



Proceedings of the
4th CVC Workshop
on the Progress of Research & Development
CVC R&D 2009

New Trends and Challenges in Computer Vision

Centre de Visió per Computador
Universitat Autònoma de Barcelona
Bellaterra, Catalonia - Spain
October 30, 2009

Copyright © 2009 by the authors in the table of contents.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage or retrieval system, without permission from the authors.

ISBN: 978-84-937261-1-9

Printed by: Gráficas Rey, S.L.

Printed in Spain

Preface

This book contains the papers presented in the *4th CVC Workshop on the Progress of Research and Development CVCR&D'09: New Trends and Challenges in Computer Vision*. The Workshop was held at the *Computer Vision Center (CVC)*, sited in the Campus of the *Universitat Autònoma de Barcelona*, Spain, on October 30, 2009. The CVC workshops provide an excellent opportunity for young researchers and project engineers to share new ideas and knowledge about the progress of their work. It sketches the state of the research and development activities in the previous year. In addition, the workshop is the welcome event for new people that have joined the Institute recently.

The CVC is an institution devoted to Research and Development in Computer Vision. The CVC was established in 1994 by the Industry Department and the CIRIT of the *Generalitat de Catalunya* and the *Universitat Autònoma de Barcelona*. The center promotes industrial development of computer vision applications as well as research collaboration in the same field. CVC brings together faculty, postdoctoral fellows, visiting scientists and students, and provides support for training and development. The main aim of the CVC is to contribute to innovation and industrial competitiveness by means of Research and Development in Computer Vision, as well as to collaborate with industry on technological projects. CVC is an organization open to innovation, creativity, and collaboration.

The program of CVCR&D aims to analyze the current research that the groups established in the CVC are investigating and the potential new lines they are facing. Under the title *New Trends and Challenges in Computer Vision*, the CVCR&D organizing committee wants to promote the discussion of the key topics and the current research lines among the whole CVC community, including Ph.D. students, engineers, CVC staff, etc. The contributions consist of papers, which may include not only the existing work, but also some preliminary results, stressing on the potential relevance of the novel lines of research proposed. The workshop event comprises both oral and poster sessions. Oral sessions are dedicated to specific topics, organized following criteria of similarity in the area of the contributions, including: Advanced Driver Assistance Systems, Color and Texture, Region Detectors and Descriptors, Document Image Analysis, Image Compression, Medical Imaging, Tracking and Motion Analysis, Object Recognition, and Sensors, Stereo Vision, and Graphics. In these oral sessions, the participants will perform a short presentation in which the main problems and suggested lines of research are presented. At the end of the session, a debate time is opened among the audience and the participant, with the moderation of the Session Chairs. In addition, a poster session will take place including those works not presented as an oral talk.

We would like to thank all the contributors and all the Session Chairs for their participation in the Workshop. The reader of this book will find that the pieces of research contained in these proceedings show the dynamic, active, and promising scientific work carried out at the CVC with the aim to tackle *trends and challenges* that Computer Vision research is facing nowadays.

This year, the CVCR&D workshop presents a record of 39 papers from more than 70 authors. We hope you all enjoy the Workshop contributions and we are looking forward to meeting you, together with new people, again next year in the 5th CVCR&D.

Bellaterra, October, 2009
The Organization Committee of the CVCR&D'09

4th CVC Workshop on the Progress of Research & Development CVCRD 2009

New Trends and Challenges in Computer Vision

Workshop Organization

GENERAL CHAIRS

Xavier Baró
Sergio Escalera
Miquel Ferrer

ORGANIZATION COMMITTEE

Josep Lladós
Montse Culleré
Helena Piulachs

PROGRAM COMMITTEE

Jaume Amores
Andrew Bagdanov
Ramon Baldrich
Pau Baiget
Simone Balocco
Robert Benavente
Alicia Fornés
Carlo Gatta
Débora Gil
Jordi González
Aura Hernández
Laura Igual
Carme Julià
Dimosthenis Karatzas

Ágata Lapedriza
Josep Lladós
Antonio López
Felipe Lumbreras
Enric Martí
David Masip
Mikhail Mozerov
Xavier Otazu
Carlos Alejandro Párraga
Daniel Ponsa
Oriol Pujol
Petia Radeva
Bogdan Raducanu
Xavier Roca

David Rotger
Marçal Rusiñol
Anna Sabaté
Gemma Sánchez
Xavier Sánchez
Ángel Sappa
Joan Serrat
Ricardo Toledo
Ernest Valveny
Joost Van de Weijer
Maria Vanrell
Fernando Vilariño
Juan José Villanueva
Jordi Vitrià

Table of Contents

<i>Preface</i>	i
<i>Workshop Committees</i>	ii

1. Advanced Driver Assistance Systems

Adaptive Model--based Road Detection using Shadowless Features.....	1
<i>José M. Álvarez and Antonio M. López</i>	
Performance of classical monocular egomotion methods in the ADAS context.....	7
<i>Diego Cheda, Daniel Ponsa and Antonio López</i>	
Automotive Applications based on Video Alignment	13
<i>Ferran Diego, Daniel Ponsa, Jose M. Álvarez, Joan Serrat and Antonio López</i>	
Synthetic Urban Development to Evaluate Pedestrian Detection.....	17
<i>Javier Marín and Antonio López</i>	
Feature matching with graphical models for night vehicle detection	23
<i>Jose Carlos Rubio Ballester and Joan Serrat</i>	
Detecting small pedestrians	29
<i>David Vázquez, David Gerónimo, Antonio López</i>	

2. Color and Texture

Object Color Alteration.....	34
<i>Shida Beigpour and Joost van de Weijer</i>	
Human and Computational Color Constancy	40
<i>Jordi Roca, C.A Párraga and Maria Vanrell</i>	
A Computational Colour Naming Model Trained on Real-Life Images	46
<i>Hany M. SalahEldeen, Robert Benavente, Maria Vanrell</i>	
Hybrid Fusion: Beyond Early and Late Fusion for Texture Classification	52
<i>Fahad Shahbaz Khan, Joost van de Weijer, and Maria Vanrell</i>	
Towards non-supervised segmentation: a comparison of goodness measures based on saliency and contrast.....	58
<i>Eduard Vazquez and Ramon Baldrich</i>	

Computational Color: Representation, Constancy and Psychophysics	64
<i>Javier Vazquez-Corral and Maria Vanrell</i>	

3. Region Detectors and Descriptors

Object Pixel-Level Categorization using Bag of Features	70
<i>David Aldavert, Ricardo Toledo</i>	
Perceptual Feature Detection	76
<i>Naila Murray, Xavier Otazu, Maria Vanrell</i>	
Coloring Laplacian-of-Gaussian Detector for Image Matching	82
<i>David A. Rojas Vigo and Joost van de Weijer</i>	
On Experimental Evaluation of Descriptors for Facial Feature Point Detection.....	88
<i>Mario A. Rojas Q., David Masip and Jordi Vitrià</i>	
Using Colour Saliency for Image Retrieval	94
<i>Juan Ignacio Toledo, Joost van de Weijer</i>	

4. Document Image Analysis

Text Segmentation in Colour Poster from the Spanish Civil War Era	100
<i>Antonio Clavelli, Dimosthenis Karatzas</i>	
A rotation invariant page layout descriptor for document classification and retrieval	106
<i>Albert Gordo, Ernest Valveny</i>	
Comparison of Seal Detection by Different Character Shape Features	112
<i>Partha Pratim Roy, Umapada Pal and Josep Lladós</i>	

5. Image Compression

Perceptual Criteria on JPEG2000 Quantization.....	119
<i>Jaime Moreno, Xavier Otazu and Maria Vanrell</i>	

6. Medical Imaging

Use of Filtered Back-projection Methods to Improve CT Image Reconstruction	125
<i>Jorge Bernal, Javier Sánchez</i>	
Towards Detection of Measurable Contractions using WCE	131
<i>Michal Drozdal, Petia Radeva, Santi Seguí, Fernando Vilariño, Carolina Malagelada, Fernando Azpiroz and Jordi Vitrià</i>	

7. Tracking and Motion Analysis

Reactive object tracking with single uncalibrated PTZ camera.....	137
<i>Murad al Haj, Andrew D. Bagdanov and Jordi González</i>	
Robust Background Subtraction Approach based on Chromaticity and Intensity Patterns.....	143
<i>Ariel Amato, Mikhail Mozerov and Jordi González</i>	
Interest Point based Human Action Recognition.....	149
<i>Bhaskar Chakraborty, Andrew D. Bagdanov and Jordi González</i>	
Adaptive Dynamic Space Time Warping for Real Time Sign Language Recognition	155
<i>Sergio Escalera, Alberto Escudero, Petia Radeva, and Jordi Vitrià</i>	
An Analysis of Theoretical and Practical Aspects of Spatio-Temporal Regular Flow (SPREF) .	161
<i>Juan Diego Gomez, Carlos Gatta, Petia Radeva</i>	
3D Human Action Recognition using Key Poses	167
<i>Wenjuan Gong, Andrew D. Bagdanov and Jordi González</i>	
Advances in Variational Optical Flow	173
<i>Naveen Onkarappa and Angel D. Sappa</i>	

8. Object Recognition

Image Description using Local Binary Patterns: Application to Scene Classification.....	179
<i>Noha Elfiky, Jordi González</i>	
Graph-based representations for Object Recognition	185
<i>Jaume Gibert and Ernest Valveny</i>	
Semantic Segmentation of Images Using Random Ferns.....	191
<i>Josep M^a Gonfaus, Jordi González, Theo Gevers</i>	
Ranking Error-Correcting Output Codes for Class Retrieval	197
<i>Mehdi Mirza-Mohammadi, Francesco Ciompi, Sergio Escalera, Oriol Pujol, and Petia Radeva</i>	
Colour Logo Recognition	204
<i>Farshad Nourbakhsh, Dimosthenis Karatzas and Ernest Valveny</i>	
Object Detection using Coarse-to-Fine relocalization.....	210
<i>Marco Pedersoli, Jordi Gonzalez, Andrew Bagdanov and Juan José Villanueva</i>	

7. Tracking and Motion Analysis

Reactive object tracking with single uncalibrated PTZ camera.....	137
<i>Murad al Haj, Andrew D. Bagdanov and Jordi González</i>	
Robust Background Subtraction Approach based on Chromaticity and Intensity Patterns.....	143
<i>Ariel Amato, Mikhail Mozerov and Jordi González</i>	
Interest Point based Human Action Recognition.....	149
<i>Bhaskar Chakraborty and Andrew D. Bagdanov</i>	
Adaptive Dynamic Space Time Warping for Real Time Sign Language Recognition	155
<i>Sergio Escalera, Alberto Escudero, Petia Radeva, and Jordi Vitrià</i>	
An Analysis of Theoretical and Practical Aspects of Spatio-Temporal Regular Flow (SPREF) .	161
<i>Juan Diego Gomez, Carlos Gatta, Petia Radeva</i>	
3D Human Action Recognition using Key Poses	167
<i>Wenjuan Gong, Andrew D. Bagdanov and Jordi González</i>	
Advances in Variational Optical Flow	173
<i>Naveen Onkarappa and Angel D. Sappa</i>	

8. Object Recognition

Image Description using Local Binary Patterns: Application to Scene Classification.....	179
<i>Noha Elfiky, Jordi González</i>	
Graph-based representations for Object Recognition	185
<i>Jaume Gibert and Ernest Valveny</i>	
Semantic Segmentation of Images Using Random Ferns	191
<i>Josep M^a Gonfaus, Jordi González, Theo Gevers</i>	
Ranking Error-Correcting Output Codes for Class Retrieval	197
<i>Mehdi Mirza-Mohammadi, Sergio Escalera, Petia Radeva</i>	
Colour Logo Recognition.....	204
<i>Farshad Nourbakhsh, Dimosthenis Karatzas and Ernest Valveny</i>	
Object Detection using Coarse-to-Fine relocalization	210
<i>Marco Pedersoli, Jordi Gonzalez, Andrew Bagdanov and Juan José Villanueva</i>	

9. Sensors, Stereo Vision, and Graphics

Calibration and Rectification of Multimodal Stereo Rigs	216
<i>Fernando Barrera, Felipe Lumbreras, and Angel Sappa</i>	
BeaStreamer-v0.1: a new platform for Multi-Sensors Data Acquisition in Wearable Computing Applications.	222
<i>Pierluigi Casale, Oriol Pujol, Petia Radeva</i>	
Quadric Surface Fitting: Orthogonal versus Estimated Distances	228
<i>Mohammad Rouhani and Angel Sappa</i>	

Author Index	235
---------------------------	-----

Adaptive Model-based Road Detection using Shadowless Features

José M. Álvarez and Antonio M. López

*Computer Vision Center and Computer Science Dept.
Universitat Autònoma de Barcelona, Barcelona, Spain
{jalvarez,antonio}@cvc.uab.es*

Abstract

Road detection is an essential functionality for autonomous driving. The key of vision-based road detection algorithms is the ability of classifying image pixels as belonging or not to the road surface. In this paper, we propose an adaptive color-based road detection algorithm which combines a physics-based illuminant-invariant color space with a model-based classifier in a frame by frame framework using a monocular camera. The novelty of our approach resides in using shadowless features to characterize road pixels. Besides, the road model is built on-line and dynamically updated based on feedback from the current detection and predictions of a Markov model. Experiments are conducted on different road sequences including different scenarios, different weather conditions, extreme shadows and the presence of other vehicles. Qualitative results validate the proposal for reliable road detection.

Keywords: Road detection, region growing, illumination invariance, color invariants, shadows.

1 Introduction

Road detection is an essential functionality for autonomous driving as well as for supporting other advanced driver assistance systems (ADAS) such

as road following or vehicle detection and tracking. The aim of vision-based road detection is detecting the road ahead a moving vehicle using a vision system such as a camera (Fig. 1). Vision-based road detection is very challenging since the road is in an outdoor scenario imaged from a mobile platform. Thus, the detection algorithm should be able to deal with a continuously changing background, the presence of different objects (e.g., vehicles, pedestrians) with unknown movement, different scenarios (e.g., urbans, highways, off-roads), different road attributes (e.g., shape, color), and different imaging conditions such as varying illumination, different viewpoints and weather conditions (Fig. 2).

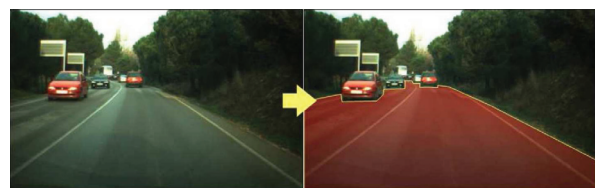


Figure 1: The aim of vision based-road detection is detecting pixels belonging to the road surface. Shoulder pixels are included since they are asphalted.

Color cues have been widely used for road detection since color provides powerful information of the road to be detected even in the absence of

shape information. In addition, color imposes less physical restrictions leading to more versatile systems. The two most popular color spaces that have proved to be robust to minor illuminant changes are *HSV* [1, 2] and normalized *RGB* [3]. However, algorithms based on these color spaces fail under wide lighting variations (strong shadows and highlights among others).



Figure 2: Major challenges for monocular vision-based road detection are the treatment of shadows and the presence of other vehicles.

Therefore, in this paper, we extend an existing color-based road detection algorithm using color invariant features proposed in [4]. This road detection algorithm exploits the lighting invariant benefits of a physics-based color space combined with a non-parametric adaptive model-based region growing algorithm in a frame by frame framework. Color invariants are usually derived from the dichromatic reflection model of Shafer [5]. This dichromatic model describes how photometric changes, such as shadows, shading, highlights and illumination, influence the RGB-values in images. Among the existing color invariants [6, 7, 8, 9] the algorithm uses the illuminant-invariant color space introduced by Finlayson [10] since it assumes Lambertian surfaces, approximately narrow band sensors and Planckian light sources. Further, once the data is characterized, a classifier determines whether a pixels belongs or not to the road class. In this way, the classifier is constructed based on the knowledge of a road model and without any consideration regarding unlabelled or non-road pixels. The contribution of this paper resides in including an on-line learning stage. In this way, the road model used is built on-line and dynam-

cally adapted based on feedback from the current detection and predictions of a Markov model. This learning stage is accomplished under the only assumption that the bottom central region of the image belongs to the road surface. In fact, the lowest row of the image corresponds to a distance of about 4 meters away from the camera placement which is a reasonable assumption most of the time.

The rest of this paper is organized as follows. First, in Sect. 2 the proposed road detection algorithm is introduced. Then, in Sect. 3, results of applying the algorithm to different road sequences are discussed. The road sequences include different scenarios and different illumination conditions. Finally, in Sect. 4, conclusions are drawn.

2 Road Detection Algorithm

The algorithm depicted in Fig. 3 has been devised to perform frame by frame road detection. The algorithm consists of three main stages: color space conversion, model-based region growing and on-line learning.

In the first stage the input image is converted onto the illuminant-invariant (shadowless) feature space [10]. This conversion consists in projecting the log-chromaticity pixel values of the incoming data onto a direction orthogonal to the lighting change line. This direction is device dependent and can be estimated off-line using calibration patterns [10]. The result is a gray-scale image \mathcal{I} where the influence of lighting variations (i.e., shadows) is greatly attenuated (Fig. 4).

In the second stage each pixel in the image is classified as road or background accordingly to a road model H , a fixed threshold λ and a connectivity criterion (Fig. 5). The road-model provides the probability of a pixel p being road according to its illuminant-invariant value $\mathcal{I}(p)$. That is, $H = P(\mathcal{I}(p)|road)$. Applying this model results in a probability (Fig. 5a) map which is binarized using λ (Fig. 5b). Then, connected components (region growing) is applied to the binary image

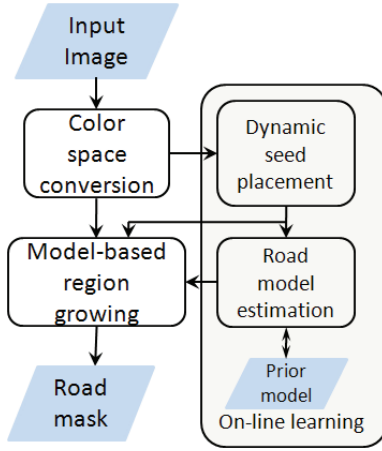


Figure 3: Road detection algorithm. Images are converted onto shadow invariant feature space and each pixel is classified as road or background accordingly a road model and a fixed threshold. Region growing starts at several dynamically placed seeds. These seeds are also used to estimate on-line the road model which is adapted over time.

starting at a set of seeds (Fig. 5c). Finally, a filling holes process using simple mathematical morphology operations is applied to obtain the final result (Fig. 5d).

The third stage is the learning process which is divided in two different parts: dynamic seed placement and the estimation of the road model. In the first part, the seeds are placed dynamically using a two steps process (Fig. 6). In the first step, robust statistic methods are applied to the surrounding region of seeds equidistantly placed at the bottom part of the image to generate a noiseless model (Fig. 6a). In the second step, the mean value of the surrounding region of each seed is projected onto this model discarding those seeds lying outside the model (Fig. 6b). In the second part of the on-line learning stage the road model is estimated. This model is built on-line for each image in the sequence using non-discarded seeds. In this way, the surrounding region of these seeds is used to build the normalized histogram of road pixel val-

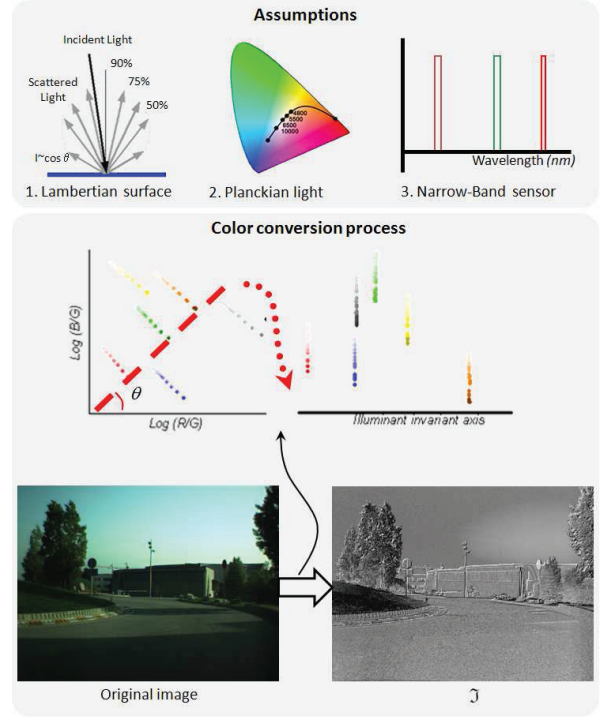


Figure 4: Under the assumptions of Planckian light, Lambertian surface and narrow-band sensors, an illuminantinvariant image which is almost shadow free is obtained projecting log-chromaticity values onto an characteristic direction θ .

ues. Further, the model is adapted on-time using the road detected in previous images. In particular, the model is adapted using a second-order Markov model where prior road knowledge and the current estimated road-model are integrated into the final model:

$$H_i(t) = (1 - \alpha)H(t - 1) + \alpha H_i^p(t), \quad (1)$$

where $H(t)$ is the adapted (final) model for the current image, $H(t - 1)$ is the model estimated using the road detected in previous image and $H^p(t)$ is the model for the current image estimated using the seeds at the bottom part of the image. Further, α is an adaptation parameter. The lower α the more persistent model. The result is a highly adaptive model which can cope with almost all sudden

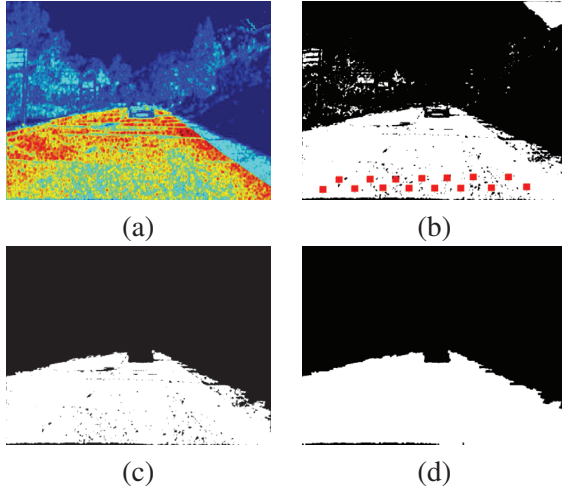


Figure 5: a) road probability map; b) binarized image with seeds overlapped; c) result applying connected components; d) final result after applying standard mathematical morphology.

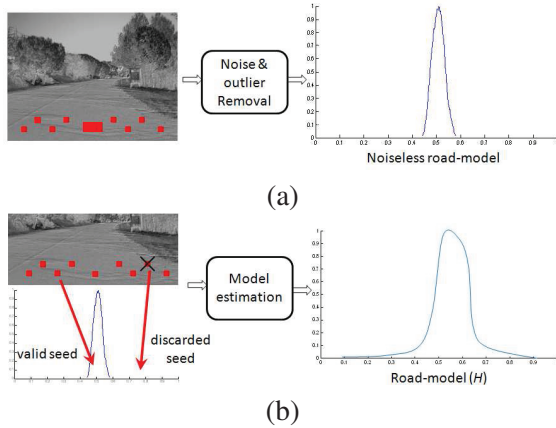


Figure 6: The dynamic seed placement is a two steps process. (a) Robust statistics on the surrounding region of a fixed set of seeds are used to estimate a noiseless model. (b) The surrounding region of those seeds lying on the noiseless model are used to estimate the model for the current image.

changes and variability in road pixels due to noise or the texture of the road.

3 Results

Experiments are conducted on different video sequences acquired using an on-board camera based on the Sony ICX084 sensor. This is a CCD chip of 640×480 pixels and 8 bits per pixel that makes use of a Bayer Pattern for collecting color information. The camera was equipped with a microlens of 6mm focal length. We use standard Bayer pattern decoding (bilinear interpolation) to obtain a 3channels color image (RGB) of 640×480 pixels per channel and 8 bits per pixel. The frame acquisition rate was 15 fps. Images cover up to approximately the nearest 80m ahead of the target vehicle. These images include different daytime, different scenarios (i.e., urban, highways and un-paved roads) complex road shapes due to intersections, unstructured roads, the presence of other vehicles and nonhomogeneous road appearance due to extreme shadows and highlights. The parameters of the algorithm are fixed using exhaustive search. In this way, a set of images is processed and evaluated using all possible values within the range of each parameter. The optimal set of parameter values is the one which maximizes the average performance. Example results are shown in Fig. 7. As shown, the road surface is well recovered most of the time, with the segmentation stopping at road limits and vehicles.

The analysis of failure reveals two limitations of the proposed method. First, the algorithm is not robust to highlights (Fig. 8a). The main reason is that the feature space used assumes Lambertian surfaces and diffuse lighting conditions. Thus, the model does not hold direct reflections. Second, the algorithm may fail recovering the whole road surface when strong lane markings are present (Fig. 8b). This is mainly due to the seed placement and the connected components algorithm since lane markings results in edges which stop the growing procedure.



Figure 7: Results of the proposed algorithm to detect roads.

4 Conclusions

In this paper we have proposed a road detection method based on the use of a physics-based illuminant-invariant feature space combined with a simple model-based classifier. The model is built on-line and adapted over time for each frame under the only assumption that the bottom part of the image shows road surface. Although this over-time adaptation, the algorithm does not use shape constraints. Hence, it can deal with complex road shapes such as urban or crowded scenarios. Provided qualitative results suggest that a reliable road detection algorithm is obtained by combining shadowless features and an adaptive model-based region growing algorithm.

As future work we want to incorporate an adaptive threshold method into the system.

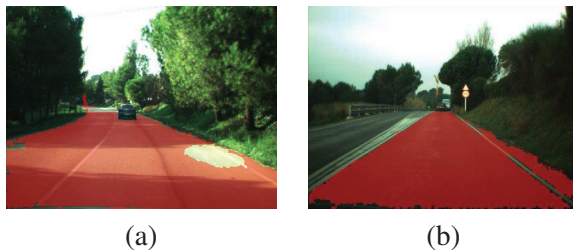


Figure 8: Example results where the road surface is not completely recovered. a) Road surface is not recovered when highlights are present. b) Strong lane markings force stopping the growing procedure.

References

- [1] C. Rotaru, T. Graf, , and J. Zhang, “Color image segmentation in hsi space for automotive applications,” *Journal of Real-Time Image Processing*, pp. 1164–1173, 2008.
- [2] M. Sotelo, F. Rodriguez, L. Magdalena, L. Bergasa, and L. Boquete, “A color vision-based lane tracking system for autonomous driving in unmarked roads,” *Auton. Robots*, vol. 16, no. 1, 2004.
- [3] C. Tan, T. Hong, T. Chang, and M. Shneier, “Color model-based real-time learning for road following,” *Procs. IEEE ITSC*, pp. 939–944, 2006.
- [4] J. M. Álvarez, A. M. López, and R. Baldrich, “Illuminant-invariant model-based road segmentation,” in *Procs. of the 2008 IEEE Intel. Vehicles Symposium (IV’08)*, Eindhoven, The Netherlands.
- [5] S. A. Shafer, “Using color to separate reflection components (A),” *Journal of the Optical Society of America A*, vol. 1, pp. 1248–+, 1984.
- [6] B. V. Funt and G. D. Finlayson, “Color constant color indexing,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 5, pp. 522–529, 1995.
- [7] T. Gevers and A. Smeulders, “Color-based object recognition,” *Pattern Recognition*, vol. 32, no. 3, pp. 453–464, March 1999.
- [8] T. Zickler, S. P. Mallick, D. J. Kriegman, and P. N. Belhumeur, “Color subspaces as photometric invariants,” *International Journal of Computer Vision*, vol. 79, pp. 13–30, 08/2008 2008.
- [9] S. Narasimhan, V. Ramesh, and S. Nayar, “A class of photometric invariants: separating material from shape and illumination,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, Oct. 2003, pp. 1387–1394 vol.2.
- [10] G. Finlayson, S. Hordley, C. Lu, and M. Drew, “On the removal of shadows from images,” *IEEE Trans. on PAMI*, vol. 28, no. 1, 2006.

Performance of classical monocular egomotion methods in the ADAS context

Diego Cheda, Daniel Ponsa and Antonio López

Centre de Visió per Computador - Universitat Autònoma de Barcelona
08193 Bellaterra, Barcelona, Spain - {dcheda, daniel, antonio}@cvc.uab.es

Abstract

This work focuses on egomotion estimation from a monocular camera (i.e., the changes in its position and orientation). Some of the different approaches proposed in the literature to compute egomotion are based on feature correspondences, and others on optical flow. In this work, we compare several egomotion methods with simulated ADAS-like sequences. From the conclusions of our experiments, we show which of the considered nonlinear and linear algorithms have the best performance under our context.

Keywords: Advanced Driver Assistance System, Egomotion estimation, Monocular camera.

1 Introduction

The development of systems to assist a driver in their driving activity has a great interest in the car industry. These systems are referred as *Advanced Driver Assistance Systems* (ADAS), and its main aim is increasing driver safety and comfort. In general, they use sensors to perceive and monitor the environment surrounding the vehicle, and based on the data obtained from it, assist the driver during the driving process to avoid unsafe situations.

An essential task in many ADAS applications (e.g., collision avoidance, autonomous driving, lane change assistance, etc.) is the estimation of the changes in the 3D position and orienta-

tion of the camera hosted in a vehicle (i.e., the camera egomotion) [4]. There is a large number of approaches for estimating camera egomotion from monocular sequences, as we will see in Sec. 3. Most of them recover the motion parameters from the observed motion on a set of detected and tracked points over an image sequence. In this sense, the goal of this work is determining the best performing methods to recover egomotion under ADAS-like sequences. To this end, we evaluate the performance of some proposals available in the literature with a synthetic dataset. This dataset was defined with the aim of simulating a typical camera motion mounted in a vehicle.

The paper is organized as follows. In Sec. 2, we study the relation between two views. In Sec. 3, we overview some egomotion algorithms. The definition of ADAS-like sequences and evaluation of different proposals to egomotion estimation using this dataset is given in Sec. 4. Finally, we present our conclusions.

2 Relation between two views

The discrete relation between two camera views is described by the well-known epipolar constraint [9], which arises from the relative motion between them. This constraint appropriately represents relatively large motion between two views. However, in our problem, the two views are related by small camera motion. For this reason, we reformulate the

epipolar constraint as a continuous one by modeling the camera motion as a rigid body motion, and considering the motion projection onto the image plane (called image flow field).

2.1 Rigid body motion

The camera motion is modeled as a rigid body motion. Rigid body motion is defined by a rotation matrix \mathbf{R} , and a translation vector \mathbf{t} at time t as

$$\mathbf{p}_t = \mathbf{R}_t \mathbf{p}_0 + \mathbf{t}_t . \quad (1)$$

This motion can be expressed by its velocity $\dot{\mathbf{p}}$, which is obtained by deriving Eq. (1) with respect to time

$$\dot{\mathbf{p}} = \boldsymbol{\omega} \times \mathbf{p} + \dot{\mathbf{t}} , \quad (2)$$

where $\dot{\mathbf{p}}$ is the velocity vector that describes the 3D motion, with an angular velocity vector $\boldsymbol{\omega}$ and a translational velocity vector $\dot{\mathbf{t}}$.

2.2 Image flow field

Assuming the camera and scene as rigid bodies, the image flow field is the projection of the 3D velocity onto the image plane. The perspective projection of a 3D point $\mathbf{p} = [p_x, p_y, p_z]^T$ onto the 2D point $\mathbf{q}_H = [q_x, q_y, f]^T$ (in homogeneous coordinates), is given by $\mathbf{q}_H = f \frac{\mathbf{p}}{p_z}$, where f is the focal length. Differentiating with respect to time and operating on this equation, the image flow vector in components is obtained as

$$\begin{bmatrix} \dot{q}_x \\ \dot{q}_y \end{bmatrix} = \begin{bmatrix} \frac{\dot{t}_x f - \dot{t}_z q_x}{p_z} + (\omega_y f - \omega_z q_y - \frac{\omega_x q_x q_y}{f} + \frac{\omega_y q_x^2}{f}) \\ \frac{\dot{t}_y f - \dot{t}_z q_y}{p_z} + (-\omega_x f + \omega_z q_x + \frac{\omega_y q_x q_y}{f} - \frac{\omega_x q_y^2}{f}) \end{bmatrix} \quad (3)$$

The image flow is the sum of two component vectors: one that only depends on $\dot{\mathbf{t}}$, and another only depends on $\boldsymbol{\omega}$. Notice that the translational component length is inversely proportional to the point depth, while the rotational component does not depend on it.

Image flow field is an ideal concept. It can only be approximated by computing the optical flow, or, partially, by tracking interest points [12, 2].

2.3 Continuous relation between two views

Under slow camera motion or high-frame rate, the relative motion between two consecutive views can be assumed as continuous. In this case, a differential epipolar constraint is defined, which finds a relation between a point \mathbf{q} and its velocity $\dot{\mathbf{q}}$.

By operating on Eq. (3), a bilinear constraint on $\boldsymbol{\omega}$ and $\dot{\mathbf{t}}$ that does not depend on the point depth is

$$(\dot{\mathbf{t}} \times \mathbf{q}_H)^T (\dot{\mathbf{q}}_H - (\boldsymbol{\omega} \times \mathbf{q}_H)) = 0 . \quad (4)$$

By algebraic manipulations, Eq. (4) is rewritten as

$$\dot{\mathbf{q}}_H^T [\dot{\mathbf{t}}]_{\times} \mathbf{q}_H + \mathbf{q}_H^T [\boldsymbol{\omega}]_{\times} [\dot{\mathbf{t}}]_{\times} \mathbf{q}_H = 0 , \quad (5)$$

which is the differential epipolar constraint.

3 Egomotion estimation

The egomotion problem concerns the estimation of the 3D rigid camera motion along a sequence. The aim is estimating $\dot{\mathbf{t}}$ and $\boldsymbol{\omega}$ from image flow observed in subsequent frames.

Egomotion methods can be classified according to the kind of used information. Depending on whether they use point correspondences or optical flow, egomotion methods are discrete [13] or differential [1]. In addition to these two classes, direct methods compute egomotion without the need of matching between views [6].

While the two first ones depend on the point matching accuracy, direct methods are based only on the brightness constraint constancy, which make them less robust in practice. Then, we study the two first kind of methods. In the following sections, the evaluated methods are described.

3.1 Methods based on epipolar constraint

The most common method is the 8-point algorithm [9] that estimates an essential matrix relating a pair of calibrated views from 8 or more point matches by solving a set of linear equations by least-squares. The original algorithm is sensitive to noise, but applying a normalization of points coordinates leads to significant improvements [5]. Algorithms handling less than 8 points have been developed, but their poor performance with respect to the 8-point algorithm have been shown in [10].

3.2 Methods assuming continuous motion

Based on the image flow field, different methods have been derived to compute $\dot{\mathbf{t}}$ and ω . Some of these are described here.

Bilinear optimization method Based on Eq. (4), $\dot{\mathbf{t}}$ and ω are computed, solving the following nonlinear system of equations by optimization [3]

$$\dot{\mathbf{t}}^T (\mathbf{A}\omega + \mathbf{b}) = 0, \quad (6)$$

where, assuming a unit focal length,

$$\mathbf{A} = \begin{bmatrix} q_y^2 + 1 & -q_x q_y & -q_x \\ -q_x q_y & q_x^2 + 1 & -q_y \\ -q_x & -q_y & q_x^2 + q_y^2 \end{bmatrix}, \text{ and}$$

$$\mathbf{b} = \begin{bmatrix} \dot{q}_y \\ -\dot{q}_x \\ -q_x \dot{q}_y + \dot{q}_x q_y \end{bmatrix}.$$

From Eq. (6), ω can be defined as a function of $\dot{\mathbf{t}}$

$$\omega = ((\dot{\mathbf{t}}^T \mathbf{A})^T \dot{\mathbf{t}}^T \mathbf{A})^{-1} (\dot{\mathbf{t}}^T \mathbf{A})^T \dot{\mathbf{t}}^T \mathbf{b}. \quad (7)$$

Substituting the estimated ω back into the Eq. (6) gives a nonlinear constraint on $\dot{\mathbf{t}}$. Then, $\dot{\mathbf{t}}$ is estimated by minimizing this nonlinear constraint in an iterative procedure, subject to be unitary since it only can be estimated up to a scale factor. Once $\dot{\mathbf{t}}$ is determined, ω is computed from Eq. (7).

Linear subspace method Based on Eq. (4), the idea of this method is reducing an initial nonlinear system of equations to a linear one by discarding nonlinear constraints [7].

Applying simple operations on Eq. (3), the image flow vector can be expressed as

$$\dot{\mathbf{q}} = \frac{1}{p_z} \mathbf{A} \dot{\mathbf{t}} + \mathbf{B} \omega, \quad (8)$$

where, assuming a unit focal length,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & -q_x \\ 0 & 1 & -q_y \end{bmatrix}, \text{ and}$$

$$\mathbf{B} = \begin{bmatrix} -q_y q_x & 1 + q_x^2 & -q_y \\ q_y^2 - 1 & -q_x q_y & q_x \end{bmatrix}.$$

The vector $\dot{\mathbf{t}}$ is estimated through an overdetermined linear system with $N - 6$ equations [7].

Once $\dot{\mathbf{t}}$ is computed, the depth from Eq. (8) can be eliminated, leaving a linear constraint for ω . To this end, a unit vector \mathbf{d} can be defined such that it is perpendicular to $\mathbf{A} \dot{\mathbf{t}}$.

Multiplying both terms of Eq. (8) by \mathbf{d}

$$\mathbf{d}^T \dot{\mathbf{q}} = \mathbf{d}^T \mathbf{B} \omega,$$

we found the least square solution for ω as

$$\omega = ((\mathbf{d}^T \mathbf{B})^T \mathbf{d}^T \mathbf{B})^{-1} (\mathbf{d}^T \mathbf{B})^T \mathbf{d}^T \dot{\mathbf{q}}. \quad (9)$$

Differential epipolar constraint methods The differential epipolar constraint formulation is parallel to the discrete case, then a similar process to the eight-point algorithm is used [8]. An overdetermined system of equations is formulated, and solved by least-squares minimization. This solution is systematically biased, then in [8] a “renormalization” step is proposed to subtract an estimate of the output bias from the solution. Other resolution schemes using this constraint are compared in [1], being [8] (with renormalization process) the best performing one.

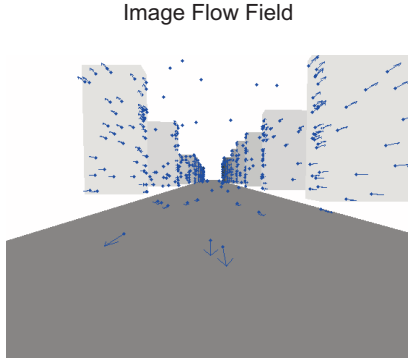


Figure 1: 3D Scene projection and image flow.

4 Experimental assessment

In this section, we report the experiments done to quantify the performance of different egomotion methods using synthetic data.

Inspired in the comparative work of [11], we adopt a similar methodology to quantify algorithms. They generate random clouds of 3D points and compute its 2D projections, and the image flow vectors corresponding to a particular 3D motion. Then, for each algorithm, egomotion is estimated in one thousand trials, quantifying its accuracy from the dissimilarity observed with respect to the ground truth motion. In our case, instead of random clouds of points, we generate random points configurations similar to the ones obtained from a camera mounted in a car, moving in a typical driving environment. The scene contains approximately 700 points randomly generated with a normal distribution, and placed on the road, objects, and plane at infinite. The road length is about 500 meters (m), and over it we generate a few number of points, because, in real situations, a road does not have sufficient texture to detect many points. We generate the most number of points in objects of different random sizes located on sides of the road, according to what we have seen experimentally. A small number of points are generated in a plane perpendicular to the road, which is placed at infinity, emulating the distant structures

observed in real sequences above the horizon line. Occlusions between different elements in the scene are managed by the z-Buffer algorithm.

