carvalho@dei.uc.pt
Carlo Gatta[2]
cgatta@cvc.uab.es

[1] Centre for Informatics and Systems
University of Coimbra
Coimbra, Portugal

[2] Computer Vision Centre
Autonomous University of Barcelona
Barcelona, Spain

We introduce a context-aware keypoint extractor, coined as CAKE, aimed at capturing the most informative image content. We find this algorithm particularly useful in tasks such as image retrieval, scene classification, and object (class) recognition, in which local features are mainly used to provide a robust and efficient image representation. We are motivated by the fact that the majority of local feature extractors are designed to respond to a reduced number of structures. Furthermore, we observe that the existent complementarity among feature sets is often neglected. Our context-aware algorithm is designed to respond to complementary features as long as they are informative. In the particular case of images with different types of structures, one can expect a high complementarity among the features retrieved by a context-aware extractor. By contrast, images with repetitive patterns will inhibit our method from retrieving a clear summarised description of the image content. Nonetheless, the extracted set of features can be complemented with a counterpart that retrieves the repetitive elements in the image. These two cases are depicted in Figure 1. The upper image shows a context-aware keypoint extraction on a well-structured scene, which retrieves the 100 most informative keypoints. This small number of features is sufficient to provide a good coverage of the content, which includes different types of structures. The lower image illustrates the advantages of combining context-aware keypoints with strictly local ones (SFOP keypoints [2]) to obtain a better coverage of images with repetitive patterns.

An information theoretic framework is used to formulate our context-aware keypoint extraction. A keypoint will correspond to a certain image location within a structure with a low probability of occurrence (high information content). For each image location $\mathbf{x}$, we consider $\mathbf{w}(\mathbf{x}) \in \mathbb{R}^D$, any viable local representation (e.g, the Hessian matrix or the structure tensor matrix) as a "codeword" that represents the neighbourhood of $\mathbf{x}$. To define the saliency measure, we regard the image codewords as samples of a multivariate probability density function. We compute the probability of a codeword $\mathbf{w}(\mathbf{y})$ using a Kernel Density Estimator [4] in which the kernel is a multidimensional Gaussian function with zero mean and standard deviation σ_k:

$$\tilde{p}(\mathbf{w}(\mathbf{y})) = \frac{1}{N\Gamma} \sum_{\mathbf{x}\in\Phi} \mathrm{e}\left(-\frac{d^2(\mathbf{w}(\mathbf{y}),\mathbf{w}(\mathbf{x}))}{2\sigma_k^2}\right), \quad (1)$$

where d is a distance function, K is a kernel, Φ is the image domain, N represents the number of pixels, and Γ is a proper constant such that the estimated probabilities are taken from an actual PDF. From Eq. (1), the saliency measure at $\mathbf{y}$ is defined as

$$m(_{\mathbf{y}}, I(\Phi)) = -\log\left(\frac{1}{N\Gamma}\sum_{\mathbf{x}\in\Phi} e\left(-\frac{d^2(\mathbf{w}(\mathbf{y}),\mathbf{w}(\mathbf{x}))}{2\sigma_k^2}\right)\right), \quad (2)$$

where I denotes the image. In this case, context-aware keypoints will correspond to local maxima of $m(\cdot, I(\Phi))$ that are beyond a certain threshold. We use the *Mahalanobis distance* as the distance function d. Since this distance is invariant under affine transformations, we can draw the following result:

Property 1. *Let $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ be image codewords such that $\mathbf{w}^{(2)}(\mathbf{x}) = T(\mathbf{w}^{(1)}(\mathbf{x}))$, where T is an affine transformation. Let $p^{(1)}$ and $p^{(2)}$ be the probability maps of $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, i.e., $p^{(i)}(\cdot) = p(\mathbf{w}^{(i)}(\cdot))$, $i = 1,2$. In this case,*

$$p^{(2)}(\mathbf{x}) \le p^{(2)}(\mathbf{y}) \iff p^{(1)}(\mathbf{x}) \le p^{(1)}(\mathbf{y}), \forall \mathbf{x},\mathbf{y} \in \Phi.$$

We propose a strategy that includes approximating the KDE computations of a D-dimensional multi-variate PDF by estimating D separate univariate PDFs, which simplifies the computation of distances. Furthermore, we reduce the number of samples: samples that are close to each other are replaced by a new one that summarises the previous ones.

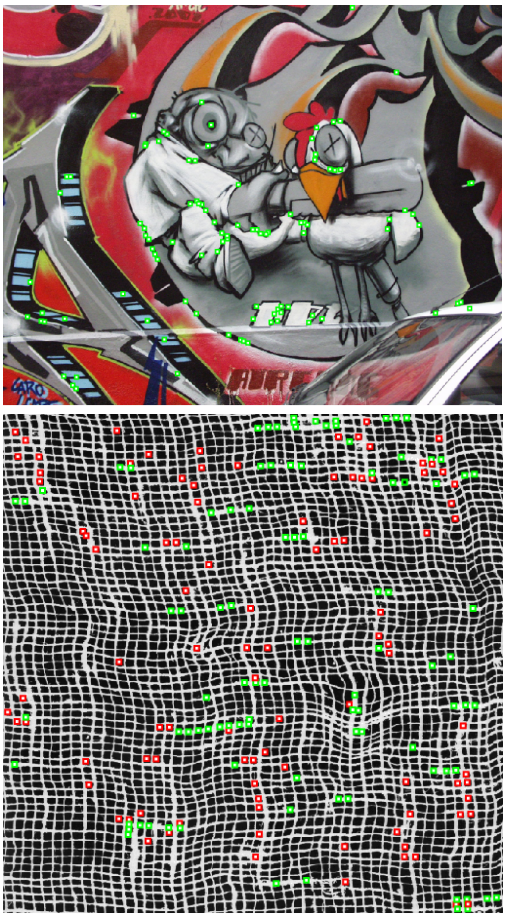


Figure 1: Proposed keypoint extraction. Upper image: Context-aware keypoints on a well-structured scene (100 most informative locations). Lower image: a combination of context-aware keypoints (green squares) with SFOP keypoints [2] (red squares) on a textured image.

The context-aware keypoint extractor is evaluated in terms of repeatability, completeness, and complementary. A multi-scale Hessian matrix is used as the codeword. Repeatability is evaluated following the standard protocol proposed by Mikolajczyk et al. [3]. Completeness and complementarity are evaluated on the benchmark proposed by Dickscheid et al. [1].

[1] T. Dickscheid, F. Schindler, and W. Förstner. Coding images with local features. *International Journal of Computer Vision*, 94(2):154–174, 2011.

[2] W. Förstner, T. Dickscheid, and F. Schindler. Detecting interpretable and accurate scale-invariant keypoints. In *IEEE International Conference on Computer Vision (ICCV'09)*, Kyoto, Japan, 2009.

[3] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A Comparison of Affine Region Detectors. *International Journal of Computer Vision*, 65(1/2): 43–72, 2005.

[4] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Person-Specific Subspace Analysis for Unconstrained Familiar Face Identification

Giovani Chiachia[1,2]
giovanichiachia@gmail.com
Nicolas Pinto[2,3]
pinto@mit.edu
William Robson Schwartz[4]
william@dcc.ufmg.br
Anderson Rocha[1]
anderson.rocha@ic.unicamp.br
Alexandre X. Falcão[1]
afalcao@ic.unicamp.br
David Cox[2]
davidcox@fas.harvard.edu

[1] Institute of Computing
University of Campinas
Campinas, Brazil

[2] Rowland Institute
Harvard University
Cambridge, USA

[3] McGovern Institute
Massachusetts Institute of Technology
Cambridge, USA

[4] Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

Present face recognition systems can surpass human performance in the task of matching unfamiliar faces from images acquired under relatively controlled conditions [2]. However, in settings where images are less controlled and where human subjects are familiar with the faces that are tested, the advantage of humans over machines is still substantial [3]. While the issue of uncontrolled variation has received increased attention in recent years, the notion of "familiarity" in face recognition systems has been relatively unexplored.

The recognition of "familiar" faces is increasingly relevant in an age where an ever-growing torrent of images of friends and family members is made available. While many current face recognition benchmark sets are organized around deciding whether two probe faces are the same or different, the problem is often instead that of recognizing which individual a given face image belongs to. This identification problem has a natural relationship to the notion of "familiarity" in human face recognition, in that a large number of past examples of a relatively small cohort of individuals are leveraged to recognize new examples.

A growing body of neuroscience and psychology research suggests that human face recognition with familiar and unfamiliar faces is substantially different, possibly even relying on qualitatively different internal representations. Indeed, while human performance with unfamiliar faces is generally poor, performance with familiar faces is excellent [1].

Inspired by the idea that humans may rely on enhanced representations for familiar individuals, in this work we explore the construction of per-individual subspaces for performing face identification. We build these subspaces from orthonormal projection vectors obtained using a person-specific configuration of partial least squares [5, 6], which we refer to as PS-PLS models. A key motivating idea for this work is that such person-specific subspaces, due to its supervised nature, can capture both those aspects of the face that are good for discriminating it from others, as well as natural variation in appearance that is present in the unconstrained images of that individual.

Partial least squares (PLS) is a class of methods primarily designed to model relations between sets $\mathbf{X}$ and $\mathbf{Y}$ of observed variables by means of latent vectors [5]. The projection vectors $\mathbf{w}$ that are the basis of our person-specific subspaces are determined iteratively by the NIPALS algorithm [6] such that

$$\max_{||\mathbf{w}||=||\mathbf{c}||=1} [cov(\mathbf{Xw}, \mathbf{Yc})]^2, \quad (1)$$

where $cov(\mathbf{Xw}, \mathbf{Yc})$ is the sample covariance between $\mathbf{Xw}$ and $\mathbf{Yc}$ and $\mathbf{c}$ is dependent on $\mathbf{w}$ in our case. For each person, we model $\mathbf{Y}$ as a set with a single indicator variable such that $\mathbf{Y}_{n\times 1} = \mathbf{y}_c$, and $y_{cs} = 1$ if sample s (out of n) belongs to class c or $y_{cs} = 0$ otherwise. In this case, obtaining projection vectors $\{\mathbf{w}\}_i$ is straightforward [5]. At each iteration i,

$$\mathbf{w}_i = \mathbf{X}i^T \mathbf{y}_c, \quad (2)$$

where $\mathbf{X}i$ is the matrix $\mathbf{X}$ deflated up to iteration i according to the NIPALS algorithm [5, 6].

We use the *Pubfig83* dataset, a subset of the *Pubfig* face dataset [3] reconfigured for the problem of unconstrained face identification [4]. We replicate the previous best results with this data set [4] and consider them as baselines. The baseline methods consist of binary linear support vector machines (SVMs) trained on different visual representations of faces in a one-versus-all setting. To compare these methods, we project feature descriptor vectors from this method into custom PS-PLS subspaces that we construct so that each binary linear SVM is trained in a different and person-specific space corresponding to the positive class (Fig. 1).

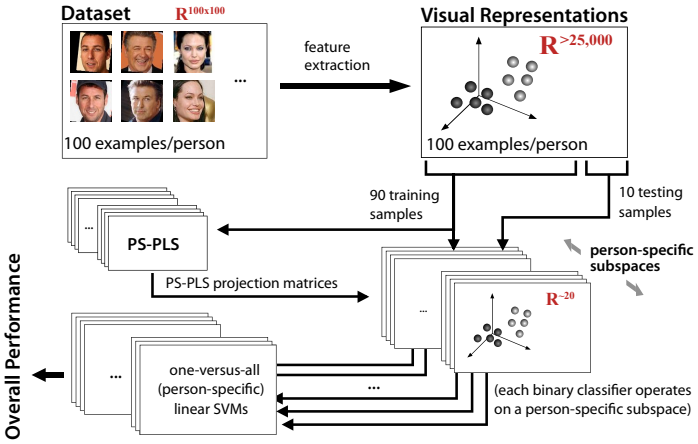


Figure 1: Our approach. From the training samples, PS-PLS creates a different face subspace for each individual. All training samples are then projected onto each subspace, so that a classification engine can be trained by considering the different representations of the samples over the subspaces. Given a test sample, an overall decision is made according to decisions made in each person-specific subspace.

We also compare our approach with subspaces built via other linear techniques. Among them, we consider person-specific principal component analysis (PCA), similar in spirit to the approach of Burton *et al.* [1], along with traditional non-person-specific PCA and linear discriminant analysis (LDA). Finally, as an additional test, we evaluate the approach on the *Facebook100* dataset [4], which is constructed from a large set of real-world face images taken from the Facebook social network.

With the use of the PS-PLS models, we could consistently get better results across the four different visual representations we consider. In the paper, we present details of the method and the results. In general, we argue that these subspaces are useful both for noise removal and for accentuating discriminative person-specific face aspects.

[1] A. M. Burton, R. Jenkins, and S. R. Schweinberger. Mental representations of familiar faces. *British Journal of Psychology*, 2011.

[2] P.J. Phillips et al. FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE TPAMI*, 2010.

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. *IEEE ICCV*, 2009.

[4] N. Pinto, Z. Stone, T. Zickler, and D. D. Cox. Scaling-up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. *IEEE CVPR*, 2011.

[5] R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. *LNCS*, 2006.

[6] H. Wold. Partial least squares. *Encyclopedia of Statistical Sciences*, 1985.

Image Classification by Hierarchical Spatial Pooling with Partial Least Squares Analysis

Jun Zhu[1]
junnyzhu@sjtu.edu.cn
Weijia Zou[1]
zouweijia@sjtu.edu.cn
Xiaokang Yang[1]
xkyang@sjtu.edu.cn
Rui Zhang[1]
zhang_rui@sjtu.edu.cn
Quan Zhou[2]
qzhou.lhi@gmail.com
Wenjun Zhang[1]
zhangwenjun@sjtu.edu.cn

[1] Institute of Image Communication and Information Processing
Shanghai Jiao Tong University
Shanghai, China

[2] Department of Electronics and Information Engineering
Huazhong University of Science and Technology
Wuhan, China

In recent image classification systems, spatial pooling is a key step to form image-level representation from patch-level local features. It captures meaningful statistical information of local feature codes over different ROIs, and achieves certain spatial invariance property to facilitate classification. On the spatial representation model, although the spatial pyramid is predominately used in image classification literature, its rigid structure may limit the resultant image representation from exploring richer spatial statistical information further.

Based on the tangram model, we construct a hierarchical ROI dictionary (called by HRD in this paper for short) for spatial pooling. Compared to rigid spatial pyramid model, it assembles the ROIs with more shape types, locations and scales, and is capable of retaining richer spatial statistical information. Besides, by taking advantage of mutual compositionality among ROIs, HRD can be inherently organized into a directed acyclic graph, and this derives an efficient hierarchical algorithm to facilitate spatial pooling. Besides, we further employ partial least squares (PLS) analysis for dimensionality reduction on the pooled features. It can capture the statistical relationship between pooled features and class labels for different visual words, and learn a more compact and discriminative image-level representation for classification. The experimental results demonstrate superiority of the proposed hierarchical pooling method w.r.t. spatial pyramid, on three benchmark datasets (Caltech-101, Caltech-256 and Scene-15) for image classification.

1 Hierarchical ROI dictionary

In this paper, a layered dictionary of shape primitives (called *tans*) is constructed to quantize the spatial configuration space. Formally, the tan dictionary $\Delta = \bigcup_{l=1}^{L} \Delta^{(l)}$ is an union of L subsets. $\Delta^{(l)} = \{B_{(l,i)} \mid i = 1,2,\cdots,N_l\}$ denotes a set of tans for the l^{th} layer, where $B_{(l,i)}$ refers to the i^{th} tan. When placing each tan onto different locations in the image lattice, one tan $B_{(l,i)}$ may produce a set of multiple instances $\{\Lambda_{(l,i,j)} \mid j = 1,2,\cdots,J_{(l,i)}\}$, which makes up a HRD $\mathcal{D}_\Delta$ for spatial pooling:

$$\mathcal{D}_\Delta = \bigcup_{l=1}^{L} \mathcal{D}_{\Delta^{(l)}},$$
$$\forall l, \quad \mathcal{D}_{\Delta^{(l)}} = \{\Lambda_{(l,i,j)} \mid i = 1,2,\cdots,N_l \text{ and } j = 1,2,\cdots,J_{(l,i)}\}.$$

Moreover, to describe the compositionality among tans, an associated And-Or graph (AOG) is accordingly built for organizing the ROIs in a deep hierarchy. The And-node represents that a tan can be composed by two smaller ones in layer belows, while the Or-node implies that it can be generated in alternative ways of shape composition. Fig.1(a) shows a 16-layer HRD for 4×4 grid, with an associated AOG illustrated in Fig.1(b).

2 Efficient Spatial Pooling in Deep Hierarchy

Based on the HRD, we can perform spatial pooling operation over the ROIs. Due to the over-completeness and increasing degree of freedom induced by recursive shape composition, the cardinality of HRD grows drastically with its granularity level so that direct spatial pooling operation on HRD is computationally demanding. However, the over-completeness and compositionality of the ROIs result in that each ROI in the HRD can be exactly composed by its child ones in the layers below. Considering

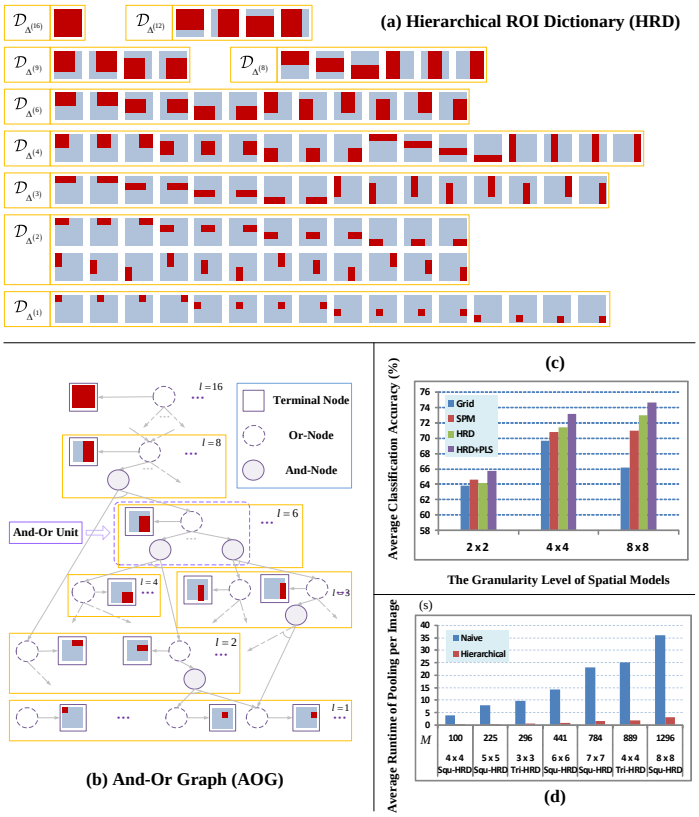


Figure 1: (a) Illustration on a 16-layer HRD. (b) Illustration of the associated AOG (only a portion of graph is shown for clarity). (c) Performance comparison on different spatial representation models (Caltech-101). (d) Comparison on runtime of spatial pooling algorithm.

the directed acyclic structure of associated AOG with deep hierarchy, we present an efficient algorithm for spatial pooling on HRD. Given a HRD and its associated AOG, the proposed pooling algorithm can be divided into two steps: I. For each ROI at the first layer, we directly compute its pooled feature from codes; II. For the other layers above, the pooled feature for each ROI is bottom-up propagated from its child nodes. Thus, most computational cost can be saved by taking advantage of recursive compositionality among the ROIs.

3 Learning Image Representation with PLS Analysis

Although the pooled features can be directly used for classification, for a HRD with large number of ROIs, it may produce a huge number of variables, which tend to be highly correlated and redundant. To obtain a more compact and discriminative image-level representation, we learn a PLS model for each visual word individually, to preserve class-specific discriminative information in the extracted representation. Finally, we perform dimension reduction via projecting pooled features onto the learned subspaces, obtaining a new image-level representation for classification.

Through-the-Lens Synchronisation for Heterogeneous Camera Networks

Evren Imre
h.imre@surrey.ac.uk
Adrian Hilton
a.hilton@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK

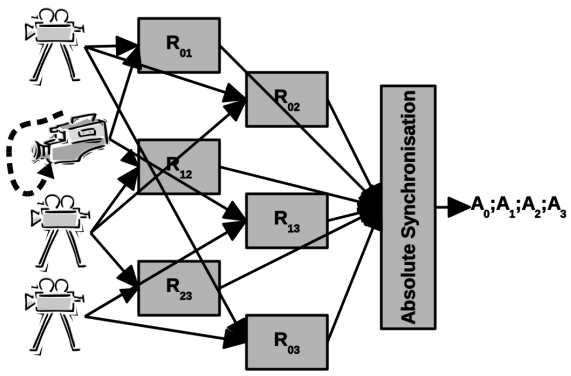


Figure 1: Overview of the algorithm, with one moving and 3 static cameras. R_{ij} blocks compute the relative synchronisation between the ith and the jth cameras. A_i is the absolute synchronisation for the ith camera.

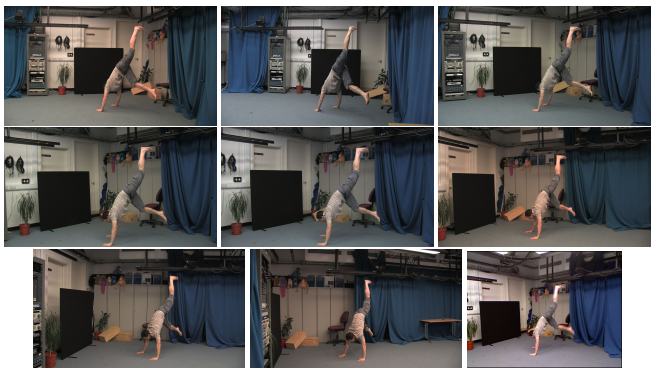


Figure 2: *Acrobatics. Top:* Cameras 0-2. *Middle:* Cameras 3-5. *Bottom:* Cameras 6-7, and the Kinect.

	0	1	2	3	4
α	1	1	1	1	1
τ	0	-11	-15	-31	-9
$\Delta\alpha$	0	5.8×10^{-7}	1.1×10^{-6}	7.6×10^{-7}	1.1×10^{-6}
$\Delta\tau$	0	5.8×10^{-4}	1.3×10^{-4}	3.7×10^{-4}	6.8×10^{-5}
	5	6	7	Kinect	
α	1	1	1	1.2	
τ	-27	-26	-32	≈ 26.8	
$\Delta\alpha$	1.2×10^{-6}	7.1×10^{-7}	1.7×10^{-6}	5.9×10^{-4}	
$\Delta\tau$	2.7×10^{-4}	1.9×10^{-5}	7.6×10^{-4}	6.2×10^{-2}	

Table 1: Absolute synchronisation estimates, with Camera 0 as the reference. α and τ are the ground-truth values of the synchronisation parameters, whereas $\Delta\alpha$ and $\Delta\tau$ indicate the *absolute value* of the estimation error. The ground-truth offset for the Kinect is not known, but manually determined from the sequence.

Camera synchronisation involves the temporal alignment of a set of video sequences, independently acquired by two or more cameras. Accurate synchronisation is crucial for a wide variety of applications requiring multi-camera setups, ranging from 3D modelling of dynamic scenes (*e.g.*, featuring a performance, or a sports event) to video surveillance and super-resolution. Conventional synchronisation methods, which typically rely on hardware or audio signals, have practical limitations, imposing constraints on the size and the span of the network [2][1]. Through-the-lens synchronisation offers a robust and flexible way to synchronise a camera network from the content it generates.

In this paper, we propose a bottom-up synchronisation algorithm to estimate a frame rate and an offset for each member of a network composed of 2 or more cameras. Our approach involves the computation of a relative synchronisation estimate between each camera pair, from which the absolute synchronisation parameters of the individual cameras are calculated (Figure 1). The algorithm can handle hybrid networks of static and moving cameras with different resolutions and frame rates, and does not require rigid objects, long trajectories or overlapping fields-of-view beyond 2 cameras. It needs a set of image features on the dynamic scene elements, and the geometric relation between the images (which can be obtained from the static background features).

Relative Synchronisation: The frame indices of the jth camera (t_j) with respect to those of ith (t_i) is defined by the line

$$t_j = \alpha_{ij} t_i + \tau_{ij}, \tag{1}$$

where α_{ij} is the relative frame rate, and τ_{ij} is the relative offset. The pair $R_{ij} = \{\alpha_{ij}; \tau_{ij}\}$ can be estimated by fitting a line to the indices of the corresponding frames via robust methods, such as RANSAC. The index correspondences are established via the Viterbi algorithm, which maximises an image similarity measure across the sequences, while enforcing the ordering constraint (*i.e.*, $\alpha_{ij} \geq 0$). In order to measure the similarity of two images, first the corresponding image features are identified, and then, their median deviation from the geometric relation (*e.g.*, epipolar constraint) is computed. The resulting integer-level correspondences can be refined to subinteger resolution, by minimising the deviation over 3-frame feature trajectories.

Absolute Synchronisation: R_{ij} is computed from the absolute synchronisation estimates for the ith and jth cameras as

$$\begin{aligned} \alpha_{ij} = \frac{\alpha_j}{\alpha_i} \Rightarrow \quad \alpha_{ij}\alpha_i - \alpha_j = 0 \\ \tau_{ij} = \tau_j - \frac{\alpha_j}{\alpha_i}\tau_i \end{aligned}, \tag{2}$$

where $\{\alpha_i; \tau_i\}$ and $\{\alpha_j; \tau_j\}$ are the synchronisation parameters for the cameras i and j, respectively. For an L camera set, L such linear independent constraints can be stacked into a linear system, whose solution yields the absolute synchronisation parameters. The sets of constraints that satisfy the linear independence requirement appear as the spanning cycles of the graph, where the nodes representing the cameras are linked by the edges representing the corresponding R_{ij}. The absolute synchronisation algorithm solves Equation 2 for each spanning cycle, and generates $\{\hat{R}_{ij}\}$, an estimate of the set of all available relative synchronisations $\{R_{ij}\}$. The absolute synchronisation is computed from the cycle, whose associated $\{\hat{R}_{ij}\}$ minimises the Euclidean distance of the index correspondences to their synchronisation lines. The result is further improved by the inclusion of all consistent R_{ij} to the minimal set, and refinement over their index correspondences.

Experimental Results: The algorithm is tested on 3 datasets, one of which is illustrated in Figure 2. The datasets feature 8 static and moving HD cameras, and a Kinect, capturing a non-rigid object undergoing rapid deformations. The results, presented in Table 1 indicate that the algorithm can achieve very high accuracy despite the challenging content.

Acknowledgements: This work is supported by the TSB project "SYMMM: Synchronising Multimodal Movie Metadata".

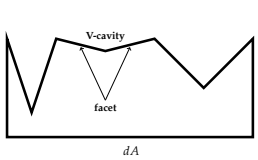
[1] N. Hasler, B. Rosenhahn, T. Thormahlen, M. Wand, J. Gall, and H.-P. Seidel. Markerless motion capture with unsynchronized moving cameras. In *Proc. CVPR*, pages 224–231, 2009.

[2] A. Miller. R&d and blue peter- ski rossendale free-viewpoint visualisation. http://www.bbc.co.uk/blogs/researchanddevelopment/2011/03/rd-and-blue-peter--ski-rossend.shtml.

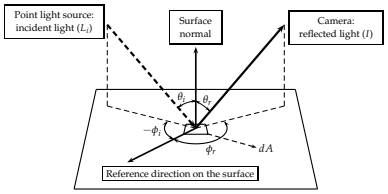
Shape from Shading for Rough Surfaces: Analysis of the Oren-Nayar Model

Yong Chul Ju
y.ju@mmci.uni-saarland.de
Michael Breuß
breuss@tu-cottbus.de
Andrés Bruhn
bruhn@vis.uni-stuttgart.de
Silvano Galliani
galliani@mia.uni-saarland.de

Vision and Image Processing
Saarland University, Germany
Institute for Applied Mathematics and Scientific Computing
BTU Cottbus, Germany
Institute for Visualization and Interactive Systems
University of Stuttgar, Germany
Mathematical Image Analysis Group
Saarland University, Germany



(a) Facet model for surface patch dA with many symmetric Lmabertian V-shape cavities.



(b) Illustration of diffuse reflectance parameters for SfS with Oren-Nayar reflectance.

Figure 1: Sketch of the Oren-Nayar surface reflection model.

Since almost five decades *Shape from Shading (SfS)* is one of the fundamental problems in computer vision. Having many interesting applications such as astronomy, terrain reconstruction, endoscopy or dentistry, the goal of SfS is to recover the surface of an object from a single input image under the assumption that a reflectance model and the light information are available. While, for a long time, research in SfS was mainly dominated by approaches based on relatively simple model assumptions such as an orthographic camera setup and a Lambertian surface model, recently more realistic concepts such as *perspective cameras* [4] and *non-Lambertian reflectance models* [5] found their way into research and led to considerable progress in the field.

One of these non-Lambertian reflectance models that seems particularly appealing is the *Oren-Nayar* reflectance model [2]. It allows to model rough materials such as concrete, plaster, clay or cloth realistically whose surface properties are considerably different from those of Lambertian surfaces. However, since the corresponding SfS models are rather involved, no theoretical analysis of such models with respect to model convexity or critical points has been performed so far. Moreover, it has not yet been investigated for which model parameters the popular and efficient fast marching (FM) method [6] can be applied safely to solve the resulting PDE of Hamilton-Jacobi type. This issue is particularly important, since the FM method has already shown impressive speed ups of up to factor 100 when applied on a pure experimental basis [7].

In our paper we perform such an in-depth analysis of the Oren-Nayar SfS model based on *Osher's criterion* [3]. This criterion allows to decide, if the FM method can be applied for solving a certain Hamilton-Jacobi equation, even if the underlying model is non-convex. By investigating this criterion for the Oren-Nayar model, we do not only succeed in providing concrete bounds for the model parameters that allow a safe application of the FM method, but we also put the findings of previous authors on a solid theoretical basis.

Figure 1 (a) illustrates the Oren-Nayar surface reflectance model that models rough surfaces by aggregating many Lambertian surface patches. In this context, the roughness is characterised by a Gaussian probability distribution of the patch slopes with standard deviation (roughness parameter) $\sigma \in [0, \frac{\pi}{2}]$. In our paper we investigate the SfS model by Ahmed and Farag [1] that additionally makes use of a light attenuation factor, that assumes the light source to be located in the centre of the camera and that models a perspective projection of the camera. This model is given by

$$H_{2D} = \mathtt{f}^2 I \frac{M+1}{A\sqrt{M+1}+BM} - e^{-2v} = 0 \tag{1}$$

with

$$M = \left[\mathtt{f}^2 |\nabla v(\mathbf{x})|^2 + (\nabla v(\mathbf{x}) \cdot \mathbf{x})^2\right] \left(\frac{|\mathbf{x}|^2 + \mathtt{f}^2}{\mathtt{f}^2}\right), \tag{2}$$

where the patch statistics σ and the different angles depicted in Figure 1 (b) enter the model via the local brightness I and the factors A and B [1].

To simplify the computations, the so-called Hamiltonian H_{2D} is expressed here in terms of the *logarithm* of the sought depth $v = \ln u$.

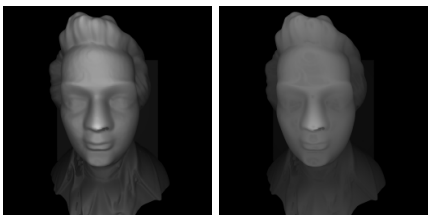
On the one hand we perform a *general* investigation of the model using Osher's criterion. In this context, we can prove that in 1-D for

$$0 \leq \sigma < \sqrt{0.3869067207} \approx 0.622$$

the FM method works by construction for the Oren-Nayar model. Moreover, we also conduct a *parameter dependent* analysis showing that the FM method can even be applied safely for a much wider range of settings that are particularly relevant for practical applications. Only strong discontinuites and very flat regions pose problems to the FM method. Based on the 1-D case we extend our theoretical findings also to the 2-D case. For the first time in the literature it becomes thus possible to *theoretically justify* the use of the FM method as solver for the SfS Oren-Nayar model which had been applied so far on a purely empirical basis only.