Once the scene is generated, it is projected onto an image plane by using a pinhole camera model. The image flow vectors are produced by a rigid body motion of the camera with a translation of 1 m/frame on Z axis and a rotation of 2° /frame on the X axis, since these are the dominant translation and angular variation in ADAS. Assuming a frame rate of 25 frame/s, these magnitudes correspond to a real-life situation with a car moving at 90 km/h, with some shifts in the camera orientation due to the effect of the suspension system. In Fig. 1 an example of image flow caused by a translational motion over Z axis is shown.

Zero-mean Gaussian noise of various amounts is added to each image flow vector, to simulate errors in the optical flow computation. The noise levels are 0.05, 0.15, 0.25, 0.5, and 0.7 pixels. These values can be seen as the localization accuracy of the point tracked between frames.

Using the same criteria than [11], to quantify the algorithm accuracy, we measure for each algorithm the mean error of the estimates in one thousand different scenes. The mean error $\mu_{\dot{\mathbf{t}}}$ in the $\dot{\mathbf{t}}$ estimation is quantified as the angle (in degrees) between the true translation direction $\dot{\mathbf{t}}$, and the average of the translation direction $\bar{\mathbf{t}}$ of all trials, that is

$$\mu_{\dot{\mathbf{t}}} = \cos^{-1}(\bar{\mathbf{t}}^T \dot{\mathbf{t}}) ,$$

since the dot product is $\bar{\mathbf{t}}^T \dot{\mathbf{t}} = |\bar{\mathbf{t}}||\dot{\mathbf{t}}|\cos(\alpha)$, and $|\dot{\mathbf{t}}| = |\bar{\mathbf{t}}| = 1$ because the translation is only recovered up to a scale factor.

To quantify the mean rotation error, we use the difference angle between the true rotation ω and the mean of estimated rotations $\bar{\omega}$. For this purpose, rotation matrices \mathbf{R} and $\bar{\mathbf{R}}$ for both ω and $\bar{\omega}$ are built. The product between \mathbf{R} and $\bar{\mathbf{R}}$ is an identity matrix when both are equal. Thus, the difference between both matrices is defined as $\Delta\mathbf{R} = \mathbf{R}^T \bar{\mathbf{R}}$. In Euler terms, $\Delta\mathbf{R}$ can be characterized by an axis unit vector and an angle. This

angle is used as the mean rotation error. Since $\text{trace}(\mathbf{R}) = 1 + 2\cos(\alpha)$, then the angle is equal to

$$\mu_{\mathbf{R}} = \cos^{-1} \left(\frac{1}{2} (\text{trace}(\Delta \mathbf{R}) - 1) \right) .$$

We choose this approach because it provides a compact error measure that facilitates the evaluation of the algorithms, i.e., a scalar error value for $\dot{\mathbf{t}}$ and ω , respectively. Another option could be the mean of each component vector of the estimated $\dot{\mathbf{t}}$ and ω , but this makes more difficult the analysis.

We evaluate the egomotion algorithms surveyed in Sec. 3, which estimate both $\dot{\mathbf{t}}$ and ω , in order to determine which algorithms have the best performance. These algorithms have been selected due to the following reasons

- With respect to discrete epipolar constraint, the results of the comparative study fulfilled by [10] indicate that for sideways motion the five-point algorithm conduces to the best estimation with respect to linear, iterative, and robust tested methods. However, during forward motion—which is of particular interesting in our case—, normalized eight-point algorithm (8pts) greatly overcomes all the methods, corroborated also by [5]. Thus, we include the 8pts algorithm.
- As representative of the linear algorithms using the differential essential matrix we test the infinitesimal eight-point algorithm (KA) and its renormalized version (KB) of Kanatani’s proposal, since these two algorithms have been shown as ones of the best performing method in [1].
- Due to its particular approach to deal with the bilinear constraint, we test the linear subspace method of Jepson and Heeger (J&H).
- The optimization-based method of Bruss and Horn (B&H) because exhibited the best performance of all (linear and nonlinear) methods tested by [11].

All algorithms are programmed in Matlab. Some of them (i.e., KA, KB, B&H, and J&H) have been taken from the comparative survey and toolbox given by [11].

Fig. 2 shows the results of this experiment. The best result is achieved by B&H since it entails an iterative optimization process. KB overcomes all linear approaches because the renormalization process removes the bias caused by noise. J&H is noise sensitive since it uses only the linear constraints. The 8pts algorithm is based on discrete epipolar constraint which does not work well when the camera motion is small.

5 Conclusions

In this work, we have reviewed current egomotion methods in the literature. Also, in order to compare the surveyed algorithms, we have defined a synthetic dataset that allow us to simulate ADAS-like sequences. Using this dataset, we have evaluated the considered algorithms to determine those that have better performance in our context. As conclusion of our experiments, we can show that the best nonlinear and linear performing methods are B&H and KB, respectively. B&H is based on bilinear constraint, and applies an iterative optimization process to find egomotion parameters by solving a nonlinear system of equations. On the other hand, KB is based on differential epipolar constraint, and egomotion is recovered by solving a linear system of equations. Based on the evaluation results, these two algorithms could be used as a performance baseline, in order to judge the validity of other egomotion methods in our future work.

6 Acknowledgment

This work has been partially funded by the Autonomous University of Barcelona, and Spanish MEC research projects Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and TRA2007-62526/AUT.

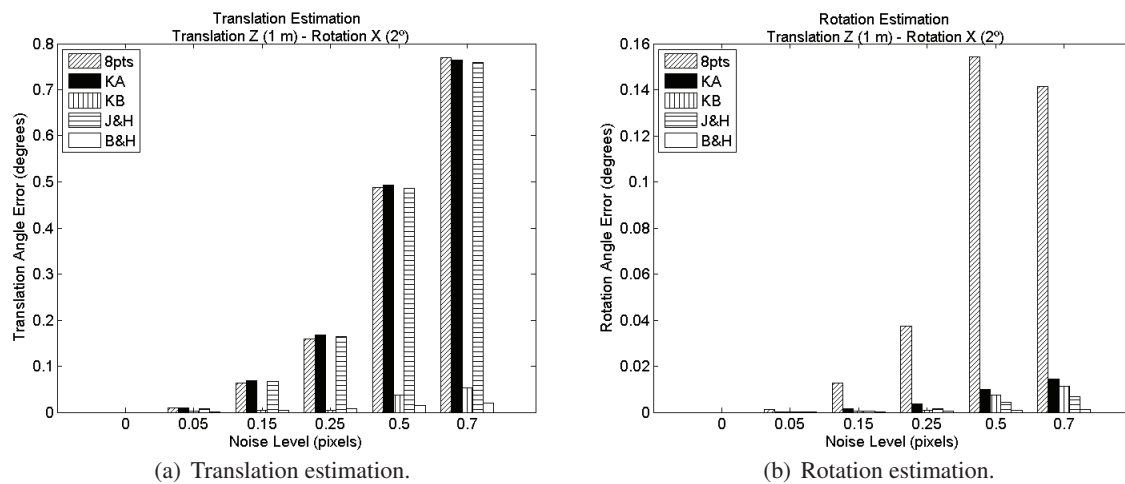


Figure 2: Results of translation and rotation estimation.

References

- [1] X. Armangué, H. Araújo, and J. Salvi. A review on egomotion by means of differential epipolar geometry applied to the movement of a mobile robot. *Pattern Recognition*, 36(12):2927–2944, 2003.
- [2] S. Beauchemin and J. Barron. The computation of optical flow. *ACM Computing Surveys*, 27(3):433–466, 1995.
- [3] A. Bruss and B. Horn. Passive navigation. *Computer Vision, Graphics, and Image Processing*, 21:3–20, 1983.
- [4] D. Cheda, D. Ponsa, and A. López. Towards improving an on-board monocular multiple vehicle 3D tracking system by exploiting egomotion. *Current Challenges in Computer Vision*, 1–6, 2008.
- [5] R. Hartley. In defense of the eight-point algorithm. *Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997.
- [6] M. Irani and P. Anandan. About direct methods. *Intl. Workshop on Vision Algorithms: Theory and Practice*, 267–277, 2000.
- [7] A. Jepson and D. Heeger. Simple method for computing 3D motion and depth. *Intl. Conference on Computer Vision*, 96–100, 1990.
- [8] K. Kanatani. 3D interpretation of optical flow by renormalization. *Intl. Journal of Computer Vision*, 11(3):267–282, 1993.
- [9] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.
- [10] D. Nistér. An efficient solution to the five-point relative pose problem. *Pattern Analysis and Machine Intelligence*, 26(6):756–777, 2004.
- [11] T. Tian, C. Tomasi, and D. Heeger. Comparison of approaches to egomotion computation. *Conference on Computer Vision and Pattern Recognition*, 315–320, 1996.
- [12] C. Tomasi and T. Kanade. Shape and motion from image streams: a factorization method - part 3 detection and tracking of point features. Technical report, 1991.
- [13] P. Torr and A. Zisserman. Feature based methods for structure and motion estimation. *Vision Algorithms: Theory and Practice, N° 1883 in LNCS*, 278–295, 2000.

Automotive Applications based on Video Alignment

Ferran Diego, Daniel Ponsa, Jose M. Álvarez, Joan Serrat and Antonio López

Computer Vision Center & Computer Science Dept., Edifici O, Universitat Autnoma de Barcelona, 08193 Cerdanyola, Spain

E-mail: {fdiego,daniel,jalvarez,joans,antonio}@cvc.uab.es

Abstract

We address the problem of synchronization of a pair of video sequences captured from moving vehicles and the spatial registration of all the temporally corresponding frames in order to compute their pointwise differences. Video synchronization has been attempted before but often assuming restrictive constraints like fixed or rigidly attached cameras, simultaneous acquisition, known scene point trajectories, etc. which limit its practical applicability. We have solved the most difficult problem of independently moving cameras which follow a similar trajectory, based only on the fusion of image intensity and GPS data information. Another novelty of our approach is the probabilistic formulation and the combination of observations from these two sensors, which have been revealed as complementary. We have employed video alignment in three automotive applications: vehicle detection, night time outdoor surveillance and road segmentation algorithms.

Keywords: ADAS, video alignment, DBN

1 Introduction

Consider the following scenario. A vehicle is driven twice through a certain circuit, following approximately the same trajectory. Attached to the windshield screen, a forward facing camera records one video sequence for each of the two rides. Imagine that, somehow, we are able to sub-

tract pixelwise the two sequences. That is, for each frame of, say, the first sequence, we get to know which is the corresponding frame in the second sequence, in the sense of being the camera at the same location. In addition, suppose we succeed in spatially aligning every such pair of frames, so that they can be properly subtracted to build the frames of the difference video. What would it display? Moving objects, objects present in only one of the video sequences, like pedestrians and on road vehicles, or changes in the scenario, which is useful for vehicle detection and night-time outdoor surveillance.

Video alignment or matching consist on the simultaneous correspondence of two image sequences both in the time and space dimension. The first part, which we refer to as synchronization, aims at estimating a discrete mapping $c(t_o) = t_r$ for all frames $t_o = 1 \dots n_o$ of the observed video, such that the reference frame at t_r maximizes some measure of similarity with observed frame at t_o , among all frames of the reference sequence. In the former scenario, we assume this will happen when the location where reference frame was recorded is the closest to that of observed frame. The second part, registration, takes all corresponding pairs and warps a frame so that it matches with its corresponding frame, according to some similarity measure and a spatial deformation model.

1.1 Previous work on Video Alignment

Several solutions to the problem of video synchronization have been proposed in the literature. Here we briefly review those we consider the most significant. This is relevant to put into context our work, but also because, under the same generic label of synchronization, they try to solve different problems. The distinction is based on the input data and the assumptions made by each method.

The first proposed methods assumed the temporal correspondence to be a simple constant time offset $c(t_o) = t_o + \beta$ [4, 10, 8] or linear $c(t_o) = \alpha t_o + \beta$ [1, 9], to account for different camera frame rates. More recent works [7, 6] let it be of free form. Clearly, the first case is simpler since just one or two parameters have to be estimated, in contrast to a curve of unknown shape.

Concerning the basis of these methods, most of them rely on the existence of a geometric relationship between the coordinate systems of frames *if* they are corresponding: an affine transform [9], a plane-induce homography [1], the fundamental matrix [8], the trifocal tensor [4], and a deficient rank condition on a matrix made of the complete trajectories of tracked points along a whole sequence [6, 10]. This fact allows either to formulate some minimization over the time correspondence parameters (e.g. α β) or at least to directly look for *all* pairs of corresponding frames. Again, the cases in which this geometric relationship is constant [1, 9, 6], for instance because the two cameras are rigidly attached to each other, are easier to solve. Other works [7, 4, 3, 5] address the more difficult case of independently moving cameras, where no geometric relationship can be assumed beyond a more or less overlapping field of view.

Each method needs some input data which can be more or less difficult to obtain. For instance, feature-based methods require tracking one or more characteristic points along the two whole sequences [1, 6, 10, 8], or points and lines in three sequences [4]. In contrast, the so-called direct methods are based just on the image intensity or color

[1, 7, 9] which in our opinion is better from the point of view of practical applicability.

Like still image registration, video alignment has a number of potential applications. It has been used for visible and infrared camera fusion and wide baseline matching [1], high dynamic range video, video mating and panoramic mosaicing [7], visual odometry [5], action recognition [9] and loop closing detection for SLAM [3].

1.2 Objective

We have employed video alignment as a tool to spot differences between two videos, recorded by on-board cameras in the context of three applications: 1) vehicle detection, 2) night time outdoor surveillance and 3) automatic ground-truthing for road segmentation. Specifically, for vehicle detection, we perform a sort of pre-detection, that is, to select regions of interest on which supervised classifiers should be applied. One of the difficulties of such classifiers have overcome is the variability of such objects in size and position within the image. In addition, recent results with state-of-the-art methods like boosting indicate that a fixed, known background greatly simplifies the complexity of the problem. To our knowledge, this is a novel approach to the problem. Night time outdoor surveillance compares directly successive sequences because their differences are potential signs of intruders or unexpected events, and it could help a private guard to notice slight differences. Road segmentation transfers a previous road segmentation to a new sequence if it drives by the same track but following the similar trajectory.

2 Video alignment

We formulate the video synchronization problem as a labeling problem. A list of n_o labels $\mathbf{x}_{1:n_o} = [x_1 \ x_t \ x_{n_o}]$ has to be estimated. Each label $x_t \in \{1 \dots n_r\}$ is the number of the frame in the reference video corresponding to the t^{th} frame of the observed sequence. To perform that, we rely on the available observations $\mathbf{y}_{1:n_o}$, namely, the frames

themselves of the observed sequence and the GPS data associated to them. We pose this task as a maximum a posteriori Bayesian inference problem,

$$\mathbf{x}_{1:n_o}^{MAP} \propto \arg \max_{\mathbf{x}_{1:n_o} \in \mathcal{X}} P(\mathbf{y}_{1:n_o} | \mathbf{x}_{1:n_o}) P(\mathbf{x}_{1:n_o})$$

The prior $P(\mathbf{x}_{1:n_o})$ assumes that the transition probabilities are conditionally independent given their previous label values. In addition, the constraint that the vehicle can stop but not reverse its motion direction in both the reference and observed sequences implies that labels x_t increase monotonically. We also assume that the likelihood of the observations $\mathbf{y}_{1:n_o}$ is independent given their corresponding label values and also between observations because they come from different sensors. From these dependencies between variables, it turns out that our problem is one of MAP inference in a Hidden Markov model. Hence, we can apply the well-known Viterbi algorithm to exactly infer $\mathbf{x}_{1:n_o}^{MAP}$. For a detailed description we refer the reader to [2].

Once a list of pairs of corresponding frame numbers (t x_t), $t = 1 \dots n_o$ is obtained as a result of synchronization, we suppose a conjugate rotation between each such pair because our assumption is that corresponding frames are recorded at the same position or very close to each other. In order to obtain the deformation parameters, we use the additive forward extension of the Lucas–Kanade algorithm using the sum of squared linearized differences (i.e., the linearized brightness constancy) as a error measure.

3 Applications

3.1 Vehicle detection

Vehicle detection using video alignment spots differences between a pair of videos which we could subtract pixelwise, but, of course, *after* they were recorded. Instead of calculating the MAP infer-

ence on a network formed by observation and hidden nodes representing all the observed video sequence as we described above, a truly dynamic Bayesian network is built on-line and a different type of inference is carried out called fixed-lag smoothing. Once we have registered and subtracted the corresponding frames, the detection is done by applying simple morphological operations and thresholds. Figure 1 shows an example of vehicle detection. However, since still pictures are not the best way to present the results, we have built a web page where original and result videos can be viewed at [11].

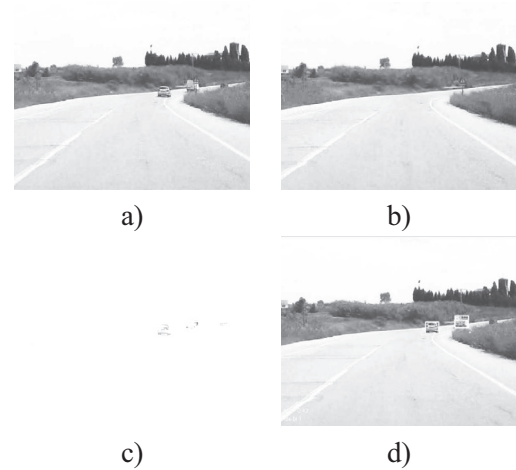


Figure 1: VD: a) observed frame, b) reference frame, c) frame spotting differences and d) an example of vehicle detection.

3.2 Night time Outdoor surveillance

Night time outdoor surveillance consists on guarding private buildings with a vehicle following a specific patrol. The main difficulty is to notice slight differences that can have been occurred between successive patrols. Therefore, this specific surveillance using video alignment consist on aligning consecutive sequences of what the driver sees along a same track in order to emphasize the differences happened. Figure 2 shows an example of synchronized surveillance frames. Again, some



Figure 2: Outdoor surveillance: a) observed frame, b) reference frame and c) differences marked by rectangles.

result videos can be viewed at [11].

3.3 Road segmentation

Road segmentation is an essential functionality for supporting advanced driver assistance systems *ADAS* such as road following or vehicle detection and tracking. Significant efforts have been made in order to solve this task using video techniques. Many approaches use ad-hoc mechanisms to perform the road segmentation. We present an alternative procedure that consists on manually labelling the frames of one video sequence and then is able to transfer this road segmentation when the vehicle drives along the same track following the similar trajectory at different time. Figure 3 shows an example of road segmentation. Again, some result videos can be viewed at [11].

Acknowledgments

This work was supported by Spanish Ministry of Education and Science (MEC) under Project TRA2007-62526/AUT, Research Program Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and MEC grant AP2007-01558 (first author).

References

- [1] Y. Caspi and M. Irani. Spatio-temporal alignment of sequences. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(11):1409–1424, 2002.
- [2] J. S. A. L. F. Diego, D. Ponsa. Video alignment for difference-spotting. In *Proceedings of the ECCV workshop on Multi-camera and Multi-*

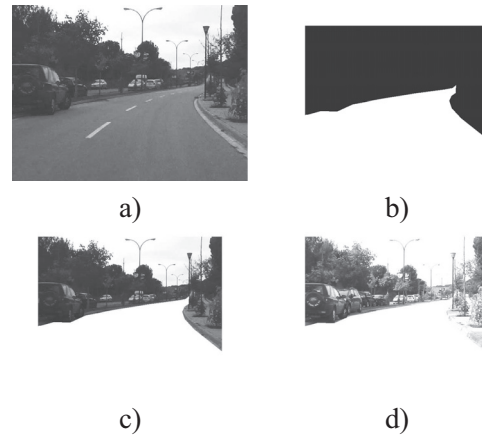


Figure 3: RS: a) reference frame, b) ground truth of the reference frame, c) fusion of reference and manual segmentation of the frame and d) fusion of the observed and transferred manual segmentation.

modal Sensor Fusion Algorithms and Applications, 2008.

- [3] K. L. Ho and P. Newman. Detecting loop closure with scene sequences. *Int. J. Computer Vision*, 74(3):261–286, 2007.
- [4] C. Lei and Y. Yang. Trifocal tensor-based multiple video synchronization with subframe optimization. *IEEE Trans. Image Processing*, 15(9):2473–2480, 2006.
- [5] A. Levin and R. Szeliski. Visual odometry and map correlation. In *Proc. Computer Vision and Pattern Recognition*, pages 611–618. IEEE Computer Society, 2004.
- [6] C. Rao, A. Gritai, and M. Shah. View-invariant alignment and matching of video sequences. In *In ICCV*, pages 939–945, 2003.
- [7] P. Sand and S. Teller. Video matching. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 22(3):592–599, 2004.
- [8] T. Tuytelaars and L. V. Gool. Synchronizing video sequences. *cvpr*, 01:762–768, 2004.
- [9] Y. Ukrainitz and M. Irani. Aligning sequences and actions by maximizing space-time correlations. In *European Conf. on Computer Vision*, volume 3953 of *Lecture Notes in Computer Science*, pages 538–550. Springer, 2006.
- [10] L. Wolf and A. Zomet. Wide baseline matching between unsynchronized video sequences. *Int. Journal of Computer Vision*, 68(1):43–52, 2006.
- [11] www.cvc.uab.es/ADAS/projects/sincro/BVMA/.

Synthetic Urban Development to Evaluate Pedestrian Detection Methods

Javier Marín and Antonio López

ADAS, Computer Vision Center, Universitat Autònoma de Barcelona, Spain
E-mail: jmarin@cvc.uab.es

Abstract

The number of driving accidents in which pedestrians are involved motivates research efforts to develop pedestrian protection systems. Using an image acquisition system together with Computer Vision techniques is a possible solution for detecting pedestrians in front of a vehicle. In this context, the research effort during the last years has focused on developing classifiers that can say if there is a pedestrian inside a given image region. A key component to develop such classifiers is the dataset in use. Our main interest in this work is to explore the possibility of evaluating pedestrian detection systems in virtual urban environments, which eventually would turn out in better classifiers. For answering such a question, the experiments are done in our own virtual scenario using one of the commonly presented framework in the literature trained utilizing the most relevant real databases publicly available. Obtained results suggest that this is an interesting framework to explore in deep.

1 Introduction

Nowadays in the Advanced Driver Assistance Systems (ADAS) research area, cameras are used for many reasons such as cost, information richness, etc. Thus, Computer Vision (CV) plays an im-

portant role in the development of intelligent systems that combine sensors and algorithms to work in real-time to be able to assist the driving activity. Some examples of assistance applications are: lane departure warning, collision warning, automatic cruise control, pedestrian protection, headlights control, etc.

As a result of the large number of accidents in which pedestrians and vehicles are involved, research on pedestrian protection systems (PPS) has become a relevant research field in the automotive industry.

Pedestrian detection represents a very challenging task not only because of the pedestrian and background diversity but also illuminance changes, weather conditions, partially occluded targets and elements, etcetera. Therefore, there are many different classification techniques to detect pedestrians as HOG+LinSVM [2], Haar+AdaBoost [9], HOG+LatSVM [5], etc. Also, they need to learn from examples, this is that the algorithm needs previous information to train, and then validate and test the classifiers in different images which have to be different from the ones used in the training process. The stage of acquiring these images and making the annotation in all the frames is an expensive work.

In the state of the art of pedestrian detection, a lot of different important pedestrian datasets have been presented like INRIA [2] and recently Daim-

ler [7].

The task of evaluating different pedestrian detection methods have been done by using real sequences in which the same database for training and testing was used. Hence, the idea of using virtual scenarios to evaluate, not just tracking methods, but new pedestrian detection methods trained in real sequences.

For reasons of cost, virtual simulation techniques are a classic in Robotics. Before selling or using a robot in the real world, the companies use first prototypes in virtual environments to see how they will evolve.

Other research areas related to surveillance systems that use synthetic techniques are: motion detection and tracking.

The use of new virtual technologies have produced a lot of advances in Medical Imaging. Magnetic Resonance Imaging (MRI) simulation; Synthetic Brain Imaging, these and other virtual techniques have become indispensable tools to streamline diagnosis and treatment time.

Realistic pedestrian models appear in modern video games. Indeed, video games have evolved from naive world representations into more realistic scenarios thanks to the use of physics, real textures, different kind of illuminances, weather conditions and artificial intelligence (AI). Also, different game engines offer the possibility of creating urban virtual scenarios where the user controls everything: people and cars movement, physics, weather conditions, illumination, number of targets in the world, etcetera.

Accordingly, we propose to explore the use of virtual urban scenarios for evaluating pedestrian detection methods. In this work we focus on model building. This virtual scenario will be created according to the necessities in order to get a realistic urban environment.

The remained of the report is as follows. In section 2 we introduce the most relevant existent pedestrian datasets, as well as one of the commonly used pedestrian detection method. In section 3, we explain in more detail the virtual devel-

opment and its requirements, and in section 4 how to obtain the synthetic data. After that, in section 5 we assess the viability of using virtual scenarios for pedestrian detection. Finally, in section 5 we summarize conclusions and draw our future work.

2 Literature Review

2.1 Pedestrian datasets

INRIA. Currently, the INRIA person dataset [2] is the most widely used in pedestrian detection. In these images people are usually standing, but appear in any orientation and on a wide variety of background image including crowds. The data set contains images from several different sources, images from pedestrian datasets, images from personal digital image collection and some images from the web.

Daimler09. During the development of this Master thesis, Enzweiler et al. [7] have published their pedestrian dataset. The test set contains a sequence with more than 21.790 images with 56.492 pedestrian labels (fully visible or partially occluded), captured from a vehicle during a 27 min drive through urban traffic, at VGA resolution (640x480, uncompressed). As such, like [3] does, the dataset is realistic and about one order of magnitude larger than other datasets. The real number of useful pedestrian in the test set is not equal to the total number of pedestrian. However, this dataset is more adapted than other ones to PPS.

2.2 Pedestrian detection method

In the state of the art of Object Classification a wealth of methods have been proposed for PPS. Thanks to the recent surveys published by M. Enzweiler and D. Gavrilu [4] in 2008, by Gerónimo et al. [6] in 2009 and the paper published by Dollár et al. [3] in 2009, the wearisome task of summarizing, comparing and evaluating the differ-

ent methodologies and datasets present in literature has been made easier.

We are going to focus our interest in one of the most commonly used, which provide a good baseline to develop classifiers in our experiments.

In 2005, N. Dalal and B. Triggs [2] introduced a new pedestrian detection method. They demonstrated that using a linear SVM based on grids of Histograms of Oriented Gradient (HOG) as descriptors, significantly outperformed existing feature sets for human detection. Besides they studied that the influence of each stage of the computation on performance, concluding that fine-scale gradients, fine orientation binning, relatively coarse spatial binning, and high-quality local contrast normalization in overlapping descriptor blocks were all important for good results.

3 Virtual scenarios development

Half life 2 is a science fiction first-person shooter computer game developed by Valve Corporation [1]. This game has been critically praised for its advances in computer animation, sound, narration, computer graphics, AI, and physics. The main features of its engine that need to be remarked are High Dynamic Range (HDR) rendering, blended skeletal animation system, including inverse kinematics, configurable AI and physics, dynamic lighting and shadowing, significant source code access for mod teams, and a map compiler.

The Source SDK is a software development kit compiled by Valve Software that is used to create maps or mods for the Source engine. This tool provides the Hammer editor, which is the Valve Software's map creation program that is helpful to create new scenarios with roads, streets, buildings, cars, pedestrians, traffic signs, etc.

3.1 Virtual World Objects

The Source engine divides the world objects in two different categories:

- **Primitives.** This category as its name suggests is the basic one, and refers to a brush-based object that conforms to a common shape. These structures include blocks, arches, cylinders, spheres, spikes, torus and wedges. The user can control the size, the shape and the texture used in it, and for the non-polyhedral shapes all the parameters can be managed by the user (wall width, number of sides, arc angle, start angle, rotation parameters, etc). All of these structures are used to develop all the basic elements in the scenario such as buildings, roads, walls, side walks, etcetera.



Figure 1: Primitive structures

- **Entities.** These are the opposite of world brushes (or primitives). This is, they do not provide any structure to the Virtual World. Objects like a door, a control panel or an elevator that interact or act with other elements in the map are entities.

3.2 Virtual World Mapping

After explaining the different elements related to the virtual scenarios, we are going to explain the mapping process sequentially.

Drawing the basic map structure. The goal here is to create the urban environment (buildings, walls, sidewalks, roads, streets, etc). The simple objects categorized as Primitives are needed to create these basic structures. Then, when the structures are made, the user selects for each one the texture (Hammer provides a huge database of textures). This step is one of the hardest, because a graphic design is needed before start. The Valve community and a lot of sites provide urban maps, but not all of them can be modified. However, the

user has the possibility of decompile (there is free software available) them and learn how are they are made.

Adding entities. After the urban structure is made, we will need more urban elements if we want to obtain a more realistic world (street lights, traffic signs, benches, trash, fences, etc). Once the urban is made, we place the pedestrians and the vehicles in the strategy positions (it is important to remember how they are going to interact with other entities before allocate them in each position).

Illumination, ambient and brightness. Maybe the easiest step. Only three entities have to be taken into account, the one that controls the illumination, `light_environment`, the one that controls the shadows, `shadow_control` (this one is needed for shade dynamic entities) and the entity `sky_camera` that plays an important role, because the sky shown in the map, the fog and the dynamic shadowing are managed by it.

Dynamic world behaviour development. Now, the AI, the entity paths and the physics for each dynamic entity involved in the map have to be defined. Due we want to reproduce an urban environment, the only dynamic entities we are going to model are the pedestrian and vehicle models.

Map compilation. This process is divided in three steps.

- Run the Valve Binary Space Partitioning (BSP or VBSP). All the brush faces are located in the map and checked in relation to how they interact with the rest in the map.
- Run the Visibility (VIS). The visibility of each player is determined in order to help the game when maximizing the rendering.
- Run the Valve Radiance (RAD). In this last process the light is properly added to the map.

High Dynamic Range (HDR) can be selected in this step.

Along all these stages it is possible to tune the process using different settings.

4 Virtual data adquisition

We can obtain virtual data from synthetic scenarios utilizing the software offered by ObjectVideo (OV [8]). OV is a private company that provides intelligent video software for security, giving also support to the advancement of CV research through ObjectVideo's Virtual Video Tool (OVVV [10]). The OVVV Tool generates realistic videos from simulated cameras in an interactive virtual world. This tool is free and is based on a modification (aka 'mod') of Half-Life 2. With this tool it is possible to get sequences or images from the Virtual World in an automatic way obtaining the foreground ([10], section 3.4) pixel by pixel of all the targets in the scene including partially occluded targets. Besides, it also provides continuous tracking of all the pedestrians in the scene (even if they are totally occluded or they disappear for a while). This software opens up a lot of possibilities in many research fields.

This Virtual Video mod uses a socket-based network protocol to receive PTZ¹ commands and send video frames to client applications (see figure 2). These are the three different ways to connect:

DirectShow Filter². You can start streaming real-time virtual video using this filter without any other modifications.

Virtual Video C Library. It allows applications to retrieve frames without using the DirectShow filter, and to send PTZ commands

¹Pan, Tilt, and Zoom (PTZ), refer merely to features of specific surveillance cameras.

²The Microsoft DirectShow application programming interface (API) is a media- streaming architecture for Microsoft Windows.

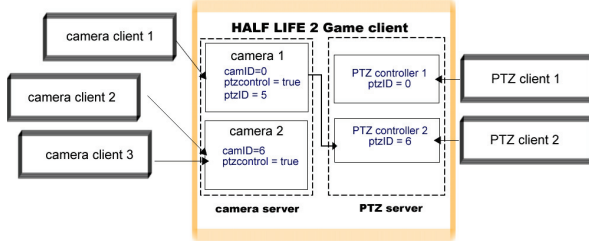


Figure 2: Status and connections example.

to the Half-Life 2 mod to control the camera view in real time.

Applications written in any programming language and on any platform that supports sockets can use these low-level protocols.

In our case, the student M. Fenés, supervised by the Dr. A. López, has developed a software application, which provides all the features introduced above, using the Virtual Video C Library.

5 Experiments and Discussion

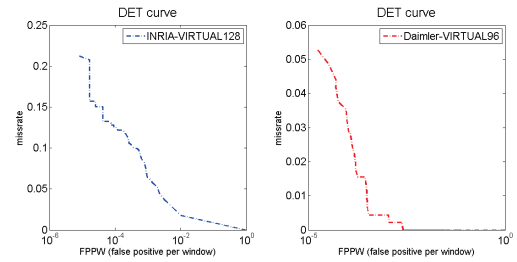
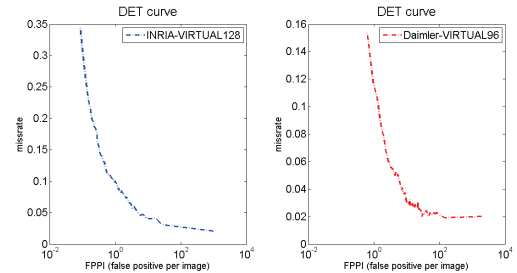
In this section we evaluate two different pedestrian detection methods. Both of them using HOG+linSVM [2] as classification method. One is trained using the INRIA dataset and the other one with the Daimler dataset.

For the virtual testing sets two different annotations are used. For the INRIA classifier, the virtual set takes into account pedestrian from 92 pixels tall. While, for the Daimler classifier, the minimum size is 80 pixels. This is because of the different parameters used for each real dataset, [2] and [7], respectively.

Virtual Evaluation using INRIA. In this evaluation we assess the classifier trained with the INRIA dataset, using the parameters introduced in [2]. For this evaluation, the virtual testing set used is named Virtual128, which has annotated pedestrian from 92 pixels tall.

Virtual Evaluation using Daimler. In this evaluation we assess the classifier trained with the Daimler dataset, using the parameters introduced in [7]. The virtual testing set used is named Virtual96, which has annotated pedestrian from 80 pixels tall.

In next figures we plot the results obtained doing the evaluation *per-window*, Fig. 3, and the evaluation *per-image*, Fig. 4.

Figure 3: Results obtained *per-window*.Figure 4: Results obtained *per-image*.

In Table 1 we summarize the most relevant numeric results extracted from the plotted curves. These results correspond to the ones used in the literature.

Virtual Evaluation		
	DET <i>per-window</i>	DET <i>per-image</i>
INRIA	0.13	0.10
Daimler	0.03	0.11

Table 1: Virtual Evaluation. DET values: the miss rate at 10^{-4} FPPW and the miss rate at 10^0 FPPI.

The results described in the Table 1, suggest that all the classifiers detect the virtual pedestrians nearly perfect, in particular the results obtained by the authors were worst than these obtained in the virtual scenario, this means that the virtual pedestrians are easier to detect. Thus, virtual pedestrians such as fat persons, tall persons, kids, women with skirt or persons carrying bikes are still necessary to increase the variability of our virtual urban scenario.

6 Conclusions and Future Work

In this paper we have explored the novel idea of using synthetic virtual scenarios for evaluating pedestrian detection methods being the main component a classifier. We have developed a synthetic urban environment and the framework to generate it.

We have evaluated two classifiers trained in different real datasets by using two different evaluation methods, *per-window* and *per-image*. Regarding to the experiments, new synthetic models have to be introduced to increase the pedestrian variability. The *per-window* and *per-image* evaluations reveal us the potential of the use of virtual scenarios.

References

- [1] Valve Software Corporation. Official site: <http://www.valvesoftware.com>.
- [2] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.
- [3] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, June 2009.
- [4] Markus Enzweiler and Dariu M. Gavrilă. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2008.
- [5] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multi-scale, deformable part model. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.
- [6] David Gerónimo, Antonio M. López, Angel D. Sappa, and Thorsten Graf. Survey on pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (in press), 2009.
- [7] S. Munder and D.M. Gavrilă. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.
- [8] ObjectVideo. Official site: <http://www.objectvideo.com>.
- [9] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *Int. J. Comput. Vision*, 38(1):15–33, 2000.
- [10] Geoffrey R. Taylor, Andrew J. Chosak, and Paul C. Brewer. Ovvv: Using virtual worlds to design and evaluate surveillance systems. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

Feature matching with graphical models for night vehicle detection

Jose Carlos Rubio and Joan Serrat

*ADAS, Computer Vision Center, Universitat Autònoma de Barcelona, Spain
E-mail:jcrubio@cvc.uab.es*

Abstract

Feature Matching consists of finding the most likely correspondences between two or more sets of points. We provide an exact solution to this problem, based on quadratic programming, and two approximations which find the maximum probability configuration in a graphical model. The first, interpreting the variables as associations between features, and the second modeling the variables as the features themselves. We finally show how the matching algorithm can be applied to multiple target tracking, carrying out several experiments with real sequences of vehicles at night.

Keywords: Feature Matching, Tracking, Belief Propagation, Driving Assistance

1 Introduction

Feature Matching (FM) is a widely studied problem in computer vision, which consists of associating a set of points extracted from one image to another set in a second image. Usually this is done by identifying certain attributes of the features such as color, gradients, or spatial coordinates, and finding the most likely correspondences between these characteristics.

Multiple Target Tracking (MTT) is the process of simultaneously track several targets, which is a critical component in many vision applications

such as video surveillance or visual navigation. The main difficulties when performing MMT are related with the events affecting the targets being followed: occlusions, a merging, a splitting, a target leaving the view field, or a new target entering in it.

This work tackles the problem of MTT by means of applying feature matching between points extracted from two consecutive frames of a video sequence. Specifically, we formulate the problem of FM probabilistically, as a Maximum a Posteriori (MAP) estimation in a graphical model. We compare two different approaches, which we call feature-oriented and association-oriented, principally inspired in the works of Caetano et. al [1, 2, 3] and Kolmogorov [4], respectively.

2 Motivation

This research was first motivated by the application developed by Antonio Lopez, namely, Nighttime Vehicle Detection for Intelligent Headlight control [5]. At nighttime, detecting vehicles using a camera requires identifying their head or tail lights. The main challenge of this approach is to distinguish these lights from reflections due to infrastructure elements. Tracking these blobs permits combining the beliefs of the lights classifier, and in the present work we take for granted that performing an accurate tracking of the blobs will increase the classification rates.



Figure 1: Example of the headlight control application

3 Past Works

Works like [4] model the matching problem as a minimization of an energy function of weighted terms where each component favors certain behavior of the matching. Eq. 1 shows the component-based objective function, which in [4] is minimized using a decomposition approach, finding sub-problems and optimizing each of them separately. The λ^{app} , λ^{occl} and λ^{geom} are scalar weights to ponder the contribution of each of the energy terms.

$$E = \lambda^{app} E^{app} + \lambda^{occl} E^{occl} + \lambda^{geom} E^{geom} \quad (1)$$

In Eq (1), the function $E^{app}(x)$ favors the appearance compatibility between two features using a distance between descriptors. The more similar the features are, the lower the cost related to the association. $E^{occl}(x)$ penalizes or favors the occlusions depending on the number of features in the source image S^{src} and the destination image S^{dst} . The last component $E^{geom}(x)$ keeps the geometric compatibility between pairs of features, based on the angle and the distance of two vectors defined by neighbor features.

In [4] the random variables are representing associations, and following this idea we developed one of our approaches to the problem, which we call *Association-Oriented*. An example of this can be seen in Figure 2. In the works [3, 1] the point correspondence is faced using graph matching. This is a versatile tool consisting of using

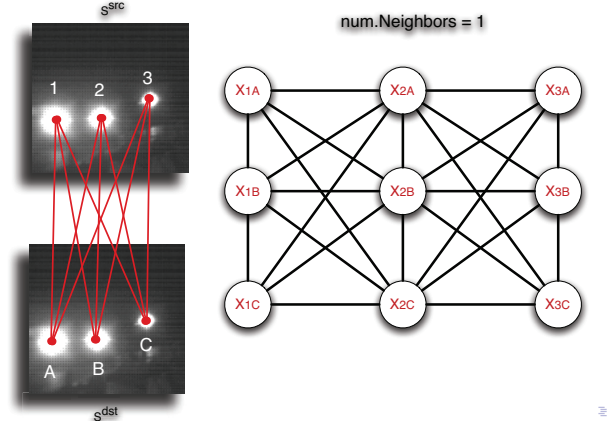


Figure 2: Example of Association-Oriented approach

a graph data structure to represent objects and its relations. In this work the random variables are modeling features from the source image, and FM is performed finding the most likely correspondence between the nodes of two graphs, namely the source image, and destination image. This approach gives name to our second model, called *Feature-Oriented* approach, and an example is represented in Figure 3.

3.1 Objective Function

The terms of Eq 1, are function one or two variables, and the cost associated. Both E^{app} and E^{occl} are function of only one variable, while E^{geom} operates with two variables, since depends on two associations to express geometric information. Distinguishing these linear and quadratic terms of the energy, we can express the energy in a compact way, as follows:

$$\arg \min_{\mathbf{x} \in \mathbf{A}} E(\mathbf{x} | \bar{\theta}) = \sum_{a \in A} \theta_a x_a + \sum_{(a,b) \in N} \theta_{ab} x_a x_b \quad (2)$$

The coefficient θ_a is the cost associated to the variable x_a , and θ_{ab} is the cost associated to the pair of variables $x_a x_b$. The point assignments are

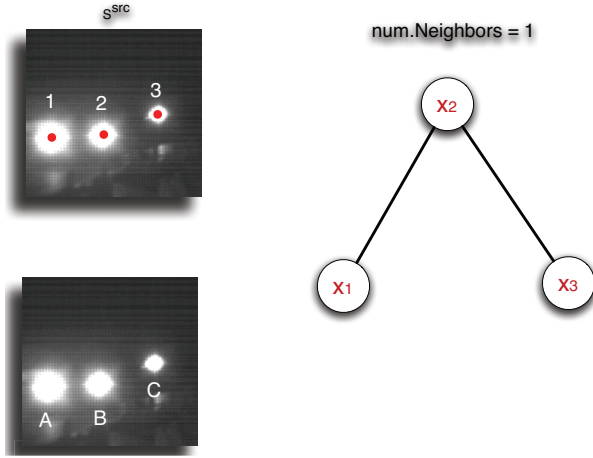


Figure 3: Example of the Feature-Oriented approach

denoted with a vector \mathbf{x} of binary variables. Each variable x_a will be active ($=1$) if the correspondence a exists, and inactive ($=0$) otherwise. The energy function is a sum of the binary variables multiplied by their respective costs θ , and after optimizing, the vector \mathbf{x} will indicate the optimal configuration of the correspondences. The set A defines all the possible correspondences, and N is the set of pairs of neighbor associations.

3.2 Quadratic Pseudo Boolean Optimization (QPBO)

In [6] a pseudo-boolean function is defined as a mapping $f : \mathbb{B}^n \rightarrow \mathbb{R}$. When the objective function has a quadratic term, and it is subject to constraints, the optimization of the function is known as quadratic pseudo-boolean programming. This, in the case of the energy function of Eq.(4.2) is a NP-hard problem.

In this work we use QPBO to provide an exact solution for the energy optimization, comparing its limitations and performance with other approximate techniques based on graphical models. The major benefits of this approach is the guarantee of reaching a global minimum, and the effortless im-

position of what we call the multiple-assignment constraints.

We have carried out a performance analysis of the QPBO method, and its limitations. We have studied the relationship between the number of blobs in the images and the number of variables resulting in the energy function. For example, with 10 features per frame, and considering 4 neighbors when creating the set N of pairs of associations, the number of variables generated is around 20000. The maximum amount of variables the algorithm is capable to handle is around 2500, which makes the QPBO approach infeasible even for the simplest of the scenarios.

4 Method

4.1 Association-Oriented (AO) vs Feature-Oriented (FO)

These two methods are both based on Graphical Models, specifically on Markov Random Fields (MRF), but they differ in the interpretation of the model variables. As we explained in section 3, the AO approach models the variables as associations, based on [4], and the FO as features, based on [3, 1]. The main consequence of these different interpretations, is the dimensionality of the variables and the structure of the graphical model. In the case of the AO approach the variables are binary, while in the FO the dimensionality of the variables will depend on the number of features of the destination image. Notice that in the FO, since the variables model source features, the possible values these variables can take will be any of the destination feature labels.

After analyzing both method's MRF structure, we can state that the AO approach generates more variables than the FO. On the other hand is capable of handling any type of multiple assignments, due to the association oriented representation, which can express any configuration of active correspondences. The FO approach generates a smaller graph but its limited when expressing

one-to-many associations, because the variables denote origin features, and can only take one destination label as a value. The graph structure and dimension of the variables will have consequences in the performance of the inference algorithm, when finding the most likely configuration of the random variables. The bigger the graphical model, the slower the convergence of the algorithm. This favors the FO approach, since produces a considerably smaller graph. On the other hand, the AO model has more expressive power, since it is able to represent any kind of assignment.

4.2 Model Constraints

In this work we are interested in letting the user specify the type of restriction to impose in the matching, depending on the application.

We can express the energy formulation and its constraints, as follows:

$$\arg \min_{\mathbf{x} \in \mathbf{M}} E(\mathbf{x} | \bar{\theta}) = \sum_{a \in A} \theta_a x_a + \sum_{(a,b) \in N} \theta_{ab} x_a x_b$$

subject to

$$M = \{\mathbf{x} \in \{0,1\}^A \mid \sum_{a \in A(p)} x_a \leq L, \forall p \in S^{src}\} \cup$$

$$\{\mathbf{x} \in \{0,1\}^A \mid \sum_{a \in A(q)} x_a \leq K, \forall q \in S^{dst}\}$$

$A(p)$ is the set of correspondences involving feature p ; The set S^{src} contains all the features of the source image, while S^{dst} contains all the features of the destination image (see Figure 4). The set M defines the valid vectors \mathbf{x} which accomplish the multiple-assignment constraint. L and K are positive integers defined by the user which specifies the maximum amount of features to associate simultaneously. Thus, the algorithm will be able to perform *one-to-0*, *1*, ..., *K* associations as well as *0*, *1*, ..., *L-to-one*.

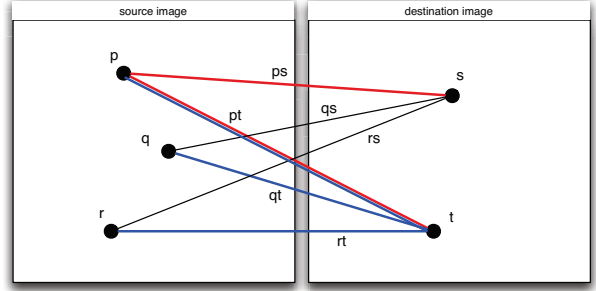


Figure 4: Red associations correspond to the set $A(p)$. Blue associations correspond to the set $A(t)$. Association pt is in the two sets

5 Experiments and Results

We have applied our algorithm to three different applications: In a synthetic sequence of a rotating object, to tracking of headlights at night, and to tracking of live cells which divide in two.

5.1 Synthetic sequence

For this experiment we consider the CMU 'House' sequence, matching the same 30 points in each frame. We explore the performance of our method as the separation between frames increases, and compare it to the work of [7]. Figure 5 shows the matching of two images with a separation of 40 frames.

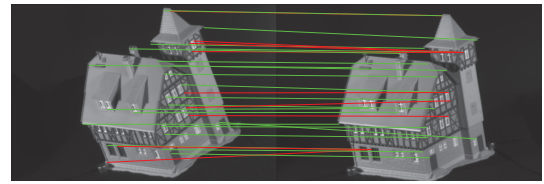


Figure 5: Points matched using the Feature Oriented Belief Propagation with 40 frames as baseline. Green lines denote the correct matchings, and red lines indicate the errors.

Figure 6 shows the performance of the method as the baseline increases, for the Cyclic Belief

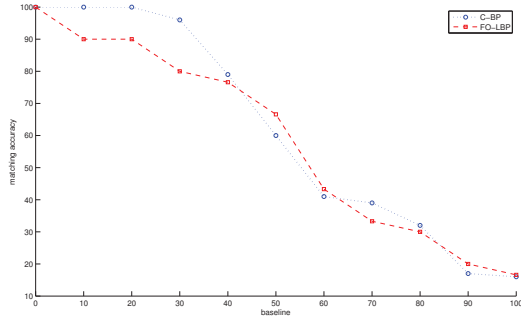


Figure 6: Matching accuracy of the feature-oriented model using the "house" data set, as the baseline varies

Propagation developed in [7] (C-BP) and our Feature-Oriented approach + Belief Propagation. The accuracy is similar in both methods, even though for a small baseline the C-BP performs better, and in frame distances around 50 frames our method presents a better accuracy. With smaller frame distances the C-BP outperforms our method.

5.2 Nighttime vehicle tracking

For this experiment we apply our tracking method to two different sequences of 100 frames of vehicles at nighttime. The sequence A has an average number of features per frame of 4.3, while the sequence B has 7.6 features per frame. The multiplicity constraint, for the association-oriented approach, limits the associations to one-to-one.

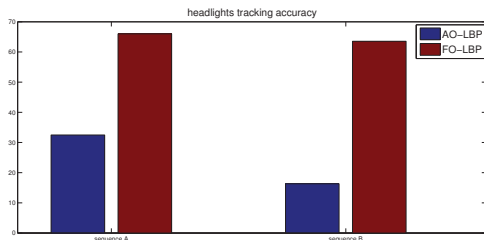


Figure 7: Accuracy percentage in the headlight sequences

Figure 7 shows the accuracy for both sequences using both AO and FO approaches. The FO model outperforms the AO in both sequences. The drop of the performance in the AO-BP is due to the greater amount of associations generated by the AO-Oriented model. Since the binary variables in this approach represent associations, the inference converges to configurations in which several matchings are active for each feature, and the accuracy drops significantly. Figure 8 shows an example of the output of the algorithm for the FO-BP.



Figure 8: example of two frames matched with FO-BP.

6 Conclusions

We have presented a comparison of three feature matching methods. One, based on an energy minimization through quadratic programming, and other two based on inference on a Markov Random Fields. We have successfully applied them

to tracking of vehicles at night and we have compared the accuracy obtained in the vehicle tracking application. We have studied the limitations of the Quadratic Programming approach which becomes computationally intractable from few number of features upwards. This justifies the necessity of a model capable of handling a large amount of features.

References

- [1] T.S. Caetano, T. Caelli, and D.A.C. Barone. Graphical models for graph matching. volume 2, pages II–466–II–473 Vol.2, June-2 July 2004.
- [2] Julian John McAuley, Tiberio S. Caetano, and Alexander J. Smola. Robust near-isometric matching via structured learning of graphical models. *CoRR*, 2008.
- [3] T. S. Caetano, T. Caelli, D. Schuurmans, and D. A. C. Barone. Graphical models and point pattern matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1646–1663, 2006.
- [4] Lorenzo Torresani, Vladimir Kolmogorov, and Carsten Rother. Feature correspondence via graph matching: Models and global optimization. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 596–609, Berlin, Heidelberg, 2008. Springer-Verlag.
- [5] Antonio López, Jörg Hilgenstock, Andreas Busse, Ramón Baldrich, Felipe Lumbreras, and Joan Serrat. Nighttime vehicle detection for intelligent headlight control. In *ACIVS '08: Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*, Berlin, Heidelberg, 2008.
- [6] Endre Boros and Peter L. Hammer. Pseudo-boolean optimization. *Discrete Appl. Math.*, 123(1-3), 2002.
- [7] Julian J. Mcauley, Tiberio S. Caetano, and Marconi S. Barbosa. Graph rigidity, cyclic belief propagation, and point pattern matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), 2008.

Detecting small pedestrians

David Vázquez*, David Gerónimo* and Antonio López*

* *Computer Science Department, Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra,, Barcelona, Spain*
E-mail:dvazquez@cvc.uab.es

Abstract

Pedestrian accidents are one of the leading preventable causes of death. In order to reduce the number of accidents, the pedestrian protection systems, a special type of advanced driver assistance systems, have been introduced. Since the appearance of pedestrians varies significantly as a function of distance to the camera. The main aim of this work is to explore if we can detect small pedestrians. We have evaluated the HOG pedestrian detector in two different datasets (INRIA and Daimler09) for two different distances (far and near). The obtained results suggest the use of different detectors for far and near distances.

Keywords: ADAS, Pedestrian detection, HOG.

1 Introduction

Pedestrian run overs represent the second largest source of traffic-related injuries. In order to reduce these road accidents, appear the Pedestrian Protection Systems (PPSs), a special type of Advanced Driver Assistance Systems (ADASs), where an on-board camera explores the road ahead for possible collisions with pedestrians in order to warn the driver or performing braking actions [2, 3]. Pedestrian detection is a challenging task. The main challenges rely on the pedestrian appearance variability due to clothes, poses or sizes and the context

where they can be found.

Pedestrian detection produced a vast amount of techniques, models, features and general architectures. Then, it is difficult to compare and study them. Recently, a pedestrian detection survey has been presented [3] which proposes a general module-based architecture and reviews different approaches with respect to the tasks defined in the proposed architecture. This work points out relevant aspects for future research in PPS. Important ones are the lack of good databases and benchmarking protocols as well as exploring the effect of the distance from pedestrians to the camera.

The aim of this work is to explore these two PPSs aspects. This idea of exploring the effect of the distance has been reinforced by Enzweiler and Gavrilu's work [2]. We want to answer the following questions: (1) which are the most adequate datasets from the latest ones? (2) for detecting far away people, which is the difference between training a system with actual small pedestrians and training it with scaled pedestrians? (3) which is the performance for each size of the pedestrians and why? (4) how do the optimum parameters of the method vary with respect to the distance?

Given that pedestrians closer to the camera are seen with more detail than the ones which are further away (see Fig. 1 and 2), a study on the benefices of training different classifier models depending on the target distance is of key interest. Since distant pedestrians are smaller and have less details, they tend to be more difficult to classify.

Closer targets present more details and their classification is easier but, the reaction time should be lower. This leads us to think that a system based on multiple classifiers, each specialized on different depths is likely to improve the overall performance with respect to a typical system based on a single general detector.

In this work we train the HOG pedestrian detector in two dataset (INRIA and Daimler09) and for two different distances (far and near) assessing their usefulness by both per-window and per-image evaluation.

The remainder of this paper is organized as follows. In Sect. 2 we overview the benchmark datasets and in Sect. 3 the detection approach needed to study the effect of the distance. The experiments, the evaluation criteria and the results are explained in the Sect. 4 which finalizes with a discussion of the results. Finally, in Sect. 5 we summarize the conclusions of this work, and draw some future work.

2 Pedestrian datasets

Existing datasets can be grouped in two types: (1) *person* datasets that contain non-occluded people in different poses and backgrounds but with a restricted point of view and, (2) *pedestrian* datasets that contain upright or partially occluded pedestrians in an urban environment and usually with motion information and more complete labellings.

Among the *person* datasets we select the INRIA one [1] that remains the most widely used and it is a spread reference in pedestrian detection. Among the *pedestrian* datasets we use the Daimler09 [2] because it is the appropriate dataset for studying the effect of the distance given that it has a lot of examples from different sizes to train and also provides a video sequence fully annotated that allows us to evaluate the experiments.



Figure 1: Images from the INRIA dataset. Top: images with pedestrians. Bottom: cropped pedestrians sorted by distance.



Figure 2: Images from the Daimler09 dataset. Top: images with pedestrians. Bottom: cropped pedestrians sorted by distance.

3 Pedestrian detector

Dalal et al. [1] proposed a pedestrian detector based on Histogram of Oriented Gradients (HOG) features inspired on SIFT and a SVM learning machine. It is the reference in the state-of-the-art of pedestrian detection. These features model the shape and appearance using normalized histograms of the image gradient orientation. The idea is to divide the image with a dense spatial grid in small regions called *cells*. A cell is represented as a histogram of its local gradients binned according to their orientation and weighted by their mag-

nitude. These cells are grouped in larger regions called *blocks*. A block is represented as a feature vector formed by concatenated and normalized histograms of its cells. The final descriptor is a feature vector formed by all the blocks attached and it is classified using a linear SVM. To compute the features we use the parameters suggested by the authors.

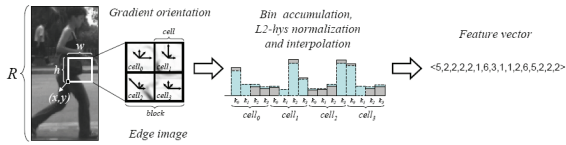


Figure 3: histogram of oriented gradients. (Reproduced from [4])

4 Experiments

Our experiments consist in evaluating the HOG classifier in the selected databases (INRIA and Daimler09) and for two different distances (near and far).

4.1 Evaluation methodology

To evaluate the experiments, there exist two established methodologies: *per-window* and *per-image* evaluation. In the *per-window* approach (Fig. 4), the detector is evaluated by classifying cropped pedestrians versus non-pedestrians crops and the performance is shown in the Detection Error Trade off (DET) curve that plots the miss-rate versus the False Positive Per analyzed Window (FPPW). In the *per-image* approach (Fig. 5), an image is given to the detector and it returns a list of Bounding Boxes (BB) with a given confidence. In this case, the detector scans the image by a sliding window approach and clusterize the detections with a Non-Maximum Suppression (NMS). The evaluation consists in performing a correspondence between the BB detections, namely BB_{dt} and the

BB groundtruth, BB_{gt} . To compare methods we employ the False Positives Per Image (FPPI).

Usually in the *per-window* evaluation to compare the results of two methods we look the miss-rate value at FPPW of 10^{-4} over the DET curves and in the *per-image* evaluation we compare the missrate values at 1 FPPI.

4.2 Results

To compare our results with the obtained by other authors we use the INRIA dataset. Figure 1a shows that our implementation of HOG gives the same results than the original HOG of Dalal and how the bootstrapping iterations improve the results. In Figure 1b we can see some images of detections and in Figure 6 we can see the obtained models for the detector.

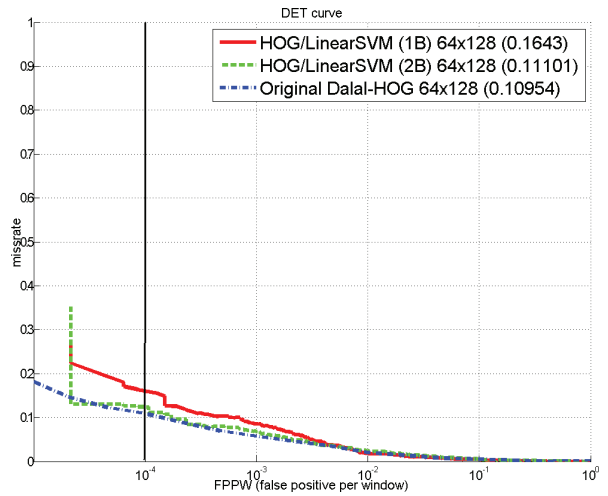


Figure 4: Per-window evaluation of the HOG and in the INRIA dataset. The missrate at the interesting point of false positives is the number inside the parenthesis.

To obtain small pedestrian samples we down-scale the images of the two datasets to a smaller size: 32x64 for INRIA and 24x48 for Daimler09. Then we train and test in the *per-window* evaluation over the scaled images. And in the *per-image* evaluation we use the classifiers trained with the

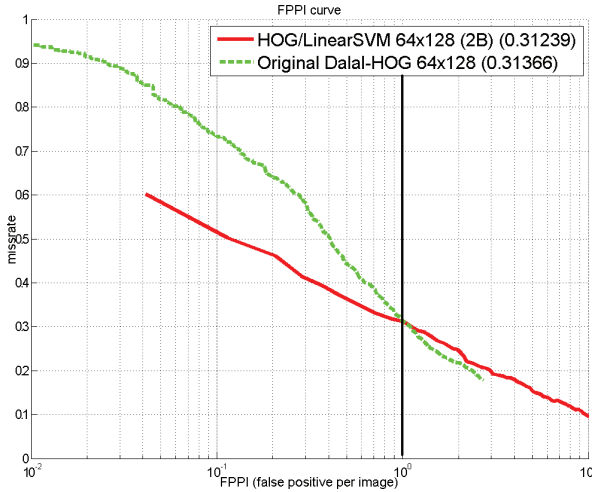


Figure 5: Per-image evaluation of the HOG and in the INRIA dataset. The missrate at the interesting point of false positives is the number inside the parenthesis.

small pedestrians to detect the big ones in order to compare the classifiers over the same set of images. Analyzing Table 1 it can be appreciated that for the HOG detector the bigger the image is the better the detector performs.

INRIA	Per-window		Per-Image	
	32x64	64x128	32x64	64x128
HOG	0.30	0.11	0.60	0.31

Daimler09	Per-window		Per-Image	
	24x48	48x96	24x48	48x96
HOG	0.60	0.23	0.32	0.23

Table 1: Performance evaluation of the HOG method in the INRIA and Daimler09 datasets.

To evaluate if there is any difference between training with these scaled pedestrians and training with actual small pedestrians, we split the Daimler09 set into a training and a testing sets and we train and test a HOG. The obtained results with the actual small images is slightly better in the per-

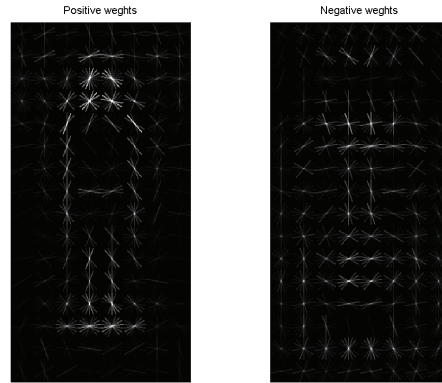


Figure 6: Models learned by the HOG method in the INRIA dataset.



Figure 7: Some detections of the HOG method in the INRIA dataset. The blue BBs are the groundtruth and the red ones are the detections.

image evaluation: a 3% of improvement.

To find the method optimum parameters we tuned the HOG parameters in the same way as Dalal et al. suggest in [1]. Among the parameters that can be optimized we optimize only the size of the cells and blocks and the orientation bins. From the results it seems that a fine binning (9 orientation) and large scale features (blocks of 2x2 cells of 8x8 pixels) are the best parameters for INRIA and Daimler09 datasets.

4.3 Discussion

At the beginning of the work raised some questions and after the experiments, some other questions appeared. Let's answer them.

• *Should a pedestrian detection system take into account the pedestrian distance?:* Experi-



Figure 8: Some detections of the HOG method in the Daimler09 dataset.

ments suggest to use multiple detectors specialized in distances is better than to have only one detector for all the distances.

- ***How does affect the size of the training images in the optimum parameters?:*** After the study of the HOG parameters we can conclude that there is a set of canonical parameters that performs well for all the cases and they are not affected by the pedestrian size.

- ***In order to learn a classifier for far away pedestrians, do we need samples with small pedestrians for training or is it enough to down-scale the big ones?:*** Results show that the classifier trained on the actual pedestrians is slightly better than the other. Thus, we expect that if we get a training set with more actual small pedestrians this performance difference could be higher.

- ***What are the differences between the performances obtained by the per-window and the per-image evaluation?:*** We have seen that the per-window performance could be not very realistic as it does not take into account the errors caused by the sliding window and the NMS.

- ***Which are the most suitable datasets for PPSs?:*** After working and analyzing the datasets we have seen that the Daimler dataset is more suitable for our purposes: it has many more examples, well labeled and at several scales. The problem of this dataset it is that the original frames from which the training samples were extracted are not available.

5 Conclusions

We have explored two interesting aspects to develop a pedestrian detector: the datasets used for learning the classifiers and the effect of the distance in the detection. To study this effect we have used two datasets (INRIA and Daimler09) and the HOG pedestrian detection method. The evaluation has been done following two different approaches (per-window and per-image) for two different distances (far and near). After the discussion we have realized that the distance is of key importance in PPSs. Therefore, we propose to build a system that combines specialized classifiers optimized for different ranges of distance to improve its performance.

References

- [1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.
- [2] Markus Enzweiler and Darius M. Gavrilă. Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2008.
- [3] David Gerónimo, Antonio M. López, Angel D. Sappa, and Thorsten Graf. Survey on pedestrian detection for advanced driver assistance systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.
- [4] David Geronimo, Angel D. Sappa, Antonio Lopez, and Daniel Ponsa. Adaptive image sampling and windows classification for on-board pedestrian detection. *Proceedings of the International Conference on Computer Vision Systems*, 2007.

Object Color Alteration

Shida Beigpour and Joost van de Weijer

Computer Science Department, Computer Vision Center
Computer Vision Center
Edifici O - Campus UAB, 08193 Bellaterra, Spain
E-mail: { Shida, Joost } @cvc.uab.es

Abstract

Color-alteration or recoloring refers to modification of the chromaticity of the object pixels in an image. We propose a novel method in order to model the change in the chromaticity of an object and illuminant light using a physics-based surface reflectance estimation. Unlike the existing methods, chromaticity of an object is estimated independent of the illuminant, while the illuminant color is estimated using the constraint of the Planckian Locus of natural illuminants. Realistic recoloring results on complex natural images captured by non-calibrated cameras clearly demonstrate that the proposed framework significantly outperforms state-of-the-art work on recoloring which disregards the underlying rules of physics.

Keywords: Dichromatic Reflection Model, Physics-based vision, Color-vision, Recoloring.

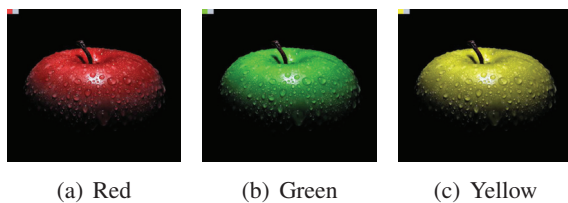


Figure 1: Image of an apple has been recolored. Can you guess which one was the original image?!

1 Introduction and related work

Recoloring or color-alteration is a term referring to the modification and adjustment of the image color appearance. Color modification methods are applied to photo montage, color correction, visual effects in movies, and also in industrial and commercial applications as a technique to visualize the final color appearance of the 3D object products before actual production in order to improve and facilitate their design [10]. Here we will focus on recoloring an object in a single snapshot.

Graphical recoloring methods often fail to properly generate a realistic perception of the recolored object (e.g. the case of non-white illuminant) as they suffer from a lack of knowledge about the physics behind it. According to the physics, the incident light not only affects the brightness of the pixels we perceive in an image, but also affects their chromaticity. Therefore we believe by making a distinction between different regions of the object surface (e.g. shading, and highlights), and decomposing the object chromaticity into the natural color of the object surface as well as the color of the illuminant, we are capable of performing the correct color modification.

Several methods developed for colorization of the gray-scale snapshots have also been used for recoloring. These methods mainly consist of par-

tial hand-coloring of regions in an image or video and propagating the colored points (known as *color markers* or *hot-spots*) to the rest of the image using a fairly complex optimization algorithm [4, 13, 6]. The main problem with such methods is their significant computational cost, intensive user assistance to assign the markers, and excluding the color information obtained by modern imaging devices. However no segmentation is required.

To improve the result of the existing color-marker based colorization, a novel method has been developed based on the idea that preserving the direction of the maximum-contrast in the image results in a more realistic colorization [1]. Despite the significant improvement in the quality of the colorization result, the method still suffers from the same drawbacks of the marker-based colorization methods used as its prior stage.

Color modification and recoloring embedded in professional photo-editing applications performs by calculating an offset in the hue, saturation and luminance between the source and destination colors. The source image is adjusted to produce the desired color [3]. Although such method is quite fast, it suffers from a lack of physical model to correctly separate object reflectance from illuminant color leading to a less realistic result, while requiring prior segmentation.

Color transfer methods extract the color characteristics (pixels color distribution) from a source image and apply it to a target image by building a transform using covariance, and mean of pixel color values [14, 8]. Although these methods provide fairly realistic results, they require information about the color distribution of the target recolored scene.

Here we propose a physics-based method to achieve a high quality re-colored image by extracting the physic-based geometrical model of the light interaction with the object surface. Yet we have managed to maintain a fairly low computational cost and, with the exception of object segmenta-

tion, virtually no user interaction is required.

2 Color Modeling and Recoloring

In order to develop a better understanding of the light interaction with the object surface, we use a physics-based model of reflection called the *Dichromatic Reflection Model* [9]. Using this model we are able to generate a snapshot of that object with the same imaging condition varying only the chromaticity of the object and illuminant light.

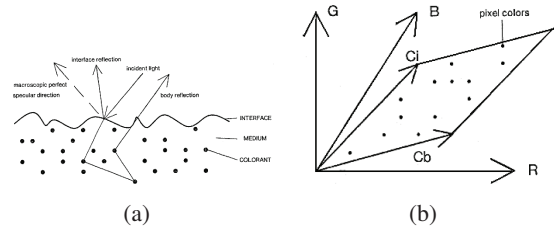


Figure 2: (a) Reflection of the light from an inhomogeneous material; (b) Pixel values for a set of points on a single surface lie within a parallelogram in color space.

2.1 Dichromatic Reflection Model (DRM)

According to Shafer, pixel values for a set of points on a single surface must lie within a parallelogram in the RGB space, bounded by RGB vectors C_i and C_b (here on we indicate vectors in bold font). These vectors represent the direction of the interface and body reflectance from the object surface respectively (Figure 2(b)). The validity of the dichromatic model has been proven for a variety of inhomogeneous dielectric materials commonly observed in natural scenes [12]. Although this model does not assume a point light or uniform illumination distribution over the scene, it requires a prior segmentation (Figure 3(b)) for multi-colored objects in order to fulfill the assumptions of the model.

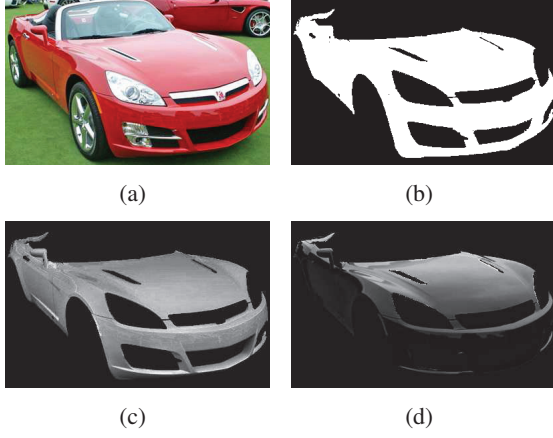


Figure 3: The DRM fitting: (a) The original image; (b) Segmentation mask; (c) and (d) are the intrinsic images for body and interface reflectances respectively.

The dichromatic model can describe the color of each pixel inside a single-colored object and illuminated by a single-colored illuminant, using images of the amount of body ("diffuse") and interface ("specular") reflections at each pixel which are called *intrinsic* images (Figure 3(c) and 3(d)). The dichromatic model in the mathematical format is demonstrated below,

$$\mathbf{f} = m_b \mathbf{C}_b + m_i \mathbf{C}_i, \quad (1)$$

where \mathbf{f} is the RGB triple defining the color of every pixel in the object surface, m_b and m_i are the intrinsic images of body and interface reflectance respectively, and \mathbf{C}_b and \mathbf{C}_i are the colors of the corresponding colors.

Several methods have been developed to approximate the dichromatic model of an object. Whether using a spatial-based approach in which the two dichromatic planes for specular and body reflectance are approximated considering the lighter and darker pixels separately [5], or with the assumption of a known illuminant [10, 11]. Later on in Section 3 we propose a novel method in which a fairly accurate approximation of the

dichromatic plane of an object under an unknown illuminant is achieved.

2.2 Intrinsic images estimation

Using the correlated RGB color space the dichromatic equation can be solved with the assumption that the \mathbf{C}_b and \mathbf{C}_i color vectors are constant for the entire object to be re-colored (single-colored object and illuminant). The material coefficients (m_b and m_i) are fixed for each pixel which means the coefficients are the same for R, G, and B values of the same pixel. Then for an image of N pixels, we would have $3 \times N$ equations (Equation 2) while the number of unknown values would be $2 \times N$ for m_b and m_i in addition to the 6 values defining the RGB triples of \mathbf{C}_b and \mathbf{C}_i color vectors. Having said that, for a large enough number of pixels, this set of equations can then be solved using an error minimization. Algorithms for approximating \mathbf{C}_b and \mathbf{C}_i are proposed in section 3

$$\begin{pmatrix} R_j \\ G_j \\ B_j \end{pmatrix} = m_{b_j} \begin{pmatrix} C_b^R \\ C_b^G \\ C_b^B \end{pmatrix} + m_{i_j} \begin{pmatrix} C_i^R \\ C_i^G \\ C_i^B \end{pmatrix} \quad (2)$$

Therefore, given the RGB values of \mathbf{C}_b and \mathbf{C}_i , and using the pixel RGB values \mathbf{f} , we are able to calculate the intrinsic image matrices m_b and m_i which minimize the fitting error of the model to the object pixels (Equation 3). Note that the pseudo-inverse notation implies the least square error minimization.

$$\begin{bmatrix} m_b \\ m_i \end{bmatrix} = pinv \left(\begin{bmatrix} C_b^R & C_i^R \\ C_b^G & C_i^G \\ C_b^B & C_i^B \end{bmatrix} \right) \mathbf{f} \quad (3)$$

2.3 Gamma Correction

Using uncalibrated RGB color images one should bear in mind that due to the *Gamma expansion* that

occurs largely in the nonlinearity of the electron-gun current-voltage curve in Cathode Ray Tube (CRT) monitor systems, image signals are *gamma encoded* prior to be shown on monitors [7]. Therefore a *Gamma Correction or decoding* process should be performed to preserve the linearity of the color signals prior to the DRM approximation. Here we have set γ to be 2.2 for sRGB color space.

2.4 Color alteration or Recoloring

The main goal of our method is changing both object and illuminant colors. After the estimation of the object reflectance model, recoloring of the object is straightforward. The entire color alteration process is demonstrated in the Equation 4, where \mathbf{f}' is the object reflectance in the new body and illuminant color (\mathbf{C}_b' and \mathbf{C}_i' respectively) specified by user.

$$\mathbf{f}' = m_b \alpha \mathbf{C}_b'' + m_i \mathbf{C}_i' \quad (4)$$

$$\mathbf{C}_b'' = \begin{bmatrix} C_i^{R'}/C_i^R & 0 & 0 \\ 0 & C_i^{G'}/C_i^G & 0 \\ 0 & 0 & C_i^{B'}/C_i^B \end{bmatrix} \mathbf{C}_b' \quad (5)$$

Note that according to the model (see Figure 2(a)), the body reflectance itself relies also on the illuminant color. Therefore, we have modeled the effect of illuminant color changes on the object surface using the term \mathbf{C}_b'' which is defined as in Equation 5. The term α simulates the desired change in the color intensity.

Despite the computational complexity of the dichromatic plane estimation, the re-coloring algorithm is simple addition and multiplication operations which, implemented in logic-gates, would perform in real-time.

3 Chromaticity estimation methods

In this section we propose two methods for the estimation of the body reflectance color \mathbf{C}_b and interface reflectance color \mathbf{C}_i (Figure 4(c)) respectively. The body reflectance color (\mathbf{C}_b) is estimated using Singular Value Decomposition (SVD) regardless of illuminant color. And the Planckian Locus concept is used to form the illuminant color (\mathbf{C}_i) estimation method.

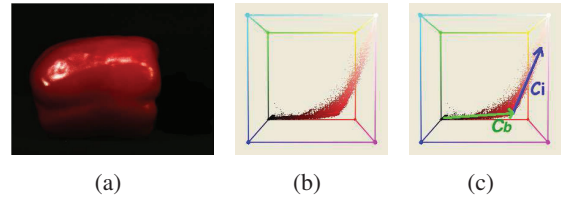


Figure 4: A single-colored object: (a) Original image; (b) RGB color histogram; (c) Expected directions for \mathbf{C}_b and \mathbf{C}_i vectors;

3.1 Body reflectance color estimation

The object pixel values for which the $m_i = 0$, form a line passing through the origin. Fitting a line through these pixels allows us to compute \mathbf{C}_b . The fitting error of an object pixel to a line given by the vector $\hat{\mathbf{v}}$ is

$$e(\mathbf{x}) = \left\| \mathbf{f}(\mathbf{x}) - \left((\mathbf{f}(\mathbf{x}))^T \hat{\mathbf{v}} \right) \hat{\mathbf{v}} \right\|. \quad (6)$$

The above Least Squares error minimization problem can be solved using SVD, where the eigenvector which corresponds to the higher eigenvalue is desired. Intuitively, having the assumption that most of the object pixels belong to the body reflectance, the higher eigenvalue is expected to indicate the \mathbf{C}_b direction.

3.2 Illuminant chromaticity estimation

The chromaticity of common light sources is limited and follows closely the Planckian locus of black-body radiators which is believed to be a function

of temperature T in Kelvins [2]. For this matter we sample the colors of the Planckian Locus (Figure 5) for the standard illuminants ($T \subset 4000 \sim 25000$ with steps of $1000 K^\circ$). Then the dichromatic equation is solved for all the pixels of the colored object using each of the possible illuminants, and m_b and m_i values are calculated. The illuminant chromaticity (C_i) which minimizes the object reconstruction error (Equation 7) would be chosen, and the corresponding m_i and m_b values for each pixel are then considered as the dichromatic model of the object.

$$E(C_b, C_i) = \sum_j ((f_j - m_{b_j} C_b - m_{i_j} C_i)^T \times (f_j - m_{b_j} C_b - m_{i_j} C_i)) \quad (7)$$

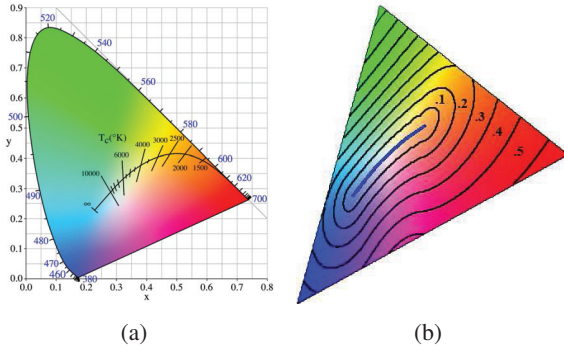


Figure 5: The chromaticity of common light sources: (a) The Planckian Locus inside the color gamut for CIE XYZ color space; (b) Color regions for which the distance to the Planckian Locus are in the same range.

4 Results

A set of natural images have been recolored using our framework. The estimated intrinsic images are demonstrated in Figure 6. The intrinsic images seem to make a good estimation of areas of highlights (see images of interface reflectance). Figure 7 illustrate the recoloring results obtained with

the proposed framework. Realistic results achieved suggest that the proposed framework outperforms previous work on recoloring in which the underlying physics rules have been disregarded.

Note that few existing methods which make use of physics-based reflectance model, have only presented the results on a set of images under laboratory restricted conditions [10]. To best of our knowledge, these are the first reported results for reflectance estimation in real-world images with complex shading and highlights.

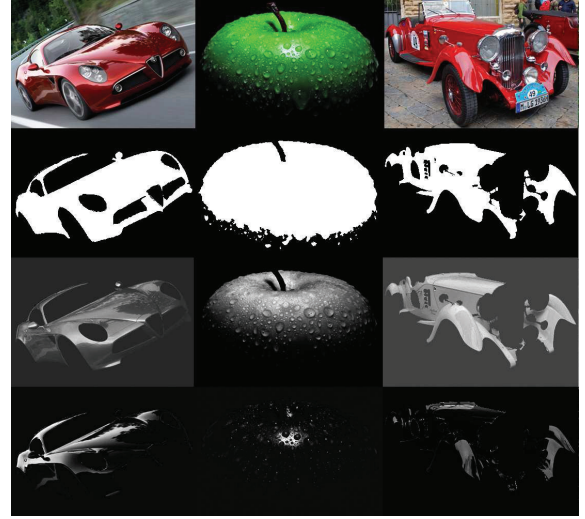


Figure 6: A set of natural images, their corresponding segmentation masks, and intrinsic images.