Numerical experiments demonstrate the validity of our theoretical analysis. They demonstrate a stable behaviour of the FM method for the predicted range of model parameters. As Figure 2 (d) shows, the FM method works reasonably well when the roughness parameter σ is within the admitted range. In constrast, in case σ is outside the range, the reconstruction shows significant problems, since the FM method fails.

Figure 2: Mozart Experiment.



(a) Input image, $\sigma = 0.5$. Fulfills theoretical bounds.

(b) Input image, $\sigma = \frac{\pi}{2}$. Exceeds theoretical bounds.

(c) Ground truth, $\mathtt{f} = 128$.

(d) Reconstruction of (a). Observation: FM works.

(e) Reconstruction of (b). Observation: FM fails.

[1] A.H. Ahmed and A.A. Farag. A new formulation for shape from shading for non-Lambertian surfaces. In *Proc. CVPR*, 2006.

[2] S.K. Nayar and M. Oren. Generalization of the Lambertian model and implications for machine vision. *IJCV*, 14(3):227–251, 1995.

[3] S. Osher. A level set formulation for the solution of the Dirichlet problem for Hamilton–Jacobi equations. *SIMA*, 24(5):1993.

[4] E. Prados and O.D. Faugeras. Perspective shape from shading and viscosity solutions. In *Proc. ICCV*, 2003.

[5] D. Samaras and D. Metaxas. Incorporating illumination constraints in deformable models for shape from shading and light direction estimation. *IEEE T-PAMI*, 25(2):247–264, 2003.

[6] J.A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 2nd edition, 1999.

[7] O. Vogel and E. Cristiani. Numerical schemes for advanced reflectance models for shape from shading. In *Proc. ICIP*, 2011.

Finding Groups of Duplicate Images in Very Large Datasets

Winn Voravuthikunchai
winn.voravuthikunchai@unicaen.fr
Bruno Crémilleux
bruno.cremilleux@unicaen.fr
Frédéric Jurie
frederic.jurie@unicaen.fr

GREYC — CNRS UMR 6072,
University of Caen Basse-Normandie,
Caen, France

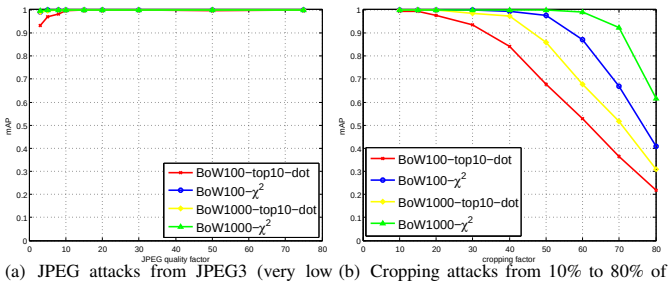
This paper addresses the problem of detecting groups of duplicates in large-scale unstructured image datasets such as the Internet. Leveraging the recent progress in data mining, we propose an efficient approach based on the search of *closed patterns*. Moreover, we present a novel way to encode the images based on bag-of-words vectors inspired by the text processing literature, that can be transformed into data mining transactions. Unlike other existing approaches, our method can scale gracefully to larger datasets as it has linear time and space (memory) complexities.

To encode the images as data mining transactions, we represent images by lists of their most top K informative visual words, using tf-idf weighting (term frequency-inverse document frequency). Tf-idf has been successful for normalizing BoW in vision tasks [2]. After representing images as transactions of items, we extract *all* closed itemsets whose length is greater than a given threshold (denoted *minlength*) as the length reflexes the similarity among the images containing the itemset. The minimum frequency (denoted *minfr*) support has to be set to 2 as two images can form a group of duplicates. Our mining strategy is based on LCM [3], which is one of the most efficient algorithms for mining frequent closed itemsets.

In our experiments, we first validate our proposed image binary representation in an image search scenario using the **Copydays dataset** [1]. Each of the 157 original image is used as a query to retrieve its corresponding attack image which is mixed in a set of 10,000 images. The average precision is computed and the mean is reported for all of the 157 queries. We compare the BoW based binary representation (with $K = 10$ and dot product similarity) to a standard BoW representation (with χ^2 distance) using two vocabularies size i.e. 100 and 1,000. We do the comparison by using the JPEG attacks (from JPEG3 to JPEG75) shown in Figure 1a and by using the cropping attacks (from 10% to 80%) shown in Figure 1b. In conclusion, this experiment demonstrate that this representation is sufficient for detecting near duplicate images, while being very compact (each image is encoded by ~13 bytes only).Furthermore, as this representation is made of lists of items, it can be used efficiently for finding frequent closed patterns.

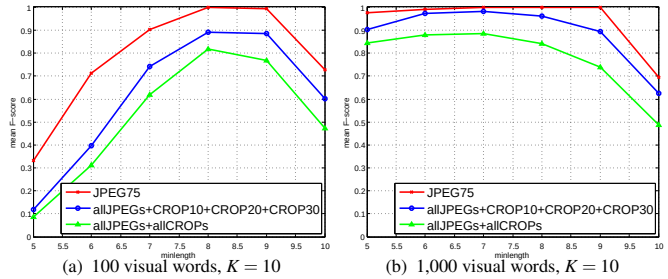
The following experiment validates the mining algorithm proposed for discovering groups of duplicate images. We use again the **Copydays dataset**, but in a different way: in these experiments, we put together the 157 original images, their corresponding attacked images, and 1,000,000 artificial image descriptors. In the ideal case, the algorithm should correctly discover the 157 groups of duplicates. The performance is evaluated by using the mean F-score which is equal to one only if the system outputs exactly 157 groups containing the original image and its transformations. Figure 2 shows the F-Score as a function of *minlength*, for representations made from 100 and 1,000 visual words dictionaries. $minlength = 7$ and 1,000 visual words dictionary give optimal results. We can see that for the light attacks, the groups of images are perfectly detected. Even for the strongest attacks the results are still very good.

Then we perform our method to detect groups of duplicates on our **One million random web images database**. We show that the computation time and the memory usage scales linearly with the size of the dataset in Figure 3b and Figure 3c. We are able to obtained more than 80 thousands groups of duplicates in less than 3 minutes. Figure 4 shows some of these groups. Beside computational efficiency, these results demonstrate the robustness against compression, scaling, slight crops, rotation, insertion/removal of small elements, brightness/contrast changes.



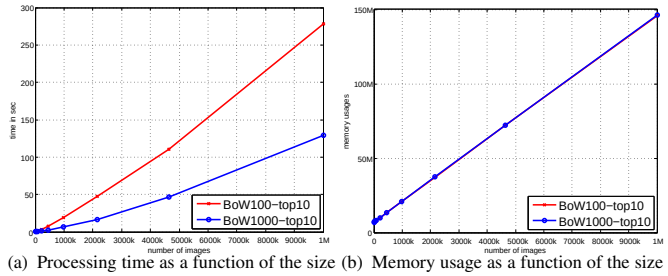
(a) JPEG attacks from JPEG3 (very low quality) to JPEG75 (typical web quality). (b) Cropping attacks from 10% to 80% of the image surface.

Figure 1: Image retrieval experiments: performance of the proposed representation and of the baseline representation, for two vocabularies (100 and 1,000 visual words).



(a) 100 visual words, $K = 10$ (b) 1,000 visual words, $K = 10$

Figure 2: Mean F-score as a function of *minlength*



(a) Processing time as a function of the size of the database (b) Memory usage as a function of the size of the database.

Figure 3: Computation time and memory usage as a function of the number of images

Figure 4: Some of the groups of duplicate/similar images found on the **One million random web images database**

[1] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid. Evaluation of gist descriptors for web-scale image search. In *CIVR*, pages 19:1–19:8, 2009.

[2] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[3] T. Uno, T. Asai, Y. Uchida, and H. Arimura. An efficient algorithm for enumerating closed patterns in transaction databases. In *proceedings of Discovery Science (DS'04)*, volume 3245 of *LNAI*, pages 16–31, Padova, Italy, 2004. Springer.

Fast and Robust Surface Normal Integration by a Discrete Eikonal Equation

Silvano Galliani
galliani@mia.uni-saarland.de
Mathematical Image Analysis Group
Saarland University, Germany

Michael Breuß
breuss@tu-cottbus.de
Institute for Applied Mathematics and Scientific Computing
BTU Cottbus, Germany

Yong Chul Ju
y.ju@mmci.uni-saarland.de
Vision and Image Processing Group
Saarland University, Germany

Since the integration of normal vectors plays an important role for reconstructing a surface, over decades it has been one of the most fundamental problems in computer vision and thereby extensively investigated by many researchers [6]. While many schemes have been proposed, there is, however, still a need for methods that combine accuracy, robustness and high efficiency. In view of efficiency, the fast marching (FM) [1, 3] method appears to be a natural candidate for an algorithmic approach, because the method gives us a complexity of $\mathcal{O}(N \log N)$, where N is the number of pixels of the computational domain, for the problems described by a static eikonal-type equation. In the work of Ho et al. [2] this strategy has been adopted, which is based on an analytic formulation of the integration task in terms of an eikonal equation. Whereas in [2] some promising results are presented, the authors also report significant problems with the robustness and accuracy of the scheme.

In this paper, we improve the scheme of Ho et al. [2] by proposing a complete discrete formulation (DEFM) in terms of a proper approximation of the underlying partial differential equation (PDE). Furthermore, by relying on pre-computed geodesic distance as a metric on the computational domain we extend our method in such a way that the DEFM can handle topologically more challenging domains, e.g. domains with holes.

From the fundamental theorem of calculus an antiderivative v in 1D is given by $\int v'(x_1)\,dx_1 = v(x_1) + c$ with a constant c. In 2D, this can be extended as

$$w(x_1,x_2) := v(x_1,x_2) + \lambda f(x_1,x_2)\,, \tag{1}$$

where $\lambda > 0$ is a constant parameter and f denotes a function. Since a function f in (1) should not change the important structure of w, specially critical points, in [2] as such a function

$$f_{\mathrm{Ho}} := x_1^2 + x_2^2 \tag{2}$$

is chosen which admits only one minimum at origin. For the deployment of FM, the expression in (1) is turned into an eikonal-type expression

$$|\nabla w| = |\nabla v + \lambda \nabla f_{\mathrm{Ho}}| = \sqrt{(v_{x_1} + \lambda\, 2x_1)^2 + (v_{x_2} + \lambda\, 2x_2)^2} \tag{3}$$

with $v_{x_1} := \frac{\partial v}{\partial x_1}$ and $v_{x_2} := \frac{\partial v}{\partial x_2}$. Since all elements on the right hand side of (3) are known, the FM method allows to compute w from the PDE $|\nabla w| = |\nabla v + \lambda \nabla f_{\mathrm{Ho}}|$. In the method of Ho et al. [2], the analytic formulation of ∇f_{Ho} in (2) is employed. However, since the analytic formulation has the same effect as the central difference method, the result by this method suffers from severe instability for solving (3) by the FM, see Figure 1(a).

In view of the properties from the underlying eikonal-type PDE and FM method, our main advancement stems from the deployment of a proper discretisation for (3) – *upwind scheme* [5]. In 1D, this upwind discretisation reads as

$$\hat{f}_x \;:=\; \max\left(D^- f, -D^+ f, 0\right) \tag{4}$$

with

$$D^- f = \frac{f_i - f_{i-1}}{\Delta x} > 0 \qquad \text{and} \qquad D^+ f = \frac{f_{i+1} - f_i}{\Delta x} < 0 \tag{5}$$

where Δx is the mesh width and f_j denotes a function value at a grid point $j \in \mathbb{Z}$. Each inequality in (5) holds for consistency since the upwind scheme chooses only one direction for the propagation of the information.

Our scheme analysis based on [4] shows that the proposed method is monotone and thereby stable if

$$\lambda \geq \varepsilon > 0\,, \tag{6}$$

where ε is a very small pre-defined constant. This suggests that the proposed method gives us no restrictions for the choice of the parameter λ in (3) in contrast to the case of Ho et al.

(a) Optimal result by the scheme of Ho et al. (b) Generic result by our method.

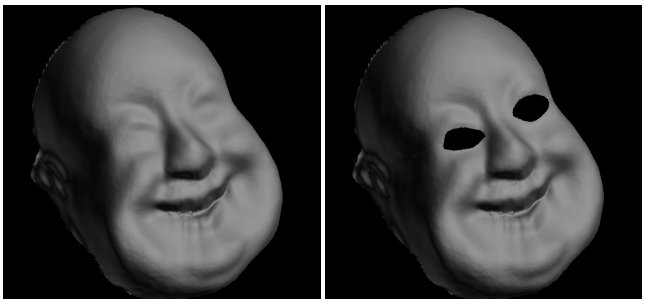
Figure 1: Reconstruction results by each method.

As shown in Figure 1 and Table 1, numerical experiments confirm our analysis in that even with very large λ values the present result outperforms in all error measures.

Table 1: Error measurements for Lena experiment shown in Figure 1.

	Mean	Median	Standard deviation
Ho et al. ($\lambda = 0.2$)	0.3060	0.2079	0.3604
Our method ($\lambda = 1000000$)	0.0785	0.0364	0.1325

Moreover, in order to deal with topologically more challenging computational domains we employ the more general *geodesic distance* for the function f in (1) instead of L_2 metric given in (2). Our numerical experiment again verifies that the geodesic measurements can handle non-trivial integrations domains accordingly as shown in Figure 2.



(a) Reconstruction without a mask. (b) Reconstruction with a mask.

Figure 2: Renderings of the Buddha face.

[1] J.A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 2nd edition, 1999.

[2] J. Ho and J. Lim and M.-H. Yang and D.J. Kriegman. Integrating surface normal vectors using fast marching method. In *Proc. ECCV*, 239–250, 2006.

[3] J.N. Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE T-Automatic Control*, 40 (9): 1528–1538, 1995.

[4] R.J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser, 1992

[5] E. Rouy and A. Tourin. *A viscosity solutions approach to shape-from-shading*. *SINUM*, 29(3): 867–884, 1992.

[6] J.-D. Durou and J.-F. Aujol and F. Courteille *Integrating the normal field of a surface in the presence of discontinuities*. In *Proc. EMMCVPR*, 261–273, 2009.

Multi-step flow fusion: towards accurate and dense correspondences in long video shots

Tomás Crivelli, Pierre-Henri Conze, Philippe Robert, Matthieu Fradet, Patrick Pérez
Technicolor
firstname.lastname@technicolor.com

With high quality editing of video shots of arbitrary duration in mind, we focus on this problem: how to construct accurate dense fields of correspondences over extended time periods using series of elementary optical flows. Highly elaborated optical flow estimation algorithms are at hand, and they were applied before for dense tracking by simple accumulation, however with unavoidable position drift. On the other hand, direct long-term point matching is more robust to such deviations, but is very sensitive to ambiguous correspondences. Why not combining the benefits of both approaches? Following this idea, we develop a *multi-step flow fusion* method that optimally generates a dense long-term displacement field by first merging several candidate estimation paths and then filtering the tracks in the spatio-temporal domain. Our approach permits to handle small and large displacements with improved accuracy and is able to recover a trajectory from temporary occlusions.

Consider a sequence of RGB images $\{I_n\}_{n:0\ldots N}$. Let $\boldsymbol{d}_{n,m} : \Omega \to \mathbb{R}^2$ be a *displacement field* defined on the continuous rectangular domain Ω, such that for every $\boldsymbol{x} \in \Omega$ it corresponds a *displacement vector* $\boldsymbol{d}_{n,m}(\boldsymbol{x}) \in \mathbb{R}^2$ for the ordered pair of images $\{I_n, I_m\}$. Given a *reference image*, say I_0, point tracking is compactly represented by $\boldsymbol{d}_{0,m}(\boldsymbol{x}) \forall m : 1 \ldots N$ (*from-the-reference* correspondences), *i.e.*, the set of displacement fields from I_0 to the subsequent frames I_m. Instead, for propagating (pulling) information present at a key reference frame to the rest of the sequence it is more natural to deal with $\boldsymbol{d}_{n,0}(\boldsymbol{x}) \forall n : 1 \ldots N$ (*to-the-reference* correspondences).

We address the problem of estimating from-the-reference as well as to-the-reference long-term displacement fields from elementary optical flow fields. Temporal integration of successive optical flow fields using classic tools such as Euler and Runge-Kutta schemes is possible (for instance in [3, 4]) but flow estimation errors are inevitably accumulated through this process. A solution would be to estimate the direct displacements between the reference frame and the other frames. However the longer the distance in time between two frames, the more ambiguous the matching process. So-called large displacement dense matching algorithms deal either with fast motions between consecutive frames [1] (but are not at all oriented to finding point correspondences along hundreds of frames) or assume parametric models [5] also constrained to limited frame distances. However, matching non-consecutive (time distant) frames can still be very useful as its accuracy much depends on inter-frame motion range: indeed one observes that for short/mid-term dense point matching, some regions of the image are better matched by concatenating consecutive motion vectors, while for others a direct matching is preferred (e.g., if displacement between consecutive frames is small). So, the idea is to consider multiple displacement fields with various inter-frame distances in order to have the best vectors available among all the candidates. The process is carried out in three phases: considering a pair $\{I_n, I_m\}$, first elementary optical flow fields with various inter-frame distances (called steps) are estimated. Then, various candidate displacement fields $\boldsymbol{d}_{n,m}$ are computed by different concatenations of the elementary fields, and finally the displacement field is obtained by merging these candidate fields. This is called Multi-Step Fusion (MSF).

Let us consider the pair $\{I_n, I_0\}$, corresponding to respectively the current and reference frames. Suppose that as an input we are given a set $\boldsymbol{M}_n$ of Q_n elementary optical flow fields $\boldsymbol{v}_{n,t}$ (between frames n and t): $\boldsymbol{M}_n = \{\boldsymbol{v}_{n,n+s_1}, \boldsymbol{v}_{n,n+s_2}, \ldots, \boldsymbol{v}_{n,n+s_{Q_n}}\}$ for image I_n. Considering any step s_k, one can compute a set of displacement vectors between I_n and I_0 resulting from the combination of the elementary vector $\boldsymbol{v}_{n,n+s_k}$ and the displacement $\boldsymbol{d}_{n+s_k,0}$ available between I_{n+s_k} and I_0:

$$\boldsymbol{d}^k_{n,0}(\boldsymbol{x}) = \boldsymbol{v}_{n,n+s_k}(\boldsymbol{x}) + \boldsymbol{d}_{n+s_k,0}(\boldsymbol{x} + \boldsymbol{v}_{n,n+s_k}(\boldsymbol{x})). \quad (1)$$

In this manner we generate different candidate displacements or *paths* (Fig. 1) among which we aim at deciding the optimal for each pixel $\boldsymbol{x}$.

The selection of the optimal path for all the points of the grid for a pair $\{I_n, I_0\}$ is achieved via an appropriate global optimization stage that fuses all the candidate fields into a single optimal displacement field. To do so, we apply the method recently presented in [2] in the context of instantaneous optical flow estimation by flow fusion.

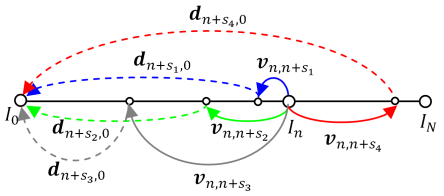


Figure 1: Multi-step point correspondence. The displacement from frame I_n to frame I_0 can be generated following different *paths* according to the available elementary motion fields (solid lines) and the previously estimated long-term displacements (dashed lines).

Figure 2: The *for rent* logo was inserted at frame I_0 by the user and was then automatically inserted in the other frames. The proposed MSF method overcomes the large occlusion by the arm while temporal integration of single step (equal to 1) optical flows (STEP1) fails.

At the end of the multi-step fusion stage, the set of forward vectors $\boldsymbol{d}_{0,n}(\boldsymbol{x})$ that give the position of point, originally at $\boldsymbol{x}$ in frame I_0, in subsequent frames I_n describe its trajectory along the sequence. We take these trajectory features taken into account in a next filtering stage comprising two steps.

First, for all pairs $\{I_0, I_n\}$, forward displacement fields $\boldsymbol{d}_{0,n}$ are spatio-temporally filtered considering spatiotemporal neighbouring forward vectors as well as the trajectories of spatial neighbouring pixels in the reference frame. To do so, the weights of our multilateral filter depend on spatial distance, colour similarity, matching cost and on a trajectory similarity that we introduce.

Until now, forward and backward displacement fields $\boldsymbol{d}_{0,n}$ and $\boldsymbol{d}_{n,0}$ have been estimated independently and carry complementary or contradictory information. In a second stage, they can be advantageously combined in a mutual refinement step. To this end, we present a joint multilateral filtering approach, both forward/backward and backward/forward.

Our experiments show that the optimal combination of short and long term matching does its job reducing the drift compared to optical-flow integration. Concerning temporary occlusions, while for single step methods (STEP1) it is impossible to estimate the trajectories of the occluded pixels after the occlusion (finally attaching all the tracks to the motion of the occluding object, which obliges to stop the trajectory), our multi-step fusion algorithm is able to circumvent the problem thanks to the long-step input displacement fields, as illustrated in Fig. 2.

[1] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *PAMI*, 33(3):500–513, 2011.

[2] V. Lempitsky, S. Roth, and C. Rother. Fusionflow: Discrete-continuous optimization for optical flow estimation. In *CVPR*, 2008.

[3] J. Lezama, K. Alahari, J. Sivic, and I. Laptev. Track to the future: Spatio-temporal video segmentation with long-range motion cues. In *CVPR*, 2011.

[4] N. Sundaram, T. Brox, and K. Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *ECCV*, 2010.

[5] J. Wills, S. Agarwal, and S. Belongie. A feature-based approach for dense segmentation and estimation of large disparity motion. *IJCV*, 68(2):125–143, 2006.

Fusing Structured Light Consistency and Helmholtz Normals for 3D Reconstruction

Michael Weinmann
mw@cs.uni-bonn.de
Roland Ruiters
ruiters@cs.uni-bonn.de
Aljosa Osep
osep@cs.uni-bonn.de
Christopher Schwartz
schwartz@cs.uni-bonn.de
Reinhard Klein
rk@cs.uni-bonn.de

Institute of Computer Science II,
University of Bonn,
Germany

For obtaining highly accurate 3d reconstructions, several methods combine positional information with normal information (e.g. [2], [1], [4]). Whereas triangulation-based 3D reconstruction techniques such as structured light or laser scanners typically suffer from noise or over-smoothing, reconstruction techniques based on normal information are capable of preserving high-frequency surface details but are prone to low-frequency drift due to the accumulation of errors. Fusing both types of information overcomes the individual problems.

In this paper, we propose a combination of normals estimated via Helmholtz stereopsis with structured light. In contrast to photometric stereo techniques, using Helmholtz normals has the advantage that it is largely BRDF-invariant. This is also true for structured light and, hence, our approach can be a applied to a wide range of materials. Furthermore, the structured light directly provides information about occlusion and shadowing that can be utilized in the Helmholtz stereopsis. We use a turntable-based setup which is capable of acquiring reciprocal image pairs as well as the structured light scans in an efficient and automated way.

Usually, when employing a turntable, the triangulations are performed independently for each turntable configuration. This way, not all cameras which have seen a certain location on the object surface at the same time are combined to compute one consistent point. Instead, several possibly contradicting point clouds have to be unified in the final surface reconstruction. To overcome this limitation and obtain one globally consistent reconstruction integrating the information over all rotations, we use a variational approach which combines a consistency term derived from the structured light with the Helmholtz normals. Here, the structured light consistency term allows us to combine several structured light measurements although the object was moved with respect to the projector.

This variational problem for the reconstruction of the object surface δV depends on a vector field of normals $\mathbf{H}$ and three scalar fields defined on the continuous volume $\mathbb{R}^3$: the consistency measure c, the outside count o and the visibility count v. At each point $\mathbf{x} \in \mathbb{R}^3$ and for all combinations (i,j,k) of rotations, cameras and projectors, we perform an independent classification. Utilizing the structured light information, we count the number of times the point is consistent ($c(\mathbf{x})$), lies before the surface ($o(\mathbf{x})$) and is thus outside of the object and the total number $v(\mathbf{x})$ of configurations in which it was visible from the camera. From these, we derive visibility-normalized versions $\hat{c} = \frac{c}{v}$ and $\hat{o} = \frac{o}{v}$.

Furthermore, the normal field $\mathbf{H}(\mathbf{x})$ is estimated via the Helmholtz principle [5]. We exploit the almost symmetrical mounting of the light sources around the cameras in our turntable-based setup. To compensate for the non-perfect symmetry between light source positions and camera positions in the setup, we relax the assumption of perfect Helmholtz stereopsis. Assuming that the BRDF is locally smooth enough allows us to use barycentric interpolation between three neighbouring light samples to approximate the brightness at the ideal position.

The resulting consistency-weighted vector field $c\,\mathbf{H}$ has a large magnitude in the vicinity of the surface, is aligned perpendicular to the surface and diminishes further away. To find the object interior $V \subset \mathbb{R}^3$, we seek to solve the following problem

$$\min_V \quad E(V) = -\lambda_1 \underbrace{\int_{\delta V} \langle c\,\mathbf{H}, \mathbf{n} \rangle \, \mathrm{d}A}_{E_1} + \lambda_2 \underbrace{\int_V \hat{o} \, \mathrm{d}V}_{E_2} + \lambda_3 \underbrace{\int_{\delta V} (\alpha - \hat{c}) \, \mathrm{d}A}_{E_3}, \tag{1}$$

where λ_1, λ_2 and λ_3 are relative weights of the individual terms and $\alpha > 1$ denotes a constant determining the minimum regularization strength within a consistent region. The first term E_1 considers the flux of the vector field $c\,\mathbf{H}$ through the object surface. This term is minimized by a surface running perpendicular to the reconstructed Helmholtz normals $\mathbf{H}$ and in regions with a high consistency c. The second term E_2 is used as an outside constraint to penalize regions of large values $\hat{o}$. This prevents the algorithm from short-cutting through concavities. The last term E_3 represents a regularization term and enforces a minimal surface. This penalty is weighted with the consistency $\hat{c}$ obtained from the structured light.

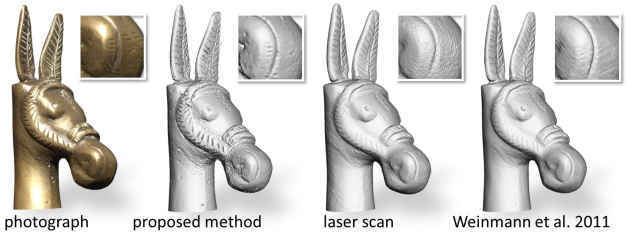


Figure 1: **Comparison of results on a glossy brass figurine.** The laser-scan was created with a high-precision line laser-scanner mounted on a measuring arm with a total accuracy of about $60\mu m$. Additionally, the results from the structured light based reconstruction exploiting projector super-resolution [3] are shown. Our reconstruction shows considerably more fine surface details.

We solve this optimisation problem via an octree-based continuous min-cut framework which is memory efficient and alleviates metrification errors. To compensate for the discretisation artefacts from the min-cut, a smooth signed distance function is then computed from the resulting binary labelling, again taking the reconstructed normals into account. Finally, the reconstruction result is derived from this smooth signed distance function obtained from the min-cut and the Helmholtz normals.

Further implementation details of the proposed approach are described in the paper. We demonstrate that the combination of structured light scanning with Helmholtz normal estimation enables the reconstructions of high-quality 3D models with a considerable amount of fine surface details. In addition, using Helmholtz normals instead of photometric stereo for normal recovery allows to handle objects with a significantly larger range of complex surface reflectance behaviour.

[1] C. Hernandez Esteban, G. Vogiatzis, and R. Cipolla. Multiview photometric stereo. *TPAMI*, 30(3):548–554, 2008.

[2] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM Trans. Graph.*, 24:536–543, 2005.

[3] M. Weinmann, C. Schwartz, R. Ruiters, and R. Klein. A multi-camera, multi-projector super-resolution framework for structured light. In *3DIMPVT*, pages 397–404, 2011.

[4] Y. Yoshiyasu and N. Yamazaki. Topology-adaptive multi-view photometric stereo. In *CVPR*, pages 1001–1008, 2011.

[5] T. Zickler, P. N. Belhumeur, and D. J. Kriegman. Helmholtz stereopsis: Exploiting reciprocity for surface reconstruction. In *IJCV*, pages 869–884, 2002.

Resolution-Aware 3D Morphable Model

Guosheng Hu
g.hu@surrey.ac.uk
Chi Ho Chan
chiho.chan@surrey.ac.uk
Josef Kittler
j.kittler@surrey.ac.uk
William Christmas
w.christmas@surrey.ac.uk

Centre for Vision, Speech and Signal Processing
University of Surrey
Guildford, UK

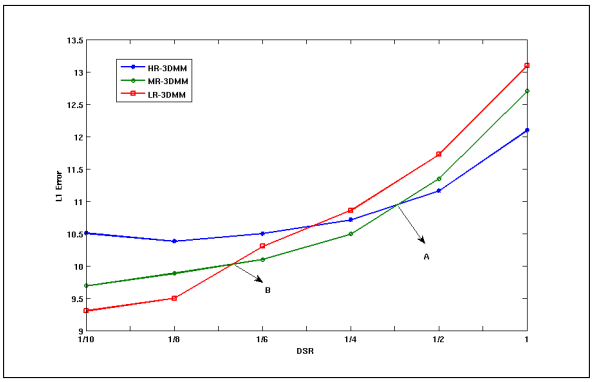
Face reconstruction models can be classified into two groups: the 2D face models and the 3D face models. The 2D-based group includes Active Appearance Models (AAMs) [4], which achieved great success. But the problem for AAMS is that the reconstruction with AAMs fails if the in-depth rotation of the face becomes large. 3D-based methods have been proposed to solve the problem. 3D morphable model (3DMM) [1] is a well-known statistical model for face reconstruction. It is very interesting to investigate how the fitting performance is affected by the resolution of 3DMM and of the input image. In this paper, the relationship between 3DMM resoluton and input image resolution is studied and a Resolution-Aware 3DMM (RA-3DMM) model is proposed. The construction of RA-3DMM is motivated by the assumption that **a high resolution 3DMM fits high resolution input images better and a low resolution 3DMM fits low resolution input images better**. Based on this assumption, a set of 3DMMs of different resolution should work better than a single 3DMM if the input images are of different resolution.

In this work, RA-3DMM consists of a selector and these 3DMMs: High Resolution 3DMM(HR-3DMM), Medium Resolution 3DMM(MR-3DMM) and Low Resolution 3DMM(LR-3DMM). The model selector automatically selects the best 3DMM to fit the input face image according to its resolution. Clearly, the selection strategies of the selector are very important for RA-3DMM. These three 3DMMs are obtained by the 4-8 mesh subdivision algorithm [5].

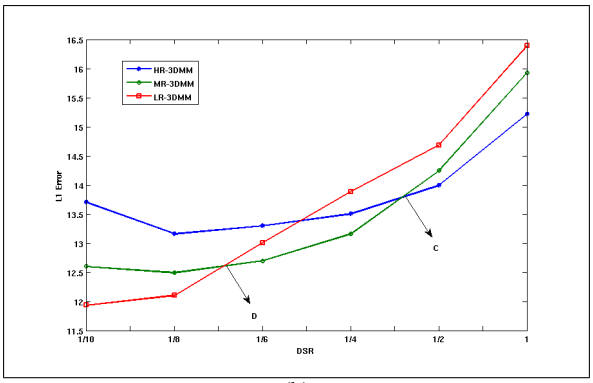
We evaluated the proposed RA-3DMM model on two face databases: XM2VTS and PIE. For building the RA-3DMM, only images with frontal pose and neutral illumination in the two databases are used. The images in XM2VTS and PIE are down-sampled to different resolutions. In this work, the down-sample rate (DSR) is 1, 1/2, 1/4, 1/6, 1/8, 1/10. All these down-sampled images were fitted by HR-3DMM, MR-3DMM and LR-3DMM. The L1-Norm is used to estimate the fitting error between the input image and the fitted image.

The fitting results on XM2VTS and PIE are shown in Fig. 1. It is clear that different models perform differently with input images of different resolutions: Obviously, HR-3DMM works best with DSR=1, 1/2; MR-3DMM with DSR=1/4, 1/6; and LR-3DMM works best with DSR=1/8, 1/10. Also it is important to know point A, B, C and D in Fig. 1 to define the selection strategies. It is not hard to conclude that the resolutions corresponding to A, B, C and D are 9397, 1344, 7422 and 985 respectively. So the selection strategies of RA-3DMM are determined as follows: under the diffused light (such as XM2VTS), HR-3DMM is selected for fitting if the input image is of high resolution (greater than 9397 pixels). MR-3DMM is selected for fitting if the input image is of medium resolution (between 1344 pixels and 9397 pixels). LR- 3DMM is selected for fitting if the input image is of low resolution (smaller than 1344). Under the point source light (such as PIE), HR-3DMM is selected for fitting if the input image is of high resolution (greater than 7422 pixels). MR-3DMM is selected for fitting if the input image is of medium resolution (between 985 pixels and 7422 pixels). LR-3DMM is selected for fitting if the input image is of low resolution (smaller than 985).

Then we apply RA-3DMM to pose correction with input images ranging from high resolution to low resolution [2]. The texture of the visible parts is extracted from the input image, and RA-3DMM automatically selects the best model to reconstruct the occluded part as discussed above. Then the Multiscale Local Phase Quantisation histogram (MLPQH) [3] descriptor is used for face verification. In this experiment, XM2VTS MPEG7 and the standard frontal datasets are used in conjugation with the Configuration 1 of the Lausanne protocol[3]. Following the Lausanne protocol, the total error rate (TER) is reported. TER is the sum of false acceptance rate (FAR) and false rejection rate (FRR) at a threshold. In our experiments the performance with RA-3DMM pose correction and without pose correction is compared. The TER of all poses with RA-3DMM pose correction is much smaller than that without pose correction for all resolutions. Even for low resolution face verification, which is a hard task in face recognition, RA-3DMM shows considerable improvement over the method without pose correction.



(a)



(b)

Figure 1: Fitting Results on (a) XM2VTS and (b) PIE

[1] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, sept. 2003.

[2] V. Blanz, P. Grother, P.J. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 454–461. IEEE, 2005.

[3] C.H. Chan, J. Kittler, N. Poh, T. Ahonen, and M. Pietikainen. (multi-scale) local phase quantisation histogram discriminant analysis with score normalisation for robust face recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 633–640. IEEE, 2009.

[4] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):681–685, jun 2001.

[5] L. Velho and D. Zorin. 4-8 subdivision. *Computer Aided Geometric Design*, 18(5):397–427, 2001.

Learning to rank images using semantic and aesthetic labels

Naila Murray[1]
nmurray@cvc.uab.es
Luca Marchesotti[2]
luca.marchesotti@xrce.xerox.com
Florent Perronnin[2]
florent.perronnin@xrce.xerox.com

[1] Computer Vision Center
Universitat Autònoma de Barcelona
Spain

[2] Xerox Research Centre Europe
Meylan, France

Most works on image retrieval from text queries have addressed the problem of retrieving semantically relevant images. However, the ability to assess the aesthetic quality of an image is an increasingly important differentiating factor for search engines. In this work, given a semantic query, we are interested in retrieving images which are semantically relevant and score highly in terms of aesthetics/visual quality. We use large-margin classifiers and rankers to learn statistical models capable of ordering images based on the aesthetic and semantic information. In particular, we compare two families of approaches: while the first one attempts to learn a single ranker which takes into account both semantic and aesthetic information, the second one learns separate semantic and aesthetic models. We carry out a quantitative and qualitative evaluation on a recently-published large-scale dataset and we show that the second family of techniques significantly outperforms the first one.

To develop our approaches we require images with semantic and aesthetic annotations. Recently, a large scale database (AVA, Aesthetic Visual Analysis [3]) containing such annotations was published. AVA contains 33 *semantic labels* in the form of textual tags. It also contains *aesthetic labels* in the form of a distribution of scores in a pre-defined range. We would like to represent the aesthetic information in a manner suitable for learning ranking models. Since we have scores distributions associated with each image, a natural approach would be to represent the information by a ranking, where the ranks would be obtained by sorting the images' mean scores. Such a ranking would assume that the difference between the mean scores of a pair of images, termed $\Delta_{i,j}$, is statistically significant. We tested the validity of this assumption and found that it is not a good one. Instead, we opted for an annotation strategy involving three labels: $''high-quality''$, $''medium-quality''$, $''low-quality''$.

We experiment with the images in AVA tagged with one of the 33 semantic tags. Each image is described using the Fisher Vector (FV) proposed in [4, 5]. To learn the semantic and aesthetic models, we employed Stochastic Gradient Descent (SGD) [1] because of its scalability.

Models for Combined Semantic and Aesthetic Retrieval We assume that we have a training set of N images $\mathcal{I} = \{(x_i, y_i, z_i), i = 1 \ldots N\}$ where $x_i \in \mathcal{X}$ is an image descriptor, $y_i \in \mathcal{Y}$ is a semantic label and $z_i \in \mathcal{Z}$ is an aesthetic label. In what follows, we assume that $\mathcal{X} = \mathcal{R}^D$ is a D-dimensional descriptor space, $\mathcal{Y} = \{0,1\}^C$ is the space of C semantic labels (where $y_{i,c} = 1$ indicates the presence of semantic class c in image i), and $\mathcal{Z} = \{1, \ldots, K\}$ is the set K aesthetic labels. In our case $K = 3$, where $3 = ''high-quality''$, $2 = ''medium-quality''$ and $1 = ''low-quality''$.

The joint ranking model (JRM): Each semantic class is treated independently in which case the label set can be simplified to $\mathcal{Y} = \{0,1\}$, i.e. semantically irrelevant or relevant. A new set of labels denoted u_i is then defined as follows: $u_i = y_i z_i$. We have $u_i \in \mathcal{U} = \{0,1,\ldots,K\}$. Hence $u = 0$ means that the image is irrelevant, $u = 1$ means that the image is relevant and that its quality is the poorest possible and $u = K$ means that the image is relevant and has the highest possible quality. Let us denote by (x^+, u^+) and (x^-, u^-) a pair of images together with their semantic and aesthetic labels in $\mathcal{U}$ such that $u^+ > u^-$. JRM learns w such that $w^\top x^+ > w^\top x^-$. A ranking SVM as proposed for instance in [2] may be trained by minimizing the following regularized loss function:

$$\sum_{(x^+,u^+),(x^-,u^-):u^+>u^-} \max\{0, \Delta(u^+,u^-) - w^\top(x^+ - x^-)\} + \frac{\lambda}{2}||w||^2 \quad (1)$$

where $\Delta(u^+,u^-)$ encodes the loss of an incorrect ranking, for instance $\Delta(u^+,u^-) = u^+ - u^-$.

For the JRM, the ranker has to deal with 4 relevance levels (the three aesthetic labels, and the semantic irrelevance level). However, these labels are very imbalanced. As a result, virtually all pairs used to train the JRM model encode semantic differences, rather than aesthetic information. To rebalance the labels we randomly draw a pair of images (i, j) subject to $u_i \neq u_j$. Then we simply multiply the probability $p_i(u)$ of drawing an image i with relevance level u_i by the probability of drawing an image j with relevance u_j: $\mathcal{W}_{i,j} = p_i(u = u_i) \cdot p_j(u = u_j)$.

At iteration t of the SGD optimization, the $\mathcal{W}_{i,j}$ weight for the sample pair modulates the degree of change of the model during the update step. With this weighting, highly probable relevance pairs are strongly penalized. We believe that a major weakness of the JRM is that it confounds both sources of variability: semantics and aesthetics. This makes the task of the linear SVM ranker more difficult. For this reason, we advocate models which treat semantics and aesthetics separately.

Independent Ranking Model (IRM): In this simple model, a set of semantic classifiers (one per class) and a single class-independent aesthetic ranker are trained. For the semantic part, we use the popular strategy of learning a set of one-vs-rest binary classifiers independently. We learn one linear classifier with parameters α_c for each class $c = 1, \ldots, C$ using the set $\{(x_i, y_i), i = 1 \ldots N\}$. We use a logistic loss: $-\log p(y_c = 1|x) = \log\left(1 + \exp(-\alpha_c^\top x)\right)$. The semantic parameters α_c are learned by minimizing the (regularized) negative log-likelihood of the data on the model which leads to the traditional logistic regression formulation:

$$-\sum_{i=1}^{N} \log p(y_{i,c}|x) + \frac{||\alpha_c||^2}{2} \quad (2)$$

For the aesthetic part, we learn a class-independent aesthetic ranker on the set $\{(x_i, z_i), i = 1 \ldots N\}$. Let us denote by (x^+, z^+) and (x^-, z^-) a pair of images with their aesthetic labels in $\mathcal{Z}$ such that $z^+ > z^-$. We learn the aesthetic parameters β by minimizing the following regularized loss:

$$\sum_{(x^+,z^+),(x^-,z^-):z^+>z^-} \log[1 + \exp(-\beta^\top(x^+ - x^-))] + \frac{\lambda}{2}||\beta||^2. \quad (3)$$

Dependent Ranking Model (DRM): In this model, we introduce an explicit dependence of the aesthetic labels on the semantic labels: $p(y,z|x) = p(y|x)p(z|y,x)$. This model is quite similar to the IRM model, but with one major difference: for class c we learn a ranker with parameters β_c using only the images of this class.

Results As table 1 shows, the best performance for three different measures is achieved by DRM. IRM performs slightly better than JRM.

	Precision(k)				nDCG(k)		
METHOD	k=10	k=20	k=50	mAP	k=10	k=20	k=50
Sem. Class.	8.538	8.284	8.270	5.810	0.230	0.227	0.224
JRM	8.760	8.254	7.762	5.602	0.234	0.228	0.217
JRM-balanced	14.272	13.104	11.574	6.980	0.253	0.244	0.227
IRM	18.128	17.000	15.450	8.806	0.255	0.247	0.236
DRM	**20.992**	**19.912**	**17.444**	**9.726**	**0.295**	**0.285**	**0.265**

Table 1: Comparison between the three learning strategies

Therefore we conclude that it is advantageous to learn semantics and aesthetics separately. We also conclude that data rebalancing is an important step to improve the ranking performance.

[1] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *NIPS*, 2007.

[2] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[3] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

[4] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, 2007.

[5] F. Perronnin, J. Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

Online Bayesian Nonparametrics for Group Detection

Matteo Zanotto
matteo.zanotto@iit.it
Loris Bazzani
loris.bazzani@iit.it
Marco Cristani
marco.cristani@iit.it
Vittorio Murino
vittorio.murino@iit.it

Pattern Analysis & Computer Vision
Istituto Italiano di Tecnologia
Via Morego 30 - 16163
Genova, Italy

Social interactions are essential in our daily activities and automatic modelling of interactional exchanges has become an active research topic over the last few years. An important step towards the analysis of social behaviour and the understanding of social interactions involves the identification of groups of people, which is the goal of this work.
We propose to perform group detection from videos in real surveillance scenarios through a Dirichlet Process Mixture Model (DPMM) [1]. Within this model, groups are represented as components of an infinite mixture model, and individuals are seen as observations generated from them. Each individual is represented by its position (x and y coordinates) and velocity (decomposed in heading direction and module) and groups are detected in an unsupervised way by performing clustering in the feature space defined above. The introduction of a "social" constraint based on proxemics rules proposed by Hall [3] allows to only maintain components associated to groups satisfying theories from social psychology.
In order to keep computation fast and compatible with real-time video analysis, we propose a sequential variational inference performing single parameters updates rather then iterating to convergence for each single frame (see Figure 1). This approach builds upon ideas by Neal and Hinton [5] and is possible because grouping configurations evolve smoothly, allowing to exploit the temporal correlation across consecutive frames to refine the detection of groups taking advantage of the evolution of the observations. The posterior distribution estimated at one time step can then act as prior knowledge for the following one, distributing inference over time. This sequential approach also allows to take advantage of the dynamics of observations evolution without explicitly modelling it, which is a valuable feature in cases where no prior knowledge on such dynamics is available to be coded in the model.

The method has been tested on two public benchmark datasets for group detection: the *BIWI* dataset [6], containing the `eth` and `hotel` sequences, and the *Crowd by Example* dataset [4], containing the `zara01`, `zara02`, `students003` sequences. In order to capture all the relevant aspects of group detection, comparison with two benchmark methods has been performed. The proposed method outperforms that by Yamaguchi *et al.* [7] on 4 out of the 5 considered sequences, and that by Bazzani *et al.* [2] on the chosen test sequence. Examples of the qualitative results obtained by our method are reported in Figure 2 along with the ground truth.

Summarising in this paper we propose a model for group detection which

- Does not require to fix a priori the number of groups to find.
- Can dynamically adapt the number of groups from frame to frame to effectively match the observed data, also coping with split and merge of groups.
- Can take advantage of the dynamics of the data without explicitly modelling it.
- Produces results through online inference.
- Can perform real-time processing up to 42 fps (bounded by the pedestrian detector performance).

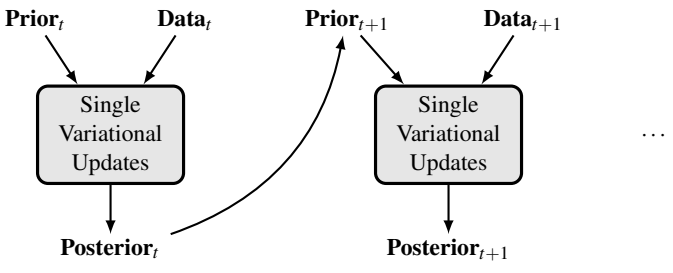


Figure 1: Pictorial representation of the proposed single-step variational inference scheme.

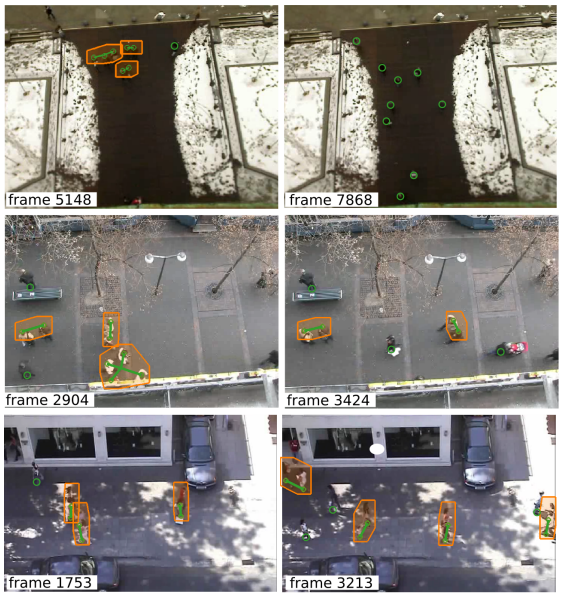


Figure 2: Qualitative results on `eth` (first row), `hotel` (second row), and `zara01` (last row). The ground truth position of individuals and groups are shown with green circles and segments, respectively. The estimated groups are depicted as orange convex hulls (best viewed in colors).

Experimental results show that our method outperforms state-of-the-art methods on evaluation metrics capturing different aspects of group detections, suggesting a better overall performance.

[1] C.E. Antoniak. Mixtures of Dirichlet Processes with Applicatioins to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6): 1152–1174, 1974.

[2] L. Bazzani, V. Murino, and M. Cristani. Decentralized particle filter for joint individual-group tracking. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[3] E.T. Hall. *The Hidden Dimension*. 1966.

[4] A. Lerner, Y. Chrysanthou, and D. Lischinski. Crowds by example. volume 26, pages 655–664, Sep 2007.

[5] R.M. Neal and G.E. Hinton. A new view of the EM algorithm that justifies incremental and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*. MIT Press, 1998.

[6] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *International Conference on Computer Vision*, 2009.

[7] K. Yamaguchi, A.C. Berg, L.E. Ortiz, and T.L. Berg. Who are you with and where are you going? In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2011.

Moving Volume KinectFusion

Henry Roth
roth@ccs.neu.edu
Marsette Vona
http://ccis.neu.edu/research/gpc

College of Computer and Information Science
Northeastern University
Boston, MA

Newcombe and Izadi et al's KinectFusion [5] is an impressive new algorithm for real-time dense 3D mapping using the Kinect. It is geared towards games and augmented reality, but could also be of great use for robot perception. However, the algorithm is currently limited to a relatively small volume fixed in the world at start up (typically a $\sim$ 3m cube). This limits applications for perception.

Here we report *moving volume KinectFusion* with additional algorithms that allow the camera to roam freely. We are interested in perception in rough terrain, but the system would also be useful in other applications including free-roaming games and awareness aids for hazardous environments or the visually impaired.

Our approach allows the algorithm to handle a volume that moves arbitrarily on-line (Figure 1).

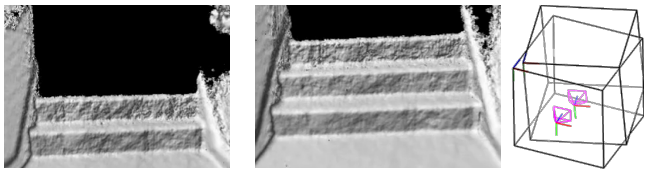


Figure 1: Remapping to hold the sensor pose fixed relative to the volume. Raycast images before and after a remapping show a third step coming into view as the volume moves forward. A reconstruction of the volume and camera poses shows that the volume-to-volume transform is calculated to maintain the camera at the rear center of the volume.

We based our implementation on the open-source *Kinfu* code that has recently been added to the Point Cloud Library (PCL) from Willow Garage [6], and we have submitted our code for inclusion there as well.

With our algorithm in place, the "absolute" camera pose C_g, a 4×4 rigid transform expressing the current camera pose in the very first volume frame, can be calculated at any time t as

$$C_g = P_0 \cdots P_{\mathrm{vf}(t)} C_t \tag{1}$$

where C_t is the current camera tracking transform from KinectFusion taking camera coordinate frame to its *parent volume frame*, each $P_{i>0}$ takes volume frame i to volume frame $i-1$, $P_0 = I_{3\times3}$, and $\mathrm{vf}(t)$ is a bookkeeping function that maps a depth image timestamp to the index of its parent volume. (Volume frames are generally sparser than camera frames.)

Moving volume KinectFusion both tracks *global* camera motion (equation 1) and simultaneously builds a spatial map of the *local* surroundings. However, this is not a true SLAM algorithm as it does not explicitly close large-scale loops and will inevitably incur drift over time. Rather, it can be considered a 6D *visual odometry* approach in that the camera pose ${}^aC_b$ at any time b relative to an earlier time a is

$${}^aC_b = C_a^{-1} P_{\mathrm{vf}(a+1)} \cdots P_{\mathrm{vf}(b)} C_b. \tag{2}$$

Of course the significant additional benefit beyond visual odometry alone is that a map of local environment surfaces is also always available.

After the Kinfu tracking phase gives the current local camera pose C_t we determine if a new volume frame is needed by calculating linear and angular camera offsets l_d, a_d relative to a desired local camera pose C_s.

$$D = \begin{bmatrix} R_d & \mathbf{t}_d \\ 0 & 1 \end{bmatrix} = C_s^{-1} C_t, \quad l_d = \|\mathbf{t}_d\|, \quad a_d = \|\mathrm{rodrigues}^{-1}(R_d)\| \tag{3}$$

A new volume frame is triggered if $l_d > l_{\max}$ or $a_d > a_{\max}$. We typically use $l_{\max} = 0.3$m, $a_{\max} = 0.05$rad, and

$$C_s = \begin{bmatrix} I_{3\times3} & \mathbf{t}_s \\ 0 & 1 \end{bmatrix}, \quad \mathbf{t}_s = \begin{bmatrix} W_m/2 \\ H_m/2 \\ -D_m/10 \end{bmatrix} \tag{4}$$

for volume W_m, H_m, D_m meters, which is the default initial camera pose for Kinfu. This keeps the camera centered just behind the volume (Figure 1, right; note that the origin of each volume frame is the upper left corner of the volume with $\hat{\mathbf{x}}$ right, $\hat{\mathbf{y}}$ down, and $\hat{\mathbf{z}}$ pointing into the page). Other strategies for determining C_s may make sense—for example keeping the camera centered in the volume, orienting the volume to task-relevant directions—but are subject to the constraint that the camera must see scene surfaces within the volume.

To introduce a new volume frame we *remap* the new volume from the old. We maintain a swap buffer in GPU memory the same size as the volume buffer for this; memory requirements for this large data structure are thus doubled but still feasible on current GPUs. After the remap the buffers are swapped and a new relative volume transform P_{n+1} is set as

$$P_{n+1} = C_t C_{t+1}^{-1} \tag{5}$$

where C_{t+1} is the new camera transform. Conceptually $C_{t+1} = C_s$, though we allow an offset in some cases.

Remapping—sometimes called *reslicing* for the 3D case—has been studied for medical images [3], but speed is often sacrificed for accuracy. Efforts have been made to improve the speed [2], but generally reslicing has not been done in real time. Here we require a fast parallel algorithm which is tuned for common-case KinectFusion data.

Our approach is hybridized in two ways. First, if $l_d > l_{\max}$ but $a_d \leq a_{\max}$ we use a fast and exact memory shift algorithm, otherwise we use a more traditional resampling based on trilinear interpolation. Second, during resampling we take advantage of the fact that in the common case much of the volume is either uninitialized or marked "empty": we do a nearest-neighbor lookup first, and only if that is within the truncation band do we continue with a more expensive interpolation.

Using a novel battery-powered Kinect we collected 18 rocky terrain datasets comprising an estimated 662m path length. (Though the Kinect cannot cope with direct sunlight it does work outdoors on a reasonably overcast day.) The richness of 3D depth features makes our approach work well on rocky terrain—no camera tracking failures were incurred, and reconstructed surfaces appear to be high quality (quantitative analysis of the geometry is future work). We present performance and tracking accuracy measurements for our algorithm on 6 datasets, comparing it with the original Kinfu implementation and with ground truth and reference results for RGB-D SLAM [1] where applicable.

While two other groups are also developing approaches to translate the KinectFusion volume [4, 7], a key distinction of our method is the ability to rotate the volume in addition to translation. Since the volume is rectilinear this can be useful to control its orientation, e.g. to maximize overlap of the camera frustum or to align the volume with task-relevant directions, such as the average ground surface normal in locomotion.

[1] N. Engelhard, F. Endres, J. Hess, J. Sturm, and W. Burgard. Real-time 3D visual SLAM with a hand-held RGB-D camera. In *RGB-D Workshop, European Robotics Forum*, 2011.

[2] J. Fischer and A. del Río. A fast method for applying rigid transformations to volume data. In *WSCG*, 2004.

[3] J. Hajnal, N. Saeed, E. Soar, A. Oatridge, I. Young, and G. Bydder. A registration and interpolation procedure for subvoxel matching of serially acquired MR images. *Journal of Computer Assisted Tomography*, 19(2):289–296, 1995.

[4] Francisco Heredia and Raphael Favier. Kinfu Large Scale in PCL. `http://www.pointclouds.org/blog/srcs`, 2012.

[5] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *UIST*, pages 559–568, 2011.

[6] R. Rusu and S. Cousins. 3D is here: Point cloud library (PCL). In *ICRA*, 2011. (`http://www.pointclouds.org`).

[7] T. Whelan, J. McDonald, M. Kaess, M. Fallon, H. Johannsson, and J. Leonard. Kintinuous: Spatially extended KinectFusion. In *RGB-D Workshop at RSS*, 2012.

6D Relocalisation for RGBD Cameras Using Synthetic View Regression

Andrew P. Gee
http://www.cs.bris.ac.uk/~gee
Walterio Mayol-Cuevas
http://www.cs.bris.ac.uk/~wmayol

Visual Information Laboratory
University of Bristol
Bristol, UK

With the advent of real-time dense scene reconstruction from handheld RGBD cameras [3], one key aspect to enable robust operation is the ability to relocalise in a previously mapped environment or after loss of measurement. Tasks such as operating on a workspace, where moving objects and occlusions are likely, require a recovery competence in order to be useful. For RGBD cameras, this must also include the ability to relocalise in areas with reduced visual texture.

Approaches from both the point cloud and monocular camera literature can be used for relocalisation on these types of densely reconstructed maps. Local feature-based methods extract distinctive geometric [4] or visual [6] features from the map and match them to features extracted from the current camera view of the world to estimate pose. In contrast, view-based methods construct geometric [5] or visual [1] descriptors for complete views of the map and match these to the current camera view.

This paper describes a view-based method for relocalisation of a freely moving RGBD camera in small workspaces. In contrast to related methods [1, 2], this method combines intensity and depth information from synthetic RGBD images to estimate full 6D pose at framerate using a regression framework.

The relocalisation problem can be formulated as a minimisation problem, where the goal is to find the set of camera pose parameters $\mathbf{x} = [\mathbf{t}, \ln(\mathbf{q})] \in \mathbb{SE}_3$, that minimises the distance measure

$$\mathbf{x} = \arg\min_{\hat{\mathbf{x}}} \|\mathbf{I}_0 - \mathbf{I}(\hat{\mathbf{x}}, \mathcal{M})\|, \tag{1}$$

where $\mathbf{t}$ is a 3D position vector, $\mathbf{q}$ is a quaternion representing rotation, $\mathbf{I}(\hat{\mathbf{x}}, \mathcal{M})$ is the synthetic view generated from the map $\mathcal{M}$ at pose $\hat{\mathbf{x}}$, and $\mathbf{I}_0$ is the true RGBD image from the camera. The j-th RGBD image $\mathbf{I}_j = [\mathbf{u}_j, \mathbf{v}_j, \rho_j, \mathbf{c}_j]$ is composed of n pixels, where $[u_{ji}, v_{ji}]$ are image coordinates, ρ_{ji} is the depth value, and c_{ji} is the grey intensity of the i-th pixel.

We treat the estimation of $\mathbf{x}$ in Eq. 1 as a general regression problem over a set of m synthetic views $\mathbf{I}_j$ and their poses $\mathbf{x}_j$, for $j = 1 \ldots m$. Using the Nadaraya-Watson estimator, we can approximate the camera pose $\tilde{\mathbf{x}}$ from the set of synthetic views as

$$\tilde{\mathbf{x}} = \frac{\sum_{j=1}^{m} \mathbf{x}_j K(\|\mathbf{I}_0 - \mathbf{I}_j\|/h)}{\sum_{j=1}^{m} K(\|\mathbf{I}_0 - \mathbf{I}_j\|/h)}, \tag{2}$$

where K is a kernel function centred on each sample with bandwidth h. In this case, we opt for a Gaussian kernel, such that

$$\tilde{\mathbf{x}} = \frac{\sum_{j=1}^{m} \mathbf{x}_j d_j}{\sum_{j=1}^{m} d_j}, \tag{3}$$

$$d_j = exp\left(-\frac{1}{n\alpha}\sum_{i=1}^{n}\left(\frac{(c_{0i} - c_{ji})^2}{\sigma_{ci}^2} + \frac{(\rho_{0i} - \rho_{ji})^2}{\sigma_{\rho i}^2}\right)\right), \tag{4}$$

where σ_c and σ_ρ are vectors of standard deviations in the intensity and depth computed per pixel over all of the sample views $\mathbf{I}_j$, for $j = 1 \ldots m$, and α is a scaling factor that controls the smoothness of the regression. The estimate $\tilde{\mathbf{x}}$ is therefore a normalised weighted sum, where the contribution of each sample view is determined by the normalised Euclidean distance between the sample view and the current camera view.

One key difference between our work and previous relocalisation systems is that, enabled by the recovered 3D map, we can generate novel synthetic views that have not been visited by the system during mapping and that are considered likely poses where relocalisation will be needed. This enhances the power of the sampling used by the regression framework but introduces the issue of knowing which views should be generated.

Here we have adopted the approach of randomly sampling poses around a pre-defined trajectory. For each of the m sampled synthetic views, a pose on the trajectory is randomly selected and a random Gaussian perturbation with 10° and 5.0cm standard deviation is applied. Synthetic views are resampled if fewer than 50% of the pixels intersect with the map. During the sampling process, the statistics for σ_c and σ_ρ, required by the regression algorithm, are also calculated and stored. It is also trivial to extend the set of synthetic views online, for example if the camera is tracked into a location not covered by the initial trajectory.

Figure 1: Examples of synthetic view regression relocalisation on four different test sequences. Images show ground-truth camera view (upper rows) and synthetic view generated from relocalised pose (lower rows).

The performance of the system is demonstrated in the paper by a comparison against visual and geometric feature matching relocalisation techniques and tested on sequences with moving objects and minimal texture. Some relocalisation results for the different test sequences are shown in Fig. 1. The results in the paper show that the method is both fast ($<$ 80ms) and accurate ($<$ 10cm, $<$ 10° median error) and able to cope with small changes to the environment and low texture workspaces. The most common failure mode occurs when the camera moves to a viewpoint outside the set of synthetic view samples.

Our conclusion is that view-based relocalisation using synthetic RGBD images provides a feasible and useful alternative to slower, feature-based methods in small workspaces. This is particularly true in areas with low texture or low geometry, where the use of visual or geometric features alone is prone to failure, and in scenarios with continuously moving cameras, where the time required to relocalise is critical.

[1] M. Felsberg and J. Hedborg. Real-time view-based pose recognition and interpolation for tracking initialization. *Journal of Real-Time Image Processing*, 2(2-3):103–115, 2007.

[2] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. IEEE and ACM Int. Symp. on Mixed and Augmented Reality*, 2007.

[3] R.A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A.J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. KinectFusion: real-time dense surface mapping and tracking. In *Proc. Int. Symp. on Mixed and Augmented Reality (ISMAR)*, 2011.

[4] R.B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3d registration. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, May 2009.

[5] R.B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, October 2010.

[6] B. Williams, G. Klein, and I. Reid. Real-time SLAM relocalisation. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2007.

Image Priors for Image Deblurring with Uncertain Blur

Daniele Perrone[1]
perrone@iam.unibe.ch
Avinash Ravichandran[2]
http://cis.jhu.edu/~avinash/
René Vidal[3]
http://cis.jhu.edu/~rvidal/
Paolo Favaro[1]
paolo.favaro@iam.unibe.ch

[1] Universität Bern,
Bern, Switzerland
[2] UCLA VisionLab
University of California,
Los Angeles, CA, USA
[3] Center for Imaging Science
Johns Hopkins University,
Baltimore, MD, USA

We consider the problem of non-blind deconvolution of images corrupted by a blur that is not accurately known. A common choice for non-blind deconvolution algorithms is to use methods that rely on an exact blur estimate. However, small errors in the blur estimate result in visible artifacts in the restored image, which may not be removed by future iterations.

We propose a novel image prior to remove artifacts introduced by blur errors. To achieve this goal we use a dictionary-based prior learned only from the input blurred image and a database of images, and propose a method to prune ambiguities in the prior due to blur.

Consider an observed degraded image g

$$g = k * f + n \tag{1}$$

where $*$ is the convolution operator, k is a blur kernel, or point spread function (PSF), f is the noiseless and sharp image, and n is additive noise generated during the acquisition process. The aim of image deblurring is to estimate the noiseless image f given the noisy image g and the kernel k.