5 Conclusion and Future Work

We have presented a physic-based object reflectance (body and interface) estimation method for pre-segmented images. Object chromaticity is estimated independent of the illuminant color. A framework for modeling the change in the object color as well as the chromaticity of its illuminant light has been developed. The experimental results on natural images taken with non-calibrated cameras indicate that realistic recoloring of an object with complex specularities and shading is achieved.



Figure 7: Recoloring a set of natural images.

We propose, as our future work, to embed into the framework a DRM-based object segmentation method, and to make use of psychophysical experiments to introduce a quantitative error measurement.

References

- [1] M.S. Drew and G.D. Finlayson. Realistic colorization via the structure tensor. *ICIP*, pages 457–460, 2008.
- [2] G.D. Finlayson and G. Schaefer. Solving for colour constancy using a constrained dichromatic reflection model. *International Journal of Computer Vision*, 42:127–144, 2002.
- [3] R. Gonsalves. Method and apparatus for color manipulation. *United State Patent 6,351,557*, Feb 26, 2002.
- [4] V. Konushin and V. Vezhnevets. Interactive image colorization and recoloring based on coupled map lattices. *Graphicon*, pages 231–234, 2006.
- [5] V. Kravtchenko and J.J. Little. Efficient color object segmentation using the dichromatic reflection model. *Communications, Computers and Signal Processing, 1999 IEEE Pacific Rim Conference on*, pages 90 – 94, 1999.
- [6] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *SIGGRAPH ACM TOG archive*, 23(3):689 – 694, 2004.
- [7] Ch. Poynton. *Digital Video and HDTV Algorithms and Interfaces*. 2003.
- [8] E. Reinhard, A.O. Akuyz, M. Colbert, C.E. Hughes, and M. O’Connor. Real-time color blending of rendered and captured video. *IITSEC, Orlando, FL*, (1502):1–9, 2004.
- [9] A. Shafer and D. Lischinski. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985.
- [10] H.L. Shen and J.H. Xin. Transferring color between three-dimensional objects. *Applied Optics*, 44(10):1969–1976, 2005.
- [11] R.T. Tan and K. Ikeuchi. Separating reflection components of textured surfaces using a single image. *Computer Vision, IEEE International Conference on*, 2:870, 2003.
- [12] S. Tominaga and B.A. Wandell. Standard surface-reflectance model and illuminant estimation. *Journal of Optical Society of America*, 6(4):576–584, 1989.
- [13] A.V. Vezhnevets. Growcut - interactive multi-label n-d image segmentation by cellular. *Graphicon conference proceedings*, 2005.
- [14] X. Xiao and L. Ma. Color transfer in correlated color space. In *VRCIA ’06: ACM international conference on Virtual reality continuum and its applications*, 2006.

Human and Computational Color Constancy

Jordi Roca, C.A. Párraga and Maria Vanrell

Computer Vision Center, Computer Science Department, Universitat Autònoma de Barcelona
E-mail: jroca@cvc.uab.es

Abstract In this paper we present a brief review of the main issues surrounding the link between human and computational color constancy. Starting with the problem statement, its definition, main attributes and measuring methods, we introduce the computational color constancy problem and link it to its human equivalent. Finally we conclude that a new computational color constancy algorithm could be built by means of mimicking the functional stages of the human color constancy.

Keywords: human color constancy, computational color constancy, memory colors.

1 Introduction

Human Color Constancy is a perceptual phenomenon that stabilizes the appearance of object's colors throughout changes in illumination, [14, 16, 19].

One possible ecological justification for color constancy in mammals is to facilitate scene object recognition. In Helmholtz's words: "*Colors are mainly important for us as properties of objects and as means of identifying objects*". Then a mechanism that preserves the color appearance of objects will serve this purpose [15].

As a perceptual phenomenon, all variables affecting color constancy lie in the content of the perceived scene, e.g. scene chromaticity, three-dimensional information, object movement and some others. All these factors are called visual cues. As we will see later, the object's color is constrained by scene context and memory context.

There are also two other alternative definitions:

- The concept of *operational/relational color constancy* refers to the constancy of the perceived relations between the colors of surfaces under such changes in illumination [4, 9]. Foster et al [8] discuss the relationship between the two previous definitions and draw conditions for their equivalence.
- In computational terms: "*color constancy is achieved when a neural process transforms the signals elicited by surfaces under a test illuminant towards the signals that would be elicited by the same surfaces under a reference illuminant*" [19].

The general context where this phenomenon occurs is scene viewing. Over this process the photons reflected by surfaces are collected by sensors in the retina. This information is then translated into electrical signals and travels to the brain. Since viewing is done across time, the continuous perceived scene is a discrete succession of electric signals. Part of this information is perceived as constant, meanwhile other part is perceived as changing. Because of that, there is a pull between two processes: adaptation and discrimination [1]. According to Jameson and Hurvich [15]: "...human visual systems are likely to have evolved a design that provides perceptual information about change as well as constancy". What is the reason for such dual behavior? What are the key factors that govern these two processes? In this general context we will focus our study in the key scene factors that grant the achievement of color constant descriptors across time.

In the following pages we will present a short review of color constancy; its definition, motivation and features. Also we will introduce and link human with computational color constancy¹. In order to understand the mechanisms underlying the phenomenon of color constancy we must give a concise definition of color constancy and try to answer the right questions: What is color constancy? Which biological purpose does it serve? Where in the human visual system is it sited? Under which conditions it works and which extent does it works? We will try to give some insights into these questions.

2 Visual Cues in different neural levels

Color constancy is composed by several mechanisms situated at different neural levels in the human visual system [13, 16, 19]. However these mechanisms may have some purpose other than color constancy. The wide range of this scope is what makes specially difficult to understand this phenomenon. In order to achieve constancy the visual system extract visual cues from the scene, ranging from physical scene statistics to scene object composition. According to Brainard [16]: *“The open question of color constancy is what aspects of complex images govern the visual system’s sensitivity”*, and in our case adaptation. Also, we are interested to find out how much each visual cue contributes to the final color constancy achieved, and in particular how these contributions vary for different complex scenes [16].

Some authors have studied the problem using scene statistics from the physical scene description. This is, using the illuminant and surface reflectance spectral function and the sensitivity functions for each photoreceptor type [3]. From this point of view they try to extract relevant physical statistics from the scene. Also there has been other studies using other conditions, as object movement [24], chroma variance and the existence of 3D objects [11]. All these factors could be classified as bottom-up because they

¹ In this paper computational color constancy is framed in computer vision context, not in computational neuroscience.

belong to the scene composition. However some other kind of factors have been identified, such as the effect of ‘*memory color*’ in familiar objects [17] (i.e. the modification of perceived colors by remembered colors or memory context.) These two approaches split the perceived color formation between two information sources. On one side we have bottom-up information and on the other side we have top-down information. Traditional studies have centered their attention on bottom-up factors but recently there have been new insights into the top-down contributions.

According to Smithson [19] a color constancy model should perform three operations:

1. Identify the type of neural transformation required to undo a color conversion across illuminants.
2. Find out the parameters that rule the transformation and how these might be set from the scene.
3. Specify where in our perceptual apparatus these transformations are implemented.

Following this schema we can split the phenomena in three main stages: sensory, perceptual and cognitive [13] each with its own scene visual cues. However some cues may belong to multiple perceptual levels at the same time.

2.1 The Sensory Stage

The eye and specially the retina is where are sited the sensory mechanisms with the action of photoreceptors and the outer retina being his main components affecting color sensitivity. In fact, there are some authors that support the theory that the main part of color constancy is achieved in this level [21]. According to Hurlbert [13] we can divide the sensory stage in three main parts, each of them with his own temporal scale. Chromatic adaptation to the mean (60 seconds), chromatic adaptation to the variance (several minutes) and spatial contrast (25 milliseconds). And so the visual cues belonging to these levels are the mean chromaticity, chroma variance and spatial contrast. One of the results that strongly supports the use of spatial contrast in the color constancy mechanism is found by Foster and

states that the ratios of cone-photoreceptor excitations produced by light from natural surfaces are statistically almost invariant under changes in illumination by natural light [7]. So they contribute to the color constancy mechanism keeping the cone-excitation ratios invariant from the surfaces across illuminant changes.

2.2 The Perceptual Stage

At this stage we require prior segmentation of the scene in order to find relevant scene elements. For instance: specular highlights, mutual reflections, movement, depth perception, 3D objects and so on. We label these kind of visual cues as ‘perceptual’. Brainard [16] studied which rate of color constancy represents each of local contrast, global contrast and specular reflections cues. But other studies [24] have been proved that movement is also a perceptual cue to improve color constancy. Also the scene depth has been proved to influence the phenomenon [11, 23]. This could be explained because scene depth facilitates subject’s illumination change recognition.

2.3 The Cognitive Stage

This level require to identify the objects prior to allow the modification of their perception by the memory processes. This influences have been studied by several authors, more recently in [17], regarding the perceived fruit color. At the end, they re-developed concept of “*memory color*”. Which was introduced by Hering and states that we see the world through the object colors that we have in our memory, which have been acquired over visual experience. In his own words:

“*All objects that are already known to us from experience, or that we regard as familiar by their color; we see through the spectacles of memory color*” [12].

There is another possible perceptual-cognitive cue, the consciousness of illuminant change as reported by several authors [2].

3 Experimental Paradigms

One way to measure the color constancy phenomenon is using psychophysical experiments. The definition of each new experiment has some common parts: the paradigm, the laboratory conditions, the stimulus and the subject’s task. For a clear exposition we will discuss separately each part but they are interrelated. In order to get a successful experiment, the correct choice of all elements is critical.

3.1 Psychophysical Paradigms

As we have seen, the color constancy effect is composed of several mechanisms. There are different psychophysical paradigms, each suited to the color constancy mechanism to be measured. The most common are *asymmetric color matching*, *achromatic color setting* and *color naming* [9, 19]. In the first paradigm two illumination conditions are presented and the subject has to match one color from one condition to the other, this matching can be across space (*Simultaneous Asymmetric Matching*) or time (*Successive Asymmetric Matching*). In the second paradigm the subject adapts under one scene illumination and then is required to adjust a test patch until

Neural Transformation	Sensory	Perceptual	Cognitive
Visual Cues	<ul style="list-style-type: none"> • Chromatic adapt. to the mean(1) • Chromatic adapt. to the variance(2) • Color Contrast(3) 	<ul style="list-style-type: none"> • 3D objects • Mutual reflectance • Chroma variance • Specularities • Scene movement 	<ul style="list-style-type: none"> • Memory Colors • Consciousness of illuminant change
Location in HVS	Retina, LGN	Cortex, V1/V2	Brain, Cortex V4
Requirements	Light	Scene Segmentation	Object Identification
Temporal Scale	(1) 60 seconds, (2) minutes, (3) milliseconds	Several minutes	Days, years

Table 1. The color constancy visual cues in the human visual system levels.

it appears achromatic. The last paradigm measures whether a surface is assigned the same color name under different illuminants. In order to report the degree of constancy achieved, all this methods measure the ‘distance’ between the subject’s stimulus perception and the physical stimulus. Each method has its own strong and weak points, for a complete discussion see [9, 19].

Recently a new color constancy experiment was developed. It measures the boundary movement between color categories across illuminant changes [10, 18]. This experiment uses one of the last three paradigms in order to measure the perception of boundary colors.

3.2 The Laboratory Conditions

There is another useful classification among experimental paradigms: the visual environment where the property is measured. In everyday’s life, color constancy is experienced in natural environments, so there is plenty of visual cues that potentially could be used. But once inside the laboratory, without all these cues, the measured color constancy is weaken [16, 19]. The big question is that we do not know exactly which are the natural relevant cues to achieve full color constancy. So there is a trade off between the naturalness of the scene and the degree of color constancy achieved and measured. When we try to translate the natural scene to a lab scene we face two problems. First we have to ensure the precise control of the illuminant spectral function and object’s reflectance, while trying to keep intact the subject perceived visual

cues. There are three main different laboratory setups:

- 1. The subject is placed in a dark room with a CRT monitor, where the stimulus is displayed [24].
- 2. The subject is placed in a real 3D scene and the stimulus is presented there [16].
- 3. The setup consists in rendering a 3D scene and present a stereo-coherent pair of images to the subject [5].

Each setting has his own requirements in order to precisely calibrate the scene colors and the subject’s response. In the ‘CRT setting’ the stimulus can be presented in 2D, or 3D. Using RADIANCE² software we can calibrate correctly the presented colors in 3D rendered scenes [5, 11].

3.3 The Stimulus

The stimulus are presented to the subject along the experiment. After that, the subject’s response is measured, usually by means of a electronically device with several buttons. In order to measure a single color constancy mechanism the stimulus selection and time exposure are critical factors.

3.4 The subject’s task

In every psychophysical experiment a good subject instruction is the key factor to succeed in measuring the perceptual property. As reported by several authors the subjects instructions could affect the results[19, 20]. In the color constancy context there are two different questions that

2 <http://radsite.lbl.gov/radiance/book/>

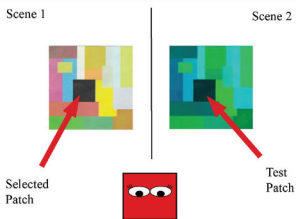
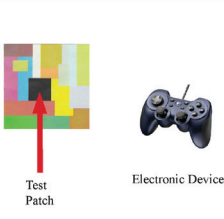
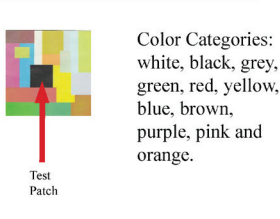
	Asymmetric Matching	Achromatic Matching	Color Naming
Lab Setting			
Subject’s Task	Match the color <i>Test Patch</i> from Scene 2 to the <i>Selected Patch</i> from Scene 1.	Transform the <i>Test Patch</i> color to achromatic using an electronic device.	Assign a color category to the <i>Test Patch</i> .

Table 2. The color constancy paradigms.

may modify the outcome. First we can ask to the subjects to make a patch on the screen ‘look as if it were cut from the same piece of paper’ or ask to match ‘hue and saturation’. In the first questions they show better color constancy [19].

4 Computational Color Constancy

The computational color constancy definition is more precise than the human one. Given one image, our central problem is to estimate the real object’s color coordinates in some color space. Then we can use these estimates for object recognition or color categorization. Illuminant change is a natural problem for the tasks described. Therefore there are two main computational approaches to fix these problem [22] :

- *Color Normalization* creates a new version of the image being independent through light changes.
- *Color Constancy* tries to estimate the illuminant color in order to transform the image to a canonical version of the scene reflectance.

In general, the ultimate goal of a color constancy algorithm is that the computed color for an image pixel is constant irrespective of the illuminant used [6].

The computational color constancy problem, that is illuminant estimation, is not solved. Over the last decades there has been many algorithms developed. It can be divided into two main groups: physical methods and statistical methods. The first ones are based on the fact that image must fulfill some physical properties. These methods use a more general model of image formation that the one used in the Statistical group. Statistical methods assume that surfaces are Lambertian and this group can also be split in three types:

1. Methods based in simple statistics: Grey World, White Patch, Shades of Grey and Grey-Edge.
2. Gamut Mapping methods as C-Rule.
3. Bayesian Methods as Color-by-Correlation and Voting Methods.

In order to evaluate color constancy algorithms there are two main ways: color-based object recognition and ground truth [6]. The first method evaluates the performance of color-based object recognition algorithms after applying to the image a color constancy algorithm. On the other hand the ground truth, when it is available, is much easier to check.

5 Conclusions

The human color constancy is a complex problem that needs further studies in order to completely understand all the phenomenon. But the localization of several mechanisms situated at different levels in the human visual system is a big step that allows to study each mechanism in isolation via the right psychophysical experiment.

The Computational and human color constancy processes begin with the same physical information but after this common departure their totally differ in information processing despite trying to solve similar problems. We must not forget that a digital image is not what a human has in his retina when is viewing scene. There are many differences between the two processes and the visual system is far more complex than any algorithm or camera device. Also we must remember that the visual system has two eyes, and so two visual approximations for the same scene, and that does not happen in the computational field where we only work with one image. So the two problems are modeled essentially in different ways.

When the final aim is the same, it seems natural to build an algorithm that mimics the human visual system. This algorithm should follow a human based design in a functional level. This means that it should have sensory, perceptual and cognitive levels. The bulk of the computational color constancy algorithms do not use these scheme, despite trying to get out the same perceptual effect. This implies the combination of multiple cues including the scene geometry, and not only the chromatic content of the images. This idea points to a new lines of research in the computational color constancy field.

Acknowledgments:

This work has been partially supported by projects TIN2007-64577, RYC-2007-00484 and the Consolider-Ingenio 2010 CSD2007-00018 of Spanish MEC (Ministerio de Educación y Ciencia).

References

1. Abrams, A. B., Hillis, J. M., and Brainard, D. H. 2007. The Relation Between Color Discrimination and Color Constancy: When Is Optimal Adaptation Task Dependent?. *Neural Comput.* 19, 10 (Oct. 2007), 2610-2637.
2. Barbur John L. , Spang K. (2007) Colour constancy and conscious perception of changes of illuminant. *Neuropsychologia*, Volume 46.
3. Brainard, D. H., Longère, P., Delahunt, P. B., Freeman, W. T., Kraft, J. M., & Xiao, B. (2006). Bayesian model of human color constancy. *Journal of Vision*, 6(11):10, 1267-1281.
4. Craven B.J. , Foster David H. (1991). An Operational Approach to Colour Constancy. *Vision research*, vol. 32, no7, pp. 1359-1366.
5. P. B. Delahunt and D. H. Brainard. Does human color constancy incorporate the statistical regularity of natural daylight? *J. Vis.*, 4(2):57-81, 2004.
6. Ebner M. (2007) Color Constancy. John Wiley & Sons Ltd, Chichester, England.
7. Foster David H. , Nascimento Sérgio M.C. (1994). Relational colour constancy from invariant cone-excitation ratios. *Proceedings: Biological Sciences*, Vol. 257, No. 1349, pp. 115-121
8. Foster David H. , Nascimento Sérgio M.C. , Craven B.J. , Linnell Karina J. , Cornelissen Frans W. , Brenner E. (1997) Four Issues Concerning Colour Constancy and Relational Colour Constancy. *Vision Research*, Vol 37.
9. Foster David H. (2003). Does colour constancy exist? *Trends in Cognitive Sciences*, Volume 7, Issue 10, 439-443
10. Hansen, T., Walter, S., & Gegenfurtner, K. R. (2007). Effects of spatial and temporal context on color categories and color constancy. *Journal of Vision*, 7(4):2, 1-15.
11. Hedrich, M., Bloj, M., & Ruppertsberg, A. I. (2009). Color constancy improves for real 3D objects. *Journal of Vision*, 9(4):16, 1-16.
12. Hering E. Outlines of a theory of light sense (L.M. Hurvich and D.Jameson, trans.) Harvard University Press, Cambridge, Mass, 1964.
13. Hurlbert A. , Wolf K. (2004). Color contrast: a contributory mechanism to color constancy. *Prog Brain Res.* 2004;144:147-60.
14. Hurlbert A. (2007) Colour Constancy. *Current Biology*. Vol 17 No 21 R906.
15. Jameson D. , Hurvich Leo M. (1989). Essay Concerning Color Constancy. *Annu Rev Psychol.* 1989;40:1-22.
16. Kraft J.M. , Brainard D.H (1998) Mechanisms of color constancy under nearly natural viewing. *Proc Natl Acad Sci U S A.* 1999 Jan 5;96(1):307-12.
17. Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision*, 8(5):13, 1-16.
18. Smithson, H., & Zaidi, Q. (2004). Colour constancy in context: Roles for local adaptation and levels of reference. *Journal of Vision*, 4(9):3, 693-710.
19. Smithson, H. E. (2005) 'Sensory, computational and cognitive components of human colour constancy.', *Philosophical transactions of the Royal Society B : biological sciences.*, 360 (1458). pp. 1329-1346.
20. Speigle Jon M. , Brainard D.H. (1996) Is color constancy task independent? in *Proceedings of the 4th Information Science and Technology/ Society for Information Display Color Imaging Conference*.
21. Vanleeuwen M.T. , Joselevitch C. , Fahrenfort I. , Kamermans M. (2007) The contribution of the outer retina to color constancy: A general model for color constancy synthesized from primate and fish data. *Visual Neuroscience* (2007), 24:3:277-290 .
22. Vazquez-Corral J. , Vanrell M. , Baldarich R. , Tous F. (2009) Color Constancy by Category Correlation. Submitted in *IEEE transactions on PAMI*.
23. Werner, A. (2006). The influence of depth segmentation on colour constancy. *Perception*. 2006;35(9):1171-84.
24. Werner, A. (2007). Color constancy improves, when an object moves: High-level motion influences color perception. *Journal of Vision*, 7(14):19, 1-14.

A Computational Colour Naming Model Trained on Real-Life Images

Hany M. SalahEldeen, Robert Benavente and Maria Vanrell

Computer Science Department, Computer Vision Center
Computer Vision Center
Edifici O - Campus UAB, 08193 Bellaterra, Spain
E-mail: { hsalah, robert, maria.vanrell } @cvc.uab.es

Abstract

Colour naming is the process done by human beings when they assign linguistic terms to objects to describe their colour. In computer vision, several colour naming models were developed and each has its own advantages and drawbacks. Models based on psychophysical data have a robust perceptual basis and obtain good results in ideal laboratory conditions, but they lack precision when applied to real images. On the other hand, models fitted using image data sets achieve better results when applied to real-life images, but they are not directly related to human perception. Thus, the goal of this paper is to merge two of these approaches to obtain a new model which includes the advantages of both methodologies. This modified model was tested against a data set of real-world uncalibrated images and the results exceeded the original model.

Keywords: Colour Naming, Context-Based, Parametric/Perceptual Model, Uncalibrated Data Set.

1 Introduction

Colour naming is one of the several visual tasks commonly done by humans involving colour. It

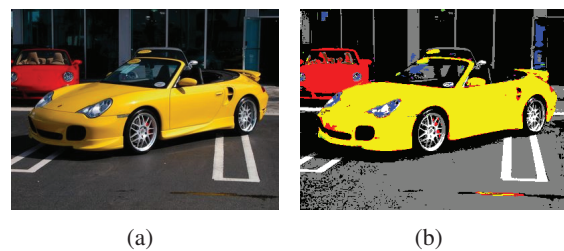


Figure 1: The Result of Color naming a picture of a car.

is the process of making a decision (in linguistic terms) about which colour best describes a given region of homogeneous hue. It is the last step in human colour-processing and it is performed in the visual cortex [8]. The aim of studying colour naming is to try to reduce the semantic gap in the task of giving names to colours in images. The semantic gap is the lack of a direct link between the low-level colour features extracted by machines and high-level semantics humans use. This gap is even more significant in applications like image retrieval where users require systems to support queries in natural languages [6]. Given so, an urgent need evolved to automate the process of colour naming and accurately imitating human perception in assigning colours. An automated colour naming model is a model that can

correctly assign a colour term to a specific pixel. Provided by any given image, a colour naming model is supposed to have the ability of analyzing each pixel in the image and successfully decide to which colour category it belongs (red, green,...etc). Several models were created to solve this problem of colour naming; some of these models will be shown shortly[7][5].

2 Related Work

On the way of shaping the current understanding of the colour naming process nowadays, a lot of experiments were done and several models were developed. Some of which are Psychophysical, Neuropsychological or Computational [1, chap. 2].

Benavente et al.[2] developed a parametric model to fit data samples based on psychophysical experiments. These samples were fitted using a Triple Sigmoid function in six lightness layers. The idea behind creating this model is to find a suitable function capable of presenting the shape of each colour in the CIELab space¹. Given a point in the colour space, it is possible to decide the membership of this point to each of the 11 basic colour terms of Berlin and Kay²[3].

$$\mu_C(p, I) = \begin{cases} \mu_C = TSE(p, parL_1) & \text{if } I \leq I_1, \\ \mu_C = TSE(p, parL_2) & \text{if } I_1 < I \leq I_2, \\ \vdots & \\ \mu_C = TSE(p, parL_N) & \text{if } I_N < I, \end{cases}$$

μ_C is the membership of p to the chromatic³ category C , I is the intensity level range and N is the number of lightness levels defined in the model while $parL_x$ are the parameters of level x . TSE stands for *Triple Sigmoid with Elliptical center*

¹The CIELab colour space is an approximately uniform colour space generated by optimal colour stimuli with respect to CIE standard illuminant D_{65}

²Pink, Red, Orange, Brown, Yellow, Green, Blue, Purple, Gray, Black and White

³Chromatics are colours Pink, Red, Orange, Brown, Yellow, Green, Blue and Purple

which is a variant of the one-dimensional sigmoid function as follows:

$$S^1(x, \beta) = \frac{1}{1 + \exp(-\beta x)},$$

where β controls the slope of the transition from 0 to 1.

Each of the chromatics are represented by a Triple Sigmoid function with an elliptical centre (TSE) (see figure 2).

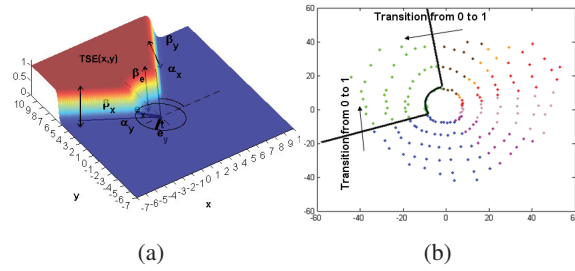


Figure 2: The TSE function fitting the chromatics.

$$TSE(p, \beta, \alpha, e, \phi, t) = DS(p, \beta, \alpha, t) \cdot ES(p, e, \phi, t),$$

where DS is the Double Sigmoid function determining the separating boundaries between chromatics as follows:

$$DS(p, \beta, \alpha, t) = S_1(p, \beta_x, \alpha_x, t) S_2(p, \beta_y, \alpha_y, t),$$

$$\gamma_1 = (x - tx)\cos(\alpha) + (y - ty)\sin(\alpha),$$

$$S_1(p, \beta, \alpha, t) = \frac{1}{1 + \exp(-\beta\gamma_1)},$$

$$\gamma_2 = (x - tx)(-\sin(\alpha)) + (y - ty)\cos(\alpha),$$

$$S_2(p, \beta, \alpha, t) = \frac{1}{1 + \exp(-\beta\gamma_2)},$$

Vector α determines the axis in which the function is oriented, p is the (x,y) point investigated and t is where the origin was translated to.

Another type of sigmoid function is used to define the middle part in each layer where the achromatic colours⁴ reside in the CIELab space. This

⁴Achromatics are colours Gray, Black and White

part takes an elliptical form and the function defining it is called the elliptic sigmoid function and it is illustrated as follows:

$$\gamma_1 = \left(\frac{(x - tx)\cos\phi + (y - ty)\sin\phi}{e_x} \right),$$

$$\gamma_2 = \left(\frac{(x - tx)(-\sin\phi) + (y - ty)\cos\phi}{e_y} \right),$$

$$ES(p, e, \phi, t) = \frac{1}{1 + \exp(-\beta_e(\gamma_1 + \gamma_2))},$$

where e is the length of the axes of the central ellipse and ϕ is the rotation of this ellipse. Furthermore, within the elliptical sigmoid centre, the three achromatic colours reside and are separated by lightness through a one-Dimensional Sigmoid function.

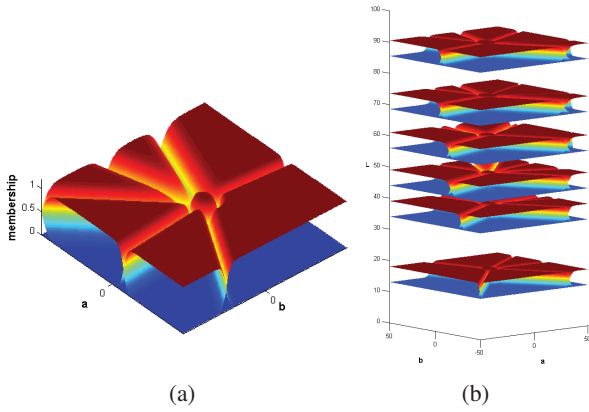


Figure 3: The Triple Sigmoid Elliptical centre model in one of the 6 intensity levels.

The Probabilistic Latent Color Naming Model was developed by van de Weijer et al.[10] based on the same concept of the *latent aspect models*. One of the most interesting characteristics of this model is that it was fitted using a data set of real-world uncalibrated images. This data set was obtained from Google images search engine and it is characterized by being weakly labelled. These weakly labelled Google images are represented by their normalized Lab histograms. These histograms form the columns of the image specific word distribution $p(w|d)$. See figure 4.

After analyzing two of the most interesting ap-

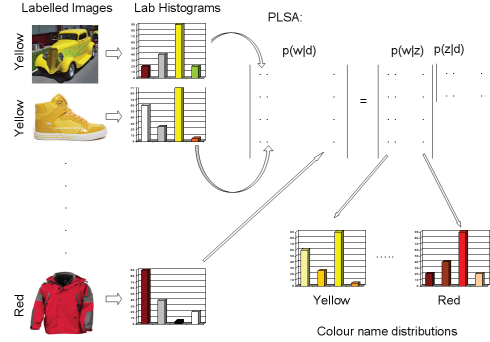


Figure 4: The PLSA color-naming model

proaches in solving the problem of colour naming automation some conclusions were drawn. Firstly, each of the two models possesses several advantages and suffers from other drawbacks. The TSE model is a parametric model, thus it is compact and easy to analyze. The compactness feature comes from the ability to fully describe a certain colour in few parameters. The ease of analysis and comparison comes from the ability of comparing different versions of the model, different colours and different intensity levels with each other by examining the corresponding parameters (see [1, chap. 4]). The PLSA model is a probabilistic model based on uncalibrated images from real-world. Thus, this model is more capable of correctly naming colours in images where acquisition conditions are unknown. The data set in which the model was trained contains built-in information about context as well. Secondly, the parametric model is powerful but lacks the ability of labelling uncalibrated images from real-world with high accuracy. These images are characterized by the variety of acquisition conditions from illuminant colour, angle of acquisition, shadows, reflectance...etc. on contrary to the psychophysical data obtained in an ideal controlled environment.

Given so, it is desired to create a model that enjoys both of the advantages of the two approaches. This model is required to be parametric and at the same time trained on context-based data from real-

world uncalibrated images. The steps of developing this model will be illustrated in the next section.

3 A Parametric Colour Naming Model for Uncalibrated Real-World Images

In the end of the previous section, it was concluded that it is needed to create a context-based parametric model. In order to achieve this goal, the parametric model of Benavente et al will be fitted using the context-based data set utilized in the probabilistic semantic model of van de Weijer et al.

3.1 Training the TSE model on Uncalibrated Data

As it is illustrated in the TSE model of Benavente et al., a data set based on psychophysical experiments is utilized. These psychophysical experiments were commenced by Seaborn et al.[9] to model human perception of colour. A fuzzy colour category map is resulted from the analysis of these experiments. This map is sampled uniformly to be used in fitting the parametric model. The samples used in the fitting are gathered in what will be mentioned in the rest of the text as Lut (*Look-up table*).

For the fitting phase, the psychophysical-based Lut will be replaced by another resulted from the analysis of Google data set images by the probabilistic model of van de Weijer et al. mentioned earlier. The resulting Lut is considered the fitting set on which the model will be fitted. As this fitting set is obtained from uncalibrated images in the real-world, thus, it contains context-embedded information.

The modified Lut was provided to the TSE model in the fitting phase. Using the same number of intensity levels and the same ranges that those levels cover, the entire Lut was used in the fitting phase. The model was fitted successfully on

the data provided as expected with some confusion areas. Fitted to the new data set, the fitting error of the model was calculated with acceptable error margin as demonstrated in table 1. The error is calculated using Mean Absolute Error (MAE).

<i>Method name</i>	<i>Num of samples</i>	<i>MAE fitting</i>	<i>% of well fitted samples</i>
Original TSE	1617	1.68%	96.60%
TSE_{uncal}	32768	3.94%	85.92%

Table 1: Statistics on the TSE model fitted to uncalibrated data.

where TSE_{uncal} is the original TSE model fitted to the uncalibrated data set alone.

3.2 Bi-Elliptic Triple Sigmoid Model

It was proved that the model can represent context-based data from uncalibrated images. After analyzing the fitting error and backtracking the misfitted samples it was noted that the colour brown achieves a high misfitting error which required a further analysis. A hypothesis was proposed that the brown colour exhibits a different behaviour than the rest of the chromatics and it is better fitted by another function due to the high error rate that it produce. This odd behaviour could be due to the non-ideal acquisition conditions of the real-life uncalibrated images and the embedded context information in the data samples.

The samples from the full Lut were gathered, divided into several lightness levels and plotted on a 3D plot (see figure 5). These samples are interpolated into a smooth surface maintaining the same characteristics. This was done to give an idea about the functions that will be needed to fit these data samples.

The analysis of the membership distribution of the brown colour confirmed the feasibility of well-fitting the data samples of brown colour using another elliptical sigmoid function to be located in

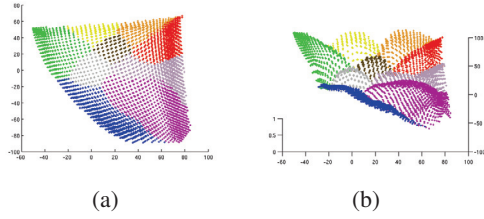


Figure 5: Samples plotted according to their memberships in space, the X-Y axes are the a-b coordinates and the Z axis is the membership value for this data sample (from 0 to 1), right figure is 2D prospective.

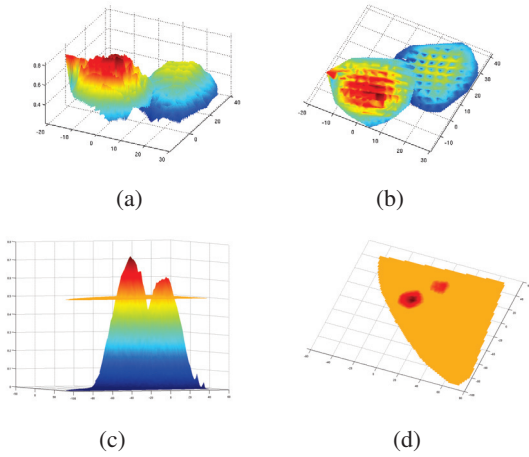


Figure 6: Analysis and surfing of the brown and achromatics-sum, right figures are 2D prospective.

the same vicinity of the achromatics-sum elliptical sigmoid. These two elliptical sigmoids would be surrounded by the triple sigmoid functions representing the rest of the chromatic colours. The rest of the chromatics maintained their expected triple sigmoid shape having the elliptic sigmoid of the achromatics to be the guiding centre. The work of Boynton[4] supports this hypothesis as well, and shows how the centroids of the colours are located in the perceptually uniform OSA space. This work illustrates how the chromatics surround the Grey-Black, White and Brown centroids uniformly.

By applying this hypothesis on the data samples

and modifying the fitting functions, the new Bi-Elliptical Triple Sigmoid (BETS) model was developed. It is a modification to the original Triple Sigmoid model with two elliptical centres instead of one. One of the elliptical centres is used to define Brown while the other is used as before to define the achromatics. Figure 7 shows the new Bi-Elliptical Triple Sigmoid model and table 2 shows the fitting error resulted from testing the new model.

<i>Method name</i>	<i>Num of samples</i>	<i>MAE fitting</i>	<i>% of well fitted samples</i>
Original TSE	1617	1.68%	96.60%
TSE_{uncal}	32768	3.94%	85.92%
BETS	32768	5.08%	88.97%

Table 2: Statistics on the Bi-Elliptical Triple Sigmoid model.

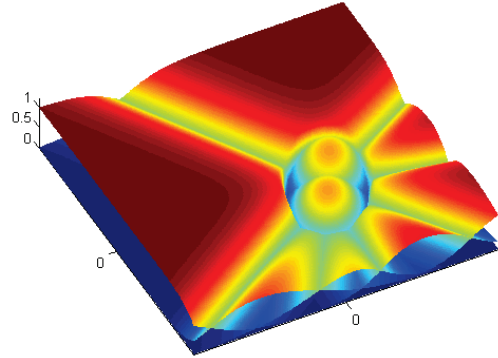


Figure 7: The Bi-Elliptical Triple Sigmoid model for an intensity level.

4 Context-Based Colour Naming in Real-World Images

After developing the new Bi-Elliptical TSE model and testing its fitting, it is needed to test it in reality as well. The fitting error is a good measure for the accuracy of the model in representing the fitting data set but it is not enough to give a whole picture about the model. It is also needed to prove that the model produces good results upon testing against

real-life images having different and unknown acquisition conditions. For these reasons the models were tested upon each step against the eBay images data set[10].

The eBay data set contains 4 categories (cars, dresses, pottery and shoes). Each category contains 10 images from each of the 11 basic colours. Along with each image there is a mask specifying the region in the image that is labelled with this colour. The testing takes place by applying the model to the pixels of each image within its mask and calculate the percentage of pixels correctly labelled with the expected colour (e.g. measuring the percentage of pixels labelled as red by the model to all the pixels inside the mask of an image containing a red car). Table 3 shows these results.

<i>Method</i>	<i>TSE</i>	<i>TSE_{uncal}</i>	<i>Bi-EllipticalTS</i>
Cars	53.34%	48.79%	51.58%
Dresses	73.25%	64.32%	78.18%
Pottery	61.5%	62.49%	71.29%
Shoes	72.19%	62.52%	70.61%
Total	65.07%	59.53%	67.91%

Table 3: Results of testing against the eBay data set.

5 Discussion and Future Work

From the experiments and analysis done so far, several conclusions could be drawn. The data from real-world uncalibrated images maintains a different shape, orientation and location from the data of psychophysical experiments. The parametric model fits the data but with minor modifications. After the analysis of this shifting and reformation, a conclusion was reached that the brown colour doesn't maintain the same behaviour of the other chromatics and tends to follow the behaviour of the achromatics in taking the shape of an elliptic sigmoid situated next to the achromatics and surrounded by the other chromatics. After successfully fitting the model to the context-based data

set it was tested against uncalibrated images. The new model achieved higher results than the original TSE model tested against the same data set separately. Thus, proving the hypothesis stated in the beginning of the possibility of creating an improved colour naming model by incorporating context-based real life data in the fitting set of a perceptual model. As for future work, a thorough analysis must be commenced for the areas between the elliptical centers and the other triple sigmoids and a search for a modification to the functions to provide better fitting should be established.

References

- [1] R. Benavente. *A Parametric Model for Computational Colour Naming*. PhD thesis, Universitat Autònoma de Barcelona, Bellaterra (Spain), Jun 2007.
- [2] R. Benavente, M. Vanrell, and R. Baldrich. Parametric fuzzy sets for automatic color naming. *Journal of the Optical Society of America A*, 25(10):2582–2593, Oct 2008.
- [3] B. Berlin and P. Kay. *Basic Color Terms: Their Universality and Evolution*. Center for the Study of Language and Inf, March 1969.
- [4] R. M. Boynton and C. X. Olson. Locating basic colors in the osa space. *Color Research & Application*, 12(2):94–105, 1987.
- [5] L. MacDonald A. Tarrant H. Lin, M. Luo. A cross-cultural colour-naming study. part iii—a colour-naming model. *Color Research and Application*, (26):270–277, 2001.
- [6] E. S. Konak, U. Gudukbay, and O. Ulusoy. A content-based image retrieval system for texture and color queries, 2002.
- [7] J. M. Lammens. *A Computational Model of Color Perception and Color Naming*. PhD thesis, University of New York, 1994.
- [8] P. Lennie. Single units and visual cortical organization. *Perception*, 27:889–935, 1998.
- [9] M. Seaborn, L. Hepplewhite, and T. John Stonham. Fuzzy colour category map for the measurement of colour similarity and dissimilarity. *Pattern Recognition*, 38(2):165–177, 2005.
- [10] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, jul 2009.