The imaging model in (1) can be written as a matrix-vector operation.

$$\vec{g} = K\vec{f} + \vec{n} \tag{2}$$

However, typically the linear system has infinite solutions due to the noise n being larger than the smallest singular values of the matrix K.

One way to obtain a unique solution is to introduce additional linear equations, which we call *image priors*, via a matrix A and a vector $\vec{b}$

$$A\vec{f} = \vec{b} \tag{3}$$

To enforce this regularity, we consider applying the same linear constraints to all pixels in patches of $L \times L$ pixels. For this purpose, we extract patches of $L \times L$ pixels centered at each pixel of the image f, rearrange the pixel intensities of each patch as a column vector and collect all such vectors into a matrix $F \in \mathbf{R}^{L^2 \times N}$, where N is the number of pixels in f. We can then write our prior as

$$F = DC, \tag{4}$$

We choose a dictionary made of a mix of both an external dictionary D_0 and the image itself ($D = [D_0\ F]$).

To learn C, we face two important challenges: The first challenge is that F is typically not available and the second is that we do not have enough equations to obtain a unique matrix C.

To deal with the first challenge, we extract noisy and blurred patches G_i from the image $g = k * f + n$. Let B be the matrix of patches extracted from the blurred noiseless image $b = k * f$. Since $B = KF = KDC$, we can express the blurred patches in B in terms of the blurred dictionary $E \triangleq KD$ using the *same* correspondence matrix C.

The second challenge, *i.e.*, the non uniqueness of the matrix C is due to the overcompleteness of the dictionary D. We introduce additional constraints on C by exploiting image self-similarities. We consider a patch at pixel i, B_i, found as a weighted average of similar patches D_j extracted from either the same image or from a dictionary of patches. Specifically, if we consider all the patches as vectors in $\mathbf{R}^{L^2}$, then we have

$$B_i = \sum_{j=1}^{M} D_j \frac{\phi(D_i, D_j)}{\sum_{\ell=1}^{M} \phi(D_i, D_\ell)}, \quad i = 1, \ldots, N, \tag{5}$$

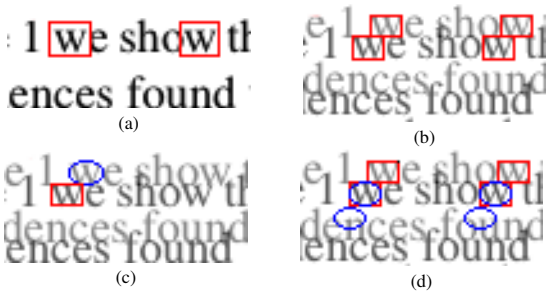


Figure 1: **Example of ambiguous correspondence correction.** In this toy example we show the correction performed by the correlation-based method. The PSF consists of only two peaks, which result in an overlap of two copies of the sharp image with two relative shifts. **(a)** Sharp image of text, and two correct correspondences (red squares) of the character 'w'. **(b)** Blurred version of the previous image which leads to additional incorrect correspondences. **(c)** Selected patch (red square) and the area corresponding to the second peak of the PSF (blue circle). **(d)** The correlated patches shown in (a) are obtained by overlapping the correspondence sets of the two patches (blue circles and red squares).

where ϕ is a positive semi-definite kernel that measures the similarity between two patches. In our work we use the following kernel,

$$\phi(D_i, D_j) = \begin{cases} 1 & \|D_i - D_j\|_2^2 \leq \varepsilon^2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where ε is proportional to the standard deviation of noise. It is easy to see that (5) can be written in matrix form as

$$B = DC^{nlm}, \tag{7}$$

where D is a matrix of patches ($D = [G\ E_0]$ in our case) and $\{C^{nlm}\}_{j,i} = \frac{\phi(D_i, D_j)}{\sum_{\ell=1}^{M} \phi(D_i, D_\ell)}$ is the correspondence matrix obtained from the procedure in eq. (5).

When we apply this procedure to a blurred image, incorrect correspondences may be generated. We distinguish two types: *false negatives*, *i.e.*, correspondences found in the sharp image, but not in the blurred one, and *false positives*, *i.e.*, correspondences present in the burred image but not in the sharp one. In Fig. 1 we provide a synthetic example to illustrate why blur generates false positives.

To reduce the false positives we suggest to use our (partial) knowledge of the blur. Formally, let $C_p = \{j \in \mathbb{Z} : ||B_p - B_{p+j}||^2 \leq \varepsilon^2\}$ be the set of correspondences for the pixel p learned from the blurred image and let $\mathcal{K} = \{i \in \mathbb{Z} : |max(k) - k_i| \leq \tau\}$ be the set of non-zero entries of the PSF k, where τ is a threshold based on blur noise. For each patch centered at p we enforce that its improved set of correspondences $\hat{C}_p$ be the intersection of all the correspondence sets of patches at pixels with relative displacement given by the main PSF peaks, *i.e.*, where the *consensus* is full: $\hat{C}_p = \bigcap_{i \in K} C_i$.

We finally pose the problem of recovering the sharp image f via the following convex optimization problem

$$\begin{aligned} \min_{f,n,e} \quad & \frac{1}{2}\|A\vec{f} - \vec{b}\|_2^2 + \beta\|\nabla f\|_2 + \frac{\lambda}{2}\|n\|_2^2 + \gamma\|e\|_1 \\ \text{subject to} \quad & g = h * f + n + e \end{aligned} \tag{8}$$

In the experimental validation our algorithm performance is overall better than the state-of-the-art methods when the blur kernel is noisy.

Improvements in Joint Domain-Range Modeling for Background Subtraction

Manjunath Narayana
narayana@cs.umass.edu
Allen Hanson
hanson@cs.umass.edu
Erik Learned-Miller
elm@cs.umass.edu

University of Massachusetts Amherst
Massachusetts, USA

Background subtraction, often a first step in segmenting moving objects in videos, is most commonly achieved by modeling the background color likelihoods at each pixel. Stauffer and Grimson [4] use a parametric Gaussian mixture model to estimate the likelihoods at each pixel. A non-parametric model was introduced by Elgammal *et al*. [1], where the likelihoods at each pixel are modeled using a kernel density estimate (KDE) by using the data samples from previous frames in history. These pixel-wise models do not allow for the observations at one location to influence the estimated distribution at a different but nearby location. By including each pixel's position information and modeling the likelihoods using a five-dimensional distribution in a joint domain-range representation, Sheikh and Shah [3] allow pixels in one location to influence the distributions in another location. They show that this sharing of spatial information leads to more accurate background subtraction. Their background model is a *single* distribution in the joint domain-range space. As we will see in this paper, their classification criterion, based on the ratio of likelihoods in this five-dimensional space, has an undesirable dependence on the size of the image. Like Sheikh and Shah, we model the foreground and background likelihoods with a KDE using pixel samples from previous video frames. However, we model the processes using a three-dimensional color distribution at each pixel. Our distributions are conditioned on spatial location, rather than being joint distributions over position and color. Our modeling avoids the dependence on the image size and yields better results.

Recent work on KDE based background modeling by Narayana *et al*. [2] has shown that adapting the kernel variance values for each pixel yields significantly better results than using a uniform kernel variance for all pixels. At each pixel location, the best kernel variance is selected from a set of candidate variances. Although we use a similar approach for adapting the kernel variance at each pixel, our background and foreground likelihood models are conceptually easier to interpret than their foreground and background *scores*. We show through both synthetic and real data examples that the adaptive kernel variance scheme is useful. With our probabilistic model, we can understand the effect of the adaptive kernel variance method of Narayana *et al*. more easily, as shown in Figures 1 and 2.

Another improvement we present over earlier approaches is the use of explicit spatial priors for the background and foreground processes. We use the foreground-background classification from the previous frame to estimate the prior probability for the processes. Our probabilistic formulation with likelihoods and a spatially dependendent prior for each process leads to a posterior distribution over the processes.

Figure 3 shows images that characterize the performance of the Sheikh and Shah model compared to ours and the effect of using adaptive kernel variance in both models. Benchmark comparisons on a standard data set show that our system's performance is comparable to the results of Narayana *et al*., which are the best reported results on our chosen benchmark. The advantage of our model over that of Narayana *et al*., is that our probabilistic model is more intuitive. The results from our model can be understood more clearly and the various constants and factors in the model can be interpreted more meaningfully.

[1] Ahmed M. Elgammal, David Harwood, and Larry S. Davis. Non-parametric model for background subtraction. In *ECCV*, 2000.

[2] Manjunath Narayana, Allen Hanson, and Erik Learned-Miller. Background modeling using adaptive pixelwise kernel variances in a hybrid feature space. In *CVPR*, 2012.

[3] Yaser Sheikh and Mubarak Shah. Bayesian modeling of dynamic scenes for object detection. *PAMI*, 2005.

[4] Chris Stauffer and W. Eric L. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1999.

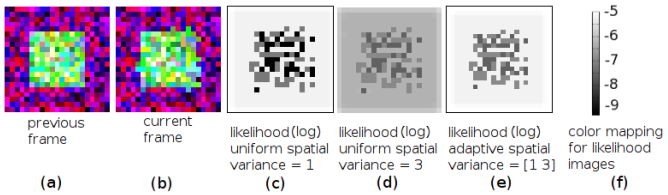


Figure 1: Consider a synthetic scene with no foreground objects, but to simulate spatial uncertainty, the colors in the central greenish part of the background in the previous frame (a) have been displaced at random by one or two pixel locations in the current frame (b). (c) Computing the background likelihoods at each location in the current frame with pixel samples from the previous frame using a small uniform variance results in low likelihoods for pixels that have moved. (d) Large uniform variance results in higher likelihoods of the moved pixels at the expense of lowering the likelihoods of stationary pixels. (e) Adaptive variance results in high likelihoods for both the moved and stationary pixels by applying a high spatial variance for pixels that have moved and a low spatial variance for pixels that have not moved.

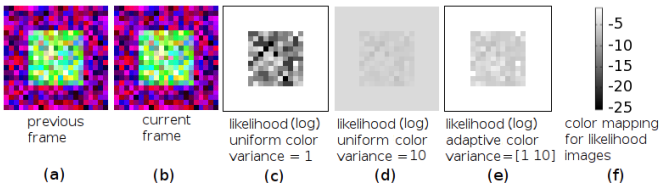


Figure 2: Uniformly sampled noise is added to the color values in the central part of image (a) to result in image (b). Color uncertainty in the central part of image (b) is best modeled by using adaptive kernel variances. (c) Small uniform variance results in low likelihoods for pixels that have changed color. (d) Large uniform variance results in higher likelihoods of the altered pixels at the expense of lowering the likelihoods of other pixels. (e) Adaptive variance results in high likelihoods for both kinds of pixels.

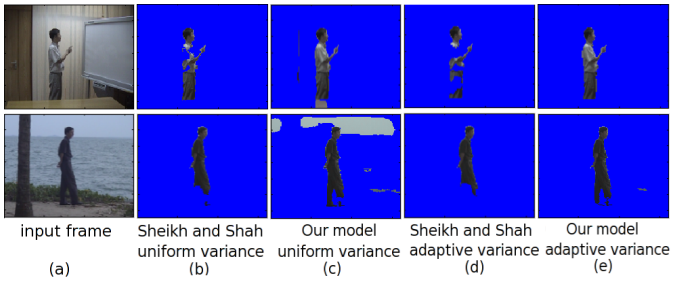


Figure 3: Comparing the effect of adaptive kernel variance for the Sheikh and Shah model versus our model. The Sheikh and Shah method has a bias towards background label (b), further exacerbated by the adaptive kernel selection (d). Our method tends to classify foreground objects well, but has more false positive foreground pixels (c). Adaptive kernel variance with our normalization yields the best results (e).The adaptive kernel for the background process, by selecting the best of the available kernel variances, in effect "tries hard" to classify each pixel as background. When a pixel is not well explained by the background model despite the selection procedure, it gets labeled as foreground.

A Multi-layer Composite Model for Human Pose Estimation

Kun Duan[1]
kduan@indiana.edu
Dhruv Batra[2]
dbatra@ttic.edu
David Crandall[1]
djcran@indiana.edu

[1] Indiana University, Bloomington, IN
[2] TTI-Chicago, Chicago, IL

Detecting humans and recognizing their body poses is a key problem in understanding natural images, since people are the focus of many (if not most) consumer photographs. Pose recognition is a challenging problem due not only to the usual complications of object recognition—cluttered backgrounds, scale changes, illumination variations, etc.—but also because of the highly flexible nature of the human body. Traditional approaches that use deformable part-based models for human pose estimation typically assume a kinematic tree structure [2, 7], capturing the kinematic constraints between parts of the body (e.g. that the lower arm is connected to the upper arm, which is connected to the torso, etc.).

In this paper, we propose a new model that addresses these problems from a different perspective. We use a composition of multiple tree-structured models with different numbers of parts and resolution scales, allowing different degrees of structural flexibility at different levels, and connect these models through hierarchical decomposition links between body parts in adjacent levels.

A visualization of our model with three layers is shown in Figure 1. Even though the composite model is a loopy graph, it can be naturally decomposed into the constituent tree-structured sub-problems within each level and the cross-model constraint sub-problem across levels, which is also tree-structured as shown in Figure 1 (right). These tree-structured sub-problems are amenable to exact inference and thus joint inference on the composite model can be performed via dual-decomposition [1].

Our approach builds on the work of Yang and Ramanan [7], which has demonstrated state-of-art performance on recent pose estimation datasets. The key innovation in their deformable parts-based model is the use of a mixture of parts, which allows the appearance of each part to change discretely between different "part types." More formally, their model consists of a set $\mathcal{P}$ of parts in a tree-structured model having edges $\mathcal{E} \subseteq \binom{\mathcal{P}}{2}$, such that $\mathcal{E}$ is a tree. Let $\mathbf{y}$ be a vector that represents a particular configuration of the parts, *i.e.* the location and type of each part. They define a function $S(I,\mathbf{y})$ that scores the likelihood that a given configuration $\mathbf{y}$ corresponds to a person in the image,

$$S(I,\mathbf{y}) = \sum_{p\in\mathcal{P}} D(I,\mathbf{y}_p) + \sum_{(p,q)\in\mathcal{E}} \Big(L(\mathbf{y}_p,\mathbf{y}_q) + T(\mathbf{y}_p,\mathbf{y}_q)\Big), \tag{1}$$

where $D(I,\mathbf{y}_p)$ is the score for part p being in configuration $\mathbf{y}_p$ given local image data (the data term), $L(\mathbf{y}_p,\mathbf{y}_q)$ is the relative location term measuring agreement between locations of two connected parts, and $T(\mathbf{y}_p,\mathbf{y}_q) = \vec{\mathbf{B}}^{t(y_p),t(y_q)}$ measures the likelihood of the observing this pair of part-types. $L(\mathbf{y}_p,\mathbf{y}_q)$ is defined as the negative Mahalanobis distance between part locations, and $T(\mathbf{y}_p,\mathbf{y}_q)$ is a part concurrence table that is learned discriminatively in the training stage.

We generalize this model to include multiple layers, each layer being similar to the base model but with a different number of parts and a different (but still tree-structured) graph structure. In particular, let $\mathcal{M} = \{(\mathcal{P}_1,\mathcal{E}_1),...,(\mathcal{P}_K,\mathcal{E}_K)\}$ be a set of K tree-structured models, let $\mathbf{y}^k$ denote the configuration of the parts in the k-th model, and let $\mathbf{Y} = (\mathbf{y}^1,...,\mathbf{y}^K)$ be the configuration of the entire multi-layer composite model. We now define a joint scoring function,

$$\hat{S}(I,\mathbf{Y}) = \sum_{k=1}^{K} S_k(I,\mathbf{y}^k) + \sum_{k=1}^{K-1} \chi(\mathbf{y}^k,\mathbf{y}^{k+1}), \tag{2}$$

where $S_k(\cdot,\cdot)$ is the single-layer scoring function of equation (1) under the model $(\mathcal{P}_k,\mathcal{E}_k)$, and $\chi(\mathbf{y}^k,\mathbf{y}^{k+1})$ is the cross-model scoring function that measures the compatibility of the estimated configurations between different layers of the model.

As Figure 1 shows, we impose a hierarchical structure on the composite model, such that each part at level k is decomposed into multiple parts at level $k+1$. We call these decomposed parts the child nodes. For a part $p \in \mathcal{P}_k$, let $C(p) \subseteq \mathcal{P}_{k+1}$ be the set of child nodes of p in layer $k+1$. The cross-model scoring function χ scores the relative location and part types of a node in one layer with respect to its children in the layer below,

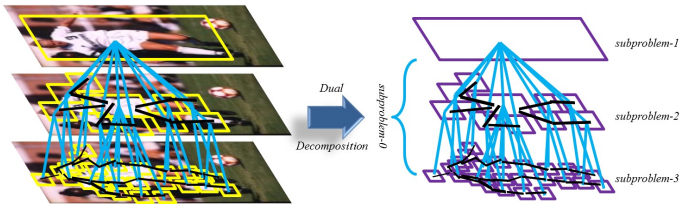


Figure 1: Illustration of our composite part-based models.

$$\chi(\mathbf{y}^k,\mathbf{y}^{k+1}) = \sum_{p\in\mathcal{P}_k} \sum_{q\in C(p)} B(\mathbf{y}_p^k,\mathbf{y}_q^{k+1}), \tag{3}$$

where $B(\mathbf{y}_p^k,\mathbf{y}_q^{k+1})$ is a measure of the likelihood of the relative configuration of a part and its child across the two submodels.

We exploit the natural decomposition of this composite model into tree-structured subproblems to perform inference using dual decomposition, where the key idea is to decompose a joint inference problem into easy sub-problems, solve each sub-problems, and then have the sub-problems iteratively communicate with each other until they agree on variable values. To learn the composite model, we stack all of the features in all of the layers together along with the cross-model features into a single feature vector, and formulate this problem as a standard structural SVM problem [5].

We evaluate our composite models on two challenging datasets: Image Parse [4] and UIUC Sport [6]. We evaluate our results using the Percentage of Correct Parts (PCP) metric as defined in [3]. We show that our composite models perform substantially better than state-of-the-art methods on both datasets, which suggests that by combining evidence across submodels, our composite models can obtain better pose estimates of body limbs.

Our model is a general framework for combining different pose estimation models. In future work, we plan to study how to capture rich cross-model constraints inside our composite model (e.g. define relative location constraints between adjacent submodels). We also plan to apply our model to related tasks like human action recognition.

[1] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, September 1999.

[2] Pedro F. Felzenzwalb and Daniel Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61 (1):55–79, 2005.

[3] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormaehlen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012.

[4] Deva Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[5] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6: 1453–1484, 2005.

[6] Yang Wang, Duan Tran, and Zicheng Liao. Learning hierarchical poselets for human parsing. In *CVPR*, 2011.

[7] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

Prioritizing the Propagation of Identity Beliefs for Multi-object Tracking

Amit Kumar K.C.
http://www.uclouvain.be/amit.kc
Christophe De Vleeschouwer
http://www.uclouvain.be/christophe.devleeschouwer

ICTEAM Institute
Université catholique de Louvain
Louvain-la-neuve, Belgium

Multi-object tracking requires locating the targets as well as labelling their identities. Inferring identities of the targets is a challenge when the availability and the reliability of the appearance features do vary along the time and the space. We see the multi-object tracking and identification as a two-stage process.

In the first stage, plausible target candidates are detected at each frame independently, and are aggregated into tracklets. The benefits obtained from such aggregation process are twofold. First, it reduces the number of entities that have to be processed later. Second, it provides more reliable and more accurate knowledge about the appearance of the target observed along the tracklet.

In the second stage, which embeds the main contributions of the paper, a graph-based belief propagation formalism is considered to estimate the identity of each tracklet. Each node in the graph corresponds to a tracklet, and is assigned a probability distribution of identities, based on the tracklet appearance, and given prior knowledge of the possible target appearances. Typically, a low confidence in the tracklet appearance measurement, or a measurement that is similar to several target appearances, both result into a flat and thus ambiguous identity distribution for the tracklet. Afterwards, belief propagation is considered to infer the identities of more ambiguous nodes from those of less ambiguous nodes, by exploiting the graph constraints. In contrast to the approaches with standard belief propagation [2], which treats the nodes in an arbitrary order, the proposed method schedules less ambiguous nodes to transmit their messages first.

From appearance features to identity distribution We assume that there are N targets, each of them being characterized by K appearance features. The feature set for the j-th target is $\mathcal{F}^{(j)} = \{\mathbf{f}_1^{(j)}, ..., \mathbf{f}_K^{(j)}\}$. Let the appearance features for a tracklet v be $\overline{\mathcal{F}}^{(v)} = \{\bar{\mathbf{f}}_1^{(v)}, ..., \bar{\mathbf{f}}_K^{(v)}\}$. Then, the probability of the tracklet v having identity j, denoted by $p_v(j)$, as

$$p_v(j) \propto \prod_{i=1}^{K} \exp\left[-\| \mathbf{f}_i^{(j)} - \bar{\mathbf{f}}_i^{(v)} \|_1 / \tau_i^{(v)}\right] \qquad \text{for } 1 \le j \le N \qquad (1)$$

where $\tau_i^{(v)}$ monitors the influence of feature i on identity assignment. It decreases as the appearance feature observation becomes more reliable. Depending on the observed appearance features and on the estimated reliability of these observations, some tracklets have less ambiguous identity distributions than others.

Graph definition The tracklets are gathered into a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes, with each node corresponding to a tracklet; $\mathcal{E}$ is a set of edges, defining the connectivity between the nodes in $\mathcal{V}$. An edge between nodes u and v implies that their identities are dependent. For example, two tracklets, which co-exist at the same time, should belong to two different physical targets. This defines a *mutex* edge between them. Additionally, if they are sufficiently close in space, time and/or appearance, they are likely to share the same identity, whereas if they are far, they should be encouraged to have different labels. This defines a *temporal* edge between them. Each node $v \in \mathcal{V}$ and each edge $(uv) \in \mathcal{E}$ is characterized by potential functions ϕ_v and ϕ_{uv} respectively. In short, $\phi_v(l_v)$ represents how likely is the label l_v be assigned to the node v. Similarly, $\phi_{uv}(l_u, l_v)$ represents the likelihood that nodes u and v have labels l_u and l_v respectively.

Belief propagation We briefly introduce how the belief propagation formalism works. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is given, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ represents the association between the nodes. The neighbourhood of node $v \in \mathcal{V}$ is denoted by $\mathcal{N}_v$. The purpose of belief propagation is to find a labelling function l that labels each node $v \in \mathcal{V}$ with a label $l_v \in \mathcal{L}$, $\mathcal{L}$ being the set of possible labels, so as to maximize the joint likelihood function:

$$p(l) \propto \prod_{v \in \mathcal{V}} \left[\phi_v(l_v) \prod_{u \in \mathcal{N}_v} \phi_{uv}(l_u, l_v)\right] \qquad (2)$$

It is done iteratively by exchanging "messages" between the nodes. Let $\mathbf{m}_{u \to v}^{(t)}$ be the message that the node u sends to a neighbouring node v at iteration t. Intuitively, $m_{u \to v}^{(t)}(l_v)$ is the belief that node u thinks about the label l_v of node v at any iteration t. Each message is initialized uniformly. Afterwards, new messages are updated (in sum-product form) at each iteration. as:

$$m_{u \to v}^{(t)}(l_v) \propto \sum_{l_u \in \mathcal{L}} \left[\phi_{uv}(l_u, l_v)\phi_u(l_u) \prod_{s \in \mathcal{N}_u \setminus v} m_{s \to u}^{(t-1)}(l_u)\right] \qquad (3)$$

After T iterations, a belief vector $\mathbf{b}_v$ is computed for node v as

$$b_v^{(T)}(l_v) \propto \phi_v(l_v) \prod_{s \in \mathcal{N}_v} m_{s \to u}^{(T)}(l_v) \qquad (4)$$

from which the most likely identity is estimated.

Construction of potential terms We briefly explain how we design the potential terms in our application scenario. The unary potential term $\phi_v(l_v)$ is defined to be the likelihood of the node $v \in \mathcal{V}$ having a label l_v. It is given as: $\phi_v(l_v) = p_v(l_v)$, $l_v \in \mathcal{L}$ (ref. Eqn 1). In case of mutex edges, u and v should have different labels. Therefore,

$$\phi_{uv}(l_u, l_v) = \begin{cases} \varepsilon & \text{if } l_u = l_v \\ 1 - \varepsilon & \text{otherwise,} \end{cases} \qquad (5)$$

We use $\varepsilon = 0.1$. We express ϕ_{uv} for temporal edges in terms of the distance d_{uv} as

$$\phi_{uv}(l_u, l_v) = \begin{cases} \exp(-d_{uv}/\tau_{\text{dist}}) & \text{if } l_u = l_v \\ 1 - \exp(-d_{uv}/\tau_{\text{dist}}) & \text{otherwise,} \end{cases} \qquad (6)$$

where τ_{dist} is a constant. If both u and v have reliable identity estimate, then the Bhattacharyya distance between the belief vectors, $\mathbf{b}_u$ and $\mathbf{b}_v$, is used to define d_{uv}. On the other hand, if one of the nodes does not have reliable identity estimate, then the computation of the Bhattacharyya distance is irrelevant. In such cases, when the nodes are close in time, the position information is used to measure their distance. In contrast, when the nodes are far in time, even the position cannot guide the definition of the distance. In this case, no message is exchanged between the nodes.

Priority scheduling of belief message exchanges To emphasize our contribution, we make two observations about the standard belief propagation: (i) nodes are arbitrarily selected to send messages, (ii) a node gathers information from all its neighbours. However, in our graph formulation, some nodes are less ambiguous about their identities than others. The messages sent by such nodes are more informative. Hence, they help the more ambiguous neighbours to disambiguate their labels [1]. Moreover, during the message construction step, since the messages coming from more ambiguous nodes are usually uninformative and even confusing, we strictly restrict gathering of messages from less ambiguous nodes as shown in Figure 1.

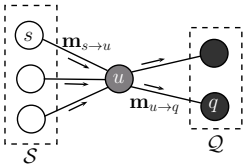


Figure 1: Message construction and dissemination at node u (in **gray**). The node u gathers information from its less ambiguous neighbors, $\mathcal{S}$ (in **white**). Afterwards, u transmits message to its more ambiguous neighbors, $\mathcal{Q}$ (in **black**).

Given the current estimate of belief vector $\mathbf{b}_v$, we use *entropy* of the belief vector to measure the ambiguity level of a node. Then, nodes are sorted in increasing order of entropies.

Results Experimental validation is performed on 10 minutes long real-life basketball video. The proposed method achieves 89% identification rate, which is an improvement of 21% and 16% compared to individual identity assignment, and to standard belief propagation, respectively. Please refer to the main paper and the supplementary paper for detailed analysis.

[1] N. Komodakis and G. Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *Image Processing, IEEE Transactions on*, 16(11):2649 –2661, nov. 2007.

[2] Wei-Lwun Lu, Jo-Anne Ting, Kevin P. Murphy, and James J. Little. Identifying players in broadcast sports videos using conditional random fields. In *CVPR*, 2011.

Face Alignment Using a Ranking Model based on Regression Trees

Hua Gao[1]
gao@kit.edu
Hazım Kemal Ekenel[1,2]
ekenel@kit.edu
Rainer Stiefelhagen[1]
rainer.stiefelhagen@kit.edu

[1] Institute for Anthropomatics
Karlsruhe Institute of Technology
Karlsruhe, Germany

[2] Faculty of Computer and Informatics
Istanbul Technical University
Istanbul, Turkey

In this work, we exploit the regression trees-based ranking model, which has been successfully applied in the domain of web-search ranking, to build appearance models for face alignment. The model is an ensemble of regression trees which is learned with gradient boosting. The MCT (Modified Census Transform) [1] as well as its unbinarized version PCT (Pseudo Census Transform) [2] are used as features due to their robustness to illumination changes. To avoid the overfitting problem and ensure quick convergence in gradient boosting, we use random trees to initialize the boosting. The Nelder Mead's simplex method is applied for fitting the learned model. We compare the proposed regression trees-based point-wise ranking model to pairwise ranking model. Experiments show that the proposed model improves both robustness and accuracy for face alignment.

Classification-based boosted appearance model using the PCT features has been proposed in [2]. However, the model has its own drawback as the positive and negative training samples are highly imbalanced. Furthermore, the learned score function does not guarantee smoothness and concavity in the neighborhood of real solution. Optimizing such a score function with a local optimizer is prone to local maxima. In [4, 5], Ranking-based Appearance Models (RAM) are investigated by boosting the score function in a pairwise ordinal classification way. This model ensures that the score function returns a higher value if the current alignment is closer to the ground truth than the others in the shape parameter space. A local optimizer benefits from such a model as the gradient of the learned score function is constrained to the same direction towards the ground truth.

Based on this idea, we propose and compare two ranking-based appearance models in this work. Both models use a generative shape model assuming 2D face shapes lie in a linear subspace. We represent a novel shape $\mathbf{s}$ with a linear combination of shape basis: $\mathbf{s} = \mathbf{s}_0 + \Sigma_{i=1}^{n} p_i \mathbf{s}_i$, where $\mathbf{s}_0$ is the mean shape, $\mathbf{s}_i$ is the i-th shape basis, and $\mathbf{p} = [p_1, p_2, \ldots, p_n]^\top$ is the shape parameter. A non-linear mapping function $\mathbf{W}(\mathbf{x};\mathbf{p})$ is defined which maps pixel $\mathbf{x}$ defined in an instance shape to the mean shape. A shape-free image (Figure 1(a)) $\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))$ is obtained after warping a face image $\mathbf{I}$ given shape parameter $\mathbf{p}$.

The first ranking appearance model learns a ranking function via pairwise ordinal classification as proposed in [4]. However, in this work, we apply the pairwise RankSVM [3] on the PCT features to build weak rankers (Eq. 1):

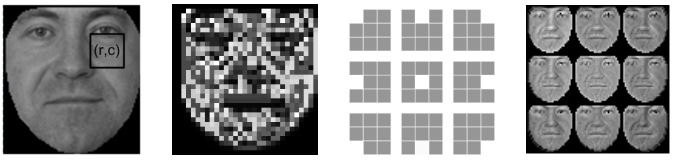
$$f_m(\mathbf{p}) = \frac{1}{\pi} atan(\mathbf{w}^{m\top} S(\varphi^m) - t^m). \quad (1)$$

A PCT feature vector φ^m is extracted at a particular location in the masked shape-free image. The function $atan()$ is used to ensure both discriminability and derivability. The $S(\cdot)$ is a sigmoid function, which normalizes the values in a raw PCT feature into a range of $(0,1)$. The projection vector $\mathbf{w}^m$ is learned with RankSVM. The threshold t^m is determined with weighted least square estimation. The final strong ranking function (Eq. 2) is combined by boosting weak rankers with gentleboost:

$$F(\mathbf{p}) \doteq \sum_{m=1}^{M} f_m(\mathbf{p}). \quad (2)$$

Fitting a learned model to a novel image is done by maximizing this score function with respect to the shape parameters using gradient ascent method.

Point-wise ordinal regression based on gradient boosted regression trees (GBRT) has gained much attention for solving ranking problem in information retrieval domain. We apply GBRT in the second ranking appearance model. The GBRT iteratively fits regression trees of certain depth to regression residues, which results in less biased estimation. As the learned ranking function is not differentiable anymore due to the hierarchical tree structure, we apply both PCT and MCT as our feature representation. Usually a small shrinkage is used for training GBRT to avoid overfitting. However, setting small shrinkage leads to slow convergence which results in enormous number of trees in the strong ranking function. This problem can be solved by initializing the GBRT properly so that the initial estimation is already close to the regression target.



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

Figure 1: (a) A shape-free face image; (b) MCT output of a shape-free image; (c) 9 PCT filter masks; (d) PCT-filter responses of a shape-free image.