Hybrid Fusion: Beyond Early and Late Fusion for Texture Classification

Fahad Shahbaz Khan, Joost van de Weijer, and Maria Vanrell

*Computer Vision Centre/Computer Science Department Building O, Campus UAB, 08193 Bellaterra (Barcelona), Spain
E-mail:fahad@cvc.uab.es*

Abstract

This paper presents a novel method for combining multiple attributes in order to classify the different categories. We start by providing a detail analysis of how to optimally fuse color and shape information for texture classification. For this reason we analyze the two existing approaches, called early and late fusion, and argue that both approaches are suboptimal for some classes. To overcome this shortcoming, we propose to merge the two approaches into a single combined early and late fusion representation of an image. We further propose to combine this new hybrid fusion with a texture representation in an efficient way. Experiments have been conducted on a large dataset of ten different image categories and the results show that all these three cues are important for the task of texture classification and our proposed method increases the overall performance significantly. *Keywords:* Color vocabulary, Texture Vo-

cabulary, Texture Categorisation.

1 Introduction and Related Work

Images play a fundamental part in our daily communication and the large amount of pictures digitally available are not manageable by humans anymore. Visual categorization is a difficult task, in-

teresting in its own right, due to large variations between images belonging to the same class. Many features such as color, texture, shape, and motion have been used to describe visual information for visual categorization. This paper focuses on the difficult problem of texture categorisation.

A still open research question within the bag-of-words context is how to optimally fuse different images cues, like color and shape, into a single bag-of-words representation. Initially many methods only used the shape feature, predominantly represented by SIFT [6], to represent the image [9], [1] and [5]. However, more recently the possibility of adding color information has been investigated [7], [12], [8]. Most of the recent works combine color and shape at an early stage focussing on the photometrically invariant properties of the color descriptors. However, none of these methods provide a thorough analysis of the problem of what is the optimal approach to fuse shape and color.

Generally, the fusion of color and shape is carried out in the visual-vocabulary construction stage. Creating a visual vocabulary is a challenging task as the vocabulary should be able to describe widely varying classes. Some classes might have very distinctive color, some very characteristic texture patterns and some might be characterized by combining both features. There exist two main approaches to fuse color and shape into the bag-of-words representation. The first approach,

called early fusion, involves fusing local descriptors together and creating one joint shape-color vocabulary. The second approach, called late fusion, concatenates histogram representation of both color and shape, obtained independently. Most of the existing methods use early fusion [7], [12], [8]. One of the few works which compares both early and late fusion for image classification is done by [10] where both early fusion and late fusion have been discussed.

To this end, the paper has been organized as follows. In section 2 Vocabularies for texture, shape and color are discussed. Afterwards, in section 3, fusing multiple vocabularies is discussed and hybrid fusion scheme along with a texture representation is proposed. Section 4 presents the experimental details like the classification algorithm, the dataset used and the classification settings. Detailed experiments are shown in section 5. Finally, we sum up the conclusions.

2 Vocabulary for Texture, Color and Shape

Visual features color, shape and texture are used to characterise visual keywords. In our approach LBP and SIFT are used to create a texture and shape vocabulary. Two options are considered to create a color vocabulary namely Hue and Color Naming values. The main essence of our work lies in the combination of these three features. In the next sections texture, shape and color vocabularies have been discussed in detail.

2.1 Texture Vocabulary

For human perception texture is an important visual category. Texture is one of the most common low level features and plays an important role for the character of region for digital images. There are many different ways of solving the problem of texture analysis. In this regard we investigate the use of LBP for creating a texture vocabulary since

it is known to yield very good performance in recent texture studies [4] and [2]. For our experiments we have investigated different variations of LBP while creating a texture vocabulary.

2.2 Shape Vocabulary

Shape is one of the most common low level features and local Shapes are often regarded as one of the most discriminant features shared by different instances of an image category. In object recognition the shape of an object plays a pivotal role in searching for similar image objects. There are many different ways of solving the problem of shape analysis. In this regard we investigate the use of SIFT for creating shape vocabulary since it is known to yield very good performance in recent studies [5], [3].

2.3 Color Vocabulary

A color vocabulary is created to represent the color aspects of an image. The measured color values vary significantly due to large amount of variations. In this work color histogram approach is used in the Hue, Saturation, Value (HSV) color space [12] and Color Naming values mentioned in the work of [11]. Given a set of cluster centers (visual words), each image is then represented by a K (The number of clusters, K , optimized for the dataset) dimensional normalized frequency histogram $n(W/I)$. Where W denotes the visual words and I denotes the set of images. For clustering K -means method is used.

3 Fusing Multiple Vocabularies

After creating the color and shape vocabulary, both vocabularies are then combined in a flexible manner to achieve better performance. The discriminative power of each vocabulary varies for different classes. Some classes are distinguished by color and some by shape. We first analyze both early and late fusion which is followed by our proposed

approach of combining both early and late fusion. We further show that combining this hybrid fusion with a texture representation improve the results significantly.

3.1 Early Fusion and Late Fusion

In early fusion, the local features of color and shape are combined before quantization. The fusion involves concatenating the color features and shape features. From these combined vectors a joint vocabulary is obtained using the K-means algorithm. A weight vector β is introduced to tune the relative weight of the color and texture in the combined vocabulary V_{sc} .

$$V_{sc} = (\beta V_c, (1 - \beta)V_s) \quad (1)$$

where V_c are the color features and V_s are the texture features. The weight vector β is learned through cross-validation on the training data.

In late fusion the two features color and shape are computed independently. The two features are then fused together in one representation. Here the different vocabularies are concatenated after quantization. A weight vector α is introduced to obtain a combined histogram $n(w|I)$ of color and shape vocabularies for an image I .

$$n(w_{s\&c}|I) = \begin{bmatrix} \alpha n(w_c|I) \\ (1 - \alpha) n(w_s|I) \end{bmatrix} \quad (2)$$

where w is the number of total vocabulary words, w_c are Color words and w_s are shape words. The weight vector α is learned through cross-validation on training data.

Figure 1 highlights the two approaches used for combining the color and shape vocabulary.

3.2 Hybrid Fusion: Combining Early and Late Fusion

The work of [10] compares early versus late fusion. Both methods have their advantages and disadvantages. Early fusion obtains a vocabulary

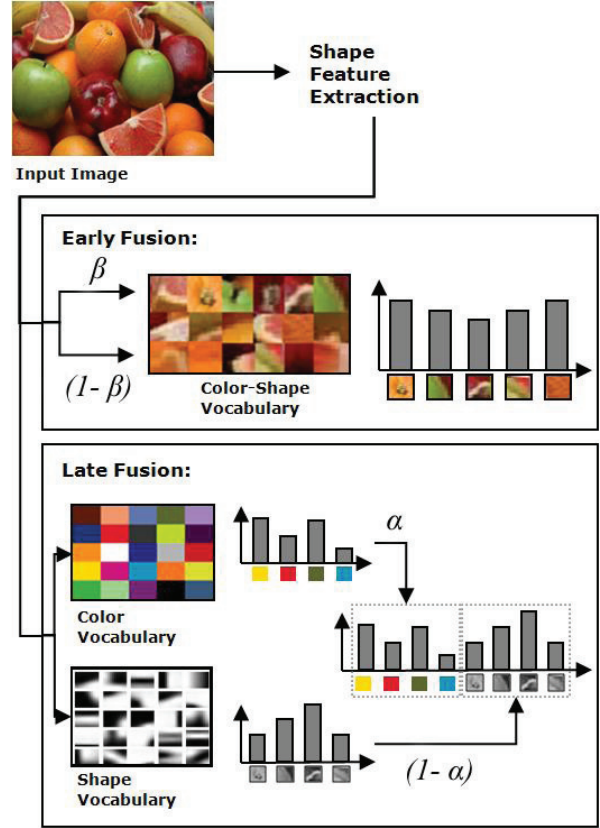


Figure 1: A Graphical explanation of early and late fusion schemes to combine color and shape information. The α and β parameters determine the relative weight of the two cues.

with a higher discriminative power, since the visual words describe both color and shape. Early fusion visual words could include red blobs, green lines, etc. . In late fusion the vocabularies are obtained by separately clustering the two cues shape and color. The image is thus represented as a distribution over shape-words and color-words. For example, if the image contains blobs and lines, and red and green features then from the late fusion representation it cannot be inferred that the image contains red lines or green blobs. In case of a class which is constant over both shape and color, early fusion representation is better. However for classes where only one of the cues is constant, late

fusion representation is preferred. To combine the advantages of the two representations we propose to combine early and shape fusion into a single histogram. This will be done in a late fusion manner as described in Eq. 2.

3.3 Combining Hybrid Fusion with Texture Vocabulary

We take one step further in fusing multiple features by combining the hybrid fusion with a texture vocabulary. In our experiments we have used LBP for the creation of texture vocabulary. As a next step, different variations of LBP has been tested since the primitive LBP representation proved unsuccessful for our data set. Thus the final histogram is a weighted combination of late and early fusion of color and shape with texture histogram.

4 Experimental Setup

The performance of combined vocabularies will be tested on a the classification task using SVM. Details of the proposed procedure are outlined in this section.

4.1 Dataset

The approach outlined above is tested on a dataset with 10 classes (Marble, Wood, Beads, Foliage, Graffiti, Lace, Clouds, Fruit, and Water) with 40 images for each class. The images in the dataset have been collected from Google, Flickr, and Corel image collection. Figure 8 shows some of the images from the dataset. The dataset is very challenging due to wide range of textures and color in it. For example, the Foliage class in our dataset has mostly green color, but there are few images in this class that has red color in it. Similarly there is a wide range of different texture patterns and colors in Marble and Graffiti class. There are a lot of variations in scale in the lace category.

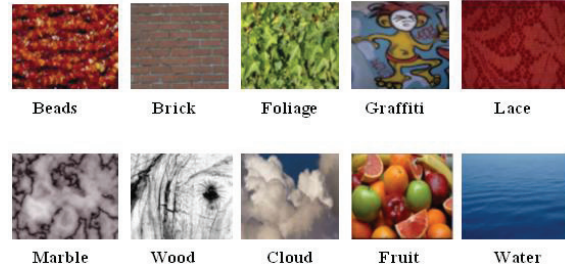


Figure 2: Typical examples of each class from the data set.

4.2 Classification Settings

The dataset has been divided into train set, validation set and test set. 25 images from each class are used for training and remaining 15 are used for testing. We believe the relative performance differences between various approaches to fuse multiple features to be independent of the detection method used. Thus an often used grid detector is employed where the patch centers lie 5 pixels apart. In our experiments multiclass non linear SVM with χ^2 kernel is used since it is known to produce best classification results [1]. To evaluate the classification performance we use the classification score. The classification score gives the percentage of correctly classified instances in the testset.

5 Experiments

This section explains in detail the creation of multiple vocabularies and the proposed methodology used for combining these vocabularies. Experiment 1 is about optimizing the individual vocabularies of texture. Experiment 2 provides an insight of color and shape vocabularies. Experiments 3 deals with combining both vocabularies in early and late fusion manner in order to optimize the classification performance. Finally in experiment 4 we combine the late and early fusion together in one representation. We further combine the hybrid and texture in one representation.

5.1 Experiment 1: Texture Vocabularies

This section provides detailed results obtained using only the texture information. As a first step a texture vocabulary has been created using local binary patterns. We explored different variations related to LBP such as Rotation Invariant LBP ($LBP_{P,R}^{ri}$), Uniform LBP ($LBP_{P,R}^{u2}$), Rotation Invariant with Uniform LBP ($LBP_{P,R}^{riu2}$), Rotation Invariant Variance LBP ($VAR_{P,R}$), and Joint distribution of Rotation Invariant Uniform LBP with its Variance ($LBP_{P,R}^{riu2}/VAR_{P,R}$).

LBP operator	P,R	Bins	Accuracy
$LBP_{P,R}^{ri}$	8, 1	36	54.16
$LBP_{P,R}^{u2}$	16, 2	243	62.30
$LBP_{P,R}^{riu2}$	8, 1	10	47.27
$LBP_{P,R}^{riu2}/VAR_{P,R}$	16, 2/8, 1	328	58.10
$LBP_{P,R}^{ri} + LBP_{P,R}^{u2}$	8, 1 + 16, 2	279	64.27

Table 1: Classification Score (percentage) using LBP.

5.2 Experiment 2: Color and Shape Vocabularies

In this experiment we evaluated individual color and shape vocabularies. The results shows that for the categories in this dataset shape is a more important cue than color.

Vocabulary	Vocabulary Size	Accuracy
<i>SIFT</i>	700	73
<i>HUE</i>	400	56
<i>ColorNaming</i>	400	58

Table 2: Classification Score (percentage) using Shape and Color Vocabularies.

5.3 Experiment 3: Early Fusion and Late Fusion

In this experiment we combined shape and color vocabularies. In Table 3 the results of these exper-

iments are summarised. The results using late fusion show a better classification score as compared to early fusion.

Vocabulary	Voc Size	Accuracy
<i>EarlyFusion(SIFT, HUE)</i>	1200	75
<i>LateFusion(SIFT, HUE)</i>	1100	77
<i>EarlyFusion(SIFT, CN)</i>	1200	77
<i>LateFusion(SIFT, CN)</i>	1100	79

Table 3: Classification Score (percentage) of Early and Late Fusion. Note that in both color cues (HUE and Color Names) late fusion performs better than early fusion.

5.4 Experiment 4: Combining Late and Early Fusion with Texture Vocabulary

The texture and shape/color are now combined by concatenating the histogram representation (late and early fusion) with LBP representation. The hybrid fusion of shape/color(color names) is denoted by *Hybrid(CN)* and shape/color(Hue) by *Hybrid(HUE)*.

Vocabulary	Vocabulary Size	Accuracy
<i>Hybrid(CN)</i>	2300	81
<i>Hybrid(HUE)</i>	2300	79
<i>Hybrid(CN)andLBP</i>	2310	83
<i>Hybrid(HUE)andLBP</i>	2310	81

Table 4: Classification Score (percentage) of hybrid fusion and combining hybrid fusion with LBP. Note that hybrid combination of Color Names with LBP provides the best results.

6 Discussion and Conclusions

The work presented in this paper is about classification on a large dataset of ten different texture categories using texture, shape and color features. We investigated the two popular approaches of combining color and shape namely early and late fusion. Our results show that both late and early

fusion have some shortcomings. The success of late fusion over early fusion and vice-versa depends on the nature of the categories in the dataset. In our case late fusion performs slightly better than early fusion. This could be due to the fact that in most of the categories only one of the cues, either color or shape, is constant. For example, foliage class is difficult to classify based on shape but relatively stable with respect to color appearance. In this case it is hard to find visual words based on early fusion that are consistent. As a next step late fusion has been analysed deeply by trying different combinations of vocabularies.

All three cues, color, shape and texture are found to be crucial to obtain a good overall classification score. Texture alone obtained 64.27%, color alone 58%, and shape alone provided 73% scores. The final combination got 83% which is obtained by combining hybrid fusion with LBP. When combining the vocabularies, color improves texture classification performance but best performance is achieved when shape has more influence than color.

Finally our final results suggest that all three cues are important for texture classification. Moreover fusing color and shape one step beyond early and late fusion improved the results. Combining our proposed hybrid fusion scheme with LBP representation further improved the overall classification score. The results also go to show that combining multiple vocabularies clearly outperforms the performance of individual vocabulary cues.

References

- [1] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A Comprehensive Study", *International Journal of Computer Vision*, 73(2): 213-238, 2007.
- [2] Topi Maenpää, Matti Pietikainen, "Classification with color and texture: jointly or separately?", *Pattern Recognition*, 37(8): 1629-1640, 2004.
- [3] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(10):1615-1630, 2005.
- [4] T. Ojala, M. Pietikinen, and T. Menp, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971-987, 2002.
- [5] S. Lazebnik, C. Schmid, and J. Ponce, "A Sparse Texture Representation Using Local Affine Regions", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.
- [6] D. G. Lowe, "Distinctive image features from scale-invariant points", *International Journal of Computer Vision*, 60(2):91-110, 2004.
- [7] A. Bosch, A. Zisserman, and J. Munoz, "Scene classification via plsa", *In Proc. ECCV*, 2006.
- [8] K. van de Sande, Th. Gevers, and C. Snoek, "Evaluation of color descriptors for object and scene recognition", *In Proc. CVPR*, 2008.
- [9] G. Dorko, C. Schmid, "Selection of scale-invariant parts for object class recognition", *In Proc ICCV*, 2003.
- [10] P. Quelhas and Jean-Marc Odobez, "Natural scene image modeling using color and texture visterms", *In Proc. CIVR*, 2006.
- [11] J. van de Weijer, C. Schmid, and J.J. Verbeek, "Learning color names from real-world images", *In Proc. CVPR*, Minneapolis, Minnesota, USA, 2007.
- [12] J. van de Weijer, C. Schmid, "Coloring local feature extraction", *In Proc. of the European Conference on Computer Vision*, volume 2, pages 334-348, Graz, Austria, 2006.

Towards non-supervised segmentation: a comparison of goodness measures based on saliency and contrast

Eduard Vazquez* and Ramon Baldrich⁺

^{**+} *Computer Science Dpt., Computer Vision Center,
Edifici O, Universitat Autònoma de Barcelona, 08193 Bellaterra (Cerdanyola), Spain*

^{*} *E-mail:eduard@cvc.uab.cat*

⁺ *E-mail ramon.baldrich@cvc.uab.cat*

Abstract

This paper analyses a combination of features to perform a correct segmentation without having a training set or benchmark. It is done by computing the called 'goodness' of a segmentation. If the proper function of goodness can be found, then is possible to automatically tune the parameters of any segmentation method, solving one of the most challenging problems in the segmentation field. In this article, the goodness is determined by using two different methods which compute features such as color distinctiveness and chromatic contrast based on Shannon's Information Theory. The main idea, present in both methods, is that a good segment (that is a region of a segmented image) have a high color distinctiveness and contrast. This article analyses this two methods and compare its performance using a benchmark. Results obtained shows that both methods are useful to compute the goodness of a segmentation.

Keywords: Color, segmentation, contrast, goodness measures.

1 Introduction

In Computer Vision, there are a great amount of algorithms which requires a good segmentation as a preprocessing step. Nonetheless, existing segmen-

tation methods require to adapt its parameters to each problem, and even each image, to yield good results. Therefore, general purpose segmentation, understood as a process without human supervision, is still a challenging problem with no solution. This paper addresses this problem presenting a combination of features which are aimed to give a measure of correctness of a segmentation when no benchmark exist. Consequently, yielding a segmentation without human supervision. The aim of this article is not to propose a solution for this problem, but to analyze the behavior of some existing methods.

One of the earliest works to decide the goodness of a segmentation is the JSEG segmentation method introduced in [2]. JSEG is a two-step segmentation schema. First, a clustering of the color space is performed. Afterwards, a criterion of *good* segmentation is applied using the spatial coherence of the image, *i.e.*, the information of the spatial relation existing between the pixels in the image space. Another schema proposes that a good segmentation region should be formed by strongly connected pixels with homogeneous colors [8]. This second approach follows a similar idea of the of the work introduced by Heidemann in [12], which uses the color distinctiveness as a measure of goodness. Other measures to describe the correctness of a segment are the

homogram proposed in [4], a calculus based in the Bhattacharyya distance [11], a probabilistic approach as explained in [9] or a graph-based method as explained in [13]. The methods studied in this article belong to those family methods that use contrast as a criteria of the goodness of a segmentation (e.g. [7][12]).

In this article, we analyze the performance of two methods. First, the one introduced in [12]. Such method proposes a goodness function for color segmentation, which allows to predict whether the segmented regions will be stable against noise, variation of lighting, and change of viewpoint. As such a measure, color saliency defined from average border contrast is proposed. The second method to be analyzed, is Color Boosting, introduced by van de Weiejer *et al.* in [10]. This method, is based on the self information of the chromatic transition (first order derivatives of the image). It is shown in [10] that Color Boosting improve the color distinctiveness in a framework of interest points detection. In the present work, we want to analyze if this characteristic can be also useful to decide the correctness of a segmentation.

To decide if these methods are able to take a correct decision, we compare their results using a benchmark. Thus, a good method of correctness should choose among different segmentations the one that have the best score in the benchmark. Such a coincidence between the non-supervised methods and the benchmark would imply that the measure of correctness applied would be a useful tool to perform a non-supervised segmentation, by automatically selecting the best set of parameters for each image.

This article is organized as follows: in section 2 we explain the method color saliency proposed in [12]. Afterwards, in section 3 we explain Color Boosting. Finally, sections 4 and 5 presents results obtained and a discussion of this article respectively.

2 Heidemann's color saliency

The approach of Heidemann introduced in [12] proposes a goodness function for color segmentation, which allows to predict whether the segmented regions will be stable against noise, variation of lighting, and change of viewpoint. Color saliency defined from average border contrast of the segmented image. This work points out that the idea to maximize contrast is segmentation has been previously considered. Nevertheless, Heidemann analyses the problem and shows how it leads to improve region stability. Experiments for three different algorithms show that the effect is independent of the particular functional principle of segmentation. Thus, the measure can be applied for the automatic and context-free optimization of segmentation parameters.

The measure proposed is based of color distinctiveness of the regions of the segmented image. Thus, as far is the (euclidean) distance between neighboring regions, the better is the segmentation. Given an image I having three chromatic channels for each pixel (x, y) , we compute a segmentation from which I is divided in N_R non-overlapped regions. The *region color* is defined as the mean color of this region in the original image.

The *region saliency* $S_R(R_i)$ is defined as the average color difference of R_i to the neighboring regions. Concretely, let the boundary of R_i be given as a set $B(R_i)$ consisting of $N_B(R_i)$ different pixels. Then $S_R(R_i)$ is calculated along the boundary as

$$S_R(R_i) = \frac{1}{N_B(R_i)} \sum_{(x,y) \in B(R_i)} \frac{1}{N_{diff}(x,y)} \times \sum_{R_j(x',y') | (x',y') \in Neigh4(x,y)} \| \overline{\cdot}(R_i) - \overline{\cdot}(R_j) \| \quad (1)$$

Here, $\| \cdot \|$ denotes the color distance measure for the particular segmentation algorithm used. For color spaces such as RGB, * * * or * * * the Euclidean distance is used.

The first sum in Eq. 1 is over all boundary pixels (x, y) . The second sum goes over the pixels (x', y') within a 4-neighborhood of (x, y) , the 4-neighborhood being denoted by $Neigh4(x, y)$. To each neighboring pixel (x', y') the corresponding region $R_j(x', y')$ has to be found, so that the Euclidean distance between the region colors $\overline{c}(R_i)$ and $\overline{c}(R_j)$ can be calculated. $N_{diff}(x, y)$ denotes the number of pixels of $Neigh4(x, y)$ that belong to a different region, not to R_i . This factor is introduced to avoid dilution of the average distance in case there is, e.g. only one neighboring pixel which belongs to a different region. $N_{diff}(x, y)$ is at least 1 since (x, y) is part of the boundary, the maximum value is $N_{diff}(x, y) = 4$ in the case that (x, y) is a region consisting of an isolated pixel.

The Saliency measure of an image I denoted by $S(I)$ is given by the average over all its regions

$$S(I) = \frac{1}{N_R} \sum_{R_i \in I} S_R(R_i) \quad (2)$$

Summarizing, $S(I)$ is a measure of the color distinctiveness of the regions of a segmented image.

In the next section we explain another way to include color distinctiveness to and contrast to decide the goodness of a segmentation.

3 Color Boosting

The second method that will be analyzed is Color Boosting, as introduced in [10]. This method, is based on the self information of the chromatic transition (first order derivatives of the image). It is shown in [10] that Color Boosting improve the color distinctiveness in a framework of interest points detection. In the present work, we want to analyze if this characteristic can be also useful to decide the correctness of a segmentation.

The color saliency method by Van de Weijer *et al.* [10] is inspired by the notion that a feature's saliency reflects its information content. Consider

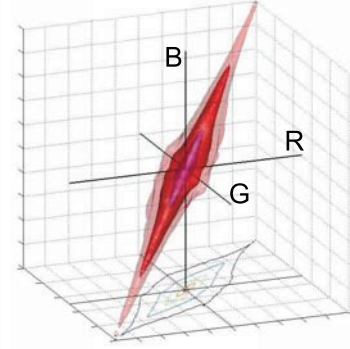


Figure 1: Statistics of the first order derivatives computed on the COREL dataset.

an image $\mathbf{f} = (R \ B)^t$. The information content, I , of an image derivative \mathbf{f}_x , according to information theory, is given by the logarithm of its probability :

$$I = - \log(\mathbf{f}_x) \quad (3)$$

Hence, color image derivatives which are equally frequent have equal information content. To map image derivatives to a saliency map, a function g is required for which the following holds:

$$(\mathbf{f}_x) = (\mathbf{f}'_x) \leftrightarrow |g(\mathbf{f}_x)| = |g(\mathbf{f}'_x)| \quad (4)$$

The saliency function g transfers color image derivatives to a space where their norm is proportional to their information content.

In Fig. 1, the distribution of color derivatives for the COREL dataset is given. The derivatives form an ellipsoid-like distribution, of which the longest axis is along the luminance direction. This indicates that equal displacements are more informative along the color directions (perpendicular to the luminance) than in the luminance direction. The saliency transformation in [10] is restricted to a transformation based on known color spaces. Now we propose a more general transformation to compute g in that it is not fixed to a pre-defined color space.

Let the distribution of the ellipsoid to be described by the covariance matrix \mathbf{M} :

$$\mathbf{M} = \overline{\mathbf{f}_x (\mathbf{f}_x)^t} = \begin{pmatrix} \overline{R_x R_x} & \overline{R_x x} & \overline{R_x B_x} \\ \overline{R_x x} & \overline{x x} & \overline{x B_x} \\ \overline{R_x B_x} & \overline{x B_x} & \overline{B_x B_x} \end{pmatrix} \quad (5)$$

where the matrix elements are computed by

$$\overline{R_x R_x} = \sum_{i \in S} \sum_{\mathbf{x} \in X^i} R_x(\mathbf{x}) R_x(\mathbf{x}) \quad (6)$$

where S is a set of images, and X^i is the set of pixels coordinates \mathbf{x} in image i . Matrix \mathbf{M} describes the derivatives energy in any direction \hat{n} . This energy is computed by $(\hat{n}) = \hat{n}^t \mathbf{M} \hat{n}$. Matrix \mathbf{M} can be decomposed into eigenvector matrix \mathbf{U} and eigenvalue matrix $\mathbf{\Lambda}$ according to $\mathbf{M} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^t$. This provides us with the saliency function g :

$$\mathbf{g}(\mathbf{f}_x) = \mathbf{\Lambda}^{-1} \mathbf{U}^t \mathbf{f}_x \quad (7)$$

Substitution of Eq. 7 into Eq. 5 yields

$$\mathbf{g}(\mathbf{f}_x) (\mathbf{g}(\mathbf{f}_x))^t = \mathbf{\Lambda}^{-1} \mathbf{U}^t \mathbf{U} \mathbf{\Lambda} \mathbf{U}^t \mathbf{U} \mathbf{\Lambda}^{-1} = \mathbf{I} \quad (8)$$

meaning that the covariance matrix of the transformed image is equal to the identity matrix. This implies that the derivative energy in the transformed space is equal in all directions.

Using this procedure, we can generate an image where the energy of the information of the first order derivatives is considered. Fig. 2c shows an example.

4 Results obtained

In this section we explain the procedure to follow in order to determine the correctness of the methods detailed in sections 2 and 3. Afterwards we show prior results obtained.

To figure out the validity of both methods, we use a widely used dataset and benchmark, introduced in [3]. First, we take an image that will be

segmented using two different methods. One of them, is the Mean Shift (MS) [5]. It have a public available version, called EDISON [5] and has been widely used and studied. As a second segmentation method, we use the one introduced in [14], as a new method which outperforms Mean Shift. The next step, is to compute an error measure. The one that we propose, is the Boundary Displacement Error (BDE) [1] since fits to a segmentation problem where there is no control about the number of segments. Having all these segmentations we compute their saliency index $S(I)$ following equations 1 and 2. Afterwards, we apply color boosting to the original image (Fig. 2a) and we generate the energy image of it (energy of its first order derivatives). The energy image is showed in Fig. 2c. Note its great similarity with human segmentation (Fig. 2b). Then, we compute the intersection between the energy image and the borders of the regions of each segmentation. That is, the overlapping of each region normalized by the number of points of the region's border $N_B(R_i)$.

Hence, if the goodness methods analyzed in this article predict that the best segmentation is the same as the one predicted using BDE and a benchmark, it would means that these methods can be effectively used to determine the goodness of a segmentation without consulting a benchmark. Table 1 summarizes results obtained. We can see how both Heidemann and Color Saliency correctly predict that the best segmentation is RAD2, as confirmed by the BDE score obtained using the benchmark.

5 Discussion and further work

In this article we have seen how the measure of goodness analyzed have a correct behavior. Heidemann computes the color distinctiveness of a region, whereas the method proposed using Color Boosting, introduces the chromatic information of every region obtained. The second difference is that whereas Heidemann uses the segmented im-

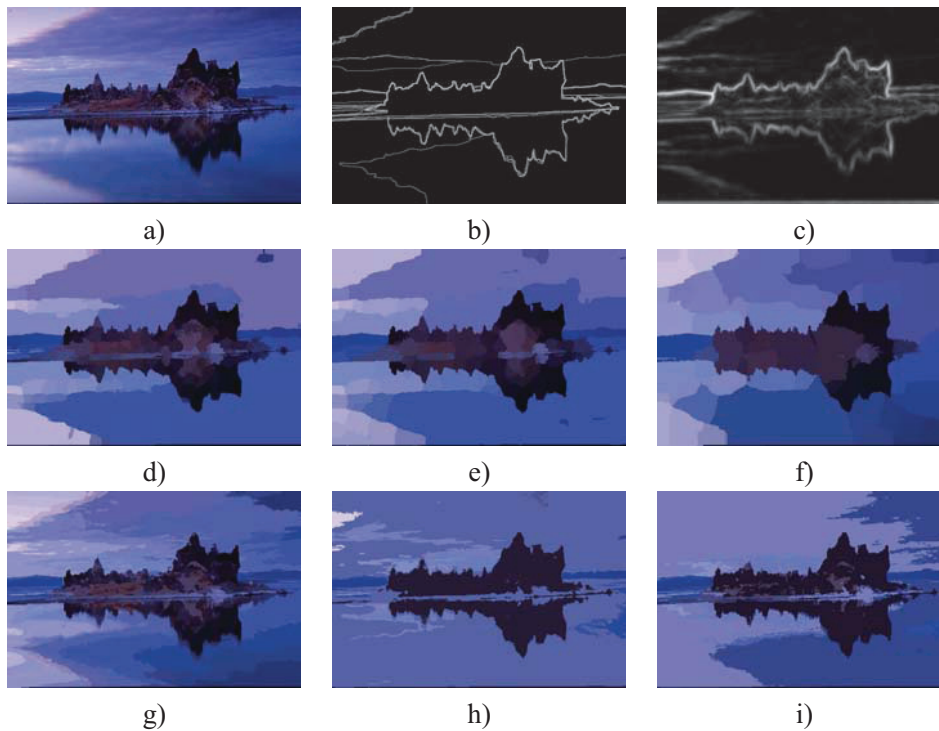


Figure 2: (a) Original image extracted from the Berkeley dataset. (b) Human Segmentation (benchmark). (c) Energy of (a) after applying Color Boosting. (d-f) Segmentations obtained with Mean Shift.(g-i) Segmentations obtained with RAD. As can be seen h) is the segmentation that looks more similar to b) being the one selected as the best segmentation for both tested methods as well as the error measure (BDE).

Table 1: Results obtained. WE can see how Heidemann and Color Boosting predict that the best segmentation is 'RAD2'. It corresponds with the BDE measure computed using a benchmark.

	Heidemann	C. Boosting	BDE
MS1	8.27E+008	1.87E-005	11.28
MS2	1.19E+009	1.91E-005	13.07
MS3	1.54E+009	1.74E-005	15.75
RAD1	9.60E+008	1.77E-005	11.82
RAD2	5.50E+010	2.92E-005	10
RAD3	1.02E+010	2.53E-005	11.15

age, the second approach uses the original image. Since both methods give correct results, the next steps is to combine both ideas in a single measure as well as to perform a further analysis of its performance.

References

- [1] Huang, Q. and Dom, B.: Quantitative methods of evaluating image segmentation. IEEE International Conference on Image Processing 3 (1995) 53–56
- [2] Deng Y. and Manjunath B.S.: Unsupervised segmentation of Color-Texture Regions in Images and Video. IEEE Transactions on Pat-

- tern Analysis and Machine Intelligence **23**(8) (2001) 800–810
- [3] Martin, D., Fowlkes, C., Tal, D., Malik, J.: A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. *Proc. Eighth Int. Conf. Computer Vision* **2** (2001) 416–423
 - [4] Cheng, H.D., Jiang, X.H., Wang, J.: Color image segmentation based on homogram thresholding and region merging. *Pattern recognition* **35**(2) (2002) 373–393
 - [5] Christoudias, C., Georgescu, B., Meer, P.: Synergism in low level vision. *International Conference on Pattern Recognition* **4** (2002) 150–155
 - [6] Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(5) (2002) 603–619
 - [7] Ge, F. and Wang, S. and Liu, T.: Image-Segmentation Evaluation From the Perspective of Salient Object Extraction. *Proc. IEEE Conference on Computer Vision and Pattern Recognition* **1**(2006) 1146–1153
 - [8] Macaire, L. and Vandenbroucke, N. and Postaire, J.G.: Color image segmentation by analysis of subset connectedness and color homogeneity properties. *Computer Vision and Image Understanding* **102**(1) (2006) 105–116
 - [9] Micusik, B., Hanbury, A.: Automatic image segmentation by positioning a seed. *Proc. European Conference on Computer Vision (ECCV)* (2006)
 - [10] van de Weijer, J. and Gevers, T. and Bagdanov, A.D.: Boosting Color Saliency in Image Feature Detection. *IEEE Transactions on Pattern Recognition and Machine Intelligence* **28**(1) (2006) 150–156
 - [11] Donoser, M. and Bischof, H.: ROI-SEG: Unsupervised Color Segmentation by Combining Differently Focused Sub Results. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on* (2007) 1–8
 - [12] Heidemann, G.: Color segmentation robust to brightness variations by using B-spline curve modeling. *Image and Vision Computing* **26**(2) (2008) 211–227
 - [13] Wattuya, P. and Jiang, X. and Rothaus, K.: Combination of Multiple Segmentations by a Random Walker Approach. *Lecture Notes in Computer Science* **5096**(214–223) (2008) 211–227
 - [14] Vazquez, E. and van de Weijer, J. and Baldrich, R.: Image Segmentation in the Presence of Shadows and Highlights. *10th European Conference on Computer Vision, LNCS* **5305** (2008) 1–14

Computational Color: Representation, Constancy and Psychophysics

Javier Vazquez-Corral and Maria Vanrell

*Computer Vision Center, Departament de Ciències de la Computació,
Universitat Autònoma de Barcelona, Edifici O, Campus UAB,
08193 Bellaterra (Cerdanyola del Valles), Catalunya
E-mail:javier.vazquez@cvc.uab.cat*

Abstract

Computational Color is a multidisciplinary research field. In this paper we will focus on its applications to Computer Vision. Color in Computer Vision has been used in different tasks: Segmentation [4], Object Recognition [12], Saliency [11], Induction [14], Naming [1], Constancy [10]... The aim of this paper is to review about some of these topics and the authors' contributions to them. In particular we will discuss three different open problems: Color Representation, Color Constancy and Psychophysical evaluation.

Keywords: Computational Color, Color and Texture analysis, Color Constancy, Psychophysics

1 Introduction

Color has been widely studied as a multidisciplinary topic for a long time. Artists, Biologists, Physics, Physiologists even Philosophers have tried to understand and to explain color from different points of view. This role of color has had an important repercussion in the study of computational models for it. In this paper we will focus on three different problems (Color Representation, Color Constancy and Psychophysical evaluation) where we

have been working during the last years. We will explain the problems and a tangential explanation of our solution to them. Then, this paper is organized as follows. First, we will explain the Color Representation problem. Later on we will move to the Constancy, and eventually we will arrive to its Psychophysical evaluation. Finally we will sum up some conclusions.

2 Color Representation

Several color spaces had been defined in color science [20], each one with a certain intention. Some of them, as RGB or CMY, trying to improve the image acquisition, visualization and results in printing devices. Others, the uniform spaces, such as, CIELAB or CIELUV, to represent perceptual similarity considering an Euclidean distance.

In computer vision, a usual way to work in color has been to extend gray-level methods to be applied on the RGB channels separately.

However, current spaces does not always preserve the features perceived in the color image to the channel representation. Two different situations can occur.

- There can be a high correlation between the three channels.

- The principal features of the color image are not represented in the channels individually, because these features are emerging from a combination of the channels (see Figure 1)

Therefore, our hypothesis is based on the fact that these situations can usually happen in color-texture images. Current spaces, therefore, are not able to correctly extract the texture information in the channels.

For this reason, in this work we propose a new color space that adapt to the image context. The main goal of our space is to facilitate the extraction of information such as blobs (where we will concentrate) in order to further represent the color-texture structure of the image.

To this end, we first extract the ridges of the color histogram of the image r_1, \dots, r_n [18]. These ridges select the essential the color information of the image. We can simplify the ridges by their main representative color point, then $r_i = (R_i, G_i, B_i)$. Then we select the three color ridges that form the biggest gamut among them

$$(r_1, r_2, r_3) = \max_{i,j,k, i \neq j, j \neq k, k \neq i} \text{Area}(r_i, r_j, r_k) \quad (1)$$

We select this three ridges in order to maximize the information. Later on, we compute a center point p where the three ridges behave as orthogonal as possible. This means

$$p = \arg \min(v_1 \cdot v_2 + v_1 \cdot v_3 + v_2 \cdot v_3) \quad (2)$$

where

$$v_1 = p - r_1, v_2 = p - r_2, v_3 = p - r_3 \quad (3)$$

From now on, the point p will be de center of the space. Finally, by using a Gram-Schmidt normalization we convert v_1, v_2, v_3 in an orthogonal space. In Figure 2 we can see the original image and the three channels found.

3 Color Constancy

The color we perceive from an object depends on three different aspects: the reflectance of the ob-

ject, the sensors of the capturing device and the illumination of the scene. Illumination, then, can substantially change the perception of an image and can disturb in many computer vision tasks such as tracking or object recognition. Then, to find an image representation independent from the illumination is useful and is the research goal in Color constancy. However, this problem is overdetermined, and, consequently, it has been tackled from different points of view.

This illuminant independency can be reached in two different ways. The first one is to create an image representation where the illuminant has been cancelled [5],[9]. This is usually called Color Normalization. The second one is based on the idea of discount the illuminant. It means, try to estimate the color of the illuminant in order to convert the image in a canonical one with the original colors of the scene. This is Color constancy.

Color constancy has usually been tackled using low-level assumptions: Image statistics, illuminant and sensors constraints, etc. Image statistics are the basis for methods as Grey-World [3], White-Patch [13], Shades of Grey [7], Grey-Edge[16]. These statistical methods are non-calibrated. On the other hand, illuminant and sensor restrictions are applied to C-Rule [8], that is, is a calibrated method.

There are other approaches, as the Bayesian Color Constancy methods. These methods are based on the Bayes' Rule. For example, Color-by-Correlation [6] or Bayesian Color Constancy [2]. And, quite related to the Bayesian Color Constancy are the Voting methods [15].

On the other hand, there are a few number of methods using High-Level hypothesis. In fact, this idea is just starting to be checked, for example in [17] where Van de Weijer based is selection in image annotation.