We use random forest as an initial estimation for the iterative GBRT training. The random forest is a combination of bagging and random feature selection. This leads to a low variance estimation and less sensitive to noise and outliers. We denote this initialized GBRT as iGBRT. The output of the final boosted ranking score function is actually the response of RF combined with the boosted regression trees:

$$T(\mathbf{p}) = F(\mathbf{p}) + \alpha \sum_{t=1}^{M_B} h_t(\mathbf{p}). \quad (3)$$

Where $F(\cdot)$ is the initial estimation from random forest. $h_t(\cdot)$ is a boosted weak regression tree. α is the shrinkage and M_B is the total number of boosted trees.

Face alignment is equivalent to maximizing Equation 3 with the constraint of the shape prior. We define the cost function as follows:

$$O(\mathbf{p}) = -T(\mathbf{p}) + \beta \sum_{i=0}^{n} \frac{p_i^2}{\lambda_i}, \quad (4)$$

where β is the parameter that we estimated from the training data. λ_i is the eigenvalue corresponding to shape parameter p_i. As it is difficult to derive the analytical gradient for the cost function, we use the Nelder-Mead simplex method to minimize Equation 4 which only requires the evaluation of the cost function.

The details of the learning the ranking appearance models are described in the paper, including data sampling and parameter setting for training the models. We evaluated the alignment on different data sets and observed that ranking-based appearance models improve the fitting performance over the classification-based model. The iGBRT model outperforms the pairwise ranking model as well as the original GBRT-based model.

[1] B. Fröba and A. Ernst. Face detection with the modified census transform. In *Proc. of 6th Int. Conf. on Automatic Face and Gesture Recognition*, pages 91–96, 2004.

[2] H. Gao, H. K. Ekenel, M. Fischer, and R. Stiefelhagen. Boosting pseudo census transform features for face alignment. In *Proc. of BMVC, Dundee, UK*, 2011.

[3] R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. *Advances in Large Margin Classifiers*, pages 115–132, 2000.

[4] H. Wu, X. Liu, and G. Doretto. Face alignment via boosted ranking model. In *Proc. of CVPR*, 2008.

[5] J. Zhang, S. K. Zhou, D. Comaniciu, and L. McMillan. Discriminative learning for deformable shape segmentation: A comparative study. In *Proc. of ECCV*, 2008.

Binary Pattern Analysis for 3D Facial Action Unit Detection

Georgia Sandbach[1]
gls09@imperial.ac.uk
Stefanos Zafeiriou[1]
s.zafeiriou@imperial.ac.uk
Maja Pantic[1,2]
m.pantic@imperial.ac.uk

[1] Department of Computing
Imperial College London
London, UK

[2] EEMCS
University of Twente
Enschede, Netherlands

Recognition of facial expressions is a challenging problem, as the face is capable of complex motions, and the range of possible expressions is extremely wide. For this reason, detection of facial action units (AUs) from the Facial Action Coding System, a comprehensive system for coding facial muscle movements, has become a widely studied area of research. The use of 3D facial geometry data and extracted 3D features for expression recognition has so far not been heavily studied. Images and videos of this kind allow a greater amount of information to be captured (2D and 3D), including out-of-plane movement which 2D cannot record, whilst also removing the problems of illumination and pose inherent to 2D data. For this reason some work has begun to employ 3D facial geometry data for facial expression recognition or facial AU detection.

We tackle this problem with the introduction of a variety of new binary pattern features that are all based on the traditional Local Binary Pattern (LBP) [6] or Local Phase Quantiser (LPQ) [5] features. In order to do this, we employ two 2D representations of the 3D facial geometry information. Firstly we utilise the depth map, that has been widely used in 3D facial analysis, and secondly we define the Azimuthal Projection Distance Image (APDI), which captures the comparative directional information of the normals in the mesh as a 2D representation. The Azimuthal Equidistant Projection (AEP) is able to project normals onto positions in a Euclidean 2D plane. For our purposes we alter the projection to create the APDI, which allows direct comparison of the projection coordinates of neighbouring points.

(a) (b) (c)

Figure 1: 2D representations of the facial mesh for subject AU20 (a) Original facial mesh (b) Depth map (b) Azimuthal Projection Distance Image

We have a regular grid of normals, defined as $\mathbf{n}(i,j) = (u_{i,j}, v_{i,j}, w_{i,j})$, to be projected relative to a set of mean normals $\hat{\mathbf{n}}(i,j)$ calculated at each point. We set the elevation and azimuth of the mean normals, $\hat{\theta}$ and $\hat{\phi}$, to be $\frac{\pi}{2}$ and 0 respectively at every point. This then allows calculation of the distances of the normals as compared to the mean $\hat{\mathbf{n}} = (1,0,0)$ which is chosen as a reference to create an image suitable for further analysis. This assumption allows the AEP projection to be simplified to:

$$x_{i,j} = k' cos\theta(i,j) sin\phi(i,j) \qquad y_{i,j} = k' cos\theta(i,j) cos\phi(i,j) \qquad (1)$$

for a point $\mathbf{p}(i,j) = (x_{i,j}, y_{i,j})$, where θ is the elevation angle, and ϕ is the azimuthal angle, of the normal at this point. The above formulation allows comparison between normal distances in Euclidean space, and this simplification also reduces the complexity of the feature extraction process. In order to employ this in the binary pattern framework, the coordinates are used to find an absolute distance from the origin $d_{i,j} = \sqrt{x_{i,j}^2 + y_{i,j}^2}$, and these values form the APDI for the facial mesh. Examples of the depth map and APDI, calculated for the facial mesh seen in Fig. 1(a), are shown in Figs. 1(b) and 1(c) respectively.

Each representation is then exploited for use with binary pattern algorithms in order to form feature types suitable for robust AU detection. Firstly, the traditional Local Binary Pattern (LBP) algorithm, which assigns a binary pattern to each point in an image by thresholding the neighbouring points on the central value, was applied directly to each representation. This forms the previously proposed 3DLBP [7], for use as a baseline test, and the new Local Azimuthal Binary Pattern (LABP) respectively. Next we utilise other methods that have been employed for

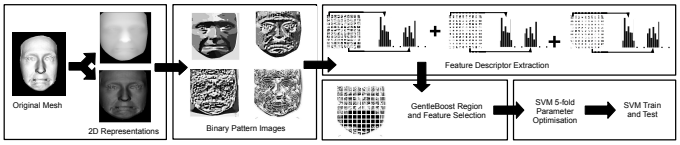


Figure 2: An overview of our proposed system.

2D feature extraction: LPQs, Local Gabor Binary Patterns (LGBPs) [4], and Histogram of Monogenic Binary Patterns (HMBPs) [3]. In the latter two cases, we extend the methods to apply them to the magnitude, phase, and, in the case of the Monogenic signal, orientation. In total this formed seven new features: (1) Local Azimuthal Binary Patterns (LABPs) (2) Local Depth Phase Quantisers (LDPQs) (3) Local Azimuthal Phase Quantisers (LAPQs) (4) Local Depth Gabor Binary Patterns (LDGBPs) (5) Local Azimuthal Gabor Binary Patterns (LAGBPs) (6) Local Depth Monogenic Binary Patterns (LDMBPs) (7) Local Azimuthal Monogenic Binary Patterns (LAMBPs). The performance of these new features is assessed as compared to the original 3DLBPs.

Feature vectors are created for each of the above descriptors through the use of histograms. First, the x-y plane of the mesh is divided into 10x10 equally-sized square blocks, and for each of these a histogram is formed from the calculated binary numbers. These histograms are then concatenated into one large feature vector. Feature selection is performed in order to reduce the dimensionality of the feature vectors before classification. The GentleBoost algorithm was used for this purpose, with two stages to the feature selection: first selection of regions, and then particular features. Support Vector Machines (SVMs) were then trained for detection of each AU, with parameter optimisation carried out using 5-fold cross-validation, and these were used for testing of all sequences. An overview of our system can be seen in Fig. 2.

Experimental testing was conducted in two ways: 10-fold cross- validation on the Bosphorus database [2], and cross-database testing with training on this database, and testing carried out on the D3DFACS database [1]. The results achieved show a definite improvement with all of the new features over the traditional 3DLBP, with a maximum cross-validation ROC AuC of 97.2 when using LDGBPs. This improvement was also seen with the depth features on the cross-database testing, though the Azimuth results in this test suggested that these features are less robust when there are large variations in smoothness of the mesh between training and testing data.

[1] D. Cosker et al. A FACS valid 3D dynamic action unit database. In *ICCV 2011*, pages 2296–2303. IEEE, 2011.

[2] A. Savran et al. Bosphorus database for 3D face analysis. *Biometrics and Identity Management*, pages 47–56, 2008.

[3] M. Yang et al. Monogenic Binary Pattern (MBP): A Novel Feature Extraction Model. In *ICPR 2010*, pages 2680–2683. Ieee, 2010.

[4] W. Zhang et al. Local Gabor binary pattern histogram sequence (LGBPHS). In *ICCV 2005*, volume 1, pages 786–791. IEEE, 2005.

[5] V. Ojansivu and J. Heikkilä. Blur insensitive texture classification using local phase quantization. *ISP*, pages 236–243, 2008.

[6] T. Ojala et al. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE PAMI*, pages 971–987, 2002.

[7] Y. Huang et al. Combining statistics of geometrical and correlative features for 3D face recognition. In *BMVC 2006*, pages 879–888. Citeseer, 2006.

A Phase Field Method for Tomographic Reconstruction from Limited Data

Russell J. Hewett
rhewett@mit.edu

Ian Jermyn
i.h.jermyn@durham.ac.uk

Michael T. Heath
heath@illinois.edu

Farzad Kamalabadi
farzadk@illinois.edu

Imaging and Computing Group
Department of Mathematics
Massachusetts Institute of Technology

Department of Mathematical Sciences
Durham University

Department of Computer Science
University of Illinois at Urbana-Champaign

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

Solar phenomena such as active regions, flares, coronal mass ejections (CMEs), and solar wind, all of which contribute to geoeffective events, collectively referred to as *space weather*, are not well understood [1], yet are of critical importance due to modern society's reliance on technologies that can be disrupted by these events. Understanding such activity requires knowledge of the electron density of the solar corona (or solar atmosphere), but such knowledge is hard to come by for local, short-lived events such as CMEs.

Direct imaging of CME electron density via tomographic reconstruction is difficult because a maximum of three unique observations of any given event are available. Such a regime requires a reconstruction algorithm that is robust to sparse data. Mumford-Shah type models have previously been proposed for CME reconstruction [2] but have not been successful due, in part, to the complex topology of CME structures.

We present a method for reconstruction from sparse data that, similarly, uses an auxiliary segmentation to constrain the density, but we represent the segmentation using the phase field level set framework, thereby eliminating important topological limitations and allowing for smooth enforcement of two different regularization regimes, while retaining robustness. We use a fast variational algorithm to compute MAP estimates, and compare our results to classical regularized tomography for synthetic CME-like images.

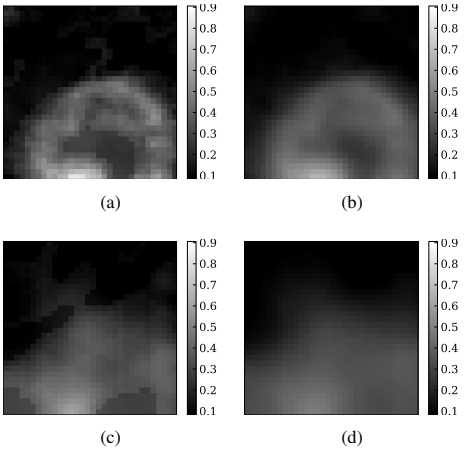


Figure 1: Comparison of joint segmentation-reconstruction with Tikhonov regularized reconstructions for CME image. (a) Joint segmentation-reconstruction and (b) Tikhonov regularized reconstruction with $\lambda = \lambda_-$ for 32 equispaced observation angles and 32 projections per angle. (c) and (d) same, for 3 equispaced observation angles.

Model We seek to infer the electron density f in a local section of the solar corona Ω and a segmentation, represented by a phase field ϕ classifying a subset $R \subset \Omega$ as part of a CME, given a set of coronagraphs Y and prior knowledge K, (*e.g.* parameter values). We compute the MAP estimate $(\hat{f},\hat{\phi})$ from $-\ln P(f,\phi|Y,K) = E(Y|f,K) + E(f|\phi,K) + E(\phi|K)$:

$$(\hat{f},\hat{\phi}) = \arg\min_{(f,\phi)} \sum_j \frac{1}{2\sigma^2}\int_\Gamma \big(y_j - h_j(f)\big)^2 + \int_\Omega \Big\{ \frac{1}{2}(\lambda_+\phi_+ + \lambda_-\phi_-)(\nabla f\cdot\nabla f) - c_4\nabla\phi\cdot\nabla f + c_1\frac{1}{2}\nabla\phi\cdot\nabla\phi + c_2(\frac{1}{4}\phi^4 - \frac{1}{2}\phi^2) + c_3(\phi - \frac{1}{3}\phi^3)\Big\}. \quad (1)$$

The first line of (1) defines $E(Y|f,K)$, representing noisy tomographic measurements.

The third line defines the segmentation energy, $E(\phi|K)$. The phase field ϕ represents the region $R = \{x \in \Omega : \phi(x) > 0\}$; c_1, c_2, and c_3 are free parameters. The last two terms are a double well potential: for $|c_3| < c_2$, local minima occur at $\phi = \pm 1$. Coupled with the smoothing effect of the first term, the potential ensures that, away from the region boundary and for fixed R, ϕ takes the values 1 in R and -1 in $\Omega \setminus R$. Near the boundary, there is a smooth transition across an interface zone of width $4\sqrt{c_1/c_2}$. The effective energy controlling R is then a linear combination of the length (area) of the boundary and the area (volume) of the interior of R for 2D (3D) regions [3].

The second line defines $E(f|\phi,K)$, which couples the density and the phase field. In the first component, $\phi_\pm = (1\pm\phi)/2$ act as pseudo-indicator functions for the CME and background regions, thus defining distinct Tikhonov regularization parameters, $\lambda_\pm$, for the interior and exterior of R. The second term favours large inward pointing ∇f on the boundary, because CMEs generally have sharply higher densities than the background. Thus, like [2], we model the background as smoother than the CME, and with a very different density, but unlike [2], the phase field defines a smooth change in the regularization parameter over the interface.

Algorithm Traditionally, energies such as (1) are optimized by split-step gradient descent methods relying on explicit finite differencing to solve the associated PDEs. However, the small time step required for stability leads to slow convergence and implicit methods allowing larger time steps are impractical due to increased computational complexity from the nonlinearity in the phase field potential. We resolve these issues by minimizing in both f and ϕ simultaneously using finite element discretization and a trust-region-based variation on Newton's method, the Levenberg-Marquardt method. In this approach, the length of the descent step is dependent upon the minimization algorithm and the local shape of the objective function, and is not explicitly constrained by the discretization.

Results We compare our results to those obtained using Tikhonov regularization (Fig. 1). We see that, even for the limited angle reconstructions, the joint segmentation-tomographic reconstructions have definition in the CME region that is not present in either of the Tikhonov regularized reconstructions. Our experiments show that our method is significantly more effective than Tikhonov regularized tomography alone, and resolves issues with CME topology and continuity of the density that affected previous work. Our model and optimization method easily extend to the full 3D CME reconstruction problem, though further work on parameter estimation is necessary to render the method automatic.

[1] M. J. Aschwanden. *Physics of the Solar Corona*. Springer-Verlag, Berlin, Germany, 2004.

[2] R. A. Frazin, M. Jacob, W. B. Manchester, H. Morgan, and M. B. Wakin. Toward reconstruction of coronal mass ejection density from only three points of view. *The Astrophysical Journal*, 695:636–641, April 2009.

[3] M. Rochery, I. Jermyn, and J. Zerubia. Phase field models and higher-order active contours. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 970–976, 2005.

Adaptive hierarchical contexts for object recognition with conditional mixture of trees

Billy Peralta
bmperalt@uc.cl
Pablo Espinace
pespinac@uc.cl
Alvaro Soto
asoto@ing.puc.cl

Computer Science Department
Pontifica Universidad Catolica de Chile
Santiago, Chile

Contextual information has emerged as an attractive option to boost the performance of single object category detectors [1][2]. Regarding to these techniques, Choi et al. [1] presents an efficient scheme to model inter-object relations using a tree-structured Bayesian network. Recently, [2] uses contextual cueing, spatial co-ocurrence, and inhibitory intra-class constraints among objects using a max-margin approach. In all these cases, contextual relations among objects are fixed and do not depend of the type of scene being analyzed. We propose that using an adaptive scheme to model contextual relations among objects can boost the performance of current object recognition techniques. We can illustrate this idea by the following example. Consider the case of the contextual relation between the presence of a person and a dog objects. Under a park scene, person and dog objects co-occur frequently, but in an office scene, they hardly co-occur, therefore modeling such relation with a fixed contextual constraint limits the flexibility of the model to fit real data.

In this work, we present a method that learns adaptive conditional relationships among objects according to the underlying scene information. To achieve this goal, we use a probabilistic model based on a conditional mixture of trees [4]. Our work is based on an extension of the scheme proposed by Choi et al. [1]. In that work, the authors use a single tree graphical model to represent dependencies among objects. In contrast, our mixture of tree allow us to model different contextual relations among object categories.

Our mixture of trees context model is built by a conditional mixture of tree-structured Bayesian Networks, each of which is an expert in some partition of the set of images. These networks have a weight that depends on the global scene information given by a Gist feature vector x_G. The model can be seen as a mixture of experts where the gate function is given by a function of the global representation, and each expert function is given by an individual Bayesian network.

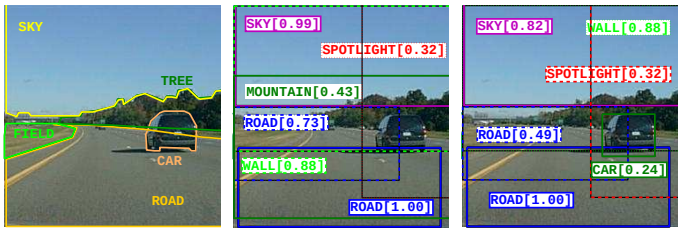
In order to incorporate global scene information in our model, we model object presence as dependent on scene type. In particular, for a training set of N images, we can construct N instance-label pairs (x_G, b), where $b \in \{0,1\}^D$ represents presence in a given image of instances of D possible object categories. Our goal is to use instances (x_G, b) to include in our model different types of contextual relations between object classes. We achieve this goal by introducing latent variable z which simplifies the analysis of the model. We refer to this latent variable as the *context variable*. We assume that there are K possible values for z, i.e., we assume the existence of K *contexts* for object classes. This is similar to a mixture of experts model with the exception that in our case b is conditionally independent of x_G given z. The context variable is assumed as a winner-take-all variable, i.e., each object class detection occurs under a specific contextual scenario. Considering K contexts, we can model the conditional density as :

$$p(b|x_G) = \sum_{i=1}^{K} p(b, z_i|x_G) = \sum_{i=1}^{K} p(b|z_i)\, p(z_i|x_G) \qquad (1)$$

Here, we have the K contexts represented by z_i with $i = \{1, ..., K\}$, where each context has its own class-conditional probability function. The two components of the mixture model given by Equation 1 are the context gate function, given by $p(z_i|x)$, and the tree experts function, given by $p(b|z_i)$. In particular, $p(z_i|x)$ is modelled as a normalized Gaussian kernel. On the other hand, $p(b|z_i)$ is modelled as a tree-structured Bayesian Network [4].

Assuming that the posterior probabilities of context gates R_{in} (*responsabilities*) for each expert i and training instance n are known, we apply the EM algorithm over the expected log-likelihood $\langle L_c \rangle$ in order to obtain the parameters:

$$\langle L_c \rangle = \sum_{n=1}^{N} \sum_{i=1}^{K} R_{in}\, log\,(\, p(b_n|z_i)\, P_i(x_n)\, \alpha_i) \qquad (2)$$



(a) Ground-Truth Image (b) Results according to [1](c) Results with Mixture of trees

Figure 1: An example of the results of our method with respect to a state-of-the-art model [1]. Our model 1(c) adaptively selects a suitable component of a mixture of trees for providing a correct detection of the car object and do not detect a phantom mountain object as in [1].

Inference is straightforward, as we separate each tree in its own partition. We make inference using message passing algorithms for each tree $(p(b, c, L/g, W, s, z))$ in an iterative fashion as in [1]. Then we obtain the final score by combining the scores of each component with its respective parameters, similar to the work of Meila and Jordan [4].

We perform an empirical evaluation of the proposed approach considering two real public datasets: (i) OUTDOOR dataset created by Oliva and Torralba [5], and (ii) SUN09 dataset publicated by Choi [1]. We employ the object detector proposed by Felzenszwalb et al. [3]. Our work use average precision-recall (APR) as a performance metric for our model. This metric corresponds to the area under the precision-recall curve.

In relation to our experiments, we achieve the best performance using 6 trees. In this case, the relative improvements in terms of APR with respect to [1] are 5.5% and 5.7% for OUTDOOR and SUN09 datasets, respectively. In terms of individual classes, for the case of 6 trees, we notice that with respect to [1] APR increases for 10 and 53 objects and decreases for 6 and 34 objects for the OUTDOOR and SUN09 datasets, respectively.

As a main conclusion, our experiments using standard object datasets indicate that the proposed model improves object recognition performance with respect to a single tree model. This validates our main hypothesis indicating the relevance of including adaptive contextual relations to boost the performance of object category detectors.

[1] M. Choi, J. Lim, A. Torralba, and A. Willsky. Exploiting hierarchical context on a large database of object categories. In *Proc. CVPR*, pages 129–136, 2010.

[2] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 95:1, 2011.

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

[4] M. Meila and M. Jordan. Learning with mixtures of trees. *JML*, 1: 1–48, September 2001.

[5] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42:145–175, May 2001. ISSN 0920-5691.

Local Shape Representation in 3D: from Weighted Spherical Harmonics to Spherical Wavelets

Cheng-Jin Du[1] , John G. Ferguson [2], Phillip T. Hawkins[2], Len R. Stephens [2], Till Bretschneider[1]

[1]Warwick Systems Biology Centre, University of Warwick
[2]The Babraham Institute, Cambridge

Numerous techniques have been proposed for shape representation, including landmarks [1], medial representation [2], spherical harmonics (SPHARM) [3], weighted SPHARM [4], and spherical wavelets (SW) [5]. Among them, both weighted SPHARM and SW have been used for local shape representation of biological structures. Questions we address in this paper are what is the relationship between them, how to derive SW from weighted SPHARM, how to formulate the derived SW for local shape representation, and which one is better in terms of performance and efficiency for a typical biological problem.

The coordinate x of a point $p(\theta,\varphi)$ on a unit sphere Ω can be represented by weighted SPHARM as the following kernel smoothing

$$x(p)=\int_{\Omega} x(q)K_t^L(p,q)d\eta(q) \qquad (1)$$

where $d\eta(q)=\sin\theta d\theta d\varphi$, $\theta\in[0,\pi]$ and $\varphi\in[0,2\pi)$ are the polar and the azimuthal angles respectively, and the symmetric positive kernel K_t^L is

$$K_t^L=\sum_{l=0}^{L} e^{-l(l+1)t}\sum_{m=-l}^{l} Y_l^m(p)Y_l^m(q)=\sum_{l=0}^{L}\frac{2l+1}{4\pi}e^{-l(l+1)t}P_l(p\cdot q) \qquad (2)$$

where Y_l^m is SPHARM with the degree $l\geq 0$ and order $|m|\leq l$. Eq. (2) is essentially the Gauss-Weierstrass kernel [6].

The term $e^{-l(l+1)t}$ in Eq. (2) can be considered as a discretized version of the continuously defined function $\varphi_0(u)=e^{-tu(u+1)}$. The dilation of φ_0 is given as

$$\varphi_j(u)=D_j\varphi_0(u)=\varphi_0(2^{-j}u) \qquad (3)$$

where D_j is called dilation operator of j-th level. A system of scale discrete scaling function can be generated via φ_0 and its dilations φ_j as

$$\Phi_j(v)=\sum_{l=0}^{\infty}\frac{2l+1}{4\pi}\varphi_j(l)P_l(v), v\in[-1,1] \qquad (4)$$

The discrete scaling function of Eq. (4) defines a "discrete approximate identity" [7] in $L^2(\Omega)$. Based on Eq. (4), scale discrete wavelets on the sphere can be introduced as the difference of two successive resolution levels

$$\Psi_j(v)=\Phi_{j+1}(v)-\Phi_j(v)=\sum_{l=0}^{\infty}\frac{2l+1}{4\pi}\left(\varphi_{j+1}(l)-\varphi_j(l)\right)P_l(v) \qquad (5)$$

which can be considered as a difference-of-Gaussian (DoG) wavelet.

As discussed in the paper, the SW derived above are poorly localized, and in fact they do not really resemble wavelets. We propose a new way to construct over-complete SW based on the group theoretic approach [8], and use the theoretical results from the work of [9] to build self-invertible filter banks, which are employed for decomposing and reconstructing images.

We construct the spherical DoG wavelet by projecting its Euclidean planar formula on to the sphere

$$DoG(\theta,\varphi)=\frac{1}{2\pi}(1+\tan^2(\theta/2))\left(\frac{1}{\sigma_1^2}e^{-\frac{2}{\sigma_1^2}\tan^2\frac{\theta}{2}}-\frac{1}{\sigma_2^2}e^{-\frac{2}{\sigma_2^2}\tan^2\frac{\theta}{2}}\right) \qquad (6)$$

The term $1+\tan^2(\theta/2)$ is to ensure the unitarity of the projection. The n[th] analysis filters $\tilde{h}_n$ of the self-invertible filter banks are the stereographic dilation [8] of Eq. (6):

$$\tilde{h}_n(\theta,\varphi)=\left(\prod_{i=1}^{n} b_i\right)D_{a_n}DoG(\theta,\varphi) \qquad (7)$$

where b_i are the amplitude scaling parameters that control the tradeoff between self-invertibility and norm-preserving dilation, and D_{a_n} is the stereographic dilation operator.

A spherical continuous wavelet transform of $x(\theta,\varphi)$ can be given in terms of a wavelet basis by the projection on to each wavelet basis function by spherical convolution

$$W_n(\alpha,\beta,\gamma)=\int_{\Omega}[R(\alpha,\beta,\gamma)\tilde{h}_n]^*(\theta,\varphi)x(\theta,\varphi)d\Omega \qquad (8)$$

where $R(\alpha,\beta,\gamma)$ is the rotation operator. To produce reconstructed surface components, the synthesis filters are used to project a function in $L^2(SO(3))$ onto $L^2(\Omega)$ by inverse convolution

$$\hat{x}_n(\theta,\varphi)=\int_{SO(3)}[R(\alpha,\beta,\gamma)h_n](\theta,\varphi)W_n(\alpha,\beta,\gamma)d\rho \qquad (9)$$

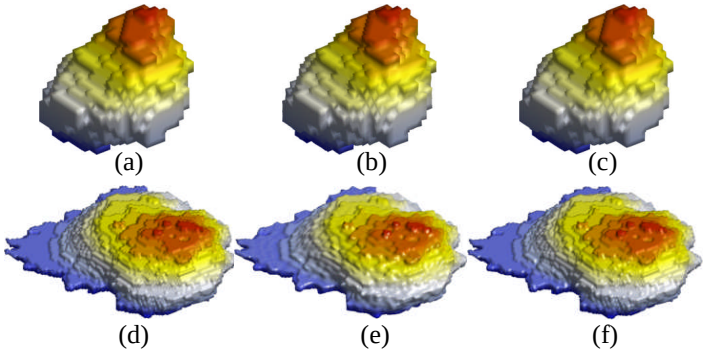


Figure 1: Shape representations of example surfaces of a left amygdala (a) and a neutrophil cell (d) via both weighted SPHARM with 78 degree ((b) and (e)) and SW with level 7 ((c) and (d)), repectively.

[1] F. L. Bookstein. *Morphometric Tools for Landmark Data*. Cambridge University Press, 2003.

[2] H. Blum. Biological shape and visual science. *J. Theor. Biol.* 38 (2), 205–287, 1973.

[3] C. Brechbuhler, G. Gerig, and O. Kubler. Parameterization of closed surfaces for 3D shape description. *Comp. Vis. Image Understanding*, 61:154–170, 1995.

[4] M. K. Chung, K. M. Dalton, L. Shen, A. C. Evans, and R. J. Davidson. Weighted Fourier series representation and its application to quantifying the amount of gray matter. Special Issue of *IEEE Transactions on Medical Imaging*, on Computational Neuroanatomy, 26, 566-581, 2007.

[5] P. Yu, P. Grant, Y. Qi, X. Han, F. Segonne, R. Pienaar, E. Busa, J. Pacheco, N. Makris, and R. Buckner, et al. Cortical surface shape analysis based on spherical wavelets. *IEEE Transactions on Medical Imaging*, 26(4), 582–598, 2007.

[6] W. Freeden, M. Schreiner, and R. Franke. A survey on spherical spline approximation. *Surv. Math. Ind.*, 7, 29-85, 1997.

[7] W. Freeden, and M. Schreiner. Orthogonal and non-orthogonal multiresolution analysis, scale discrete and exact fully discrete wavelet transform on the sphere. *Constr. Approx.*, 14, 493-515, 1998.

[8] J.P. Antoine, and P. Vandergheynst. Wavelets on the 2-sphere: A group-theoretical approach. *Appl. Comput. Harmon. Anal.*, 7(3), 262–291, 1999.

[9] B.T.T. Yeo, W. Ou, and P. Golland. On the Construction of Invertible Filter Banks on the 2-Sphere. *IEEE Transactions on Image Processing* 17(3), 283–300, 2008.

Learning discriminative space-time actions from weakly labelled videos

Michael Sapienza
michael.sapienza-2011@brookes.ac.uk
Fabio Cuzzolin
fabio.cuzzolin@brookes.ac.uk
Philip H.S. Torr
philiptorr@brookes.ac.uk

Brookes Vision Group
Oxford Brookes University
Oxford, UK
cms.brookes.ac.uk/research/visiongroup

Current *state-of-the-art* action classification methods extract feature representations from the entire video clip in which the action unfolds, however this representation may include irrelevant scene context and movements which are shared amongst multiple action classes. For example, a waving action may be perfackormed whilst walking, however if the walking movement and scene context appear in other action classes, *then they should not be included* in a waving movement classifier. In this work, we propose an action classification framework in which more discriminative action *subvolumes* are learned in a weakly supervised setting, owing to the difficulty of manually labelling massive video datasets.

The learned models are used to simultaneously *classify* video clips and to *localise* actions to a given space-time subvolume. Each subvolume is cast as a bag-of-features (BoF) instance in a multiple-instance-learning framework, which in turn is used to learn its class membership. We demonstrate quantitatively that even with single fixed-sized subvolumes, the classification performance of our proposed algorithm is superior to the *state-of-the-art* BoF baseline on the majority of performance measures, and shows promise for space-time action localisation on the most challenging video datasets.

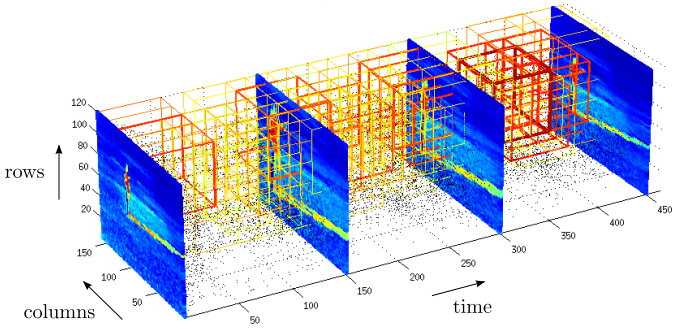


Figure 1: *A boxing video sequence taken from the KTH dataset [2] plotted in space and time. Overlaid on the video are discriminative cubic action subvolumes learned in a max-margin multiple instance learning framework [1], with colour indicating their class membership strength. Since the scene context of the KTH dataset is not discriminative of the particular action, only subvolumes around the actor were selected as positive instances.*

The contributions of this work are as follows:
i) We cast the conventionally supervised BoF action clip classification approach [3] into a weakly supervised setting, where clips are represented as bags of histogram instances with latent class variables. In this way, *more discriminative action parts may be selected which most characterise those particular types of actions*. An example of learned action subvolumes is shown in Fig. 1.
ii) In order to learn the subvolume class labels, we apply multiple instance learning (MIL) to 3D space-time videos, as we maintain that actions are better dackefined within a subvolume of a video clip rather than the whole video clip itself.
iii) Finally we propose a mapping from *instance* decisions learned in the mi-SVM approach to *bag* decisions, as a more robust alternative to the current bag margin MIL approach of taking the sign of the maximum margin in each bag. This allows our MIL-BoF approach to learn the labels of each individual subvolume in an action clip, *as well as the label of the action clip as a whole*.
The resulting action recognition system is suitable for both clip classification and localisation in challenging video datasets, without requiring the labelling of action part locations.

The proposed action recognition system is composed of three main building blocks:
i) The description of space-time video blocks via histograms of Dense Trajectory features [4], which captures the trajectory's shape, appearance, and motion information.
ii) The representation of a video clip as a "bag of subvolumes" illustrated in Fig. 2, and the learning of positive subvolumes from weakly labelled training sequences within a max-margin MIL framework [1].
iii) The mapping of instance/subvolume scores to bag/clip scores by learning a hyperplane on instance margin features. Further details of the action recognition system are discussed in the methodology section of the paper.

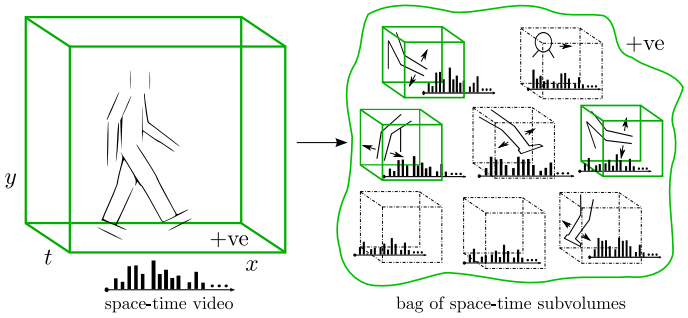


Figure 2: *Instead of defining an action as a space-time pattern in an entire video clip (left), we propose to define an action as a collection of space-time action parts contained in video subvolumes (right). The labels of each action subvolume are initially unknown. Multiple instance learning is used to learn which subvolumes are particularly discriminative of the action (solid-line cubes), and which are not (dotted-line cubes).*

In order to validate our action recognition system, we evaluated its performance on ftour challenging action datasets, namely the **KTH** (6 classes), **YouTube** (11 classes), **Hollywood2** (12 classes) and **HMDB** (51 classes).

In conclusion, we proposed a novel MIL-BoF approach to action clip classification *and* localisation based on the recognition of space-time subvolumes. By learning the subvolume latent class variables with multiple instance learning, more robust action models may be constructed and used for action localisation in space and time or action clip classification via our proposed mapping from instance to bag decision scores. The experimental results demonstrate that the MIL-BoF method achieves comparable performance or improves on the BoF baseline on the most challenging datasets. In the future, we will focus on generalising the MIL-BoF approach by learning a *mixture of subvolume primitives* tailored for each action class, and incorporating geometric structure by means of *pictorial star models*.

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 561–568, 2003.

[2] C. Schüldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *IEEE Int. Conf. on Pattern Recognition*, pages 32–36, 2004.

[3] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. British Machine Vision Conference*, pages 124.1–124.11, 2009.

[4] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action Recognition by Dense Trajectories. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3169–3176, 2011.

Spatio-Temporal Convolutional Sparse Auto-Encoder for Sequence Classification

Moez Baccouche[1]
moez.baccouche@orange.com
Franck Mamalet[1]
franck.mamalet@orange.com
Christian Wolf[2]
christian.wolf@liris.cnrs.fr
Christophe Garcia[2]
christophe.garcia@liris.cnrs.fr
Atilla Baskurt[2]
atilla.baskurt@liris.cnrs.fr

[1] Orange Labs R&D
4 rue du Clos Courtel
F-35510, France

[2] Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR 5205
F-69621, France

We address in this paper the problem of task-independent video sequence classification. We aim to introduce a generic model which differ from the highly problem-dependent dominant methodology that relies on so-called *hand-crafted* features. We propose a learning-based model with two main steps: the first one aims to automatically learn spatio-temporal features instead of hand-crafting them. These learned features are *sparse-overcomplete*, i.e. their dimension is larger than the input one, but only a small number of components are non-zero. The second step consists in labelizing the entire video sequence considering the temporal evolution of the learned features. The first learning step is performed in an unsupervised way, and is based on spatio-temporal convolutional sparse auto-encoders (which will be introduced hereafter), and the second consists in a supervised classification using recurrent neural networks.

The feature learning process is based on an auto-encoder scheme: an encoder which builds a non-sparse code vector representing the spatio-temporal salient information contained in the input, and a decoder which learns to reconstruct the input from a sparse version of the obtained code (see Figure 1). The model takes as input small space-time patches in order to reduce the diversity of the content to be encoded, since the patterns are locally less variable than if the full frame was considered. The encoder is a convolutional neural network with $2D+t$ convolution kernels (each one having the same size than the input patch). The decoder consists in a set of output neurons fully connected to the sparse code layer. The sparsity is obtained using the *sparsifying logistic* proposed by Ranzato *et al.* [2], placed between the encoder and the decoder. This model is associated to a global objective function, which is the sum of two terms, representing respectively the encoder prediction and the decoder reconstruction mean square errors, and which is minimized during training.

In order to handle the spatial and temporal shift-invariance of the learned representations, a "best shift search" module is introduced before the auto-encoder (see Figure 1). The idea is to represent the spatio-temporal neighbourhood of a given input patch by a single "shifted" patch, which is the one minimizing the objective function, given the current set of parameters. To that aim, an additional hidden variable is introduced, the translation vector, on which the optimization is done.

To avoid encoding non-relevant patterns (e.g. colour and texture), the model is trained only with the patches containing significant spatio-temporal information (according to a motion-based selection critirea). This plays the same role as the saliency detectors in the case of the *hand-crafted* features.

Entire sequences are finally labeled with a particular recurrent neural network classifier, namely *Long Short-Term Memory Recurrent Neural Network* (LSTM) [1], in order to take benefits of its ability to use the temporal evolution of features for classification. The LSTM classifier takes as input a sequence of feature vectors, each one corresponding to the concatenated responses of the patches placed at the grid of possible locations in each frame.

Aiming at verifying the genericity of the proposed model, experiments were carried out on two different problems: human actions and facial expression recognition. For the first experiment, we used the standard KTH human actions dataset [3]. To our knowledge, our method obtains the best results among the methods using automatically learned features, both on the two versions of the KTH dataset (95.83% for KTH1 and 93.74% for KTH2). More generally, we obtain the second best result for KTH1, and the third for KTH2, even when compared with approaches relying on hand-crafted features designed for the KTH dataset.

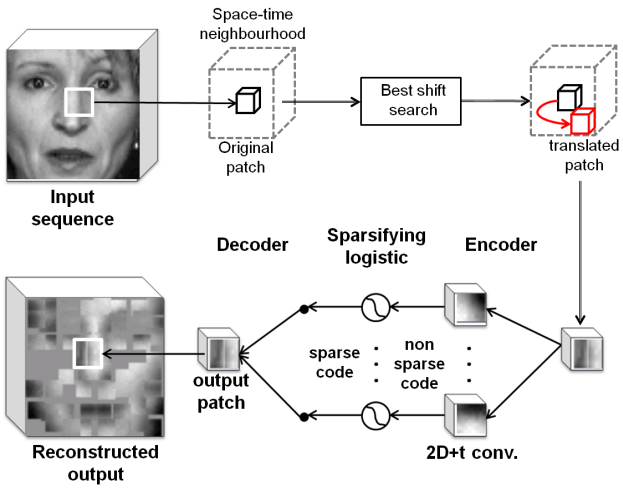


Figure 1: Overview of the proposed spatio-temporal convolutional sparse auto-encoder: Illustration on a sample from the GEMEP-FERA facial expressions dataset.

For facial expression recognition, we used the recent GEMEP-FERA dataset [4]. Obtained results are superior to the state of the art (87.57% for the overall classification rate), with a significative performance improvement particlularly for the person-independent configuration (with a recognition rate of 80.75%), which is a positive evidence of the high generalization of our approach, and eliminating the person-specific effect and capturing the facial expression salient information.

To conclude, we have proposed a neural model for video sequence classification, with a fully automated learning-based feature construction process. We have introduced the spatio-temporal convolutional sparse auto-encoder architecture, and its corresponding training procedure. We have also presented a novel approach for handling shift-invariance of the representation. Finally, we have shown how the temporal evolution of these features is used to classify the sequences, using a recurrent neural network model. Experimental results on two different problems confirms the high genericity of the model since it achieves the best results among related works. Future work will address scale invariance and applications to different problems.

[1] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[2] M.A. Ranzato, F.J. Huang, Y.L. Boureau, and Y. Lecun. Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[3] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *International Conference on Pattern Recognition*, volume 3, pages 32 – 36, 2004.

[4] M.F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *International Conference on Automatic Face & Gesture Recognition*, pages 921–926, 2011.

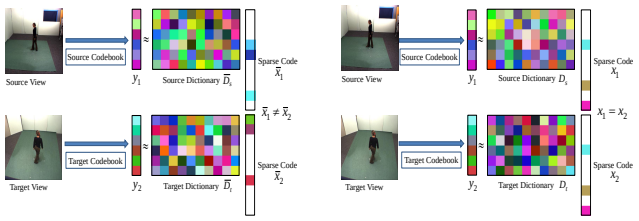
Cross-View Action Recognition via a Transferable Dictionary Pair

Jingjing Zheng[1]
zjngjng@umiacs.umd.edu
Zhuolin Jiang[2]
zhuolin@umiacs.umd.edu
P. Jonathon Phillips[3]
jonathon.phillips@nist.gov
Rama Chellappa[1]
rama@umiacs.umd.edu

[1] Department of Electrical and Computer Engineering and the Center for Automation Research, UMIACS
University of Maryland
College Park, MD, USA

[2] UMIACS, University of Maryland
College Park, MD, USA

[3] National Institute of Standards and Technology
Gaithersburg, MD, USA



(a) Independent dictionary pair (b) Transferable dictionary pair

Figure 1: **Independent dictionary pair versus Transferable dictionary pair.** (a) Based on the BoVW feature representation, the source and target dictionaries are learned individually using videos taken from two different views of the same action. (b) Based on the same BoVW feature representation, we simultaneously learn the source and target dictionaries by forcing the shared videos taken from two views to have the same sparse representations.

In this paper, we propose a novel approach for cross-view action recognition by transferring sparse feature representations of videos from the source to target view. The first step is to construct a separate codebook for each view, where the first view is the source domain and the second is the target domain. Each codebook is constructed by the k-means clustering algorithm. Each video is modeled as a Bag of Visual Words (BOVW) using the corresponding codebook from the same view. Although each pair of videos records the same action from two views, the feature representations of an action in the two views is different because each view has its own codebook. The next step is to learn a dictionary pair $\{D_s, D_t\}$, with D_s corresponding to the source view and D_t the target view. The dictionaries are designed to have sparse codes that are the same for each pair of videos that records the same action across the two views. In this way, videos across different views of the same action are encouraged to have similar sparse representations. This procedure enables the transfer of the sparse feature representations of videos in the source view to the corresponding videos in the target view. There is no reason to assume that two separate dictionaries that are learned independently for each view will have a view-invariant feature representation. The difference between learning a dictionary pair individually and our transferable dictionary pair learning can be seen in Figure 1.

Furthermore, we consider two types of actions: *shared* actions, that are observed in both *source* and *target* views, and *orphan* actions that are observed only in the source view. Orphan action labels are available only in the source view. For the shared actions, we consider two scenarios: (1) shared actions in both views are not labeled; (2) shared actions in both views are labeled. We refer them as the unsupervised and supervised settings respectively and propose corresponding unsupervised and supervised approaches for learning the transferable dictionary pair. Note that under both settings only videos of shared actions across different views are used for learning the dictionary pair, which means that the dictionary pair is not affected by videos of orphan actions.

In the unsupervised setting, our goal is to transfer orphan action models from the source view to the target view. For this purpose, we construct a transferable dictionary pair denoted by $\{D_s, D_t\}$, such that each pair of videos of the same action taken from the source and target views have the same sparse representations. Let $Y_s, Y_t \in \mathbb{R}^{n\times M}$ denote the feature representations of M videos of shared actions in source and target views. The objective function for learning a transferrable dictionary pair is given by:

$$arg \min_{D_s,D_t,X} ||Y_s - D_s X||_2^2 + ||Y_t - D_t X||_2^2 \quad \text{s.t.} \quad \forall i, ||x_i||_0 \leq s. \qquad (1)$$

In supervised setting where action categories of shared action videos are available in both views, we leverage this category information to learn a discriminative transferrable dictionary pair. Here the key idea is to partition the total dictionary items into disjoint subsets and each subset is responsible for representing videos of one action. The intuition behind this idea is that action videos from the same class tend to have same features and each action video could be well represented by other videos from the same class. On the contrary, videos from different classes tend to have different features and thus should be well represented by disjoint subsets of other videos. In order to achieve the above goal, we incorporate a label consistent regularization term introduced in [1] to the objective function in Eq. 1. Now the objective function for dictionary pair construction is given by:

$$arg \min_{D_s,D_t,A,X} ||Y_s - D_s X||_2^2 + ||Y_t - D_t X||_2^2 + \lambda ||Q - AX||_2^2 \quad \text{s.t.} \quad \forall i, ||x_i||_0 \leq \qquad (2)$$

where λ controls the tradeoff between the reconstruction error and label consistent regularization. The elements of matrix $Q = [q_1, ..., q_N] \in \mathbb{R}^{K\times N}$ are made of the ideal "discriminative" sparse codes of shared action videos in both views. The vector $q_i = [q_i^1, ..., q_i^K] = [0...1, 1, ..0] \in \mathbb{R}^K$ is a discriminative sparse code corresponding to one shared action video pair $\{y_{s,i}, y_{t,i}\}$ and the non-zeros values of q_i occur at those indices where the shared action video pair $\{y_{s,i}, y_{t,i}\}$ and the dictionary item d_k share the same label. Thus matrix A is a linear transformation matrix which transforms the original sparse code X to be most discriminative in sparse feature space $\mathbb{R}^K$.

In order to handle the situation where videos of shared actions across multiple source views are available, we propose to learn a set of view-dependent dictionaries by forcing videos of shared actions in all views to have the same representations when encoded using the corresponding view-dependent dictionary. Suppose there are p source views $\mathcal{V}^s$ and one target view $\mathcal{V}^t$, the corresponding objective function is given by:

$$arg \min_{\{D_{s,i}\}_{i=1}^p, D_t, X} \sum_{i=1}^{p} ||Y_{s,i} - D_{s,i} X||_2^2 + ||Y_t - D_t X||_2^2 \quad \text{s.t.} \quad \forall i, ||x_i||_0 \leq s. \qquad (3)$$

Given the learned view specific dictionaries, we obtain the sparse representation of each video in each view using the corresponding view-dependent dictionary. Videos of orphan actions in different views will have similar sparse representations when encoded using the corresponding view-dependent dictionary. This is because dictionaries are learned by forcing different sets of videos of shared actions in different views to have the same sparse representations. Thus, the action model learned in one view can be directly applied to classify unlabeled test videos in another different view.

We have extensively tested our approach on the publicly available IXMAS multi-view dataset [2]. The resulting performance clearly confirms the effectiveness of our approach for cross-view action recognition.

[1] Zhuolin Jiang, Zhe Lin, and Larry S. Davis. Learning a discriminative dictionary for sparse coding via label consistent K-SVD. In *CVPR*, 2011.

[2] Daniel Weinland, Edmond Boyer, and Rémi Ronfard. Action recognition from arbitrary views using 3D exemplars. In *ICCV*, 2007.

A Videography Analysis Framework for Video Retrieval and Summarization

Kang Li*[1]
kangli@buffalo.edu
Sangmin Oh*[2]
sangmin.oh@kitware.com
A. G. Amitha Perera[2]
amitha.perera@kitware.com
Yun Fu[3]
raymondyunfu@gmail.com

[1] Department of CSE
State University of New York
Buffalo, NY, USA

[2] Kitware, Inc.
Clifton Park, NY, USA

[3] Department of ECE and College of CIS
Northeastern University
Boston, MA, USA

Overview: In this work, we focus on developing features and approaches to represent and analyze videography styles in unconstrained videos. By unconstrained videos, we mean typical consumer videos with significant content complexity and diverse editing artifacts, mostly with long duration. We present an approach for *unsupervised videography analysis* for unconstrained videos. Intuitively, each videography can be understood as a camera director's direction on a movie script, *e.g.*, "capture the running actress by panning the camera, to have her face appear at 20 percent size of the video". The idea is that different classes of video content will have different styles—the videography style of a wedding video should be different from a sports video—and so, the videography style should provide a valuable signal for automated content analysis.

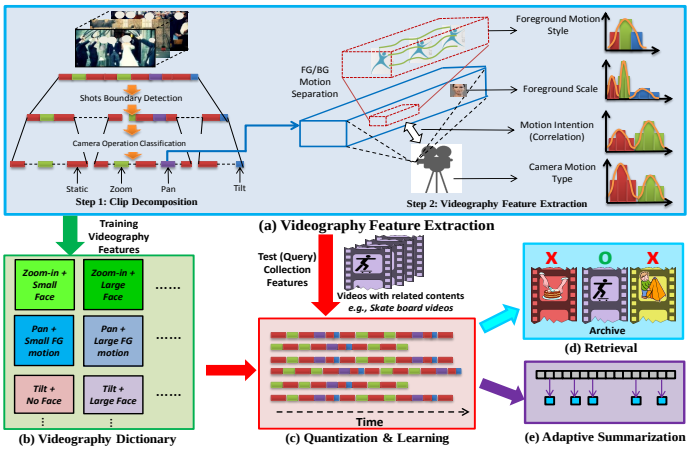


Figure 1: Framework for videography analysis and applications.

Videography Analysis: The overall framework of our approach is illustrated in Fig. 1(a). First, a two-level motion analysis is conducted to decompose long clips into sequences of segments with coherent motion types (S/P/T/Z). Second, multiple features related to motion and scale patterns are measured from every segment, which are used to characterize videography. Throughout this work, we utilize densely computed KLT tracks over the entire clips as main basis for the derived features.

We assume that there are diverse videography styles in unconstrained videos, which are discovered as a *videography dictionary* via unsupervised clustering on proposed features. Then, a video clip can be represented as a series of segments with varying videography words. For the underlying videography features, we extend conventional features such as camera motion and foreground (FG) object motion (e.g., [1]) by incorporating two novel features: *motion correlation* and *scale* information.

Once videography features are obtained from segments, they are used to build *videography dictionary* (VD) shown in Fig. 1(b). The computed VD will be used to quantize video clips into sequences of videography words (VWs), as shown in Fig. 1(c). Our analysis shows that there are regularized patterns in the videography used in the unconstrained Internet videos, and correlations between the exhibited videography styles and video contents. Such observation on discriminative correlations suggests that videography analysis can actually be used for challenging tasks such as content-based retrieval and content summarization.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

* Indicates equal contributions.

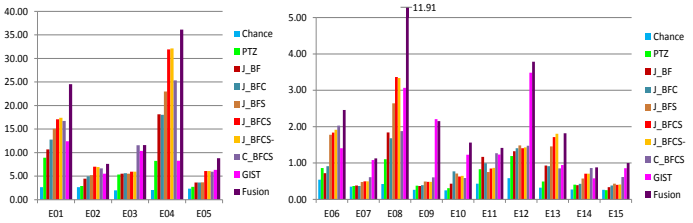


Figure 2: Average Precision (%) of video retrieval results on MED corpus, for 15 events: ***(E01) Board trick***, *(E02) Feeding animal, (E03) Fishing,* ***(E04) Wedding***, *(E05) Working wood project,* ***(E06) Birthday party***, *(E07) Change vehicle tire,* ***(E08) Flash mob***, *(E09) Getting vehicle unstuck, (E10) Groom animal, (E11) Make sandwich,* ***(E12) Parade***, ***(E13) Parkour***, *(E14), Repair appliance, and (E15) Sewing project.*

Video Retrieval: For retrieval, we computed bag-of-word representations based on the videography word sequences and employed them as the basis for content-based video retrieval tasks. We have conducted experiments on a large TRECVID '11 MED dataset where we tried diverse variations of the proposed approach as well as using more conventional features such as GIST. Our results indicate that the proposed videography features effectively improve the retrieval performance and are complimentary to traditional appearance features such as GIST, improving performance further when both features are used jointly. Figure 2 shows the list of video event classes and the extent of conducted retrieval experiments as well as summarized performance profiles. Event classes that show the most benefits by videography-based analysis are marked in bold.

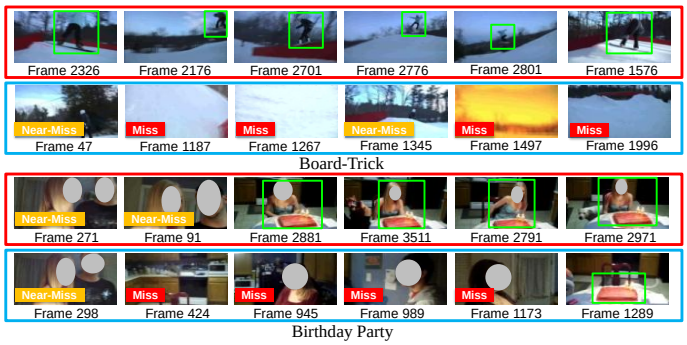


Figure 3: Videography-aware adaptive summarization. Three summarization results by this work (red rows) and baseline (blue rows). Detected FG regions (green) and human judgements on relevance of key frames (good:none, near-miss: yellow, miss: red) are marked on each image.

Video Summarization: We also show that the proposed videography analysis can be used to provide videography-aware adaptive summarization method. For example, Fig. 3 shows example summarization results for different events where the segments with distinctive videography styles for particular events are highlighted in the summaries, e.g., board tricks during snowboarding and candle blowing during a birthday party. Summarization produced by our proposed approach is shown in red and results by baseline approaches of using color histogram changes are shown in blue.

[1] Xingquan Zhu, Ahmed K. Elmagarmid, Xiangyang Xue, Lide Wu, and Ann Christine Catlin. InsightVideo: Towards hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, 7(4):648–666, 2005.

Scene Text Recognition using Higher Order Language Priors

Anand Mishra[1]
http://researchweb.iiit.ac.in/~anand.mishra/
Karteek Alahari[2]
http://www.di.ens.fr/~alahari/
C.V. Jawahar[1]
http://www.iiit.ac.in/~jawahar/

[1] CVIT
IIIT Hyderabad
Hyderabad, India
[2] INRIA - WILLOW
ENS
Paris, France

The problem of recognizing text in images taken in the wild has gained significant attention from the computer vision community in recent years. The scene text recognition task is more challenging compared to the traditional problem of recognizing text in printed documents. We focus on this problem, and recognize text extracted from natural scene images and the web. Significant attempts have been made to address this problem in the recent past, for example [1, 2]. However, many of these works benefit from the availability of strong context, which naturally limits their applicability. In this work, we present a framework to overcome these restrictions. Our model introduces a higher order prior computed from an English dictionary to recognize a word, which may or may not be a part of the dictionary. We present experimental analysis on standard as well as new benchmark datasets.

The main contributions of this work are: (1) We present a framework, which incorporates higher order statistical language models to recognize words in an unconstrained manner, *i.e.* we overcome the need for restricted word lists. (2) We achieve significant improvement (more than 20%) in word recognition accuracies in a general setting. (3) We introduce a large word recognition dataset (atleast 5 times larger than other public datasets) with character level annotation and benchmark it.

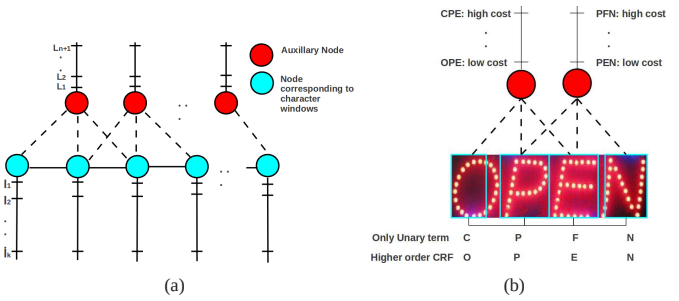
Method Overview. We propose a CRF based model for recognizing words. The CRF is defined over a set of random variables $x = \{x_i | i \in V\}$, where $V = \{1, 2, ..., n\}$. Each random variable x_i denotes a potential character in the word, and can take a label from the label set, $L = \{l_1, ..., l_k\} \cup \varepsilon$. The label set L is the set of English characters and digits, and a null label (ε) to suppress weak detections, similar to [1]. The most likely word represented by the set of characters x_i is found by minimizing the energy function, $E : L^n \rightarrow \mathbb{R}$, corresponding to the random field. The energy function $E(\cdot)$ can be typically written as sum of potential functions:

$$E(x) = \sum_{c \in \mathcal{C}} \psi_c(x_c), \quad (1)$$

where $\mathcal{C}$ represents a set of subsets of V, *i.e.* cliques, and x_c is the set of random variables included in a clique c. The set of potential characters is obtained by a sliding window based character detection step. The neighbourhood relations among characters, which determine the structure of the random field, are based on the spatial arrangement of characters in the word image. The character detection step provides us with a large set of windows potentially containing characters within them. Our goal is to infer the most likely word from this set of characters. We formulate this problem as that of minimizing the energy in (1), where the best energy solution represents the ground truth word we aim to find.

The energy function (1) is composed of unary, pairwise and higher order terms. The unary and pairwise terms are computed as described in [1]. For introducing higher order, we add an auxiliary variable x_c^a for every clique $c \in \mathcal{C}$. This auxiliary variable takes a label from the label set L_e. In our case the extended label set L_e, for a CRF of order h, contains all possible h-gram combinations present in the lexicons and one additional label (to account for h-grams that do not occur). We define a very high cost for an auxiliary variable to take a label which is not present in the dictionary. Increasing the order of the CRF allows us to capture a larger context. An illustration of our model is shown in Figure 1.

Results and Discussions. Our method outperforms [1] not only on the (smaller) SVT and ICDAR 2003 datasets, but also on the IIIT 5K-Word dataset[1]. We compare the word recognition performance of our method with pairwise CRF in Table 1. We achieve a significant improvement



(a) (b)

Figure 1: *The proposed graphical model (a model of order 3 is shown here) and an example word image to illustrate its use: Tri-grams like OPE, PEN have very high frequency in an English dictionary (> 1500), and thus are assigned a low cost, whereas unlikely tri-grams, such as CPE and PFN, are assigned a high cost.*

Method	SVT-WORD	ICDAR	IIIT 5K-word
Pairwise CRF [1]	23.49	45	20.25
Proposed Higher Order	**49.46**	**57.92**	**44.30**

Table 1: Word recognition accuracy without using an image specific small word list. Lexicon priors are computed from a large size lexicon with 0.5 million words.

Figure 2: *A few sample images from the* IIIT *5K-word dataset where our method is successful. We see that the dataset contains images with variations in font, style, background, orientation etc.*

of around 25%, 12% and 22% on SVT, ICDAR 2003 and IIIT 5K-word datasets respectively. We also show few sample images from the 5K-word dataset in Figure 2.

Our method differs from other related approaches, such as [1], as detailed below. We address a more general problem of scene text recognition, *i.e.* recognizing a word without relying on a small size lexicon. Note that recent works [1, 2, 3] on scene text recognition, recognize a word with the help of an image-specific small size lexicon, of about 50 words per image. Our method computes the prior from an English dictionary and by-passes the use of edit distance based measures. In fact, we also recognize words missing from the given dictionary. One of the main reasons for the improvements we achieve is the use of n-grams extracted from the dictionary.

In summary, we proposed a powerful method to recognize scene text. The proposed CRF model infers the location of true characters, as well as the word as a whole. We evaluated our method on publicly available datasets and a large dataset introduced by us.

[1] A. Mishra, K. Alahari, and C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. In *CVPR*, 2012.
[2] K. Wang and S. Belongie. Word spotting in the wild. In *ECCV*, 2010.
[3] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *ICCV*, 2011.

[1] Our new dataset available at: http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/

Data-Driven Scene Understanding from 3D Models

Scott Satkin
ssatkin@ri.cmu.edu
Jason Lin
jasonli1@andrew.cmu.edu
Martial Hebert
hebert@ri.cmu.edu

Carnegie Mellon University
The Robotics Institute
Pittsburgh, Pennsylvania

http://cmu.satkin.com/bmvc2012/

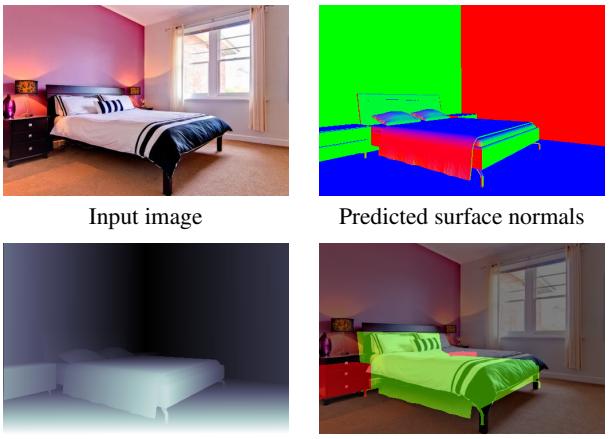


Figure 1: From a single image, we estimate detailed scene geometry and object labels.

In this paper, we propose a data-driven approach to leverage repositories of 3D models for scene understanding. Our ability to relate what we see in an image to a large collection of 3D models allows us to transfer information from these models, creating a rich understanding of the scene. We develop a framework for auto-calibrating a camera, rendering 3D models from the viewpoint an image was taken, and computing a similarity measure between each 3D model and an input image. We demonstrate this data-driven approach in the context of geometry estimation and show the ability to find the identities and poses of object in a scene. Additionally, we present a new dataset with annotated scene geometry. This data allows us to measure the performance of our algorithm in 3D, rather than in the image plane.

Recently, large online repositories of 3D data such as Google 3D Warehouse have emerged. These resources, as well as the advent of low-cost depth cameras, have sparked interest in geometric data-driven algorithms. At the same time, researchers have (re-)started investigating the feasibility of recovering geometric information, *e.g.*, the layout of a scene. The success of data-driven techniques for tasks based on appearance features, *e.g.*, interpreting an input image by retrieving similar scenes, suggests that similar techniques based on *geometric* data could be equally effective for 3D scene interpretation tasks. In fact, the motivation for data-driven techniques is the same for 3D models as for images: real-world environments are not random; the sizes, shapes, orientations, locations and co-location of objects are constrained in complicated ways that can be represented given enough data. In principle, estimating 3D scene structure from data would help constrain bottom-up vision processes. For example, in Figure 1, one nightstand is fully visible; however, the second nightstand is almost fully occluded. Although a bottom-up detector would likely fail to identify the second nightstand since only a few pixels are visible, our method of finding the best matching 3D model is able to detect these types of occluded objects. This is not a trivial extension of the image-based techniques. Generalizing data-driven ideas raises new fundamental technical questions never addressed before in this context: What features should be used to compare input images and 3D models? Given these features, what mechanism should be used to rank the most similar 3D models to the input scene? Even assuming that this ranking is correct, how can we transfer information from the 3D models to the input image? To address these questions, we develop a set of features that can be used to compare an input image with a 3D model and design a mechanism for finding the best matching 3D scene using support vector ranking. We show the feasibility of these techniques for transferring the geometry of objects in indoor scenes from 3D models to an input image.