To deal with the color constancy ill-posed problem we are working on a new high-level method that suggest the use of Semantic Categories to solve the Color constancy Problem. The main hypothesis underlying here is that illuminants allow-

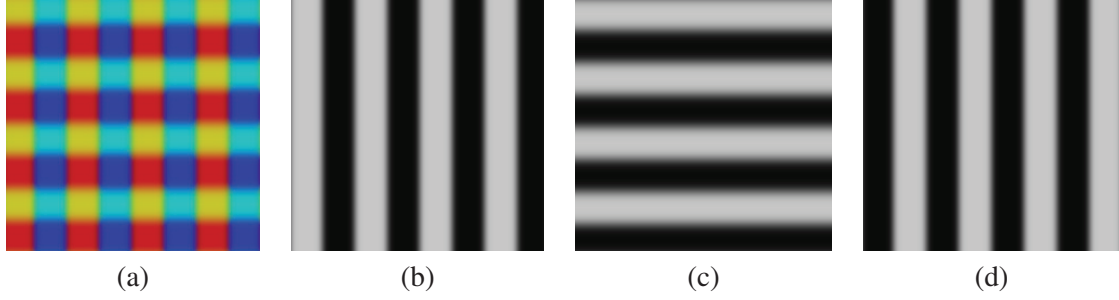


Figure 1: RGB channels of image (a), where (b) is the red channel, (c) the green channel, and (d) the blue channel

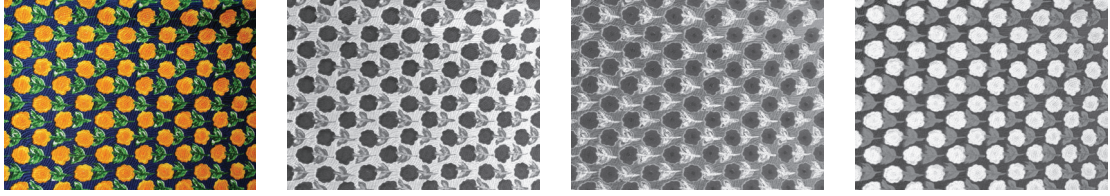


Figure 2: Examples of the results in the new color space: Original image (left) and channels found

ing a high association degree between image colors and semantic categories are the most plausible. We have tested this method by using the Categories of the Focals from Color Names extracted from [1]. Our results are achieving the current state-of-the-art in color constancy. Some results are shown in Figure 3 and they are summed up in Table 1. This results have been computed in the same Real World Dataset used in [16], learning with the 33% of scenarios. The error measure used is the angular error defined as

$$e_{ang} = \arccos \left(\frac{p_w \hat{p}_w}{\|p_w\| \|\hat{p}_w\|} \right) \quad (4)$$

where p_w is the actual white point of the scene illuminant, and \hat{p}_w is the estimation of the white point given by the method.

4 Psychophysical evaluation

Computing an error measure in Computational Color Constancy is a controversial topic. It is quite

Method	RMS
Our method	14.60°
Grey-Edge	14.73°
Max-RGB	16.28°
no-correction	20.54°

Table 1: Angular error results on the 150 image subset of the Real World Image Dataset

extended the angular error measure, that, as it has explained before, is, the angle between the physical solution of the image and the solution given by the method. Formally,

$$e_{ang} = \arccos \left(\frac{p_w \hat{p}_w}{\|p_w\| \|\hat{p}_w\|} \right) \quad (5)$$

where p_w is the actual white point of the scene illuminant, and \hat{p}_w is the estimation of the white point given by the method.

Anyway, we have shown that this measure does not always agree with the observers choice when asking them to select the most natural image. In

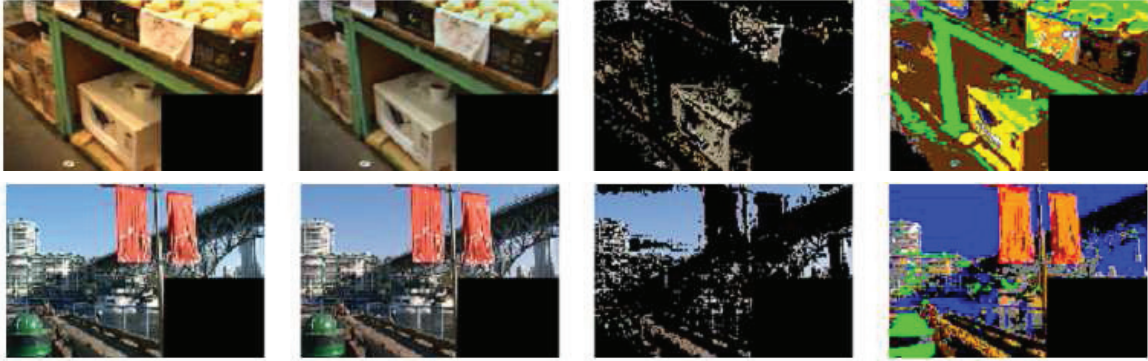


Figure 3: Examples of the method: Original image (left), corrected image (center-left), classified values (center-right), semantic interpretation of the solution(right)

fact, in less than the 50% of the cases, the observers select the image with minimal angular error. we proved this hypothesis in [19]. Regarding this problem, we have define a new measure, the perceptual angular error distance able to cope with the human preferences. this measure is defined as follows

$$e_{ang_{perc}} = \arccos \left(\frac{p_{w_{perc}} \hat{p}_w}{\|p_{w_{perc}}\| \|\hat{p}_w\|} \right) \quad (6)$$

where $p_{w_{perc}}$ is the psychophysical natural illuminant selected by the observers, and \hat{p}_w is the estimation of the white point given by the method.

All these results come from the experiment performed in [19]. Here, we mimic the explanation of this experiment. We used a set of 83 images from a new image dataset built for this experiment. The camera calibration allows us to obtain the CIE1931 XYZ values for each pixel and consequently, we converted 83 images from CIE XYZ space to CIE sRGB. Following this, we replaced the original illuminant by D65 using the chromaticity values of the grey sphere that was present in all image scenes.

Once the original illumination was digitally removed, 5 new pictures were created by re-illuminating the scene with 5 different illuminants,

totaling 415 images. Afterwards, three color constancy algorithms were applied on these newly created images (the selected algorithms are Grey-World [3], Shades-of-Grey [7] and MaxName (our new algorithm)). Consequently, we obtain one solution per test image and per algorithm, totaling 1245 different solutions. These solutions were converted back to CIE XYZ to be displayed on a calibrated CRT monitor (Viewsonic P227f) using a visual stimulus generator (Cambridge Research Systems ViSaGe). The experiment was conducted in a dark room.

The experiment was conducted on 10 nave observers recruited among university students and staff (none of the observers had previously seen the picture database). All observers were tested for normal color vision using the Ishihara and the Farnsworth Dichotomous Test (D-15). Pairs of pictures (each obtained using one of two different color constancy algorithms) were presented one on top of the other on a grey background (31 Cd/m²). The order and position of the picture pairs was random. Each picture subtended 10.5 x 5.5 degrees to the observer and was viewed from 146 cm. This brings us to 1245 pairs of observations per observer.

For each presentation, observers were asked to select the picture that seemed most natural, and to

make the selection by pressing a button on an IR button box. The set up (six buttons) also allowed observers to register how convinced they were of their choice (e.g. strongly convinced, convinced, and marginally convinced). There was no time limit but observers took an average of 2.5 seconds to respond to each choice. The total experiment lasted 90 minutes approximately (divided in three sessions of 30 minutes each)

5 Conclusion

In this paper we have explained the work we have done during the last years. This work can be roughly split in three different topics. Firstly, we have created a color space which adapts to the image content. This space extracts the different features of the image in different channels, therefore, it can improve the color-texture description of the image. Secondly, we have formulated a Color Constancy method from a new point of view: the use of categories to correct the illuminant. This method achieves current state-of-the-art results in Color Constancy by introducing high-level prior knowledge. Finally, we have shown that the usual error measure in Color Constancy does not correlate with the human preferences. For this reason we have defined a new error measure to better cope with the human preferences.

Current research lines are focusing on the extension of the categories used in the Color Constancy method, and the application of the proposed color space to object segmentation and saliency.

6 Acknowledgments

This work has been partially supported by projects TIN2007-64577 and Consolider-Ingenio 2010 CSD2007-00018 of Spanish MEC (Ministry of Science)

References

- [1] R. Benavente, M. Vanrell, and R. Baldrich. Parametric fuzzy sets for automatic colour naming. *Journal of the Optical Society of America A*, 25(10):2582–2593, 2008.
- [2] D. H. Brainard and W. T Freeman. Bayesian color constancy. *Journal of the Optical Society of America A*, 14:1393–1411, 1997.
- [3] G. Buchsbaum. A spatial processor model for object colour perception. *J. Franklin Inst.*, 310:126, 1980.
- [4] H.D. Cheng, X.H. Jiang, Y. Sun, and J. Wang. Color image segmentation: advances and prospects. *Pattern Recognition*, 34(6):2259–2281, 2001.
- [5] G. Finlayson and M. Drew. White-point preserving color correction. In *G.D. Finlayson and M.S. Drew, White-point preserving color correction, Proc. IST/SID 5th Color Imaging Conference*, pp. 258-261, 1997., 1997.
- [6] G.D. Finlayson, S.D. Hordley, and P.M. Hubel. Color by correlation: A simple, unifying framework for color constancy. 23(11):1209–1221, November 2001.
- [7] G.D. Finlayson and E. Trezzi. Shades of gray and colour constancy. In *Color Imaging Conference*, pages 37–41, 2004.
- [8] D.A. Forsyth. A novel algorithm for color constancy. *International Journal of Computer Vision*, 1990.
- [9] T. Gevers and A. W. M. Smeulders. Color based object recognition. *Pattern Recognition*, 32:453–464, 1999.
- [10] S. D. Hordley. Scene illuminant estimation: past, present, and future,. *Color Research and Application*, 31(4):303–314, 2006.

- [11] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, Nov 1998.
- [12] F. Shahbaz Khan, J. Van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *Proceedings of the ICCV, Kyoto, Japan, 2009*.
- [13] E. H. Land and J. J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, 1971.
- [14] X Otazu, M Vanrell, and C.A Párraga. Multiresolution wavelet framework models brightness induction effects, Feb 2008.
- [15] G. Sapiro. Color and illuminant voting. *PAMI*, 21(11):1210–1215, November 1999.
- [16] J. van de Weijer, Th. Gevers, and A. Gijsenij. Edge-based color constancy. *IEEE Transactions on Image Processing*, 16(9):2207–2214, 2007.
- [17] J. van de Weijer, C. Schmid, and J. Verbeek. Using high-level visual information for color constancy. In *International Conference on Computer Vision*, oct 2007.
- [18] E. Vazquez, J. van de Weijer, and R. Baldrich. Image Segmentation in the Presence of Shadows and Highlights. In *Proceedings of the 10th European Conference on Computer Vision: Part IV*, pages 1–14. Springer-Verlag Berlin, Heidelberg, 2008.
- [19] J Vazquez-Corral, C.A Párraga, M Vanrell, and R Baldrich. Color constancy algorithms: Psychophysical evaluation on a new dataset, May-June 2009.
- [20] G. Wyszecki and W.S. Stiles. *Color science: concepts and methods, quantitative data and formulae*. John Wiley & Sons, 2nd edition, 1982.

Object Pixel-Level Categorization using Bag of Features

David Aldavert and Ricardo Toledo

Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra, SPAIN

E-mail: {aldavert, ricard}@cvc.uab.cat

Abstract

In this paper we present a pixel-level object categorization method suitable to be applied under real-time constraints. Since pixels are categorized using a bag of features scheme, the major bottleneck of such an approach would be the feature pooling in local histograms of visual words. Therefore, we propose to bypass this time-consuming step and directly obtain the score from a linear Support Vector Machine classifier. This is achieved by creating an integral image of the components of the SVM which can readily obtain the classification score for any image sub-window with only 10 additions and 2 products, regardless of its size.

Keywords: Object Recognition, Bag of Features, Integral Images.

1 Introduction

A method for robustly localizing objects is of major importance towards creating smart image retrieval tools able to search in digital image collections. In the last years, object recognition in images has seen impressive advances thanks to the development of robust image descriptors [3] and simple yet powerful representation method such as the *bag of features* [7, 10].

In this work we propose a new method for fast pixel-wise categorization based on the bag of fea-

tures object representation. Given that the method will have to be applied at every pixel of the image, it is essential to optimize it to perform in the least possible time. Although different bag of features approaches have been proposed, all of them consist on four basic steps. Namely, feature extraction from the image, feature quantization into visual words, accumulation of visual word into histograms and classification of the resulting histogram. The accumulation step is an even more critical bottleneck for an object localization using bag of features. Since no geometrical information is used, many sub-windows of the image have to be evaluated. Some authors addressed this problem by reducing the number of evaluated windows, either by using a pre-processing step or by searching the best sub-window as in an optimization problem. This is done by defining an upper bound on the SVM classification, and using branch and bound to discard uninteresting areas. Other authors have focused on accelerating the accumulation step. In the approach by Fulkerson et al. [10], the authors speed up the accumulation step using integral images in a sliding windows based analysis of the image. For this speed-up measure to be effective, it is important to use small dictionaries. In order to compress the dictionary without losing classification accuracy, they propose to use Agglomerative Information Bottleneck (AIB) to create a coarse-to-fine-to-coarse architecture that is optimized for discrimination of object versus non-object. This approach shares some similitudes with

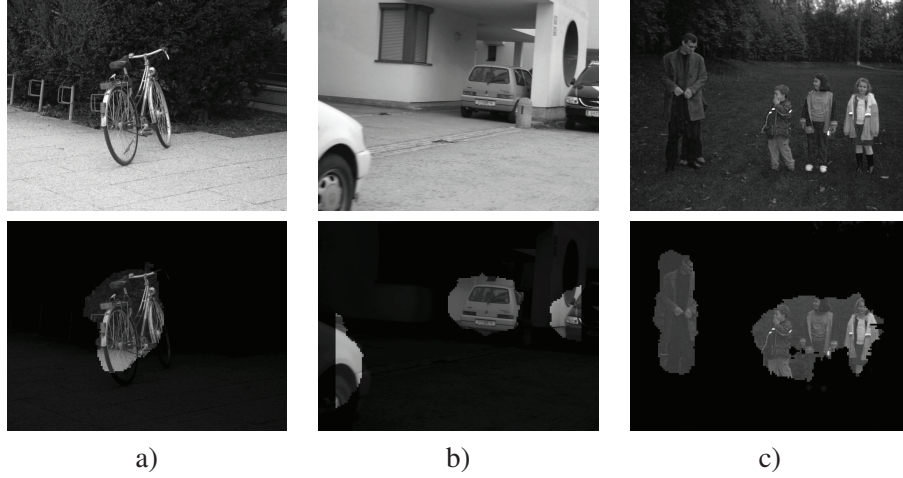


Figure 1: Examples of pixel-level categorization results obtained with our method for a) bikes, b) cars and c) person in the Graz02 database.

the one presented here. However, In contrast to Fulkerson et al. , we propose to bypass the descriptor accumulation step, and make every computed feature vote directly with its classifier score in an integral image to reduce the computational cost of classifying an image sub-window to only 10 additions and 2 multiplications, regardless of its size.

The rest of the paper is organized as follows: In Section 2 the proposed methodology for object classification and localization in images is described. Then, in Section 3, our proposed method is evaluated with the Graz02 dataset [2] and results are presented and discussed. Finally, in Section 4 the contributions and conclusions of this work, as well as future research directions, are summarized.

2 Pixel-level categorization

Our method uses an efficient categorization algorithm to assign a category label to each pixel of an image: First, region descriptors are densely sampled from the image and quantized into visual words using a codebook. Then, a sliding window scheme is used to assign a category label to each pixel of the image. Visual words within a win-

dow are accumulated in a histogram, which is later classified using a linear Support Vector Machine (SVM). In Fig. 1 some pixel-level categorization results obtained using the proposed method are shown. Categorizing all the pixels from an image with this brute-force approach in a reasonable time requires each of the previous steps to be executed in a very efficient way.

2.1 Dense features

As previously mentioned, we use densely sampled image descriptors as input data. Dense sampling has several advantages when compared to keypoint-based approaches, such as extracting more information from the underlying image, and avoiding the time-consuming keypoint detection step. Furthermore, if robust descriptors can be computed in an efficient way, it can even become faster than the keypoint-based alternative despite the larger number of descriptors computed.

With this in mind, we have decided to use the Integral Histograms of Oriented Gradients (IHOG) descriptor [6]. The IHOG is an approximation to the SIFT descriptor [3], which speeds up the descriptor extraction using integral images. Unlike

the SIFT descriptor, the IHOG descriptor is incompatible with the Gaussian mask and the tri-linear interpolation to weight the contribution of the gradient module in the spatial bins of the descriptor used in SIFT. Nevertheless, despite all these simplifications, the performance of the IHOG descriptor is only slightly worse than that of the SIFT descriptor. Moreover, IHOG descriptor is rotation invariant. However, according to Zhang et. al. [4], the use of rotation invariant descriptors has a negative effect in the performance of bag of features approaches.

2.2 Codebook Generation

Once all descriptors have been computed from the image, it is necessary to quantize them into visual words using a codebook. The computational cost of quantizing a D -dimensional descriptor using linear codebook of V visual words is $O(DV)$. From the various alternatives that have been proposed to reduce this computational cost, in this work we have evaluated two: the Hierarchical K-Means (HKM) and the Extremely Randomized Forest (ERF).

The HKM defines a hierarchical quantization of the feature space. Instead of k being the final number of visual words of the codebook, it determines the branch factor (number of children of each node) of a tree. Given a set of training descriptors, an HKM is generated as follows: First, the k -means algorithm is used to split the training data into k groups. Then, this clustering process is recursively applied to the groups from the previous level until a maximum depth is reached. This recursive method creates a vocabulary tree (i.e. codebook) with a reduced computational cost both in the training and descriptor quantization phases. The computational complexity of quantizing a D -dimensional descriptor using a HKM with V visual words is $O(Dk \log_k V)$.

The ERF [1] uses a combination of several random K-D trees in order to quantize the feature space. Given a set of labeled training descriptors

(i.e. descriptors with a category label associated), the K-D trees of the ERF are built recursively in a top-down manner as follows: Every node of the K-D trees splits the training descriptors from the previous level in two disjoint sets with a boolean test in a random descriptor vector position. The boolean test consists in dividing the descriptors in two groups according to a random threshold θ_t applied at descriptor vector dimension D_t , also chosen randomly. For each node, the random boolean test is scored using the Shannon entropy until a minimum value S_{min} is attained or a maximum number of trials T_{max} has been reached. Then, the selected random boolean test is the one that has a highest score. Parameter S_{min} can be used to select the randomness of the obtained K-D trees. For instance $S_{min} = 1$ creates a highly discriminant tree while $S_{min} = 0$ creates a completely random tree. The main advantage of the random K-D tree compared to other quantization methods is its low computational cost. Quantizing a D -dimensional descriptor vector using a random K-D tree with V visual words is $O(\log_2 V)$. Since a random K-D tree usually has less discriminative power than other clustering methods, like the k-means or the HKM, several K-D trees are combined together to obtain a more discriminative codebook. Finally, the resulting histogram of the ERF is created by concatenating the histograms generated by each K-D tree of the forest.

2.3 Integral Linear Classifiers

The “integral image” representation has been first introduced by Viola and Jones to quickly extract Haar-wavelet type features [8]. Since then, integral images have been applied to many different tasks like invariant feature extraction, local region descriptors, to compute histograms over arbitrary rectangular image regions or to compute bag of feature histograms. Inspired by these previous works, we propose the use of an integral image to quickly calculate the output score of the linear classifier which is applied to bag of features his-

tograms.

To categorize a V dimensional histogram of visual words, we use a linear classifier with weight vector \vec{W} and bias b . Then, the output score of the linear classifier is:

$$\frac{1}{\|\vec{X}\|} \sum_{i=0}^V x_i w_i + b > 0 \quad (1)$$

where x_i is the frequency of the i -th visual word of the codebook, $\|\vec{X}\|$ is the norm of histogram \vec{X} and w_i is the i -th component of the linear classifier weight vector \vec{W} . If all components of \vec{W} are positive, then, the sum of the previous equation can be calculated using an integral image. Therefore, we define the classifier weight vector \vec{W} components as:

$$\tilde{w}_i = w_i - W_m \quad (2)$$

where W_m is the w_i component with the lowest value. Then, replacing \vec{W} by $\vec{\tilde{W}}$ in Eq. 1 the output score of the linear classifier is:

$$\frac{1}{\|\vec{X}\|} \sum_{i=0}^V x_i \tilde{w}_i + \frac{W_m}{\|\vec{X}\|} \sum_{i=0}^V x_i + b > 0 \quad (3)$$

We normalize the histogram \vec{X} using L1 norm (i.e. the amount of visual words that casted a vote in the histogram) since it is fast to compute using an integral image. Then, Eq. 3 becomes:

$$\frac{1}{N} \sum_{i=0}^V x_i \tilde{w}_i + W_m + b > 0 \quad (4)$$

where N is the L1 normalization of histogram $\|\vec{X}\|$. Once all $\vec{\tilde{W}}$ components are positive, the integral image can be used to calculate the sum in Eq. 4. For each linear classifier c , let $L_c(x, y)$ be the sum of components \tilde{w}_i^c corresponding to the visual words at pixel (x, y) . Then, each image L_c is transformed into an integral image I_c , so that, the sum of Eq. 4 of a rectangular image region R can be calculated using the integral image I_c :

$$H_R = I_c(x_u, y_u) + I_c(x_b, y_b) - I_c(x_u, y_b) - I_c(x_b, y_u) \quad (5)$$

where (x_u, y_u) and (x_b, y_b) are respectively the upper left and bottom right corner coordinates of region R . Then, the output score of a linear classifier applied to any rectangular image region can be calculated as follows:

$$\frac{1}{N} H_R + W_m + b > 0 \quad (6)$$

Using integral images, the computational complexity of classifying any rectangular image region is reduced to 8 memory access, 10 additions and 2 products, independently of the size of rectangular region.

3 Experiments

We have evaluated the performance of our pixel-level categorization method on the Graz02 database [2]. The Graz02 database is a challenging database consisting on three categories (bikes, cars and people) where objects have an extreme variability in pose, orientation, lighting and different degrees of occlusion. The Graz02 annotation only provides a pixel segmentation mask for each image, so that, it is impossible to know how many object instances are present in the image. In consequence, to evaluate the performance of our pixel-level categorization method we use the pixel-based precision-recall curves as in [9]. We have taken the odd images as train and the even as test as in [9, 10]. However, due to the high variation we observed in the results depending on the train/test sets, we decided to also use random selection to split half of the images for train and half for test. The final result of a test when using the random sampling is the mean of a 1,000-repetitions experiment to ensure statistical invariance of the selected train/test sets.

3.1 Parameter setting

The results were obtained using the same parameters in each experiment. The IHOG descriptors

HKM	117.3ms
ERF 1-tree	95.8ms
ERF 3-trees	83.4ms
ERF 5-trees	72.6ms
ERF 7-trees	62.1ms
ERF 9-trees	56.7ms

Table 1: Mean time spent evaluating an image for the ERF and the HKM.

have been densely sampled each four pixels. Descriptors that have a low gradient magnitude before normalization are discarded as in [10]. Each IHOG descriptor is extracted from a 40×40 pixels patch and it has 8 orientation bins and 4 positional bins (i.e. a 32 dimensional descriptor). Therefore, as Graz02 images have a regular size of 640×480 , a maximum of 16,500 descriptors are extracted per image. Then, bag of features histograms are computed accumulating the visual words that are inside a region of 80×80 pixels. Later, those histograms are categorized using a SVM. The SVM has been trained using logistic regression (LR-SVM) with the LIBLINEAR software package [5]. Finally, the shown times results have been obtained using laptop with an Intel T7700 Core Duo CPU and 2Gb of RAM.

3.2 Parameters of the ERF

The performance of the ERF depends on the K-D tree randomness parameter and the amount of trees in the forest. Therefore, we wanted to evaluate which combination of those parameters gives better results for our categorization method. Experimental results shows that the performance for the ERF largely depends on the amount of trees, while the randomness factor has little, if any, effect in the performance. For the remaining experiments, we have selected a randomness factor of 0.05 (i.e. a completely random forest) and 5 trees, which are a good compromise between performance and computational cost.

3.3 Comparison between HKM and ERF

To compare the performance of the HKM and the ERF, dense features have been computed for all the 450 training images, resulting in about 6,000,000 training features. For the HKM we have selected a branch factor of 10 as in [10] to generate a codebook with 200,000 visual words in average. For the ERF, using the parameters selected in the previous section, we have obtained a codebook of 150,000 visual words in average. We have done two different tests: the “Single” test only uses the images containing objects from the tested category, and the “Multi” test uses all test images. As can be seen in Table 2 the precision-recall values obtained at EER show that the ERF performs slightly better than the HKM, both in the “Single” and the “Multi” tests.

3.4 Time cost evaluation

Finally, regarding the computational cost of the categorization approach, the average time needed to construct the HKM is of 5 minutes, while that of the ERF depends on the randomness factor and the number of trees used, ranging from 100 milliseconds for a completely random ERF with a single tree, to 12 minutes for a highly discriminative ERF with 9 trees. The cost of training a linear classifier using the LR-SVM is of about 2 minutes for the histograms generated with HKM codebook, and from 2 to 5 minutes for those generated with the ERF. In table 1 we can see the average time needed to categorize all image pixels using a HKM and ERF codebooks. Although being a bit slower in the training phase, the ERF is faster than the HKM in the categorization phase, where speed is truly essential. Using a ERF with 5 K-D trees we can process about 13 images per second.

4 Conclusions

In this paper we have presented an efficient method to assign a category label to each pixel of an image. Our main contribution is the introduction

Sampling	Test	Method	Bikes	Cars	Persons
Even/Pair	Single	HKM	73.77% \pm 0.20%	63.90% \pm 0.17%	61.80% \pm 0.40%
		ERF	73.63% \pm 0.32%	65.68% \pm 0.60%	62.59% \pm 0.30%
	Multi	HKM	63.42% \pm 0.24%	47.09% \pm 0.45%	44.36% \pm 0.31%
		ERF	63.27% \pm 0.32%	47.62% \pm 1.45%	46.34% \pm 0.43%
Random	Single	HKM	74.33% \pm 1.21%	65.70% \pm 1.45%	62.13% \pm 1.30%
		ERF	74.17% \pm 1.22%	68.39% \pm 1.44%	63.27% \pm 1.38%
	Multi	HKM	64.55% \pm 1.26%	49.60% \pm 1.61%	42.78% \pm 1.86%
		ERF	63.98% \pm 1.28%	51.30% \pm 2.03%	44.30% \pm 2.12%

Table 2: Comparison of the precision-recall values obtained at equal error rate on the Graz02 database both using odd/even and random train/test sampling.

of integral linear classifier, which is used to bypass the accumulation step and directly obtain the classification score for an arbitrary sub-window of the image. Besides, we have compared the performance of the Hierarchical K-Means (HKM) and Extremely Randomized Forest (ERF). The obtained results show that the ERF and HKM performs similarly and that the proposed method using the ERF is suitable for real-time applications.

References

- [1] F. Moosmann, E. Nowak, F. Jurie, "Randomized clustering forests for image classification", *IEEE Trans. on Pat. Anal. and Machine Intel.* 30(9):1632-1646, 2008.
- [2] A. Opelt, et. al., "Generic object recognition with boosting", *IEEE Trans. on Pat. Anal. and Machine Intel.* 28(3):416-431, 2006.
- [3] D. Lowe, "Distinctive image features from scale-invariant keypoints", *Int. Journal of Computer Vision* 60(2):91-110, 2004.
- [4] J. Zhang, et. al., "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study", *Int. Journal of Computer Vision* 73(2):213-238, 2007.
- [5] R.E Fan, et. al., "LIBLINEAR: A Library for Large Linear Classification", *J. Mach. Learn. Res.* 9:1871-1874, 2008.
- [6] Q. Zhu, et. al., "Fast Human Detection Using a Cascade of Histograms of Oriented Gradients", *Proc. of Comp. Vision and Pat. Recog.* 1491-1498, 2006.
- [7] D. Nister, H. Stewenius, "Scalable recognition with a vocabulary tree", *Proc. of Comp. Vision and Pat. Recog.* 2161-2168, 2006.
- [8] P. Viola, M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", *Proc. of Comp. Vision and Pat. Recog.* 511-518, 2001.
- [9] M. Marszalek, C. Schmid, "Accurate Object Localization with Shape Masks", *Proc. of Comp. Vision and Pat. Recog.* 1-8, 2007.
- [10] B. Fulkerson, A. Vedaldi, S. Soatto, "Localizing Objects with Smart Dictionaries", *Proc. of Euro. Conf. on Comp. Vision* 179-192, 2008.

Perceptual Feature Detection

Naila Murray, Xavier Otazu and Maria Vanrell

Colour in Context Group

Computer Vision Centre, Edifici O - Campus UAB, 08193 Bellaterra, Spain

naila, xotazu, maria@cvc.uab.cat

Abstract

Currently there exists no application-independent or general theory of feature detection. In this work, a brightness induction wavelet model (BIWaM) is extended with the long-term aim of developing a principled model for generic local feature detection. This detector, the Feature Induction Wavelet Model (FIWaM), uses the same “featureness” measure for a range of local features such as blobs, bars and corners. FIWaM is a wavelet-based computational model that attempts to use the perceptual processes involved in visual brightness induction to enhance and detect these features. The model uses two center-surround mechanisms in sequence to detect features - a Gabor-like mother wavelet followed by an explicitly-defined center-surround region mechanism. These center-surround regions are feature-specific and introduce the only variation in the detection schema between features. Preliminary results have shown that this mechanism is effective in detecting features and achieves a repeatability performance in line with current state-of-the-art detection methods.

Keywords: Brightness Induction, Feature Detection, Wavelet Transform.

1 Introduction

Feature detection has an essential role in many important computer vision tasks, including image

matching and registration, object recognition and tracking and scene classification. Consequently there has been a plethora of research devoted to developing efficient and effective feature detection techniques. Currently, state-of-the-art detectors use very different methods and, as a result, their performance differs widely depending on the data sets they are used to analyse. To date, there exists no application-independent or general theory of feature detection. Therefore, determining which feature detector to use on any specific application tends to require *a priori* information about the data set, and a subjective judgement on the most suitable method for feature detection.

This paper extends the perceptual processes present in a low-level human visual system (HVS) model of brightness induction, (BIWaM) [7], with the long-term aim of developing a principled model for generic local feature detection. Several successful detectors [4, 6] have been modelled using “biologically plausible architecture” [3] related to the HVS with much success. The motivation for using the BIWaM is to incorporate many relevant attributes of the HVS with the aim of combining the advantages of detectors which use these attributes separately.

2 Related Research

As mentioned in the previous section, biologically-inspired frameworks have been employed successfully in local feature detection. Lowe’s SIFT algo-

rithm [4] and Mikolajczyk & Schmid's Hessian-Affine and Harris-Affine algorithms [6] are all based on multi-scale image decomposition. In addition, the Difference of Gaussian (DoG) kernel used in the SIFT scale decomposition has a center-surround profile and thus can be thought of as a centre-surround response mechanism.

Collins & Ge [2] employ the multi-scale concept, as well as an explicit centre-surround mechanism, for feature extraction. Centre and surround regions were defined using a Laplacian of Gaussian kernel. The positive (circular) region of the kernel corresponds to the central local region while the negative (annular) region of the kernel corresponds to the surround local region. For the center, the kernel is used to weight values around a location in a Gaussian fashion. A distance measure is calculated for the centre and surround regions and compared to that of neighbouring pixels. As with SIFT, local extrema are selected as candidate features.

Agrawal et al. [1] adopt a similar approach to that of Collins & Ge [2], but the DoG kernel is simplified to a Difference of Boxes (DoB) kernel. Also, bi-level center and surround boxes are used, i.e. with values of 1 or -1, in order to enable an extremely simple and fast computation. Finally, extrema are extracted at different scales not by blurring and down-sampling the image but by changing the scale of the kernels.

3 The BIWaM Model

The BIWaM [7] modifies a visual stimulus in order to reproduce the brightness induction performed by the HVS. Brightness induction refers collectively to

- brightness assimilation, where the brightness of a visual target (considered the center region) becomes more similar to that of the surrounding region, and
- brightness contrast, where the brightness of a

visual target becomes less similar to that of the surrounding region.

The model is based on three main assumptions, derived from known psychophysical phenomena:

1. Induction is higher between features of similar **spatial scale**. Because image features are isolated by spatial scale in wavelet planes, this is achieved using the image decomposition. As Figure 1(a) illustrates, induction is strongest between features within one octave of each other in scale space.
2. Induction is higher between features of similar **spatial orientation**. Inhibition is strongest when orientations are identical, while facilitation is strongest when orientations are orthogonal (see Figure 1(b)). This is also inherent in the wavelet decomposition.
3. Induction is modulated by the **stimulus-surround relative contrast**. For increasing surround contrast there is increasing inhibition and vice-versa, as can be seen in Figure 1(c).

3.1 The Wavelet Decomposition

The wavelet decomposition of the image is a key point of the model. Images are decomposed into a series of new images (wavelet planes) with respect to spatial scale s and orientation o (vertical, horizontal and diagonal), which is inspired by parvocellular spatial frequency channels and cortical orientation-selective receptive fields in the HVS. The wavelet planes, w_s^h , w_s^v and w_s^d , contain the response of the image intensities at that orientation to the wavelet kernel corresponding to the scale, s .

The image, I , is reconstructed as:

$$I = U_{s=1} \quad (1)$$

where U_s is the s -th element of a recursive series of images

$$U_s = (U_{s+1} \uparrow 2) + d_{s+1} \quad (2)$$

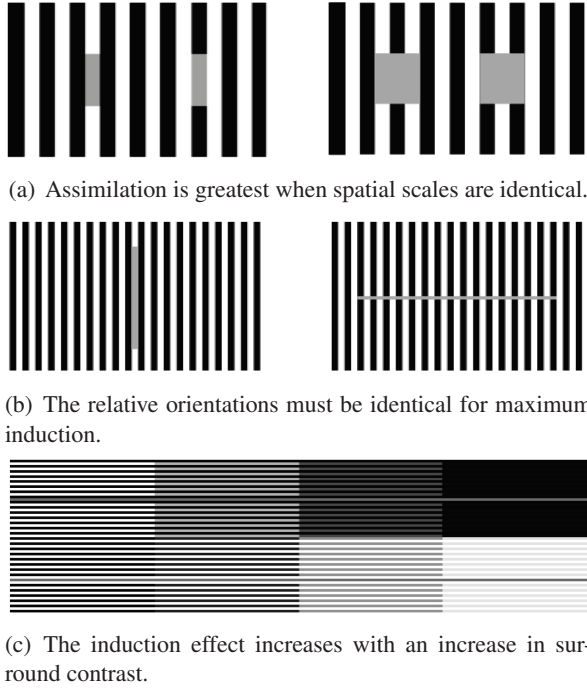


Figure 1: The assumptions of the model.

and d is the sum of the oriented wavelet planes:

$$d_s = \sum_{o=v,h,d} w_s^o, \quad (3)$$

where $\uparrow 2$ denotes up-sampling by a factor of 2.

3.2 Construction of the Perceived Image

Equation 2 describes the reconstruction of the original image from the wavelet decomposition. The perceived image is obtained from this reconstruction with the simple introduction of a weighting function α , which is designed according to the assumptions of the induction model. The modified image recovery defined by Equation 4:

$$I' = U_1^\alpha = (U_{s+1}^\alpha \uparrow 2) + \alpha \cdot d_{s+1} \quad (4)$$

introduces α , thereby generating the perceived image, I' .

3.3 The α Weighting Function

The weighting function α can be seen as a generalisation of the psychophysically-determined Contrast Sensitivity Function, (CSF), C_d . It has been shown that the HVS is very sensitive to mid-range frequencies, and to a lesser extent to low frequencies. It is important to note that frequencies are relative to the distance, d , between the viewer and the visual stimulus. This concept of viewer distance was incorporated into the definition of C_d (see Appendix A of [7]). The weighting function is based on this CSF but has been modified to introduce the effect of surround contrast and is defined as

$$\alpha(s, z_{ctr}) = z_{ctr} \cdot C_d(s) + C_{min} \quad (5)$$

The z_{ctr} term defined by

$$z_{ctr} = \frac{r^2}{1 + r^2} \quad (6)$$

where $r = \frac{\sigma_{cen}}{\sigma_{sur}}$, introduces relative contrast energy implicitly. The standard deviation, σ , of a region is used as a measure of its self contrast. Therefore the ratio r is the relative contrast energy of the center and surround regions. The r term is dependent on orientation o . To avoid null α values, the C_{min} term was introduced.

4 Perceptual Feature Detection

The brightness induction framework of the BI-WaM can be modified to accomplish feature induction by substituting α for a suitably designed weighting function. As such we present what we term the Feature Induction Wavelet Model (FIWaM), with a new weighting function β , that modifies Equation 4 as follows:

$$I' = U_1^\beta = (U_{s+1}^\beta \uparrow 2) + \beta \cdot d_{s+1} \quad (7)$$

It is apparent from Equation 7 that the modified image recovery is identical to that of the BIWaM except for the new weighting function, β .

However, to detect features using the FIWaM, β is used as a feature detection measure, rather than a weighting function and is defined using two hypotheses:

1. Features are present within a bounded range of scale space.
2. If a stimulus region's response to a feature's characteristic shape is appropriately large, the stimulus region contains that feature.

With these hypotheses in mind, we define β as:

$$\beta(s, z_{ctr}) = \gamma \cdot z_{ctr} \cdot C_{det}(s). \quad (8)$$

The new CSF, $C_{det}(s)$, is an ideal band-pass filter that bounds the range of scale space in which features are detected.

The z_{ctr} term is defined as previously. However, the center and surround regions are now defined differently for each feature, in order to reflect the feature's characteristic shape. Therefore, z_{ctr} measures the stimulus's response to a specific feature's shape. The median contrast energy term, γ :

$$\gamma = |median_{cen} - median_{sur}| \quad (9)$$

is the difference between the median intensity values of the stimulus and stimulus-surround regions and quantifies the strength of the wavelet response of the stimulus. Together, z_{ctr} and γ measure the type and the strength of a detected feature, respectively. Therefore, β constitutes a "featureness" measure, that is, the degree to which a stimulus corresponds to a feature.

4.1 Characterisation of Feature Shapes

We have investigated detection with respect to four features - blobs, bars, corners and terminators. These regions have a size that corresponds to the minimum size of interest of the feature. They also reflect the appearance of the feature decompositions in the wavelet plane, as shown in Table 1.

Feature	Feature representation	Wavelet plane representation	Center	Surround
Blob				
Bar				
Corner				
Terminator				

Table 1: Wavelet decompositions of features along with their center and surround region definitions.

4.2 Feature Selection

To select a stimulus region as a feature, the feature must have a β value that is a local extremum in (x, y, σ) -space, where σ signifies scale-space. In addition, β must be over a certain threshold to ensure the feature is strong, that is, more repeatable.

For each image, detection is performed separately for the 4 types of features described, in all possible orientations. The aggregation of the features from these separate detections comprises the final feature set for an image.

One sometimes finds that different features, for example both a blob and a bar, are detected at the same spatial location. In such cases, the different featureness responses are compared and the feature with the highest featureness response is selected.

5 Experiments

To assess the detector's performance, the repeatability of the detected features was tested using the experimental framework constructed by Mikolajczyk et al. [5]. In this experiment, feature detection is performed on a sequence of images of the same scene. One of these images is considered to be the reference image and there exists a known homographic relationship between the reference image and the other images in the sequence. Therefore, for an image in the sequence, the regions of the

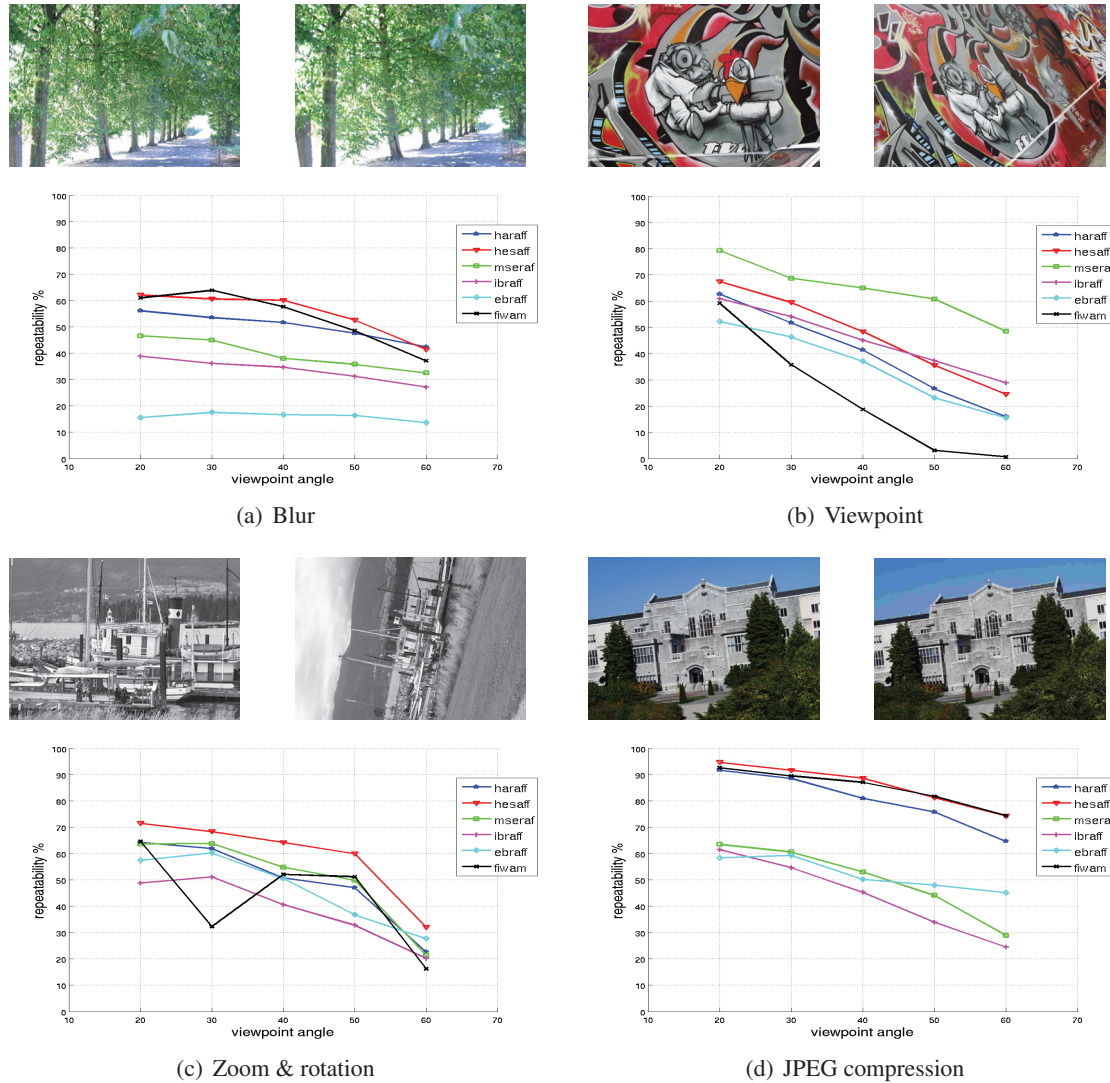


Figure 2: Repeatability for different transformations: The FIWam plot is shown in black. Examples of data set images are shown above the repeatability plots.

scene contained in both the image and the reference image are known. A feature is considered to be repeated if it is found in a region common to both images and it has been detected in both images. Additionally, the spatial overlap of the regions defined by the features must be greater than a user-defined threshold. An overlap threshold of 40% is typical and was used here. The experiment was conducted on 4 sequences of six images with homographic variances with respect to

Gaussian blurring, viewpoint, zoom & rotation and JPEG compression. The transformation increases in severity along the sequence of images.

The repeatability results are shown in Figure 2. For comparison, data for five state of the art feature detectors have been included: Harris-Affine (HARRAFF), Hessian-Affine (HESAFF), Maximally Stable External Region (MSERAF), Intensity Extrema-Based Region (IBRAFF) and Edge-Based Region (EBRAFF) [5].

One should note that repeatability is not the only criteria available for evaluating a detector's performance. However, repeatability has shown itself to be a good general indicator of performance, and so it was used here.

5.1 Discussion

It is evident that the FIWaM detector performs comparably with respect to state-of-the-art detectors, except in cases of severe affine transformation, such as in the graffiti sequence (Figure 2(b)). This is unsurprising given that the FIWaM detector has coarse affine estimation. Firstly, the wavelet's scale decomposition may be too coarse for accurate scale localisation of features. For a typical image decomposition of 7 octaves, 5 octaves are analysed due to the nature of the CSF. This means that features have only 5 possible scales. Secondly, the elliptical estimation is derived from the shape of the center regions, not the shape of the stimulus itself, resulting in an imprecise estimation.

6 Conclusions

In this paper, a brightness induction wavelet model (BIWaM) was extended with the long-term aim of developing a principled model for generic local feature detection. It has been shown that the brightness induction model can be modified successfully to create an effective local feature detector.

However, there are many avenues for further exploration. Most promising is improving the center and surround regions for several features so that they more closely resemble the appearance of these features in the wavelet decomposition. This would improve the problem of affine variance, as would using a more refined scale-space decomposition.

In addition, in this work, there has been no discussion on incorporating colour information. However, there exists a straight-forward extension of the BIWaM to colour, namely the Colour Induction Wavelet Model (CIWaM). Incorporating

colour information may allow the detection of features in channels other than intensity, such as in the opponent colour space.

References

- [1] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas. Censure: Center surround extremas for realtime feature detection and matching. In David A. Forsyth, Philip H. S. Torr, and Andrew Zisserman, editors, *ECCV (4)*, volume 5305 of *Lecture Notes in Computer Science*, pages 102–115. Springer, 2008.
- [2] Robert T. Collins and Weina Ge. Csdd features: Center-surround distribution distance for feature extraction and matching. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, pages 140–153, Berlin, Heidelberg, 2008. Springer-Verlag.
- [3] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.
- [4] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.
- [5] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.
- [6] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.
- [7] X Otazu, M Vanrell, and C.A Párraga. Multiresolution wavelet framework models brightness induction effects. *Vision Research*, 48:733–751, February 2008.

Coloring Laplacian-of-Gaussian Detector for Image Matching

David A. Rojas Vigo and Joost van de Weijer

Computer Science Department, Computer Vision Center, Edifici O, Campus UAB, Barcelona, Spain

E-mail: david.rojas@uab.es

Abstract

This work introduces a novel color-based local feature detector. Our algorithm is an extension of the Laplacian-of-Gaussian detector to color. The extension to color is done by using the color saliency boosting algorithm. The performance is evaluated through matching score tests using a matching-image benchmark and show that outperform the original detector.

Keywords: Laplacian-of-Gaussian detector, Color saliency boosting, image matching.

1 Introduction

Vision is one of the most important sensory modalities for intelligent living organisms as well as machines, and is traditionally regarded as processing primarily *achromatic* information. Both, observations that information on shape at dominant points having high curvature and perceptual experiments confirm the importance of local features in object recognition.

However, *color* perception is a central component primate vision. Experimental evidence has been shown that objects in colored scenes are more easily detected, more easily identified, more easily grouped, and more easily remembered than objects in black-and-white scenes.

Studies on early visual processing suggest that

color is processed not in isolation, but together with information about luminance and visual form, by the same neural circuits, to achieve a unitary and robust representation of the visual world. Color and local features have an essential role in computer vision as well. Local features have been used very successfully for object representation due to their robustness with respect to occlusion and geometrical transformations.

Although, it is well recognized that while color is a useful cue, its reliable use as a feature is hampered by various practical difficulties. Reverting to gray-value and postponing the use of color to when computing the descriptor is an attractive alternative. Though, the conversion to gray-value has a number of side-effects that are particularly undesirable for reliable local feature detection. It is well recognized in neurobiology and computer graphics that gray-value versions of color images do not preserve chromatic saliency. Regions that exhibit chromatic variation often lose their distinctiveness when naively mapped to scalars based on isoluminance.

In this work, we investigate the impact of color saliency boosting algorithm in local feature detection. We propose an extension of the Laplacian-of-Gaussian to color. We compare our detector to the original using an experimental framework and show through matching score tests that our detector perform better. Our experiments show that our approach provides more discriminative regions in comparison with the original detector.

2 Local Feature Detection

This section provides a brief overview of some properties of local features, detectors and descriptors relevant to our work. According to their invariance model, local features can be classified as: multi-scale, scale-invariant and affine-invariant features. Our aim is not to cover the full theory, but provide sufficient information for the technique that are used in next sections.

2.1 Multi-scale features

A multi-scale representation consists of a stack of images at different discrete levels of scale, it is crucial for many applications, and especially for extracting local features. These features exist only in a limited range of scales between the *inner* and *outer* scale, that is the smallest and the largest scale of interest. Koenderink [1] showed that space-scale satisfies the diffusion equation for which the solution is a convolution with a unique Gaussian kernel, which has also been confirmed in other studies. Images on coarse scales are obtained by smoothing images on finer scales with a circularly symmetric kernel and parameterized by one scale factor σ .

The semi-group property simplifies the scale-space representation in several ways. In order to accelerate the operation one can sample the coarser scale image by the corresponding scale factor after every smoothing operation. It is important to be careful here choosing the scale and the sampling factor as it may lead to aliasing problems. Moreover, additional relations have to be introduced in order to find the corresponding point locations at different scale levels. This makes any theoretical analysis more complicated, but computationally efficient. This representation is often referred as the scale-space image pyramid. When an interest operator is applied on multiple scales we call the detections multi-scale interest points or regions.

2.2 Scale-invariant features

Instead of extracting interest point or regions for every scale level, automatic scale selection technique determine one of a few characteristic scales at each location. This, scale-invariant features are obtained by performing automatic spatial and scale selection. The Laplacian detector extracts image regions whose locations and characteristic scales are given by scale-space maximum of the Laplace operator. A desirable property for a scale-space differential operator is that it should always produce the same response to an idealized scale-invariant structure. But, we cannot just take a blurred derivative because we will obtain weaker responses at larger scales. This motivates the definition of scale-normalized differential operators, whose output remains constant if the image is scaled or resized by an arbitrary factor. One particular useful normalized differential quantity is the scale-normalized Laplacian, which is one of the simplest scale selection operators.

$$\sigma^2 \left| \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right| \quad (1)$$

One local feature is scale-invariant whether it simultaneously achieves a local maximum with respect to scale parameter and the spatial variables.

3 Color and Local Feature Detection

There are many local feature detectors available in the literature and a wide variety of criteria for classification. Traditionally, due to the nature of the extracted local features, these have been classified as: corner, blob and region detectors. One of the most representative detectors known is the Laplacian-of-Gaussian (a blob detector).

3.0.1 Laplacian-of-Gaussian Detector

One of the first scale-invariant detectors is the Laplacian-of-Gaussian developed by Lindeberg

[2]. Detections are obtained on blob-like image structures. In [3], the author evaluated different scale selection criteria for scale-invariant image matching environments. Apart from the Laplacian they study the squared image gradients, the Difference-of-Gaussians and the Harris function. Their evaluation shows that the Laplacian function selects the highest percentage of correct characteristics scales.

3.0.2 Color-based Detectors

Color provides additional information which can be used in the process of feature detection. Salient point detection based on color distinctiveness has been proposed by [4]. Salient points are the maxima of the saliency map, which represents distinctiveness of color derivatives in a point neighborhood. Color ratios between neighboring pixels are used to obtain derivatives independent of illumination, which results in color interest points that are more robust to illumination changes.

A simple extension from luminance to color for local feature detection was introduced by [5] and known as Color Harris. Also, Color boosted Color-Opponent Harris Laplace detector was described by [6]. Corso and Hager [7] find extrema in Difference-of-Gaussians responses defined from three linear projections of the RGB space. The detected regions are represented as image-axis-aligned ellipses, and they are thus only scale and translation invariant. This detector gave a lower repeatability than the gray-value Difference-of-Gaussians pyramid detector.

In [8] the authors tested two illuminant invariant scalar functions, one invariant to independent scaling of the RGB channels, and one invariant to a full 3x3 perturbation of the RGB space. They proceed by detecting scale and rotation invariant Laplacian-of-Gaussian points. They found that the version only invariant to independent scaling of the RGB channels performs best under rotation and scale changes, and more importantly it performs better than Laplacian-of-Gaussian on a gray-value

image. They did not evaluate this detector under view changes. In [9] the author introduced a color-based affine invariant region detector, which is an extension of the Maximally Stable Extremal Region to color, this detector outperforms the original. In [10] the authors proposed and tested a new detector based on histograms, which includes color information.

4 Color saliency boosting

This section presents a mathematical analysis of the first order derivatives in an input color image based on the proposed algorithm. Let \mathbf{f} be a color image and $\mathbf{f}_x = (R_x \ x \ B_x)^T$ $\mathbf{f}_y = (R_y \ y \ B_y)^T$ its corresponding spatial derivatives. Considering those derivatives as a random vector with a multivariate normal distribution, is possible to characterize this anisotropic distribution using the sample covariance matrix, which is symmetric, positive and semidefinite. From the geometric properties of the color derivatives distribution we know that is centered at the origin of coordinates, i.e. mathematically $[\mathbf{f}_x] = 0$. Thus, this 3x3 matrix is defined by

$$\Sigma_x = [\mathbf{f}_x \mathbf{f}_x^T] \quad (2)$$

From this matrix we can estimate the derivative energy as

$$\xi(\mathbf{f}_x) = t \ ce(\Sigma_x) \quad (3)$$

Applying the PCA, we can determine the principal axes of the distribution and their corresponding squared relative half-lengths λ_1 λ_2 and λ_3 by eigendecomposition

$$\Sigma_x = U \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} V^T \quad (4)$$

where U is a 3x3 matrix whose k th column is the k th eigenvector of Σ_x , and $U = V$ because Σ_x is symmetric. These eigenvectors form a orthonormal basis aligned with the major axis of the isosaliency ellipsoid. In order to convert this anisotropic

pattern into an isotropic one (whitening, or also called sphering), we apply a linear transformation based on the inverse square root matrix.

$$\Lambda \mathbf{f}_x = \Sigma_x^{-\frac{1}{2}} \mathbf{f}_x \quad (5)$$

Therefore the desired energy-normalized *color boosting function* can be obtained by

$$g(\mathbf{f}_x) = \tau \Lambda \mathbf{f}_x \quad (6)$$

Replacing (5) in (3) and (4),

$$g(\mathbf{f}_x) = \tau U \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{\lambda_3}} \end{pmatrix} V^T \mathbf{f}_x \quad (7)$$

where τ is the rate of derivative energy and is defined as

$$\tau = \sqrt{\frac{\xi(\mathbf{f}_x)}{\xi(\Lambda \mathbf{f}_x)}} \quad (8)$$

This ensures the energy conservation, i.e. $\xi(\mathbf{f}_x) = \xi(g(\mathbf{f}_x))$ and the idempotent property $g(g(\mathbf{f}_x)) = g(\mathbf{f}_x)$. However, in real-world application we need a distinctiveness and signal-to-noise trade-off. For this purpose the parameter is proposed, which allows for choosing between best signal-to-noise characteristics $\alpha = 0$, and best information content, $\alpha = 1$:

$$g^\alpha(\mathbf{f}_x) = \tau \Lambda^\alpha h(\mathbf{f}_x) \quad (9)$$

For $\alpha = 0$ this is equal to color gradient-based salient point detection. Similar equations hold for (\mathbf{f}_y) . The theory can be straightforwardly be extended to higher image structures. The impact of this parameter will be studied in our performance evaluation experiments.

The Laplacian-of-Gaussian detector will be extended to multiple channels combining responses of individual channels in the color image based

on a simple generalization of the scale normalized Laplacian operator.

$$\sigma^2 \|\tau \Lambda^\alpha (\mathbf{L}_{xx} + \mathbf{L}_{yy})\| \quad (10)$$

where $\mathbf{L}_{xx} = (\mathbf{R}_{xx} \mathbf{G}_{xx} \mathbf{B}_{xx})^T$ and $\|\cdot\|$ is the vector norm. This simple extension leads a scale-space representation which includes the contributions of luminance and chromatic components in a scalar-valued representation.

5 Experimental Setup

We adopt the evaluation framework of [3]. They aimed at evaluating the different stages of an object recognition framework, by decomposing the benchmark in the separate evaluation of interest point detection and descriptors. They evaluate the discriminative power and invariance over various imaging conditions. Discriminative power for any detector and descriptor combinations is evaluated over: illumination intensity, viewpoint changes, blurring and JPEG compression. This software compares detected features in two views of a scene that differ by a known view change and checks whether corresponding features are extracted in both views.

This framework measures the repeatability and accuracy of the detectors. In order to obtain quantitative results use the ground truth information, which was provided by mapping the regions detected on the images in a set to a reference image using homographies. The basic measure of accuracy is the relative amount of overlap between the detected region in the reference image and the region detected in the other image, projected onto the reference image using the homography relating the images. This gives a good indication of the chance that the region can be matched correctly.

5.1 Repeatability Measure

Two regions are deemed to correspond if the *overlap error*, defined as the error in the image area

covered by regions is sufficiently small. These areas are computed numerically. The repeatability rate for a given pair of images is computed as the ratio between the number of region-to-region correspondences and the smaller of the number of regions in the pair of the images.

5.2 Matching score

This measure is computed as before between a reference image and the other images in a dataset. The matching score is computed in two steps:

1. A region match is deemed correct if the overlap error is minimal and less than 40%. This provides the ground truth for correct matches. Only a single match is allowed for each region.
2. The matching score is computed as the ratio between the number of correct matches and the smaller number of detected regions in the pair of images. A match is the nearest neighbour in the descriptor space. The descriptors are compared with Euclidean distance.

This test gives an idea on the distinctiveness of features. The results are rather indicative than quantitative. If the matching results do not follow those of the repeatability test for a particular feature type that means that the distinctiveness of these features differs from the distinctiveness of other detectors.

6 Experimental Results

Figures from 1 to 4 show the performance evaluation based on the matching score and the number of correct matches. The results show that boosting obtains an improvement for four sequences. For the graffiti sequence after an initial improvement the results drop for increased boosting.

The most significant changes are obtained in the ubc sequence, where the images have variable

amount of JPEG compression. The dramatic drop in performance is caused by the fact that color is significantly more compressed than the luminance signal. Applying boosting to these images amplifies the JPEG artifacts. From this it can be concluded that boosting should not be applied on significantly compressed images. Apart from this sequence boosting improves the matching score in four of six remaining sequences, and only slightly deteriorates for the graffiti sequence. The optimal amount of boosting varies for each sequence. This indicates that more research is needed to automatically select optimal settings.

7 Conclusions

In this work we have extended the Laplacian-of-Gaussian to the color domain based on color saliency boosting theory. Experiments on matching applications show that color might outperform Laplacian-of-Gaussian detector according to color information content in images. However, color boosting was found to be very sensitive (and hence unusable) to JPEG compression. For variations in scale, blur and lightening color boosting improves the results.

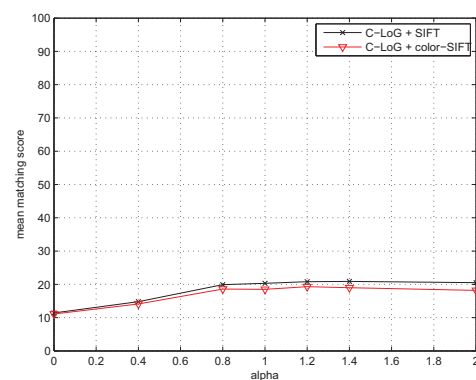


Figure 1: Evaluation of color impact on performance (bark sequence).

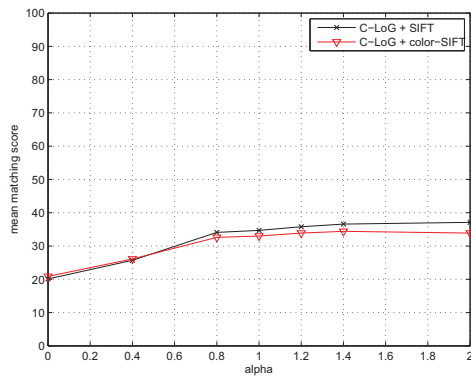


Figure 2: Evaluation of color impact on performance (bikes sequence).

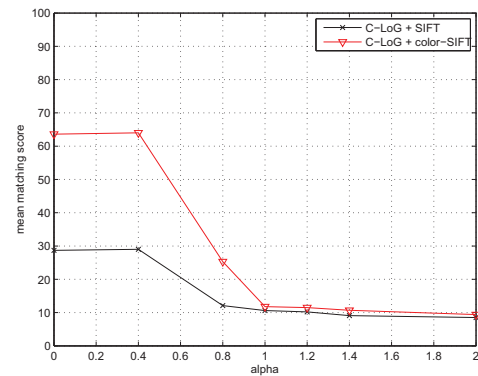


Figure 4: Evaluation of color impact on performance (ubc sequence).

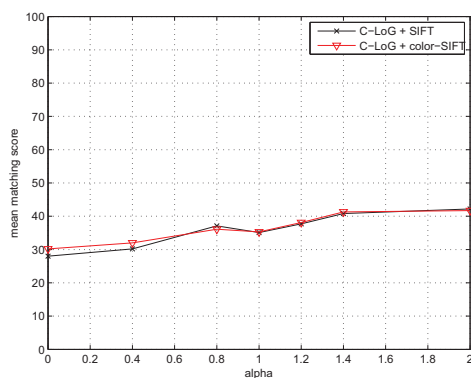


Figure 3: Evaluation of color impact on performance (leuven sequence).

References

- [1] J. Koenderink, "The structure of images", *Biological Cybernetics*, 50(5):363-370, 1984.
- [2] T. Lindeberg, "Feature detection with automatic scale selection", *International Journal of Computer Vision*, 30:79-116, 1998.
- [3] K. Mikolajczyk, "Detection of local features invariant to affine transformations", *PhD thesis, Institute National Polytechnique de Grenoble*, 2002.
- [4] J. van de Weijer, "Boosting color saliency in image feature detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):150-156, 2006.
- [5] P. Montesinos, et al., "Differential invariants for color images", *Fourteenth International Conference on Pattern Recognition*, 1998.
- [6] N. Sebe, et al., "Evaluation of intensity and color corner detectors for affine invariant salient regions", *Conference on Computer Vision and Pattern Recognition Workshop*, 2006.
- [7] J. Corso, G. Hager, "Coherent regions for concise and stable image description", *Conference on Computer Vision and Pattern Recognition Workshop*, 184-190, 2005.
- [8] U. Ranjith, H. Martial, "Extracting scale and illuminant invariant regions through color", *17th British Machine Vision Conference*, 2006.
- [9] P. Forseen, "Maximally stable colour regions for recognition and matching", *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.
- [10] W. Lee, H. Chen, "Histogram-based interest point detectors", *Conference on Computer Vision and Pattern Recognition*, 2009.

On Experimental Evaluation of Descriptors for Facial Feature Point Detection

Mario A. Rojas Q., David Masip* and Jordi Vitria⁺

*Computer Vision Center, Edifi O Bellaterra, 08193, Barcelona, Spain
E-mail: mrojas; davidm; jordi.vitria@cvc.uab.es*

** Universitat Oberta de Catalunya, Barcelona, 08018, Spain
E-mail: dmasipr@uoc.edu*

*⁺ Department de Matematica Aplicada i Analisi, Universitat de Barcelona, 08007, Barcelona, Spain
E-mail: jordi.vitria@ub.edu*

Abstract

The classification of facial expressions from natural images can be a valuable information cue for posterior processes in the human-computer interaction field. In this context, one of the initial steps in most of the facial expression analysis is the location of specific facial salient points, which are used for subsequent classification. This problem has been tackled using different approaches, being the use of machine learning techniques for multi-class classification one of the most used. It is customary for a training set to be constructed and a set of features extracted from manually annotated points to train the classifier. Nevertheless, the choice of the most adequate feature extraction technique is not a straightforward task. In this paper we present firstly an experimental evaluation of different feature descriptors for facial salient point location (Gabor filtering, SIFT, Haar-based features and NDA descriptors), in a GentleBoost classification scheme; secondly an eye detector was implemented following a cascaded framework. The study has been performed on a publicly available face database (Cohn-Kanade), and the results show

an average accuracy higher than 91%.

Keywords: Facial Expression, Boosting, Gabor filtering, Haar filters, NDA.

1 Introduction

Automatic face classification techniques have been explored in security related applications, and also in the design of human-computer interaction systems ([1, 4, 6]). Usually, these processes share the detection of specific points of interest that can be used in a subsequent classification layer. Although the detection of salient points has been studied in the object recognition literature [8, 10, 12], generic algorithms based on Laplacian-of-Gaussian pyramids [11], Difference-of-Gaussians [13], or Harris detectors [15] usually extract extra points that are not useful for face classification. In addition, facial salient points located on non salient regions of the image are usually not detected by standard methodologies.

Most of the recent works on facial expression recognition neglect this preprocessing step

[1, 2], and algorithms are normally experimentally validated using manually annotated images [16]. Nevertheless, real-time applications dealing with Human-Computer Interaction require a robust automatic estimation of the facial salient points. As pointed out in [18], most of these detection techniques do not locate all the interesting facial salient points, or the error interval in the location is exceedingly large to be applied in real applications (about 30% of the inter-ocular distance).

To solve these drawbacks, Vukadinovic and Pantic proposed in [18] to learn each salient point as a category in a multi-class classification framework. The authors used a set of Gabor filters and a GentleBoost classifier in an heuristic approach to achieve an average recognition rate of 93% of 20 facial points. On the other hand, recent works in computer vision show that other feature extraction methods for salient points description obtain promising results on visual data, algorithms such as SIFT [13], linear feature extraction techniques such as NDA, or filter banks like Haar-based filters have been successfully used on face detection algorithms [17] and can provide several advantages.

In this paper we propose to experimentally evaluate a set of different descriptors with a specific classification framework for robust detection of facial salient points. We evaluate some of the existing descriptors i.e. Gabor, SIFT, Haar and NDA via a the GentleBoost classifier [5]. On top of the former comparison, we implement an eye detection algorithm following the cascade framework proposed in [17].

Our evaluation procedure consists of four main phases: First we apply a face detection – off-the-shelf technique, using the detected face coordinates we apply the implemented eye and mouth detectors, from which we as a third step, compute the regions of interest that allow the extraction of the descriptors which are in the fourth step, classified. The remainder of the document is organized as follows in section 2 we describe the afore mentioned steps, section 3 describes the experiments performed and the results obtained and section 4

concludes this paper.

2 Methodology

In this experimental evaluation, we have followed a similar preprocessing scheme to the one in [18] (figure 1) where in an initial stage, the Viola and Jones face detector has been run in order to locate the face in the image.

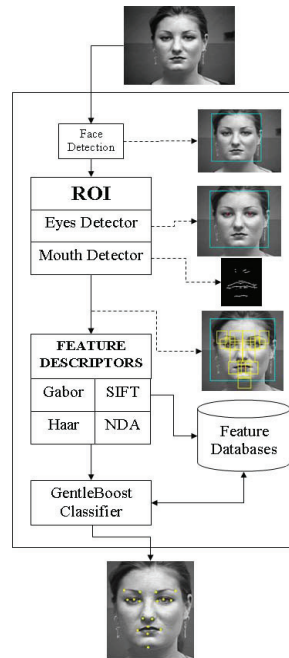


Figure 1: Diagram of the feature extraction and classification process followed.

2.1 Eye and Mouth Detector

Once the region of the face is located, the implemented eye detector is used to compute the rotation angle and the scaling factor of the face. Subsequently we normalize the image so that the interocular distance is constant among the images, and the line joining the eyes has an angle of zero degrees with the horizontal. Taking advantage of

the bounding box retrieved by the face detector, we search the coordinates of the center of the eye in the upper half of the face.

In the training stage, the coordinates of the center of each eye are used to extract a patch around them as positive examples and to generate randomly distributed patches around two different radii from the center. The extracted data patches are normalized and projected to the NDA subspace before they are passed to the classifier (for NDA see section 2.2.1). To generate a more robust detector a hierarchical cascaded approach is followed, discarding first the patches that are less likely to be of the eye class i.e. those with uniform intensity.

Faces were scaled making the interocular distance 80 pixels. The patches used are adjusted to 22x34 pixels to cover the entire region of the eye. The classifiers were trained using 630 images from the Cohn Kanade database [9], which were manually annotated. A cross validation scheme was used dividing the image set in 7 groups.

Results show that the majority of the error is clustered below the 7 pixel error, hence this value will be used as threshold for the validation of the facial salient point algorithm. Large errors account for images with very limited dynamic range (either under or over exposed). The percentage of eyes with smaller normalized error than the interocular distance value shows that over 96% of the test samples are localized with less than 20% of the normalized error.

The next step is the localization of the mouth, first we use the horizontal coordinates of the detected eyes and the bottom half of the facial bounding box to generate a search area covering the nostrils and the mouth. Within this region a combination of a horizontal edge map and the integral image, are combined and used to obtain the coordinates of the center point of the mouth.

These three reference points (eyes and mouth coordinates) serve as a base to generate the 17 regions that constitute the centre of the searching areas for the salient points as depicted in the top image of the right column in figure 2.

Each salient point receives a specific label and a binary classifier is trained to relegate the representative sample of each salient point from the surrounding region.

2.2 Descriptors

Robust classification of facial salient points requires the use of adequate feature descriptors from the local neighborhood of each pixel in the face. In this study, we use the SIFT algorithm, the Gabor filters, the Haar-based features and the NDA descriptors, which have been widely used in the object recognition literature. In this section we briefly describe the main characteristics of each algorithm and the corresponding parameters used in this work.

2.2.1 Nonparametric Discriminant Analysis

Intuitively, Discriminant Analysis aims to find the most discriminative features by maximizing the ratio of the determinant of the between-class scatter matrix to that of the within-class scatter matrix. As mentioned in [3] in the classic FDA the within class scatter matrix is usually computed as a weighted sum of the class-conditional sample covariance matrices. In NDA, the between class scatter matrix emphasizes the samples near the boundary and not only the class means, hence it makes use of the entire training set by weighting inter-class pairs according to their contribution to discrimination, i.e. it is obtained from vectors locally pointing to another class.

As stated in [3] NDA's completely nonparametric nature implies no assumption on the class-conditional distributions, thus its performance does not degrade in the presence of non-gaussian distribution. In this study, we used a 300 dimensional NDA descriptor to represent each salient point.

2.2.2 Scale Invariant Feature Transform

The SIFT presented by Lowe [13] obtains key points in the scale space that are local maxima or minima of the Difference of Gaussian function. These key points are used to extract the image gradient and orientation at those locations by means of a Gaussian-weighted window whose σ is a function of the smoothing scale being used. The product with the thresholded gradient values are stored in the orientation histogram that represents the final descriptor (a 128-dimensional feature vector in the basic implementation).

2.2.3 Gabor Filters

The main advantages of using Gabor filters for texture characterization is their performance regarding repeatability and information content as measures for saliency of a point in an image. Moreover, their use has increased due to the fitness of the real part of the complex Gabor function with the receptive field of vertebrate's visual cortex [7]. Movellan [14] presents a succinct and clear description on Gabor filters.

In our implementation we considered the orientation of the Gaussian and the sinusoid carrier as having equal value and followed the parameters suggested in [14] for the ratio of the orientation and the standard deviation of the Gaussian function. We used 6 scales and 8 orientations to construct the 48-dimensional vector of each point and the values of the intensity patch, resulting in a descriptor of $13 \times 13 \times (48 + 1)$ i.e. 8281 dimensions.

2.2.4 Haar Based filters

The *Haar descriptors* are based on the Haar basis wavelet functions and represent a basic but fast analysis tool for feature extraction. We use the basic features as mentioned in [17] i.e. the *Two, three and four Rectangle* features and the reflexions around the vertical axis. In the same way we use 6 scales starting at a base resolution of 12×12

and a factor of 2, hence each point is represented by a 40 feature vector.

3 Experiments and Results

The Experimental validation presented here was tested on a portion of 500 manually annotated grayscale images of the Cohn-Kanade database, including both genders and different races. The faces are frontal expressionless portraits of the person against a uniform background. The system was tested using a cross validation scheme, where the image database was divided in 5 sets so each time, 4 were used to train and the remaining one to test. The evaluation of the performance of each descriptor was done by comparing the coordinates of the automatically located points with the ground truth points by means of the euclidian distance with a threshold of 7 pixels, which represents less than 9% of the interocular distance.

Results for the mean accuracy results per point for the 17 points across the 5 trials as well as the mean accuracy for the descriptors for all the points are shown in the table in the left side of figure 2. To attest the reliability of these results the confidence interval computed as

$$I = 1.96 \cdot \sigma / \sqrt{5} \quad (1)$$

is shown in parenthesis.

Overall results show a 91.5% as the highest average recognition rate for the Intensity-Gabor descriptor (middle image of figure 2) which is around 4 percentual points higher than the rest. It can be seen that point 17 has by far the lowest detection rate of all, which can be explained by the lack of saliency of the region (tip of the chin) in many of the examples (bottom image in figure 2), specially of female images. In the same way, points 5, 8 and 16 have a low performance that can be attributed to the same reason as before and to the lack of consistency in the annotation of the ground truth points.

If the former point is considered and point 17 is discarded, the performance of the Gabor filters

Point	Gabor	SIFT	Haar	NDA
P-1	99 (0.17)	98 (0.76)	84 (1.26)	98 (0.06)
P-2	99 (0.09)	97 (1.02)	89 (1.13)	97 (0.09)
P-3	99 (0.14)	92 (0.86)	80 (0.83)	93 (0.07)
P-4	93 (0.14)	89 (1.04)	88 (1.02)	86 (0.07)
P-5	79 (0.17)	78 (1.12)	81 (0.70)	72 (0.06)
P-6	85 (0.17)	91 (0.73)	93 (0.93)	79 (0.06)
P-7	91 (0.09)	93 (1.04)	91 (1.04)	84 (0.09)
P-8	87 (0.09)	78 (1.46)	82 (1.47)	60 (0.09)
P-9	93 (0.17)	90 (1.15)	86 (1.12)	94 (0.06)
P-10	99 (0.17)	95 (1.44)	88 (1.01)	95 (0.06)
P-11	99 (0.09)	92 (1.85)	80 (1.34)	87 (0.09)
P-12	98 (0.09)	91 (1.89)	95 (1.11)	92 (0.09)
P-13	99 (0.14)	97 (0.49)	94 (0.54)	96 (0.07)
P-14	97 (0.14)	97 (0.79)	97 (0.98)	94 (0.07)
P-15	97 (0.14)	89 (1.06)	88 (0.82)	92 (0.07)
P-16	79 (0.14)	76 (1.13)	88 (0.65)	88 (0.07)
P-17	60 (1.18)	42 (1.69)	60 (1.22)	75 (0.10)
Mean	91.5 (0.20)	87.35 (1.15)	86.12 (1.01)	87.18 (0.07)

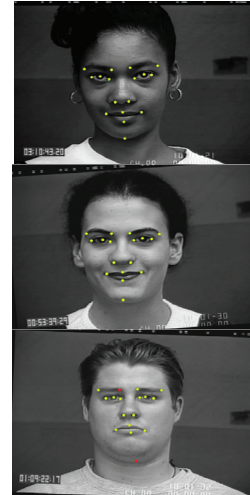


Figure 2: Left: Study Results. Mean accuracy results for each point of interest using the Gabor, SIFT, Haar and NDA descriptors with Gentleboost classification rule. Right Top: Example of the locations of the points provided with the data base, Middle: Accurate detection, Bottom: Inaccurate detection

and the SIFT descriptors is quite similar reaching a 93% for the former and a 91.5% for the later, which still favors the Gabor Jets as computational cost are also lower for it.

As for the Haar descriptor, their strength lies in the simplicity of their concept and hence its computational (low) cost, which allows for them to be combined in large amounts of ways to generate a single descriptor. In this study, only the 3 basic shapes and only 2 types of orientation were used, which could lead to their comparative poor performance.

4 Conclusion

In the present document we have presented a comparative, experimental study for the Gabor, SIFT, Haar and NDA descriptors. The method employed a GentleBoost classifier and was tested on the Cohn-Kanade database by means of a 5 fold cross validation technique.

The intensity-Gabor jet descriptor achieved the highest average recognition rate at a 91.5%. Some of the proposed salient points represent a challenge or even may not represent saliency at all, as is the case for point 17 if considered alone.

We implemented an eye detector system, following a cascade scheme and taking advantage of the characteristics of the discriminant analysis, reaching a performance of 96.2% of accuracy for a 7 pixel radius, less than 9% of the normalized interocular distance.

References

- [1] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real Time Face Detection and Facial Expression Recognition: Development and Applications to Human Computer Interaction. In *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, 2003.

- [2] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, 2004.
- [3] M. Bressan and J. Vitria. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters*, 24(15):2743–2749, 2003.
- [4] R. El Kaliouby and P. Robinson. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *Proc. Intl Conf. Computer Vision & Pattern Recognition*, volume 3, page 154. Springer, 2004.
- [5] J. Friedman, T. Hastie, and R. Tibshirani. Special invited paper. additive logistic regression: A statistical view of boosting. *Annals of statistics*, pages 337–374, 2000.
- [6] A. Goneid and R. El Kaliouby. Enhanced Facial Feature Tracking of Spontaneous and Continuous Expressions. *Systems, Social and Internationalization Design Aspects of Human-computer Interaction*, 2001.
- [7] J. Jones and L. Palmer. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6):1233–1258, 1987.
- [8] T. Kadir, A. Zisserman, and M. Brady. An Affine Invariant Salient Region Detector. *LECTURE NOTES IN COMPUTER SCIENCE*, pages 228–241, 2004.
- [9] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000. Proceedings*, pages 46–53, 2000.
- [10] V. Lepetit and P. Fua. Keypoint Recognition Using Randomized Trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1465–1479, 2006.
- [11] T. Lindeberg. Scale-space theory: a basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(1):225–270, 1994.
- [12] E. Louprias, N. Sebe, S. Bres, and J. Jolion. Wavelet-based salient points for image retrieval. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 2, 2000.
- [13] D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, volume 2, pages 1150–1157. Kerkyra, Greece, 1999.
- [14] J. Movellan. Tutorial on Gabor Filters. *Tutorial paper* <http://mplab.ucsd.edu/tutorials/pdfs/gabor.pdf>.
- [15] F. Schaffalitzky and A. Zisserman. Multi-view Matching for Unordered Image Sets, or” How Do I Organize My Holiday Snaps?”. *LECTURE NOTES IN COMPUTER SCIENCE*, pages 414–431, 2002.
- [16] N. Sebe, M. Lew, Y. Sun, I. Cohen, T. Gevers, and T. Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.
- [17] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2002.
- [18] D. Vukadinovic and M. Pantic. Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 2, 2005.

Using Colour Saliency for Image Retrieval

Juan Ignacio Toledo and Joost van de Weijer

Computer Science Department, Computer Vision Center, Edifici O - Campus UAB, 08193 Bellaterra, Spain

E-mail: JuanIgnacio.Toledo@uab.es, Joost@cvc.uab.es

Abstract

Visual saliency has grabbed a lot of attention in the computer vision field and different approaches have appeared to model it. But to our knowledge no attempts have been made to use visual saliency in real world applications. We will show how colour saliency can help in a task like image retrieval. For this purpose we will use a bag of words approach with features combining shape and colour information. We will also show how to incorporate colour salience information into a difference of Gaussians feature detector. Results show that colour saliency allows for a more compact image representation. Furthermore we demonstrate that the theoretical optimum from information theory for colour boosting agrees with optimum in real world applications.