Naturally, we cannot compare 3D models directly to a 2D image. Thus, we first estimate the intrinsic and extrinsic parameters of the camera and use this information to render each of the 3D models from the same view as the image was taken from. We then compute similarity features between the models and the input image. Lastly, each of the 3D models is ranked based on how similar its rendering is to the input image using a learned feature weighting. See Figure 2 for an overview of this process. Please read our full paper for a detailed explaination of our data-driven geometry estimation algorithm and results.

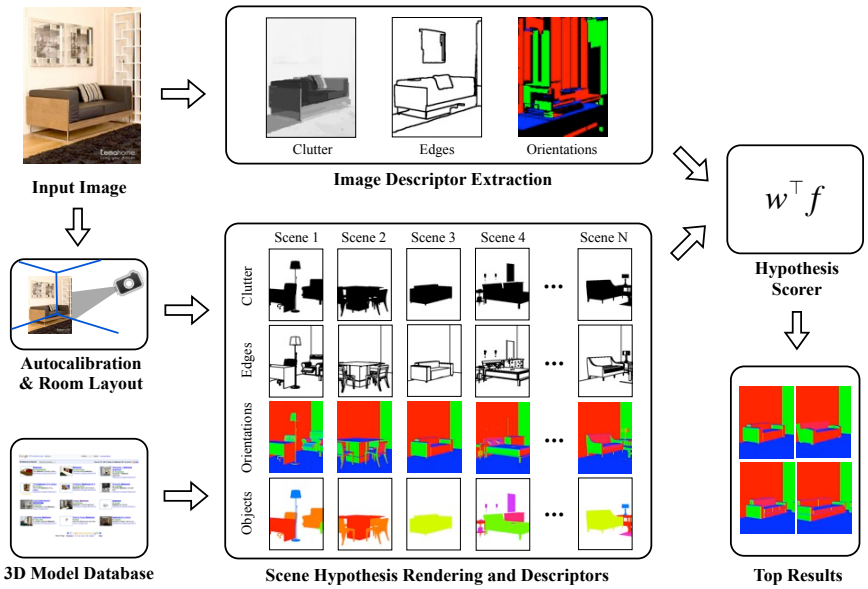


Figure 2: Overview of our approach for matching a 3D model with a monocular image.

Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification

Thomas Berg
tberg@cs.columbia.edu
Peter N. Belhumeur
belhumeur@cs.columbia.edu
Columbia University
New York, NY

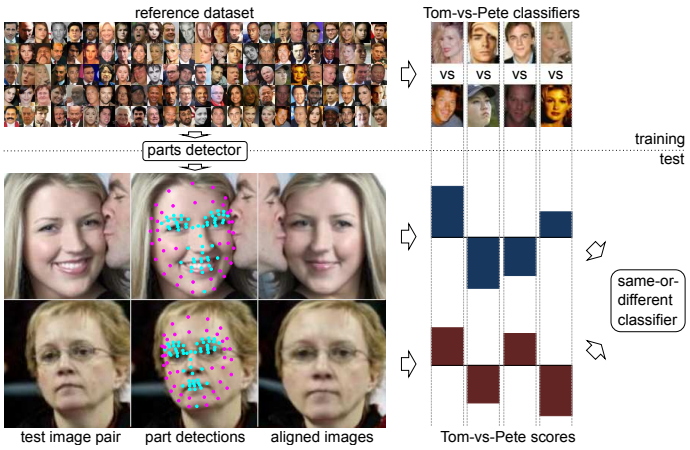


Figure 1: Overview. A reference set of images labeled with parts and identities is used to train a parts detector and a large number of binary "Tom-vs-Pete" classifiers. Given a test image pair, parts are detected and used to perform an "identity-preserving" alignment. The Tom-vs-Pete classifiers are run on the aligned images, and the results are passed to a same-or-different classifier to produce a decision.

In face verification, we are given two face images and must determine whether they are the same person. In this paper, we present a method for face verification that uses a reference dataset of images of other (non-test) people in two novel ways. First, we use part labels on the faces in the training set to perform an *identity-preserving alignment* that reduces differences due to pose and expression, but preserves identity-related differences such as nose width and lip thickness. Second, we train a large set of linear *Tom-vs-Pete classifiers* that are likely to be able to find differences between almost any two people. The outputs of these first-stage classifiers are used as features for a second-stage same-vs-different classifier that makes the verification decision. An overview of the process is shown in Figure 1.

The reference dataset consists of 20,639 images of 120 people, downloaded from the internet. Half the images are from the PubFig [3] "development set," with the rest collected online. In addition to the identity labels, each face is labeled with the locations of 95 parts, including 55 "inner" parts at well-defined points such as the corners of the eyes and mouth, and 40 less well-defined "outer" points around the boundary of the face.

Identity-preserving Alignment

Our alignment procedure is based on a set of part locations on the face. Given a test image, we first find the fifty-five inner parts using the detector of [1]. We then look to the reference dataset, and for each of the 120 reference people, find the image whose inner part locations are closest to the detected parts. This gives a set of 120 images of different people with nearly the same pose and expression. We take the avearage of all ninety-five part locations on these images to be the "generic part" locations for this image – where the parts would be on an average face with the pose and expression of the test face.

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

Figure 2: Alignment. (a) Original images. (b) Global affine alignment. (c) Alignment using detected parts. (d) Alignment using generic parts.

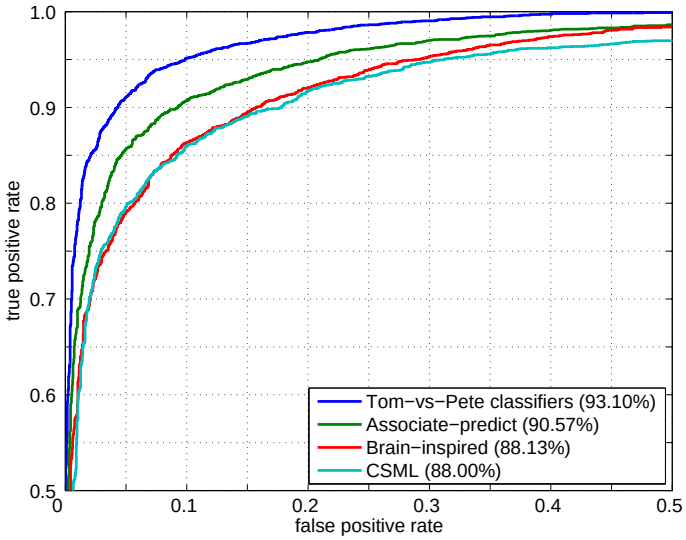


Figure 3: Our results on LFW, compared with the best previously published results [4, 5, 6].

Each part has a canonical location, where it occurs in an average, frontal face with neutral expression. To align the image, we perform a piecewise affine warp that takes the generic parts to the canonical part locations. While this sort of alignment using the original detected points can produce a "too-aligned" image in which important, identity-corelated differences like nose width have been removed, we demonstrate in the paper that using the generic parts preserves these differences. This effect is visible in Figure 2, where (c) appears anonymized relative to (b) or (d).

Tom-vs-Pete Classifiers and Verification

Each first-stage classifier is a linear SVM trained on SIFT features from some small region of the face to separate two people. We define eleven such regions, allowing us to train $11 \cdot \binom{120}{2} = 78{,}540$ classifiers. We call them "Tom-vs-Pete" classifiers to emphasize that each is trained on just two individuals. Intuitively, these classifiers represent a large number of ways in which two people can differ. We describe an adaboost-based heuristic to find a subset of these classifiers that will complement each other and will generalize well to other people, and use this subset to extract features for the second-stage classifier as shown in Figure 1.

We evaluate our system on the Labeled Faces in the Wild (LFW) [2], image-restricted benchmark, obtaining a mean accuracy of $93.10\% \pm 1.35\%$, a 26.86% reduction in error rate relative to the best previously published result [6]. Results are shown in Figure 3.

[1] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[2] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, UMass-Amherst, 2007.

[3] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[4] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *ACCV*, 2011.

[5] N. Pinto and D. D. Cox. Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition. In *Face and Gesture (FG)*, 2011.

[6] Q. Yin, X. Tang, and J. Sun. An associate-predict model for face recognition. In *CVPR*, 2011.

Let the Shape Speak - Discriminative Face Alignment using Conjugate Priors

Pedro Martins
http://www.isr.uc.pt/~pedromartins
Rui Caseiro
http://www.isr.uc.pt/~ruicaseiro
João F. Henriques
http://www.isr.uc.pt/~henriques
Jorge Batista
http://www.isr.uc.pt/~batista

Institute of Systems and Robotics,
University of Coimbra
Portugal

This work presents a novel Bayesian formulation for aligning faces in unseen images. Our approach is closely related to Constrained Local Models (CLM) [2] and Active Shape Models (ASM) [6], where an ensemble of local feature detectors are constrained to lie within the subspace spanned by a Point Distribution Model (PDM).

Fitting a model to an image typically involves two steps: a local search using a detector, obtaining response maps for each landmark (likelihood term) and a global optimization that finds the PDM parameters that jointly maximize all the detections. The global optimization can be seen as a Bayesian inference problem, where the posterior distribution of the PDM parameters (and pose) can be inferred in a *maximum a posteriori* (MAP) sense. We present a novel Bayesian global optimization strategy, where the prior is used to encode the dynamic transitions of the PDM parameters. Using recursive Bayesian estimation we model the prior distribution of the data as being Gaussian. The mean and covariance were assumed to be unknown and treated as random variables.

The Shape Model: The shape of a PDM is represented by the 2D locations of a mesh $\mathbf{s} = (x_1, y_1, \ldots, x_v, y_v)^T$ (v landmarks). Applying PCA on training examples, results in the parametric model $\mathbf{s} = \mathbf{s}_0 + \Phi\mathbf{b} + \Psi\mathbf{q}$, where $\mathbf{s}_0$ is the mean shape, Φ is the shape subspace matrix (n eigenvectors), $\mathbf{b}$ is a vector of shape parameters, $\mathbf{q}$ the pose parameters vector and Ψ holds four special eigenvectors that linearly model the 2D pose [4].

Goal: Given a $2v$ vector of observed positions $\mathbf{y}$, the goal is to find the optimal set of parameters $\mathbf{b}$ that maximizes the posterior probability of being aligned. Using an Bayesian approach, the shape parameters are

$$p(\mathbf{b}|\mathbf{y}) \propto \left(\prod_{i=1}^{v} p(\mathbf{y}_i|\mathbf{b}) \right) p(\mathbf{b}|\mathbf{b}_{k-1}) \tag{1}$$

where $\mathbf{y}_i$ is the i^{th} landmark coordinates and $\mathbf{b}_{k-1}$ is the previous optimal estimate of $\mathbf{b}$. The prior encodes how the shape/pose parameters change.

The Likelihood Term: is the following convex energy function:

$$p(\mathbf{y}|\mathbf{b}_k) \propto \exp\left(-\frac{1}{2} (\underbrace{\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b})}_{\Delta\mathbf{y}})^T \Sigma_{\mathbf{y}}^{-1} (\mathbf{y} - (\mathbf{s}_0 + \Phi\mathbf{b})) \right) \tag{2}$$

where $\Delta\mathbf{y}$ is the difference between the observed and the mean shape and $\Sigma_{\mathbf{y}}$ is the uncertainty of the spatial localization of the landmarks ($2v \times 2v$ block diagonal covariance matrix).

The response maps can be nonparametrically approximated by using a Kernel Density Estimator (KDE) [5]. Maximizing over the KDE is typically performed by the mean-shift algorithm. Let $\mathbf{z}_i = (x_i, y_i)$ be a candidate to the i^{th} landmark, being $\mathbf{y}_i^c$ the current landmark estimate, $\Omega_{\mathbf{y}_i^c}$ a $L \times L$ patch centered at $\mathbf{y}_i^c$, $\mathbf{I}$ the target image and $p_i(\mathbf{z}_i)$ the probability $\mathbf{z}_i$ is aligned. The i^{th} mean-shift landmark update and its uncertainty are

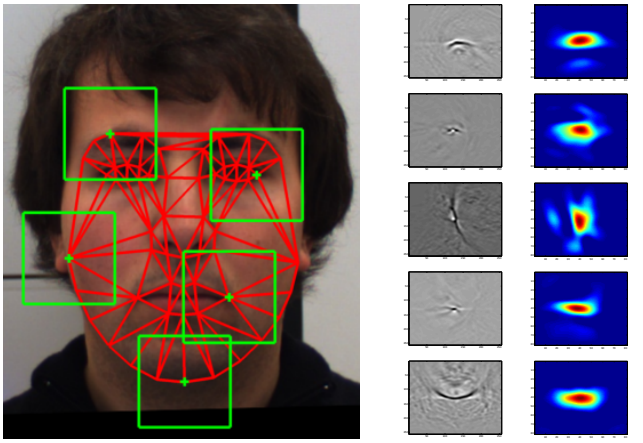
$$\mathbf{y}_i^{\mathrm{KDE}(\tau+1)} \leftarrow \frac{\sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} \mathbf{z}_i \; p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}{\sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i) \; \mathcal{N}(\mathbf{y}_i^{\mathrm{KDE}(\tau)} | \mathbf{z}_i, \sigma_{h_j}^2 \mathbf{I}_2)}, \tag{3}$$

$$\Sigma_{\mathbf{y}_i}^{\mathrm{KDE}} = \frac{1}{d-1} \sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i)(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})(\mathbf{z}_i - \mathbf{y}_i^{\mathrm{KDE}})^T, \quad d = \sum_{\mathbf{z}_i \in \Omega_{\mathbf{y}_i^c}} p_i(\mathbf{z}_i), \tag{4}$$

with $\mathbf{I}_2$ a 2D identity matrix and $\sigma_{h_j}^2$ the decreasing bandwidth.

The Prior Term: $p(\mathbf{b}_k|\mathbf{b}_{k-1}) \propto \mathcal{N}(\mathbf{b}_k|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$ follows a Gaussian distribution. Mean $\mu_{\mathbf{b}}$ and covariance $\Sigma_{\mathbf{b}}$ of the data are assumed to be unknown and modeled as random variables [1]. Recursive Bayesian estimation can be applied to infer the parameters of the prior distribution. Defining $\mathbf{b}$ as an observable vector, the joint posterior can be written as

$$p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b}) \propto p(\mathbf{b}|\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}) p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}). \tag{5}$$



(a) Local search regions. (b) Detectors [3] (c) Responses $p_i(\mathbf{z}_i)$

Figure 1: The Bayesian global optimization strategy jointly combines all detectors scores (MAP sense), explicitly modelling the prior distribution.

The joint prior $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}})$ follows a normal-inverse Wishart distribution, assuming $p(\mu_{\mathbf{b}}|\Sigma_{\mathbf{b}})$ a Gaussian (the conjugate prior for a Gaussian with known mean is an inverse Wishart). The joint posterior density $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b})$ follows an normal-inverse Wishart distribution with hyperparameters [1]:

$$\upsilon_k = \upsilon_{k-1} + m, \quad \kappa_k = \kappa_{k-1} + m \tag{6}$$

$$\theta_k = \frac{\kappa_{k-1}}{\kappa_{k-1} + m} \theta_{k-1} + \frac{m}{\kappa_{k-1} + m} \overline{\mathbf{b}} \tag{7}$$

$$\Lambda_k = \Lambda_{k-1} + \frac{\kappa_{k-1} m}{\kappa_{k-1} + m} (\overline{\mathbf{b}} - \theta_{k-1})(\overline{\mathbf{b}} - \theta_{k-1})^T \tag{8}$$

where θ_{k-1} is the prior mean, κ_{k-1} is the number of prior measurements, $\overline{\mathbf{b}}$ the mean of the new samples, m number of samples, υ_{k-1} and Λ_{k-1} are the degrees of freedom and scale matrix for the inv-Wishart distribution.

Marginalizing $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b})$ with respect to $\Sigma_{\mathbf{b}}$ gives the marginal posterior distribution for the mean $p(\mu_{\mathbf{b}}|\mathbf{b})$, that follows a multivariate Student-t distribution. Using the expectation of $p(\mu_{\mathbf{b}}|\mathbf{b})$ as the update at instance k we get $\mu_{\mathbf{b}_k} = E(\mu_{\mathbf{b}}|\mathbf{b}) = \theta_k$. Similarly, marginalizing $p(\mu_{\mathbf{b}}, \Sigma_{\mathbf{b}}|\mathbf{b})$ with respect to $\mu_{\mathbf{b}}$ gives $p(\Sigma_{\mathbf{b}}|\mathbf{b})$ that follows an inverse Wishart distribution. By the expectation of $p(\Sigma_{\mathbf{b}}|\mathbf{b})$ we get $\Sigma_{\mathbf{b}_k} = E(\Sigma_{\mathbf{b}}|\mathbf{b}) = (\upsilon_k - n - 1)^{-1} \Lambda_k$.

Global Alignment (MAP): The recursive posterior distribution is Gaussian, and takes the form of $p(\mathbf{b}_k|\mathbf{y}_k, \ldots, \mathbf{y}_0) \propto \mathcal{N}(\mathbf{b}_k|\mu_k, \Sigma_k)$ with

$$\Sigma_k = \left((\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1} + \Phi^T \sum_{m=1}^{M} \left(\Sigma_{\mathbf{y}_{(m)}}^{-1} \right) \Phi \right)^{-1} \tag{9}$$

$$\mu_k = \Sigma_k \left(\Phi^T \sum_{m=1}^{M} \left(\Sigma_{\mathbf{y}_{(m)}}^{-1} \Delta\mathbf{y}_{(m)} \right) + (\Sigma_{\mathbf{b}_k} + \Sigma_{k-1})^{-1} \mu_{\mathbf{b}_k} \right) \tag{10}$$

where $\Delta\mathbf{y}_{(m)}$, $\Sigma_{\mathbf{y}_{(m)}}$ are the multiple likelihood observations.

[1] *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edition, 2004.

[2] D.Cristinacce and T.F.Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[3] D.S.Bolme, J.R.Beveridge, B.A.Draper, and Y.M.Lui. Visual object tracking using adaptive correlation filters. In *IEEE CVPR*, 2010.

[4] I.Matthews and S.Baker. Active appearance models revisited. *IJCV*, 60(1): 135–164, 2004.

[5] J.Saragih, S.Lucey, and J.Cohn. Face alignment through subspace constrained mean-shifts. In *IEEE ICCV*, 2009.

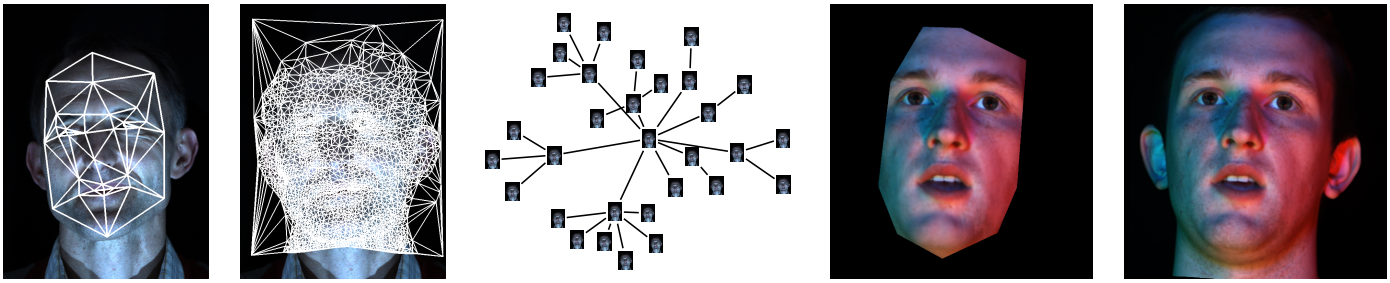
[6] T.F.Cootes, C.J.Taylor, D.H.Cooper, and J.Graham. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995.

Dense Active Appearance Models Using a Bounded Diameter Minimum Spanning Tree

Robert Anderson[1]
ra312@cam.ac.uk
Bjorn Stenger[2]
bjorn.stenger@crl.toshiba.co.uk
Roberto Cipolla[1]
cipolla@eng.cam.ac.uk

[1] Department of Engineering
Cambridge University
Cambridge, UK

[2] Cambridge Research Laboratory
Toshiba Research Europe
Cambridge, UK



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

Figure 1: A sparse AAM (a), containing 37 vertices, is automatically refined to give a much denser AAM (b), containing 1000 vertices, by bringing all training images into joint alignment using a Bounded Diameter Minimum Spanning Tree (c). Synthesis of novel sequences using a sparse AAM (d) produces a blurred output that does not capture the same details as a dense AAM (e).

Active Appearance Models (AAMs) can provide an efficient method for synthesising novel video sequences. Currently AAMs are built using only a small number of vertices (<100) due to the time required to hand label training images, fig.1(a). This small number of vertices means that fine detail is not captured by the model as non-rigid deformation of the target occurs at a smaller scale than the model's mesh. In order to produce high quality synthesis the model must be densified by adding a much larger number of vertices, fig.1(b).

Previous attempts at automatic construction of AAMs [1, 7] have been encouraging but have not been demonstrated to capture the fine detail necessary for high resolution models. The key problem in this task is one of joint image alignment, as all training images must be brought into a dense correspondence. State-of-the-art results on this problem are currently achieved by methods such as that of Cootes *et al*. [2], however registering to a mean image results in some features not being correctly aligned. We propose a registration method which is based on a bounded diameter minimum spanning tree (BDMST), fig.1(c). In a similar way to other tree-based registration methods [3, 5] our approach aims to make each pairwise registration as accurate as possible by only aligning between similar images.

Our densification process consists of the three key components outlined below;

1. a joint alignment approach using a bounded diameter minimum spanning tree,
2. a method of optical flow refinement suited to regions which contain fine texture, and
3. a method for densifying an AAM given dense correspondences between training images.

Registration through a BDMST: To register all training images to a common frame we build a BDMST with one node for each image and where edge weights are determined by the residual error between two images after pairwise alignment. Pairwise registration is calculated between each pair of connected images in the tree in the form of an optical flow field. Each image is then registered to the base image (the image at the root of the tree) by concatenating flow fields along the path from the image to the tree root. This ensures that each pairwise registration is made between similar images, increasing the quality of the registration. Experimentally we find that a suitable diameter for the tree is 4, which trades off the maximum length of path from any image to the base image against the amount of difference between images between which pairwise registrations are computed.

Optical flow computation: To calculate pairwise registrations between images in the tree we use optical flow. To find an approximate initial flow we use the implementation of Liu [6]. This fails to align some fine detail in the images and we propose a refinement step to correct for this. We assume that the initial registration is approximately correct and so limit the range of displacements in our refinement to ± 15 pixels. We minimise an energy function consisting of a photoconsistency term and a regularisation term. To find a globally optimal solution we use a Markov Random Field formulation and in order to keep the number of labels for each pixel in the MRF to a reasonable number (31, instead of 961) we optimise horizontal and vertical displacement separately.

Mesh densification: Given a set of registered images we wish to densify the original AAM by adding additional vertices. We wish to add vertices in such a way that we can model the observed deformations between each image and the base image using as few vertices as possible. To do so we follow a similar approach to that used in the construction of digital terrain models. Instead of trying to minimise the difference between a scalar field (height) and its approximation given by interpolating over a mesh we wish to minimise the difference between a vector field (the flow from each training image to the base image) and its approximation interpolating on the mesh. To achieve this we use a greedy point insertion algorithm based on that of DeFloriani [4].

Results are shown which demonstrate that dense AAMs built using the proposed approach have an improved texture model compactness over the original AAMs. We also demonstrate qualitatively the improvement in synthesis resulting from using dense AAMs over sparse ones (fig.1(e) versus fig.1(d)). More information can be found in the paper and accompanying video.

[1] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *PAMI*, 26 (10):1380–1384, 2004.

[2] T. Cootes, C. Twining, V. Petrović, K. Babalola, and C. Taylor. Computing accurate correspondences across groups of images. *PAMI*, 32 (11):1994–2005, 2010.

[3] D. Cristinacce and T. Cootes. Facial motion analysis using clustered shortest path tree registration. *MLVMA Workshop (ECCV)*, 2008.

[4] L. De Floriani. A pyramidal data structure for triangle-based surface description. *IEEE Comput. Graph. Appl.*, 9(2):67–78, 1989.

[5] J. Hamm, D. Hye Ye, R. Verma, and C. Davatzikos. Gram: A framework for geodesic registration on anatomical manifolds. *Medical Image Analysis*, 14(5):633–642, 2010.

[6] C. Liu. Beyond pixels: Exploring new representations and applications for motion analysis. *Doctoral Thesis, MIT*, 2009.

[7] K. Ramnath, S. Baker, I. Matthews, and D. Ramanan. Increasing the density of active appearance models. *CVPR*, 2008.

PMBP: PatchMatch Belief Propagation for Correspondence Field Estimation

Frederic Besse[1]
f.besse@cs.ucl.ac.uk
Carsten Rother and Andrew Fitzgibbon[2]
{carrot,awf}@microsoft.com
Jan Kautz[1]
j.kautz@cs.ucl.ac.uk

[1] University College London
London, UK
[2] Microsoft Research Cambridge
Cambridge, UK

This paper draws a new connection between two existing algorithms for estimation of correspondence fields between images: Belief Propagation [4] and PatchMatch [1]. Correspondence fields arise in problems such as dense stereo reconstruction, optical flow estimation, and a variety of computational photography applications such as recoloring, deblurring, high dynamic range imaging, and inpainting. By analysing the connection between the methods, we obtain a new algorithm which has performance superior to both its antecedents, and in the case of stereo matching, represents the current state of the art on the Middlebury benchmark at sub-pixel accuracy. The first contribution of our work is a detailed description of PatchMatch and belief propagation in terms that allow the connection between the two to be clearly described. Our second contribution is in the use of this analysis to define a new algorithm: PatchMatch Belief Propagation (PMBP) which, despite its relative simplicity, is more accurate than PatchMatch and orders of magnitude faster than PBP.

Belief propagation (BP) is a venerable approach to the analysis of correspondence problems. The correspondence field is parametrized by a vector grid $\{\mathbf{u}_s\}_{s=1}^n$, where s indexes *nodes*, typically corresponding to image pixels, and $\mathbf{u}_s \in \mathbb{R}^d$ parametrizes the correspondence vector at node s. We shall consider a special case of BP, viewed as an energy minimization algorithm where the energy combines *unary* and *pairwise* terms

$$E(\mathbf{u}_1,\ldots,\mathbf{u}_n) = \sum_{s=1}^{n} \psi_s(\mathbf{u}_s) + \sum_{s=1}^{n} \Big[\sum_{t\in N(s)} \psi_{st}(\mathbf{u}_s,\mathbf{u}_t) \Big], \qquad (1)$$

with $N(s)$ being the set of *pairwise neighbours* of node s. The unary energy $\psi_s(\mathbf{u}_s)$, also called the *data term*, computes the local evidence for the correspondence $\mathbf{u}_s$. On a continuous space, a natural representation using particles presents itself, closely related to Max Product Particle BP (PBP)[3]. With each node s, we associate a set of K *particles* $P_s \subset \mathbb{R}^d$, where each particle $p \in P_s$ is a candidate solution for the minimizing correspondence parameters $\mathbf{u}_s^*$. BP is a *message-passing* algorithm, where messages are defined as functions from nodes to their neighbours. Before defining the messages, which are themselves defined recursively, it is useful to define the *log disbelief* at node s as

$$B_s(\mathbf{u}_s) := \psi_s(\mathbf{u}_s) + \sum_{t\in N(s)} M_{t\to s}(\mathbf{u}_s), \qquad (2)$$

in terms of which the messages, using particles, are defined as

$$M_{t\to s}(\mathbf{u}_s) := \min_{\mathbf{u}_t \in P_t} \psi_{st}(\mathbf{u}_s,\mathbf{u}_t) + B_t(\mathbf{u}_t) - M_{s\to t}(\mathbf{u}_t). \qquad (3)$$

We note that this definition is in terms of a continuous $\mathbf{u}_s$, not restricted to the current particle set P_s, but the minimization over $\mathbf{u}_t$ is a discrete minimization over the particles P_t. At convergence, $\hat{\mathbf{u}}_s := \operatorname{argmin}_{\mathbf{u}} B_s(\mathbf{u})$ is the estimate of the minimizer.

The **PatchMatch** algorithm (PM) [1] was initially introduced as a computationally efficient way to compute a *nearest neighbour field* (NNF) between two images. In terms of energy minimization, the NNF is the global minimizer of an energy comprising unary terms only ($\psi_{st} = 0$). PM computes good minima while being very efficient. With such a powerful optimizer, more complex unary terms can be defined, yielding another class of state-of-the-art correspondence finders, exemplified by the recent introduction of PatchMatch Stereo [2]. Using the same particle notations as BP, the set P_s are initialized uniformly at random. One PM iteration then comprises a linear sweep through all nodes in an order defined by a *schedule function* $\phi(s)$, so that s is visited before s' if $\phi(s) < \phi(s')$. At node s, two update steps are performed: *propagation* and *resampling*:

- In the **propagation** step, the particle set is updated to contain the best K particles from the union of the current set and the set C_s of already-visited neighbour candidates, where "best" is defined as minimizing the unary cost $\psi_s(\cdot)$.
- The local **resampling** step (called "random search" in [1]) perturbs the particles locally according to a proposal distribution which we model as a Gaussian $\mathcal{N}(0,\sigma)$. The second step of the PM iteration updates P_S with any improved estimates from the local resampling set, for m resampling steps.

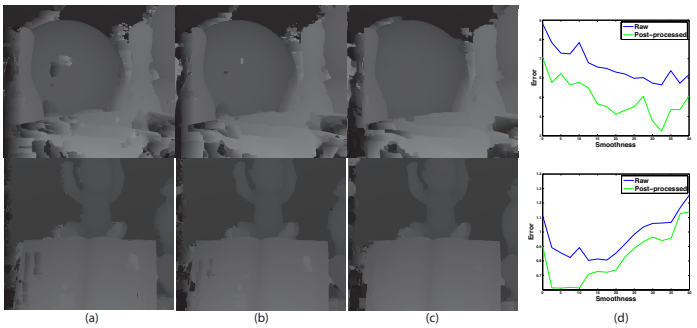


Figure 1: Evolution of the disparity map (before post-processing) with different weightings of the smoothness: (a) $\beta = 0$ (PatchMatch stereo). (b) $\beta = 5$. (c) $\beta = 17.5$. (d) Corresponding disparity error, for both raw and post-processed outputs.

After several alternating sweeps, the best particle in each set typically represents a good optimum of the unary-only energy.

PatchMatch Belief Propagation (PMBP) can be defined as a combination of the PM and PBP algorithms. We shall consider PBP our base, as the goal is to minimize a more realistic energy than PM, that is to say, an energy with pairwise terms encouraging piecewise smoothness.

First, PM resamples P_s from the neighbours of node s, while PBP's resampling is only via MCMC from the elements of P_s. The samples are evaluated using B_s, so this is a resampling of the particle set under the current belief, as proposed in PBP, but with a quite different source of particle proposals. Thus PMBP augments PBP with samples from the neighbours.

Second, PBP uses an MCMC framework where particles are replaced in P_s with probability given by the Metropolis acceptance ratio, while PM accepts only particles with higher belief than those already in P_s. This non-Metropolis replacement strategy further accelerates convergence, so it is included in PMBP.

Making these two modifications yields a powerful new optimization algorithm for energies with pairwise smoothness terms. In the case of a zero pairwise term $\psi_{st} = 0$, PMBP exactly yields PM. Conversely, running PMBP with a nonzero pairwise term is a strict generalization of PM, allowing the incorporation of an explicit smoothness control which directly addresses the deficiencies of PM while retaining its speed.

We apply our algorithm to the stereo matching case. The effect of adding a realistic pairwise term to the PatchMatch stereo algorithm under our PMBP framework can be seen in figure 1.

[1] C. Barnes, E. Shechtman, D. B. Goldman, and A. Finkelstein. The generalized PatchMatch correspondence algorithm. In *Proc. ECCV*, 2010.

[2] M. Bleyer, C. Rhemann, and C. Rother. PatchMatch Stereo—Stereo matching with slanted support windows. In *Proc. BMVC*, 2011.

[3] R. Kothapa, J. Pachecho, and E. B. Sudderth. Max-product particle belief propagation. Master's thesis, Brown University, 2011.

[4] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

Deformable 3D Reconstruction with an Object Database

Pablo F. Alcantarilla
pablofdezalc@gmail.com
Adrien Bartoli
adrien.bartoli@gmail.com

ISIT-UMR6284 CNRS
Université d'Auvergne
Clermont-Ferrand, France

Deformable 3D reconstruction from 2D images requires prior knowledge on the scene structure. Template-free methods [1, 2, 5, 6, 9, 14] use *generic* prior knowledge such as piecewise smoothness but require multiple images with significant baseline. Template-based methods [4, 10, 13] require only one image but handle only one object for which they need specific prior knowledge, namely a 3D template.

In this paper, we propose a novel method that alleviates the strong assumptions of both the template-free and template-based methods: our method uses multiple templates to achieve deformable 3D reconstruction from only one image and for multiple objects. It uses object recognition to automatically discover what objects are visible in the input image and to select the appropriate templates for deformable 3D reconstruction. The object database is built offline. Crucially, this database does not only contain appearance descriptors as in existing object recognition frameworks [7, 8, 11], but also material properties to facilitate deformable 3D reconstruction.

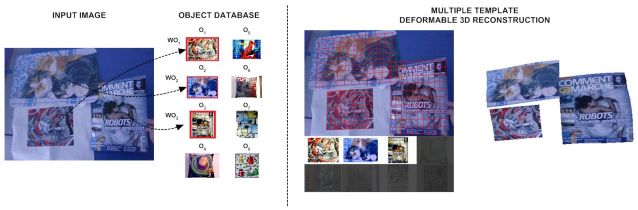


Figure 1: Given an input image, we first perform object recognition to detect database objects in the input image. Then, we compute 2D image warps that model the deformation of each particular object between the input image and a 2D parameterization of the 3D template. Finally, using the estimated warps we perform template-based isometric surface reconstruction. The detected objects from the database are highlighted, whereas non-detected objects are depicted in a darker color.

At runtime we use object recognition to automatically discover what objects are visible in the current input image and to select the appropriate templates for deformable 3D reconstruction. For this purpose, we perform wide-baseline image matching between the stored templates in the database and the input image that contains the deforming surfaces. We use an outlier rejection method [12] to obtain a set of clean-up matches between each detected template and present objects in the 2D input image. For those objects that have a number of clean-up matches higher than a defined threshold, we compute an image warp [3] that encodes the particular deformation of an object in the image. Finally, given the estimated warps we perform deformable 3D reconstruction for the detected objects [4].

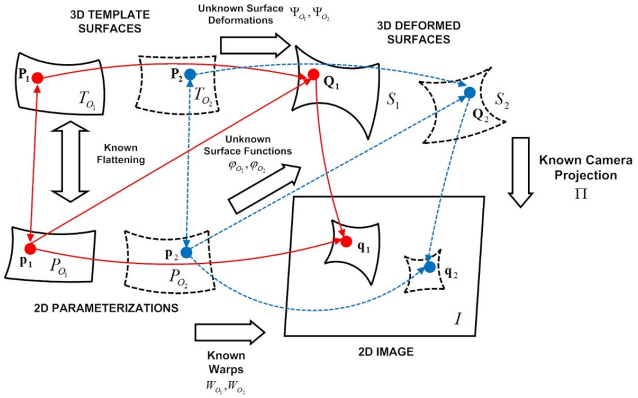


Figure 2: Geometric modeling of monocular multiple-template based reconstruction.

Our approach is the first to use an object database to aid deformable 3D reconstruction. In terms of genericity, it lies between existing template-based and template-free methods, as it assumes that strong priors on the world can be modeled but is not object-specific. We show successful deformable 3D reconstruction results of multiple objects from a single image. The objects in the database are made of different materials such as paper, cloth and plastic, and the database contains both developable and non-developable objects. Our work opens a whole new area of approaches that can benefit from using strong priors encoded in a versatile object database.

[1] A. Agudo, B. Calvo, and J.M.M. Montiel. Finite Element based Sequential Bayesian Non-Rigid Structure from Motion. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, USA, 2012.

[2] I. Akhter, Y. Sheikh, S. Khan, and T. Kanade. Trajectory space: A dual representation for nonrigid structure from motion. *IEEE Trans. Pattern Anal. Machine Intell.*, 2011.

[3] A. Bartoli, M. Perriollat, and S. Chambon. Generalized thin-plate spline warps. *Intl. J. of Computer Vision*, 88(1):85–110, May 2010.

[4] A. Bartoli, Y. Gérard, F. Chadebecq, and T. Collins. On template-based reconstruction from a single view: Analytical solutions and proofs of well-posedness for developable, isometric and conformal surfaces. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Providence, Rhode Island, USA, 2012.

[5] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3D shape from image streams. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[6] A. Del Bue, X. Lladó, and L. Agapito. Non-rigid metric shape and motion recovery from uncalibrated images using priors. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, USA, 2006.

[7] D.G. Lowe. Object recognition from local scale-invariant features. In *Intl. Conf. on Computer Vision (ICCV)*, pages 1150–1157, Corfu, Greece, 1999.

[8] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[9] M. Paladini, A. Bartoli, and L. Agapito. Sequential non-rigid structure from motion with the 3D-implicit low rank shape model. In *Eur. Conf. on Computer Vision (ECCV)*, Crete, Greece, 2010.

[10] M. Perriollat, R. Hartley, and A. Bartoli. Monocular template-based reconstruction of inextensible surfaces. *Intl. J. of Computer Vision*, 95(2):124–137, 2011.

[11] J. Pilet and H. Saito. Virtually augmenting hundreds of real pictures: An approach based on learning, retrieval, and tracking. In *IEEE Virtual Reality (VR)*, Waltham, MA, USA, 2010.

[12] D. Pizarro and A. Bartoli. Feature-based deformable surface detection with self-occlusion reasoning. *Intl. J. of Computer Vision*, 97 (1):54–70, March 2012.

[13] M. Salzmann and P. Fua. Reconstructing sharply folding surfaces: A convex formulation. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Miami, USA, 2009.

[14] L. Torresani, A. Hertzmann, and C. Bregler. Non rigid structure-from-motion: Estimating shape and motion with hierarchical priors. *IEEE Trans. Pattern Anal. Machine Intell.*, 30(5):878–892, May 2008.

Incremental Light Bundle Adjustment

Vadim Indelman
indelman@cc.gatech.edu
Richard Roberts
richard.roberts@gatech.edu
Chris Beall
cbeall3@gatech.edu
Frank Dellaert
dellaert@cc.gatech.edu

College of Computing,
Georgia Institute of Technology,
Atlanta, GA 30332, USA

Fast and reliable bundle adjustment is essential in many applications such as mobile vision, augmented reality, and robotics. Two recent ideas to reduce the associated computational cost are structure-less SFM (structure from motion) [1, 5, 6, 7] and incremental smoothing [3, 4]. The former formulates the cost function in terms of multi-view constraints instead of re-projection errors, thereby eliminating the 3D structure from the optimization. The latter was developed in the SLAM (simultaneous localization and mapping) community and allows one to perform efficient incremental optimization, adaptively identifying the variables that need to be recomputed at each step.

In this paper we combine these two key ideas into a computationally efficient bundle adjustment method, and additionally introduce the use of three-view constraints to remedy commonly encountered degenerate camera motions.

The optimized cost function in light bundle adjustment (LBA) is defined, similarly to [5, 6], as

$$J_{LBA}(\hat{\mathrm{x}},\mathrm{p}) \doteq \sum_{i=1}^{N_h} \|\mathrm{h}_i(\hat{\mathrm{x}},\mathrm{p})\|^2_{\Sigma_i} \tag{1}$$

with $\hat{\mathrm{x}}$ the estimated poses for all cameras, $\hat{\mathrm{p}}$ all image observations across all views, Σ the measurement covariance, and where $\|\mathrm{a}\|_\Sigma \doteq \mathrm{a}^T\Sigma^{-1}\mathrm{a}$ denotes the squared Mahalanobis distance. The parameter N_h represents the number of multi-view constraints h_i derived from the feature correspondences in the given sequence of views. Each constraint h_i is a function of several camera poses and the image observations in the corresponding images. The applied multi-view constraints are a combination of two- and three-view constraints [2], that, as opposed to using only two-view constraints, allow consistent motion estimation in a straight-line camera motion.

We formulate the optimization problem in terms of a factor graph, and incrementally update a directed junction tree which keeps track of the current best solution [3, 4]. The factor graph defines a factorization of the function $f(x)$ as

$$f(x) = \prod_\alpha f_\alpha(\mathcal{X}_\alpha), \tag{2}$$

where $\mathcal{X}_\alpha \subset x$ is the set of all camera poses x_j connected by an edge to factor f_α. Each factor f_α represents a single multi-view constraint between the appropriate views. A simple example of a factor graph using two- and three-view constraints is shown in Figure 1a.

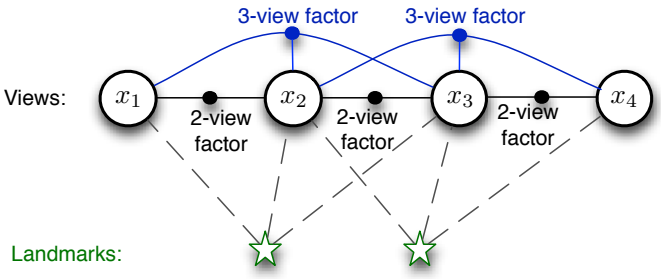
The optimization process corresponds to adjusting all the camera poses x to obtain a maximum a posteriori estimate

$$\hat{x} = \arg\max_x f(x) = \arg\min_x (-\log f(x)). \tag{3}$$

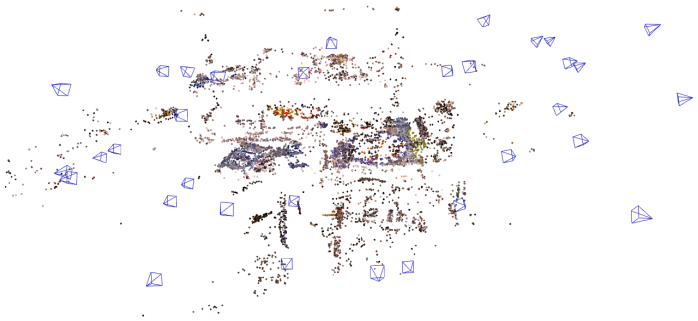
Assuming a Gaussian distribution, the above formulation is equivalent to a non-linear least-squares optimization of the cost function (1). Typically, only a small fraction of the camera poses are recalculated in each optimization step, leading to a significant computational gain. Although only the camera poses are optimized in LBA, if desired, all or some of the observed 3D points can be reconstructed after the optimization convergence.

We present a performance evaluation of incremental LBA (iLBA), i.e. applying incremental smoothing for optimizing the cost function J_{LBA}, using several datasets. Figure 1b shows the optimized camera poses and the reconstructed structure for one of the datasets. The structure reconstruction was performed based on the LBA-optimized camera poses.

Comparing iLBA to previous structure-less BA methods [1, 5, 6, 7] and to conventional bundle adjustment reveals significantly better timing performance and similar accuracy levels.



(a)



(b)

Figure 1: (a) A factor graph representation for a simple example of 4 cameras observing 2 landmarks. Two-view and three-view factors are added instead of projection factors. Landmark observations are denoted by dashed lines. (b) Optimized camera poses and reconstructed structure in the *cubicle* dataset.

[1] V. Indelman. Bundle adjustment without iterative structure estimation and its application to navigation. In *IEEE/ION Position Location and Navigation System (PLANS) Conference*, April 2012.

[2] V. Indelman, P. Gurfil, E. Rivlin, and H. Rotstein. Real-time vision-aided localization and navigation based on three-view geometry. *IEEE Trans. Aerosp. Electron. Syst.*, 48(3):2239–2259, July 2012.

[3] M. Kaess, V. Ila, R. Roberts, and F. Dellaert. The Bayes tree: An algorithmic foundation for probabilistic robot mapping. In *Intl. Workshop on the Algorithmic Foundations of Robotics*, Dec 2010.

[4] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert. iSAM2: Incremental smoothing and mapping using the Bayes tree. *Intl. J. of Robotics Research*, 31:217–236, Feb 2012.

[5] A. L. Rodríguez, P. E. López de Teruel, and A. Ruiz. Reduced epipolar cost for accelerated incremental sfm. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3104, June 2011.

[6] A. L. Rodríguez, P. E. López de Teruel, and A. Ruiz. GEA optimization for live structureless motion estimation. In *First Intl. Workshop on Live Dense Reconstruction from Moving Cameras*, pages 715–718, 2011.

[7] R. Steffen, J.-M. Frahm, and W. Förstner. Relative bundle adjustment based on trifocal constraints. In *ECCV Workshop on Reconstruction and Modeling of Large-Scale 3D Virtual Environments*, 2010.

Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding

Marco Bevilacqua[1]
marco.bevilacqua@inria.fr
Aline Roumy[1]
aline.roumy@inria.fr
Christine Guillemot[1]
christine.guillemot@inria.fr
Marie-Line Alberi Morel[2]
marie_line.alberi-morel@alcatel-lucent.com

[1] INRIA
Campus universitaire de Beaulieu,
35042 Rennes Cedex, France

[2] Alcatel-Lucent, Bell Labs France
Route de Villejust,
91620 Nozay, France

Single-image super-resolution (SR) is the problem of generating a single high resolution (HR) image, given one low resolution (LR) image as input. In this paper we propose a low-complexity and yet efficient algorithm that reconstruct the HR image in one pass (instead, e.g. of 7 passes for a magnification factor of 4 in [3]). The proposed algorithm falls into the family of *example-based SR*. Taking inspiration from machine learning, it aims at learning the mapping from the LR image(s) to the HR image by using a dictionary: the learning process is performed locally, by trying to infer the HR details through the use of small "examples". For general SR purposes the examples used are patches (sub-windows of image); the dictionary is formed by pairs of LR and HR patches.

In particular, our method adopts the Neighbor Embedding (NE) approach [1, 2], that assumes a local similarities between the LR and HR spaces. For each LR input patch $\mathbf{x}_t^i$, we construct the corresponding HR output by following three steps.

1. *Nearest neighbor (NN) search*: find among the LR patches of the dictionary $\mathcal{X}_d = \{\mathbf{x}_d^j\}_{j=1}^{N_d}$ the K NN.
2. *LR patch estimation*: find a weighted combination of the selected neighbor that approximates $\mathbf{x}_t^i$.
3. *HR patch reconstruction*: keep the same weights with the corresponding HR patches to reconstruct the HR output patch $\mathbf{y}_t^i$.

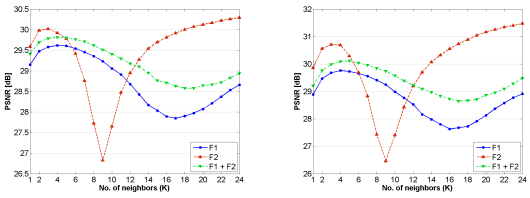
The whole procedure is carried out in a feature space, i.e. each patch is represented by a vector of features computed on its pixels. In previous NE-based SR algorithms [1, 2], the weights of each linear combination (Step 2) are the result of the following constrained least squares (LS) minimization problem (*SUM1-LS*):

$$\mathbf{w}^i = \arg\min_{\mathbf{w}} \|\mathbf{x}_t^i - X_d^i \mathbf{w}\|^2 \quad \text{s.t.} \quad \mathbf{1}^T \mathbf{w} = 1 \,. \qquad (1)$$

In the wake of the NE-based approach, we want to design a low-complexity and competitive algorithm for single-image SR. For this purpose, we analyze three key aspects:

1. the *features* used to represent the LR and HR patches
2. the method used to *compute the weights* of the patch combinations
3. the nature of the *dictionary* (external or "internal")

As for the feature representation, we propose to use centered luminance values (the straight luminance values of the pixels after subtracting the mean value of the patch) as unique features. This choice is in order to have consistency in the representation of the LR and HR patches, and to meet the low-complexity requirement (in fact, we have only one value per pixel). We show that, while keeping the usual *SUM1-LS* as a method to compute the weights, centered luminance values turn out to be good features, as they outperform gradient features (even more costly), if we observe the results for different values of the number of neighbors K.



head bird

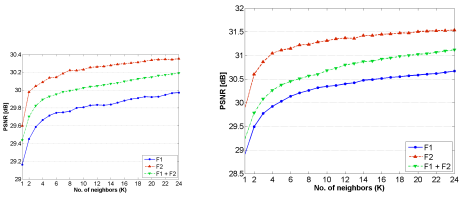
Figure 1: Comparison between LR features (*SUM1-LS*).

However, all the curves present a fall in the performance, that is even dramatic for our selected features. We explain this criticality with the fact that, for certain values of K, the LS solution is over fitted on the LR data and thus generates bad HR reconstruction. To overcome this problem, we propose another strategy to compute the weights, still based on the LS approximation of the input vectors, but with a more relaxed non-negativity constraint. The problem in (1) thus becomes the following non-negative LS problem (*NNLS*):

$$\mathbf{w}^i = \arg\min_{\mathbf{w}} \|\mathbf{x}_t^i - X_d^i \mathbf{w}\|^2 \quad \text{s.t.} \quad \mathbf{w} \geq 0 \,. \qquad (2)$$

We show that the combination of centered luminance values and *NNLS* weights gives the best results and presents monotonically increasing PSNR values (Figure 2). As for the dictionary issue, as we want to realize a single-step upscaling without iterating the algorithm for small scale factors, the external training is shown to be the obvious solution. By deriving the patch correspondences from a "self-pyramid" in the way of [3], in fact, we run the risk of having extremely poor (in size) dictionaries. Moreover, the external solution offers the possibility to build a dictionary in advance, so reducing the online running time.

Our method presents much better results than other one-pass algorithms (the original NE-based algorithm of Chang et al. [2] and the Kernel Ridge Regression method of Tang et al. [4]), but it presents comparable results w.r.t. the multi-pass (and therefore more complex) algorithm of [3]. As future work, we plan to investigate other strategies for neighbor search (i.e. other metrics), in order to select "better neighbors" for the HR reconstructions and so improve the performance.



head bird

Figure 2: Comparison between LR features (*NNLS*).

		Our algorithm		Chang et al.		Glasner et al.	
Image	*Scale*	*PSNR*	Time	*PSNR*	Time	*PSNR*	Time
bird	2	34.69	18	32.94	110	34.42	406
head	2	32.88	18	32.34	145	32.68	367
woman	2	30.91	15	29.43	114	30.61	410
bird	3	31.37	9	29.71	47	32.16	281
head	3	31.46	12	30.82	68	31.69	370
woman	3	27.98	12	26.45	37	28.79	248
bird	4	28.99	6	27.37	21	30.07	475
head	4	30.26	6	29.57	26	30.86	379
woman	4	25.66	5	24.25	17	26.79	401

Table 1: Results (*PSNR* and running time in sec.) for different images.

Our algorithm Chang et al. [2] Glasner et al. [3]

[1] Tak-Ming Chan, Junping Zhang, Jian Pu, and Hua Huang. Neighbor embedding based super-resolution algorithm through edge detection and feature selection. *Pattern Recognition Letters*, 4 2009.

[2] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-Resolution Through Neighbor Embedding. In *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 2004.

[3] Daniel Glasner, Shai Bagon, and Michal Irani. Super-Resolution from a Single Image. In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, 10 2009.

[4] Yi Tang, Pingkun Yan, Yuan Yuan, and Xuelong Li. Single-image super-resolution via local learning. *International Journal of Machine Learning and Cybernetics*, 2011.

Fast Non-uniform Deblurring using Constrained Camera Pose Subspace

Zhe Hu
zhu@ucmerced.edu
Ming-Hsuan Yang
mhyang@ucmerced.edu

Electrical Engineering and Computer Science
University of California at Merced

Introduction: Camera shake during exposure time often results in non-uniform blur across the entire image. Recent algorithms [1, 3] model the non-uniform blurry image B as a linear combination of images observed by the camera at discretized poses θ, and focus on estimating the time fraction w_θ positioned at each pose,

$$B = \sum_{\theta \in S} w_\theta (K_\theta L) + n, \tag{1}$$

where K_θ is the matrix that warps latent image L to the transformed copy at a sampled pose θ and S denotes the set of sampled camera poses. While these algorithms show promising results, they entail high computational cost as the high-dimensional camera motion space and the latent image have to be computed during the iterative optimization procedures.

In this paper, we propose a fast single-image deblurring algorithm to remove non-uniform blur. We first introduce an initialization method that facilitates convergence and avoid local minimums of the formulated optimization problem. We then propose a new camera motion estimation method which optimizes on a small set of pose weights of a constrained camera pose subspace at a time rather than using the entire space. We develop an iterative method to refine the camera motion estimation and introduce perturbation at each iteration to obtain robust solutions. Fig. 1 summarizes the main steps of our method.

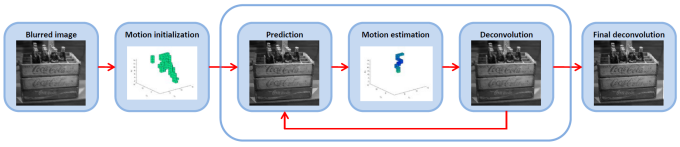


Figure 1: Algorithm overview

Initialization: Motivated by the backprojection techniques in image processing which are used to reconstruct the 2-D signal from its 1-D projections, we develop a method to reconstruct the camera motion from multiple blur kernels (the projection of the camera motion trajectory). From multiple estimated blur kernels of different regions, the direct approximation of the camera motion is to use the inverse transformation by duplicating the 2-D blur kernels across the camera motion space. That is, for each entry in blur kernel k_l, we determine the possible camera poses whose projection $p(\theta, l)$ at this site l is the interest entry and then duplicate the weight of this entry across all the possible camera poses. This procedure is called backprojection commonly used in image reconstruction and we denote $bp(k_l, l)$ as the backprojected value by

$$bp(k_l, l) = \sum_i \sum_{\{\theta | p(\theta, l) = i\}} k_l(i)\Gamma(\theta), \tag{2}$$

where $\Gamma(\theta)$ is an indicator function of camera poses with value 1 for pose θ and 0 for others. To obtain better results, we formulate the initialization estimation with an optimization problem and enforce sparse constraints of the weighs to get better reconstruction results due to the fact that the camera motion trajectory is sparse in the motion space. Moreover, we assign each backprojection function $bp(k_l, l)$ a confidence value $a(l)$ based on the distance of the site l to the optical center which can usually be assumed to be the center of the image. Thus, the initial estimation of weights W is formulated as,

$$\hat{W} = \arg\min_W ||W - \sum_l a(l) bp(k_l, l)||^2 + ||W||_1. \tag{3}$$

Weight Estimation on Constrained Pose Subspace: We formulate the weight estimation problem as

$$\min_W \sum_{\partial_*} \alpha_* || \sum_{\theta \in S} w_\theta \partial_* (K_\theta \widetilde{L}) - \partial_* B||^2 + \beta ||W||^2, \qquad \text{s.t.} \quad W \geq 0. \tag{4}$$

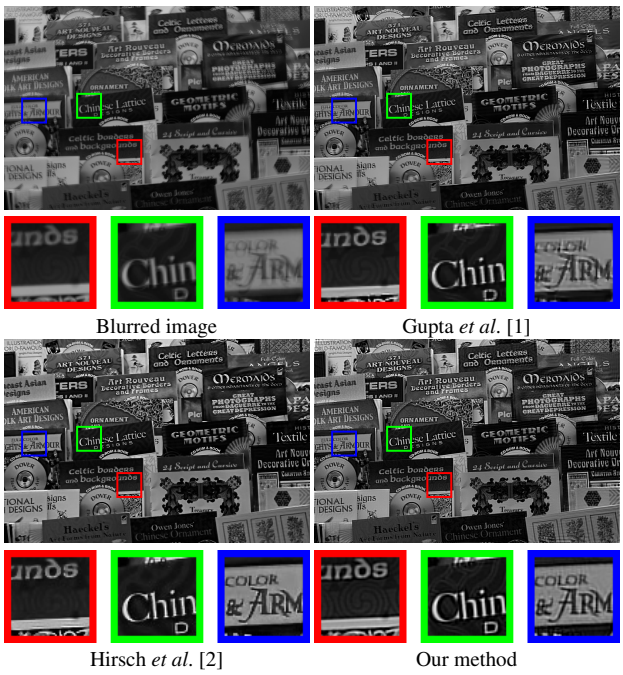


Figure 2: Comparison with state-of-the-art single image deblurring methods for spatially variant blur.

where $\partial^* \in \{\partial_x, \partial_y, \partial_{xx}, \partial_{xy}, \partial_{yy}\}$ denotes the partial derivative, and $\alpha_* \in \{\alpha_1, \alpha_2\}$ is the weight for partial derivative of different order. In this formulation, a direct optimization on the whole pose set S is computationally expensive. In this work, we propose to optimize Eqn. 4 only on a sparse subset of poses referred as an active set $A \subset S$. The active set represents the set of poses which are likely to lie on the motion trajectory.

We use the threshold strategy with a threshold ε to determine the active set at each iteration. Suppose $A^{(i)} = \{\theta_1, \ldots, \theta_{p+q}\}$ is the active set at iteration i, and the corresponding weights $W_A^{(i)} = \{w_1, \ldots, w_{p+q}\}$ is sorted in the descending order $w_1 \geq \ldots \geq w_p \geq \varepsilon \geq w_{p+1} \geq \ldots \geq w_{p+q}$. At iteration $i+1$, we set the new active set to be

$$A^{(i+1)} = \{\theta_1, \ldots, \theta_p, \hat{\theta}_{p+1}, \ldots, \hat{\theta}_{p+q}\}, \tag{5}$$

by deleting the smallest q weights $\{\theta_{p+1}, \ldots, \theta_{p+q}\}$ and adding new poses $\{\hat{\theta}_{p+1}, \ldots, \hat{\theta}_{p+q}\}$ as perturbation. We obtain the new poses by sampling based on the previous active set $A^{(i)}$ using a Gaussian distribution (with small variance). Once we determine the active set A, we assign the weights for the poses of the inactive set to be 0, and estimate the weights of poses in the active set by replacing S with A in Eqn. 4. Since the active set is already sparse with respect to the whole space, it is not necessary to introduce a sparse regularization term in the above optimization problem.

Experimental Results: We evaluate the proposed algorithm against several state-of-the-art single image deblurring methods for spatially variant blur (Fig. 2 as an example). Our MATLAB implementation is about 1.5 times faster than the MATLAB implementation of [2], the most efficient single-image non-uniform deblurring method to the best of our knowledge.

[1] A. Gupta, N. Joshi, L. Zitnick, M. Cohen, and B. Curless. Single image deblurring using motion density functions. In *ECCV*, pages 171–184, 2010.

[2] M. Hirsch, C. J. Schuler, S. Harmeling, and B. Scholkopf. Fast removal of non-uniform camera shake. In *ICCV*, pages 463–470, 2011.

[3] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce. Non-uniform deblurring for shaken images. In *CVPR*, pages 491–498, 2010.

Workshop Keynote

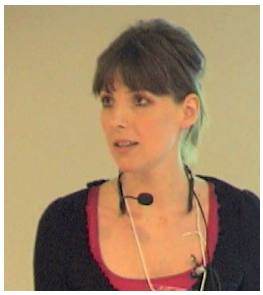
Latent Variable Models for Content-Based Image Retrieval and Structure Prediction

Ariadna Quattoni

Universitat Politecnica de Catalunya

In the first part of the talk I will present recent work on learning latent variable models for content-based image retrieval. To learn a function that predicts the relevance of a database image to an image query all that we need is some form of feedback from users of the retrieval system. For example, we can obtain triplet constraints specifying that relative to some query Q, an image A should be ranked higher than an image B. When such feedback is available ranking SVMs can be used to induce the retrieval function. I will describe an extension of this framework where instead of learning a single relevance function we learn a mixture of relevance functions. Intuitively, given a query we first compute a distribution over ”coarse” latent classes and then compute the relevance function for queries of that class. I will present a simple learning algorithm that induces both the latent classes and the parameters of each model.

In the second part of the talk I will describe some of my current work on developing efficient learning algorithms for structure prediction with latent variables. These algorithms are based on using an algebraic representation that exploits directly the Markovianity of the distribution.



Ariadna received her PhD from MIT in 2009 under the advise of professors Michael Collins and Trevor Darrell . Her dissertation was about transfer learning models for image classification. She has also worked on latent variable models for structure prediction with applications to Computer Vision and Natural Language Processing. After graduation she joined the Technical University of Catalunya as a research scientist. She is also the cofounder of dMetrics, a company specializing on NLP technologies for user generated content. She has published in all the major Computer Vision and Machine Learning Conferences. She has recently received the best paper award at the main European Natural Language Processing Conference for her work on spectral methods for latent variable dependency parsing.

Workshop Keynote

Monocular SLAM and Real-Time Scene Perception

Andrew Davison

Imperial College London

We have seen great advances in real-time 3D vision in recent years, enabled by algorithmic improvements, the continuing increase in commodity processing power and better camera technology. Research in Monocular SLAM (Simultaneous Localisation and Mapping), where a single agile camera moves through a mostly static scene, was for a long time focused on mapping only enough of a scene to enable robust real-time motion estimation of the camera itself. Attention is now turning however to gradually improving the quality of scene reconstruction which can be achieved in real-time. I will speak about how early work on feature-based SLAM is now being surpassed by methods which aim to map dense scene structure, and how this is leading towards ever-more general 3D scene modelling and understanding.

Andrew Davison received a BA in physics and the D.Phil. degree in computer vision from the University of Oxford in 1994 and 1998, respectively. He undertook his doctorate with with Prof. David Murray at Oxfords Robotics Research Group, where he developed one of the first robot simultaneous localisation and mapping (SLAM) systems using vision. He then spent two fantastic years as a European Union (EU) Science and Technology Fellow at the National Institute of Advanced Industrial Science and Technology (AIST), Japan, where he continued to work on visual robot navigation. In 2000 he returned to the University of Oxford as a Postdoctoral Researcher working with Ian Reid and was awarded a five-year Engineering and Physical Sciences Research Council (EPSRC) Advanced Research Fellowship in 2002. During this time he developed the well known MonoSLAM algorithm for real-time SLAM with a single camera. He joined Imperial College London as a lecturer in 2005, where he teaches robotics in the Department of Computing and leads the Robot Vision Research Group. In 2008, he was awarded a five year European Research Council (ERC) Starting Grant. The group's work continues to focus on the challenges in real-time, real-world 3D vision, expanding on the core problems of localisation and mapping with cameras towards a more general real-time model-based scene understanding agenda. The wide applicability of this research in robotics and beyond into areas like augmented reality, gaming, mobile devices and automotive has been proven by strong industrial interest and the group has ongoing links with companies in several different sectors. Recent work has been recognized with best paper awards at ICRA 2010 and ISMAR 2011, and the best demonstration award at ICCV 2011.

Author Index

Notes

BMVC 2012 would like to thank our Sponsors

Gold

Microsoft® Research

TOSHIBA
Leading Innovation

STEMMER® IMAGING

Silver

3dMD SCORPION VISION LIMITED

Springer
the language of science