Keywords: Color Information, Image Retrieval, Visual Saliency, Applications.

1 Introduction and related work

Visual saliency is what makes certain areas in an image grab our attention first. It is a key factor in the ability of the Human Visual System to understand a scene in real time as shown in [3].

Several successful attempts have been made to build different models of saliency, [2, 4]. However, the question is: can we use saliency to improve results in a real world application? Until now image

retrieval systems have ignored saliency theory [9].

Saliency is attributed to several cues, such as orientation, shape, motion and colour [3]. We will focus on colour saliency. Colour contrast is known to attract human attention. In [10] *colour boosting* method, is proposed which allows to focus feature detectors on image features with high colour information content. Information theory teaches that rare events have a higher information content than frequent events. Therefore, rare colours have a higher information content. Based on the statistics of real-world images the authors compute a theoretical optimal transformation to obtain colour salient features. The paper fails however to answer two questions: 1. Can the proposed colour saliency method be used in a real world application. 2. Is the theoretical transformation also optimal for real computer vision applications?

We will answer both questions for the task of image retrieval on a large benchmark dataset [7]. Until relatively lately image retrieval focussed mainly on global image representations [1]. Recently bag of words models, initially applied to object and category recognition [9, 8], were shown to obtain very good results [7] because they are robust to image occlusion and several geometric transformations.

Can we use visual saliency to improve the results in an image retrieval experiment? We will prove that in a bag-of-words image retrieval experiment, using David Lowe's SIFT [6, 5] and Van de Weijer's Hue Descriptor [11] good results can be obtained with a reduced number of features by se-

lecting them according to how salient they are.

1.1 Saliency

Complex biological systems need a real time scene understanding. Identifying any and all interesting targets in a scene has an excessive computational complexity. A common solution to this problem is to restrict the complex object recognition process to a small area and serially process interesting areas in the image. While this may be a solution to the inability of the brain to fully process all areas of a scene in parallel, it introduces the issue of which areas should be examined firstly. In human visual system, saliency is the key factor to solve this. Visual saliency is a bottom-up stimulus-driven signal that announces “this location is sufficiently different from its surroundings to be worthy of your attention”. This bottom-up deployment of attention can be triggered by properties such as shape, orientation or movement of the objects in the scene

Typical saliency detection methods, as proposed in [3], transform an input image into a saliency map, that is, gray level values showing how salient each location in the scene is. See Figure 1. Many approaches exist in the literature to compute saliency maps. Some of them based in spatial information (shape+orientation)[2] or colour. Despite this there has not yet been a test on how saliency can help on a real computer problem such as object recognition, image retrieval, robot navigation, etc...

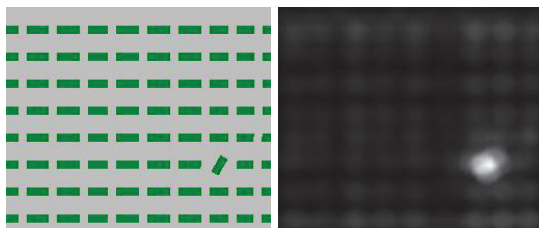


Figure 1: *Example of a saliency map*

1.2 Bag of Words based image retrieval

Bag of Words approach has been recently adapted to be used in images getting interesting results. The first step to use Bag of Words with images is to find meaningful words, that is the feature detection and description step (think of a feature as a “visual” word). Since images have noise, slightly different features should be grouped to represent the same word, we call this “Vocabulary Building”. This can be done by any clustering method, for instance, K-means. See [9].

Each image can then be represented as a histogram of the visual words in the vocabulary. Retrieval of similar images is done by using some distance measure in which the “closer” two histograms are, the more similar the images they describe will be. An efficient and high accuracy retrieval can be achieved with this framework [8].

Scale-Invariant Feature Transform:SIFT SIFT was proposed by David Lowe in [6] and reviewed in [5]. The SIFT algorithm is able to find stable image features that are invariant to image translation, scaling and rotation and partially invariant to changes in lighting and affine or 3D projection, making them suitable for object recognition.

Scale space theory is used to find interesting keypoints where the descriptors will be computed using gradient information to get a high-dimensional feature. This descriptor is formed from a vector containing the values of all the orientation histogram entries. Our experiments were done using a 4x4 array of histograms with 8 orientation bins in each. SIFT feature descriptors will be a $4 \times 4 \times 8 = 128$ element vector for each keypoint. The experiments in this report use the publicly available code by Gyuri Dorko.

Hue Descriptor The Hue descriptor was proposed by Van de Weijer in [11] and is used to describe the colour in an image area around a certain location. A rectangular grid is built around an image point and the hue of each bin then obtained. Finally we build a histogram on how many bins have

a certain hue. The descriptor is then normalised using the grey energy to make it invariant from light intensity changes.

Hue descriptor's information content is not as important as other shape descriptors like SIFT but it has been shown to be a great complement to SIFT. Both descriptors can be combined, getting a robust representation on both shape and colour of an image interest point, by simply concatenating both histograms (and applying a constant multiplicative factor on the hue descriptor to tune its influence):

$$F = (SIFT, \lambda Hue) \quad (1)$$

2 Colour Boosted Image Retrieval

We propose a new approach to improve the accuracy in image retrieval while reducing the number of features needed to represent the images using colour saliency. From information theory, the information content of an event follows:

$$I(e) = -\log(p(e)) \quad (2)$$

This means that events with lower probability are more informative. Furthermore, in the real world, most events are dominated by luminance changes [10]. An ellipsoid can be fitted to the derivative histograms. It's major axis will match the luminance while the two minor axis will contain the colour information. This anisotropic distribution can be transformed into an isotropic one by using colour saliency boosting. This will result in an increase of the energy of colour events that will attract the detector towards more salient colour features.

In this report we want to Colour Boost the Laplace blob detector. To this aim, we fit an ellipse in to the second order matrix in the RGB space.

Which is given by:

$$M = \begin{pmatrix} \overline{R_{ww}R_{ww}} & \overline{R_{ww}G_{ww}} & \overline{R_{ww}B_{ww}} \\ \overline{G_{ww}R_{ww}} & \overline{G_{ww}G_{ww}} & \overline{G_{ww}B_{ww}} \\ \overline{B_{ww}R_{ww}} & \overline{B_{ww}G_{ww}} & \overline{B_{ww}B_{ww}} \end{pmatrix} \quad (3)$$

where the matrix elements are computed as follows

$$\overline{R_{ww}R_{ww}} = \sum_{i \in I} \sum_{x \in X^i} R_{ww}(x)^i R_{ww}(x)^i \quad (4)$$

Where I is the set of all images, X^i is the set of all pixels in image i and ww indicates the second derivative in the gradient direction. The matrix M can then be decomposed into an eigenvalue and eigenvector matrices $M = V\Lambda V'$. From [10] it is known the first eigenvalue corresponds with the luminance axis. Therefore, because of information theory, the other two axis are more informative. To exploit this we propose to apply a boosting matrix which transforms second-order derivatives of equal information content into second-order derivatives of equal strength. The new boosting matrix is equal to $M_{boost} = \Lambda^{-1}V'$ and can be applied to an image to transform it in such a way that its second order matrix will be isotrope.

Whereas in the original work [10] the boosting matrix was computed only once for the whole dataset (*global saliency*) we also analyze the case of *local saliency* where the boosting matrix is computed for each image. In Figure 2 we see how, before boosting, the detector was mainly focused in black and white events (many completely white patches, with little shape and colour information, were detected) while, after boosting, the detector shifted towards more colourful events, and almost no completely white patches are found to be interesting.

2.1 Colour Laplace Framework

In this section we will describe how to extend the laplace scale-space theory to colour signals.

The laplace is defined by

$$Lap^\sigma = \sigma^2(L_{xx} + L_{yy}) \quad (5)$$

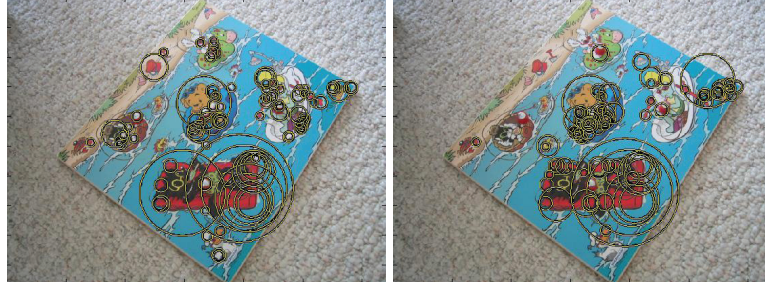


Figure 2: Features with higher response for the traditional (left) and the boosted (right) detector .

where σ is the scale of differentiation. Maxima and minima of $||Lap^\sigma||$ function can be considered scale invariant features. Similarly, the colour laplace can be defined by

$$Lap^\sigma = L_{xx} + L_{yy} \quad (6)$$

where L is a vector image, e.g $L = \{R, G, B\}$ and the vector differentiation is done according to $L_{xx} = (R_{xx}, G_{xx}, B_{xx})^T$. If the vector norm is defined as $||L|| = \sqrt{L^T L}$ then the maxima in $||Lap^\sigma||$ are the scale invariant colour features.

Then boosting can be incorporated into the colour laplace with:

$$L'_{xx} = \Lambda^{-1} V' L_{xx}. \quad (7)$$

Also intermediate boosting versions can be computed with:

$$L_{xx}^\alpha = (\alpha \Lambda^{-1} + (1 - \alpha) I) V' L_{xx}. \quad (8)$$

From a theoretical point we expect to get the best results with $\alpha = 1$, since this is the optimum derived from information theory.

3 Experimental Setup

In order to prove that colour saliency can help in a real world application we performed an image retrieval experiment. Using the dataset presented in [7] which consists of 10200 pictures organised in sets of four pictures for each object from different

viewpoints. The goal is to find the four images of the same object. Thus, the score is measured by C from 0 to 4.

Each visual word was a combination of a 128 dimensions SIFT feature and a 36 dimension Hue descriptor. We have arbitrarily selected a vocabulary size of 1000 to test the influence of colour boosting in the results. The vocabulary was built by a simple K-means clustering among the features with a higher response to the detector. Each image was represented by a histogram of 1000 bins of the occurrences of each visual word. Retrieval consisted in finding the four nearest neighbours of that image. We used Manhattan distance since it produced better results than Euclidean while having a smaller computational cost.

4 Experiments and Results

In this section we will review the different experiments performed as well as their results.

4.1 Shape and Colour Vocabulary Combination

The combination of the two features can be tuned by a constant λ that multiplies the Hue Descriptor histogram increasing the importance of the colour information in relation with the shape information from the SIFT descriptor. A first experiment was to see the impact of this factor in the results. In

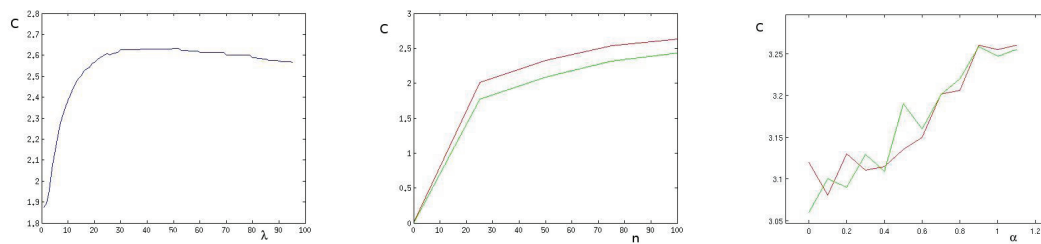


Figure 3: Results as a function of the number of top features selected (left). Results depending on the vocabulary combination constant λ (center). Local (green) vs global (red) boosting (right).

Figure 3 (left) we can see that adding the colour information can greatly increase the performance (up to a 35%) in the experiment. The higher score is reached around $\lambda = 50$ and then the score starts to drop slowly.

4.2 Boosting VS Non-Boosting

Our main goal was to show that boosting can help in image retrieval. Thus, performed an experiment with the whole dataset starting with a low number of features per image.

In figure 3 (center) we see in red the results of applying full colour boosting and in green the results without boosting. An improvement of a 14% is achieved when using only the top 25 salient features per image and slowly decreases to a 9% when using 100 features. We can also see that using colour saliency boosting allows for a more compact representation of the images with the same results. Note that earlier results on this dataset ignored colour description [7].

4.3 Optimal amount of boosting

In this third experiment we want to check if the optimum amount of boosting in a real world application is equal to the theoretical optimum. In Figure 3 (right) we can see the influence of the amount of boosting (α) and we notice that the maximum is near the theoretical optimum which

is one. An $\alpha = 1$ will produce an isotropic distribution which maximises the information content.

We also compare Global boosting and Local boosting. In Global boosting, the histogram matrix M is built using all the images in the dataset. While in Local boosting, we built a different histogram matrix M for each image which is probably closer to the behaviour of the HVS. In figure 3 (right) we can see how both approaches get similar results. We have to notice that there is a small random noise in the results due to the k-means clustering. If local and global boosting have similar results, it is less time consuming to use global boosting, since you can build a histogram matrix only once.

4.4 Shape saliency VS Colour Saliency

In this final experiment, we investigate if we can also obtain a performance gain by using shape saliency. To do so we combined the work of [2] with our method but it produced no improvements in the score. The way to do it was to multiply the maxima in the scale-space with the normalized saliency map. We tried this method with and without colour boosting but in both cases it did not change the results. This is explained by the fact that, when using boosting, the features are already located into a salient object. Also, without boosting, by selecting only the features with a higher response to the detector, they are also mostly in salient areas and shape saliency does not have that effect of avoiding low-informative white patches.

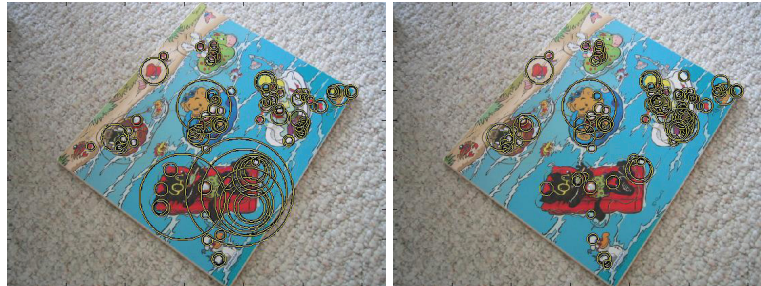


Figure 4: *Boosted detector response(left). Boosted detector modulated by a shape saliency map(right).*

5 Conclusions and Future Work

From the experiments performed we can draw the following conclusions:

a) We proved that colour saliency boosting is a good tool to increase the effectiveness of the descriptors for an image retrieval experiment, allowing for a more compact representation with the same results.

b) We showed that adding colour information to a SIFT descriptor with the Hue descriptor is a good way to improve the information content, and thus the discriminative power of the descriptor.

c) The theoretical optimum amount of boosting matches gets best results in real world.

d) Global and Local boosting both obtain similar results with real world images.

e) Using only shape saliency does not improve the results.

Further work could be done to check if local boosting outperforms global boosting for computer generated images. Work could be also done in finding a clever distance measure for combined colour and shape histograms. We could also test if Late Fusion of is better than our Early Fusion to build the vocabular Furthermore it would also be very important to test if saliency can help in other tasks, such as object recognition or robot navigation.

References

- [1] Th. Gevers and A. Smeulders. "Color based object recognition." *Pattern Recognition*, 1999.
- [2] Xiaodi Hou and Liqing Zhang. "Saliency detection: A spectral residual approach." *CVPR*, 2007.
- [3] Laurent Itti, Christof Koch, and Ernst Niebur. "A model of saliency-based visual attention for rapid scene analysis." *PAMI*, 1998.
- [4] Tie Liu, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. "Learning to detect a salient object." *CVPR*, 2007.
- [5] David G. Lowe. "Distinctive image features from scale-invariant keypoints." *IJCV*, 2004.
- [6] David G. Lowe. "Object recognition from local scale-invariant features." *ICCV*, 1999.
- [7] David Nister and Henrik Stewnius. "Scalable recognition with a vocabulary tree." *CVPR*, 2006.
- [8] Josef Sivic and Andrew Zisserman. "Video google: A text retrieval approach to object matching in videos." *ICCV*, 2003.
- [9] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. "Discovering object categories in image collections." *Proc ICCV*, 2005.
- [10] J. van de Weijer, Th. Gevers, and A.D. Bagdanov. "Boosting color saliency in image feature detection." *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2006.
- [11] Joost van de Weijer and Cordelia Schmid. "Coloring local feature extraction." *Proc. ECCV2006*, 2006.

Text Segmentation in Colour Poster from the Spanish Civil War Era

Antonio Clavelli and Dimosthenis Karatzas

Computer Vision Center, Univesitat Autònoma de Barcelona (UAB), 08193, Bellaterra, Barcelona, Spain

E-mail: {aclavelli, dimos}@cvc.uab.es

Abstract The extraction of textual content from colour documents of a graphical nature is a complicated task. The text can be rendered in any colour, size and orientation while the existence of complex background graphics with repetitive patterns can make its localization and segmentation extremely difficult. Here, we propose a new method for extracting textual content from such colour images that makes no assumption as to the size of the characters, their orientation or colour, while it is tolerant to characters that do not follow a straight baseline. We evaluate this method on a collection of documents with historical connotations: the Posters from the Spanish Civil War

Keywords: color segmentation, text extraction

1 Introduction

Among the many collections of historical documents that are routinely the focus of mass digitisation efforts, there are some collections of significant value that are typically overlooked due to their increased difficulty as far as it concerns the automatic extraction of their textual content. The presence of colour and the overlapping of text and graphical regions are usual reasons for neglecting such documents. One such collection is the Catalan Posters from the Spanish Civil War kept by the National Archive of Catalonia in Spain. The collection comprises about 10000 posters (1200 already scanned) with political, ideological and propagandistic content produced in the first half of the 20th century. The text is in Catalan and Spanish, and it is combined with photographic or hand drawn graphical content (see Figure 1).

The automatic extraction and interpretation of the semantic content in these images has a

considerable scholarly value. Moreover, this collection is a good example of numerous other poster collections produced over the last century that share the same characteristics. Techniques devised for these images should therefore be generally applicable.



(a)



(b)

Figure 1. Spanish Civil War era posters.

Technically speaking the extraction of the textual content from these images is a complicated process for a number of reasons. The text can be rendered in any colour, size, orientation and style, while complex graphical backgrounds give rise to an excessive amount of noise components during segmentation.

In this paper we propose a method for extracting textual content from colour documents of a graphical nature that makes no assumption as to the size, orientation or colour of the characters, while it is tolerant to characters that do not follow a straight line. In Section 2 an overview of existing work on the topic is presented. Section 3 details our method for text extraction, while Section 4 presents results over a set of images from the Spanish Civil War poster collection. Section 5 concludes this paper.

2 Literature Review

Contrary to classical document image analysis, where the problem of segmentation is equivalent to identifying regions of text (layout analysis) on a black and white image, in the case of colour images the problem is pixel-level segmentation, namely to produce such a black and white image on which character recognition can then take place.

The segmentation of text in colour images has received increasing attention over the last decade. Ad hoc solutions for various types of colour text containers have been reported including colour paper documents [1-3], text in Web images [4], real scenes [5-7], video sequences [8] etc. Generally the reported methods have limited applicability outside their restricted focus due to the special characteristics of each image type.

In bottom-up segmentation schemes, results are generally represented as connected components, which have to be subsequently classified as belonging to the textual content or not. Neither colour segmentation nor component classification is a trivial problem.

Retornaza and Marcotegui [7] construct connected components using contrast information in a mathematical morphology framework. Component classification is performed on the basis of 27 features while a subsequent merging process is used to group them into horizontal text lines.

Jung et al. [6] use a stroke filter to perform local region analysis. Stroke-like structures are identified in all possible orientations sizes and positions. The hypothesis here is that characters have uniform thickness.

A typical problem with the classification of the resulting connected components is the sheer number and variability of components corresponding to parts of the image background. It is frequently the case that the ratio between

foreground and background components in an image is between 1:102 and 1:103 depending on the complexity of the graphical content. Therefore, attempts to classify individual components yield limited results unless large numbers of features are used [7]. Generally, identifying groups of similar components is instead used to extract whole text lines. Even at this level problems might arise, as complex graphical regions generally give rise to patterns that resemble text.

A number of explicit or implicit assumptions are usually made when working with complex colour images. In most of the cases, text is assumed to have a constant colour [1-3, 5-8]. Also more often than not, text is supposed to be arranged on straight text lines, and quite frequently only horizontal straight lines are looked for [1, 5-8].

3 Text Extraction Method

The text extraction method follows three steps: colour segmentation, individual connected component filtering and connected component grouping into text-lines. These are explained in detail in the next sections.

3.1 Colour Segmentation

In the specific document collection we are examining here characters are rendered in uniform colour. This is an advantage as it ensures a mostly error-free segmentation. Due to this a relatively simple colour segmentation method has been employed here.

The segmentation process used here is a one-pass algorithm which creates 8-connected components of similar colour from the input image. The algorithm processes the image in a left-to-right, top-to-bottom fashion, and for each pixel calculates its colour similarity with the four of its neighbours that are already assigned to a connected component. The pixel then is either assigned to one of the existing components if their colour difference is below a set threshold, or it is used as the seed for a new component. Additional

checks are performed in case a pixel is similar to more than one existing components, in which case components might be merged. The algorithm is further described in [9].

Colour similarity is assessed in the RGB colour space. A better choice would be a perceptually uniform colour space such as CIELab or CIELuv, but since text here is rendered in a uniform colour working with RGB results in a considerable faster algorithm without seriously affecting its performance. The colour threshold was set to 75 after experimentation (see section 4). The result of the segmentation process is a set of connected components that comprise pixels of similar colour. An example is shown in Figure 2.



Figure 2. Segmentation result on the images of Figure 1.

3.2 Component Filtering

In order to reduce the number of components that correspond to areas of the background, an initial size filtering is performed. If the characters were placed on a straight horizontal line, then the most appropriate size metric to use for filtering would be their height. In reality text can have any direction, so the diagonal of the bounding box of the components is instead used here as it is much less dependent on rotation. Components that have a diagonal of less than five pixels are deemed too small to represent characters and are discarded. Similarly, components with a diagonal bigger than half the minimum dimension of the image are deemed too big and are discarded. This filtering step is very conservative and only discards components that are extremely small or too big to

be characters. Nevertheless it reduces a lot the complexity of the following steps.

3.3 Text Line Identification

The remaining components correspond to characters as well as to parts of the background. The question posed at this point is how to select groups of components that correspond to the characters of a text line without making any a-priori strong assumption. The single assumption that we will have to make is that the characters that belong to the same word are similar. However, defining in what way they are similar is not a trivial issue. Although the colour of individual characters is constant, this is not always the case when whole words or text lines are examined. For example there are words that comprise letters of different colour (see Figure 3). As a result, colour similarity is not used here as a feature for text line identification.

Similarity is instead defined purely on spatial and topological characteristics. Specifically, characters of the same word are required to have similar height and thickness, and to be placed at uniform intervals along the text line.

As will become clear next, the notion of height is dynamic, as it is always defined based on the direction of the text line being assessed. In addition, a text line is not assumed to be a straight line, instead it dynamically follows the character components.

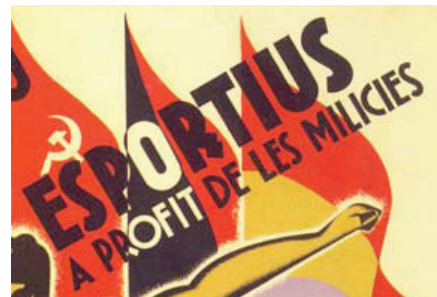


Figure 3. Example of text comprising different coloured characters.

3.3.1 Seed Components Selection

Our starting hypothesis is that every connected component can potentially correspond to a character. If this is the case, then there has to exist a second component also corresponding to a character in close proximity to the first one that shares similar height and thickness characteristics. For the purpose of this work, thickness is defined as the ratio of the area (number of pixels) to the perimeter (number of external pixels) of a component. This is not a precise definition but it serves well for our purposes.

A donut shaped area with $R_1 = 0.2 \cdot D$ and $R_2 = 2 \cdot D$ (where D is the diagonal of the original component) around the centre of the bounding box of the original component is searched, and components that satisfy the following equations are identified.

$$\frac{T_h}{1.5} \leq T_j \leq 1.5 \cdot T_k \quad (1)$$

$$\frac{H_h}{1.5} \leq H_j \leq 1.5 \cdot H_k \quad (2)$$

Where T_i and H_i are the thickness and height of component i , k is the original component, and j is the component being examined. The height of a component is defined as the projection of the component on the orientation of the line connecting the two components as shown in Figure 4.

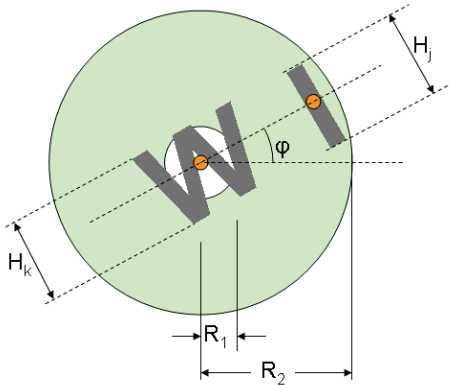


Figure 4. Search space and definition of height.

3.3.2 Text Line Building

Each pair of components identified in the first step is used as a seed for a potential text line. Given the initial line segment specified by the

centres of the two components, the algorithm searches symmetrically on both sides of the line segment on the orientation defined by the segment for more components that could belong to the text line. The search space is defined based on the distance of the two initial components (d_0), and the orientation of the line segment as shown in Figure 5. An angle threshold of $\theta_0=35^\circ$ and a maximum distance of $1.5 \cdot d_0$ is used to specify the search space.

All the components that lie within the search area specified are checked and the ones that have dissimilar height or thickness are discarded. For this test, the same equations (1) and (2) are used as before. The components that pass this test are ranked based on how far they lie from the expected location and the expected direction based on the following equation.

$$S = W_\theta \cdot \frac{\theta}{\theta_0} + W_d \cdot \frac{d}{d_0} \quad (3)$$

The weights used are $W_\theta=0.7$ and $W_d=0.3$. The component that yields the lowest score is chosen as the next component of the line, and the search process is repeated, this time centred on the new component as shown in Figure 5.

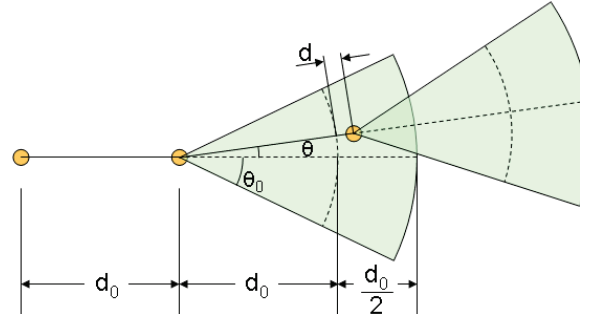


Figure 5. Adding components to the text line.

3.3.3 Text Line Selection

Following the above process a number of potential text lines can be identified for each starting component (one for each pair of components identified in step 3.3.1). These text-lines are examined and either one or none is selected as the most probable text line for the component (since each character can only belong to a single line).

This is achieved by first filtering all the text lines that comprise less than 3 components. A further filtering is then performed based on the inter-component distance; that is the distance between successive components on the line. We ask that the maximum inter-component distance is no more than 1.9 times the minimum inter-component distance. A factor of less than 2 was deliberately chosen as otherwise it is possible to obtain lines that jump over characters (select every second character on the line).

Finally, of all the lines that passed the filters above we choose the one that has the minimum average inter-component distance.

This process (steps 3.3.1 to 3.3.3) is repeated for every component in the image, and the final set of text-lines produced by the algorithm is returned as the textual content of the image.

4 Results and Discussion

We defined pixel-level ground-truth information (each pixel labelled as text or background) for 50 of the images of our dataset. The result of the algorithm was assessed based on the ground-truth and metrics for precision, recall and fall-out were calculated based on the following equations.

$$P = \frac{|retrieved \cap relevant|}{|retrieved|} \quad (4)$$

$$R = \frac{|retrieved \cap relevant|}{|relevant|} \quad (5)$$

$$F = \frac{|retrieved \cap relevant|}{|relevant|} \quad (6)$$

“Retrieved” is the set of the pixels that belong to the components of the text lines identified by our algorithm. “Relevant” is the set of pixels that belong to characters as specified in the ground-truth while “Non Relevant” the set of pixels of the background.

We repeated the experiment for different values for the colour threshold used for segmentation (see section 3.1) and in each case we calculated the precision, recall and fall-out values. The results are plotted in Figure 6.

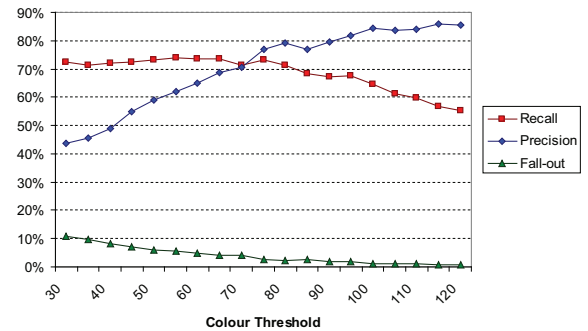


Figure 6. Precision, Recall and Fall-out for different colour thresholds.

As the threshold increases, less noisy segmentations are produced and the algorithm is not tricked by combinations of noise components that resemble text lines, therefore precision increases correspondingly and fall-out drops. After a given threshold though, characters start being merged with the background, and recall therefore drops. The best threshold seems to be around 75, which is the value we have used here. For this threshold we obtain $P=76.9\%$, $R=73.2\%$ and $F=2.6\%$. Some image results are shown in Figure 7.



Figure 7. Final results for the images of Figure 1.

The described algorithm is able to identify individual text lines. Nevertheless, the evaluation performed here focuses on the segmentation aspect, and disregards the text line information. It is reasonable to question this choice, as alternatively we could attempt to OCR each individual text-line and report on the OCR success. This is not as straightforward as it seems though as further post-processing has to take place (e.g. unwrapping curved text lines, joining words that have been identified as separate “text-lines”, updating the OCR dictionary with appropriate

Catalan terms and acronyms required by the historical context of this collection etc.). This is intended future work.

5 Conclusion

A method was described to extract the textual content from complex colour images. The proposed method makes no assumptions as to the size, orientation or colour of the characters, while it is tolerant to characters that do not follow a straight baseline. We evaluated our method on a dataset of poster images from the Spanish Civil War and achieved promising segmentation results with precision and recall above 70%.

6 Acknowledgements

This work was partially supported by the Spanish projects TIN2008-04998, CONSOLIDER INGENIO 2010 (CSD2007-00018) and the fellowship 2006 BP-B1 00046. The images used were kindly provided by the National Archive of Catalunya.

References

- [1] K. Sobottka, H. Kronenberg, T. Perroud, and H. Bunke, "Text extraction from colored book and journal covers ", *International Journal on Document Analysis and Recognition*, Springer, Vol. 2, 2000, pp. 163-176.
- [2] H. Hase, M. Yoneda, S. Tokai, J. Kato and C.Y. Suen, "Color segmentation for text extraction", *International Journal on Document Analysis and Recognition*, Springer, Vol. 6. 2004, pp. 271-284.
- [3] C. Strouthopoulos, N. Papamarkos and A.E. Atsalakis, "Text extraction in complex color documents", *Pattern Recognition*, Elsevier, Vol. 35, 2002, pp. 1743-1758.
- [4] D. Karatzas and A. Antonacopoulos, "Colour text segmentation in Web images based on human perception", *Image and Vision Computing*, Elsevier, Vol. 25(5), 2007, pp. 564-577.
- [5] V. Wu, R. Manmatha and R.M. Riseman, "TextFinder: An automatic system to detect and recognize text in images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21(11), 1999, pp. 1224-1229.
- [6] C. Jung, Q. Liu and J. Kim, "A new approach for text segmentation using a stroke filter", *Signal Processing*, Elsevier, 2008, Vol. 88, pp. 1907-1916.
- [7] T. Retornaz and B. Marcotegui, "Scene text localization based on the ultimate opening", *Proceedings of the 8th International Symposium on Mathematical Morphology*, 2007, pp. 177-188.
- [8] H.K. Kim, "Efficient Automatic Text Location Method and Content-Based Indexing and Structuring of Video Database", *Journal of Visual Communication and Image Representation*, Elsevier, Vol. 7, 1996, pp. 336-344.
- [9] D. Karatzas, "Text Segmentation in Web Images Using Colour Perception and Topological Features", PhD Thesis, University of Liverpool, UK, 2002.

A rotation invariant page layout descriptor for document classification and retrieval

Albert Gordo and Ernest Valveny

Computer Vision Center, Universitat Autònoma de Barcelona, Spain
 {agordo, ernest.valveny}@cvc.uab.es

Abstract

Document classification usually requires of structural features such as the physical layout to obtain good accuracy rates on complex documents. This paper introduces a descriptor of the layout and a distance measure based on the cyclic Dynamic Time Warping which can be computed in $\mathcal{O}(n^2)$. This descriptor is translation invariant and can be easily modified to be scale and rotation invariant. Experiments with this descriptor and its rotation invariant modification are performed on the Girona Archives database and compared against another common layout distance, the Minimum Weight Edge Cover. The experiments show that these methods outperform the MWEC both in accuracy and speed, particularly on rotated documents.

Keywords: Document Analysis, Document Retrieval, Structural Features, Rotation Invariance, Cyclic Dynamic Time Warping.

1 Introduction

Document classification is an important task in document management and retrieval. It is usually based on the extraction of some features from the document image. These document features can be of different types. In [3] three categories are pro-

posed (adapted from the four proposed in [2]): *image features*, extracted directly from the image or from a segmented image (e.g. the density of black pixels of a region), *structural features* or relationships between blocks in the page, obtained from the page layout, and *textual features*, based on the OCR output of the image. A classifier may combine these features to get better results.

Structural features are necessary to classify documents with structural variations within a class. For this task, several methods exist. Some of them define a distance between blocks and pages based on some criteria like the Minimum Weight Edge Cover [5] or the Earth's Mover Distance [8, 9]. Others represent the layout as a graph or tree and define some graph distances between them [2, 6]. Sometimes we can calculate the probability that a given graph is generated by a class representative graph generator [1]. Finally, documents can also be classified based on sets of rules [4].

In this paper we present a method to represent and classify document layouts based on a graph representation of the regions, which is later flattened into a cyclic sequence, obtaining a vector representation of the document layout. The comparison of these vector representations is done using the cyclic dynamic time warping. This approach allows a $\mathcal{O}(n^2)$ approximate comparison, which is fast compared to the typical exponential cost of some graphs methods or the $\mathcal{O}(n^3)$ of those

based on the assignment problem. Moreover, this representation is invariant to translation and can be easily made invariant to scale and rotation.

In section 2 we introduce this representation along with the rotation invariant alternative. We also define the distance measures to compare two different page representations. Section 3 deals with the experimentation, and, finally, in section 4 we summarize the obtained results.

2 The cyclic polar page layout representation

The cyclic polar page layout is a cyclic sequence representation of the page layout. This approach uses an auxiliary complete bipartite graph, constructed with the centroids of the regions on one side and, on the other, the center of mass of all the regions. Figure 1 illustrates two sample pages with their layout and corresponding graph. Nodes of the graph are labeled with the area and the type of the region while edges are labeled with their length and either the angle or its delta increment. Then, the graph is converted into a vector representation obtaining a cyclic sequence. This sequence will later be compared by means of the cyclic dynamic time warping. This representation is translation invariant, and can be made scale invariant normalizing the mass and length. Moreover, it can be made rotation invariant disregarding the angles and keeping only their delta increments.

2.1 Formulation

Let $Z = \{z_1, z_2, \dots, z_m\}$ represent the m physical layout non-rectangular regions or zones of a segmented page, where each z is a 3-tuple $z_i = (C_i, A_i, T_i)$, C being the set of centroids of the zones, A their area and T their type (in our experiments, *text* or *non text*). Let also R be the center of mass of Z , where the mass M_i of z_i is equal to A_i .

Let the 3-tuple $G = (U, V, E)$ be a $K_{1,m}$ complete bipartite graph where $U = \{R\}$, $V =$

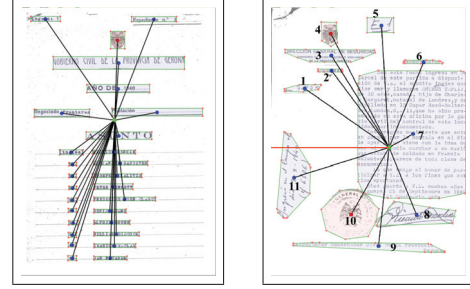


Figure 1: Two different layouts and their corresponding graphs.

C and $E = U \times V$, and, particularly, $E = \{e_1, e_2, \dots, e_m\}$ where $e_i = \overline{RC_i}$. Let the vertices be labeled with their mass (area) and type, and let the edges be labeled with their angle θ and their euclidean length L .

Let N be a sequential combined representation of the nodes and edges of G , $N = \{n_1, n_2, \dots, n_m\}$ where n is a 5-tuple $n_i = (\theta_i, L_i, M_i, D_i, T_i)$, where D_i is a vector of densities $D_i = \{D_{i1}, D_{i2}, D_{i3}, D_{i4}\}$ and where N is sorted as a function of θ .

A cyclic shift σ of a sequence $A = \{a_1, a_2, \dots, a_m\}$ is a mapping $\sigma : \Sigma^* \rightarrow \Sigma^*$ defined as $\sigma(\{a_1, a_2, \dots, a_m\}) = \{a_2, \dots, a_m, a_1\}$. Let σ^k denote the composition of k cyclic shifts and let σ^0 denote the identity. Two sequences A and A' are cyclically equivalent if $A = \sigma^k(A')$, for some k . The equivalence class of A is $[A] = \{\sigma^k(A) : 0 \leq k < m\}$ and it is called a cyclic sequence.

Finally, let the equivalence class $[N]$ be our cyclic layout representation of the page.

2.2 Invariance

The above representation has the interesting feature of being translation invariant. The representation can also be made scale invariant normalizing the length and mass of N . Nonetheless, this should be avoided unless the database actually contains significant scale variations within a class.

Also, replacing θ_i with Δ_i , where Δ_i is the absolute minimum angle formed by θ_i and $\theta_{(i+1) \bmod n}$ makes the representation rotation invariant. We will represent this invariant alternative as ΔN and its equivalence class as $[\Delta N]$. As we will prove through experimentation, this representation obtains better results on sets with rotations while keeping the good results on rotationless sets.

2.3 Distance measure

We must now define a distance measure between two layout representations $[N^a]$ and $[N^b]$ (or $[\Delta N^a]$ and $[\Delta N^b]$), cyclic sequences of sizes m and n . The very first step will be defining a distance D between sequences without considering the shiftings, i.e., between N^a and N^b . For this task, two obvious choices arise: the Edit Distance (ED) and the Dynamic Time Warping (DTW). The ED has the advantage of easily modelling some variations within a class like stamps or signatures that only appear sometimes and can be handled as insert or delete operations. On the other hand, DTW does a better job on common merge and split situations (due to errors in the layout segmentation or within class variations) thanks to the one-to-many correspondence. Preliminary experiments show that the DTW (coarsely normalized with the sum of the size of the sequences) offers better results than ED, and thus will be the one we will be using. For the DTW we must also define the cost function $\gamma(y, x)$ between nodes n_y^a and n_x^b . This can be simply defined as a linear combination of its parameters:

$$\begin{aligned} \gamma(y, x) = & k_1 \text{AngleDiff}(\theta_y^a, \theta_x^b) + \\ & k_2 \text{LengthDiff}(L_y^a, L_x^b) + \\ & k_3 \text{MassDiff}(M_y^a, M_x^b) + \\ & k_4 \text{DensityDiff}(D_y^a, D_x^b) + \\ & k_5 \text{TypeDiff}(T_y^a, T_x^b), \end{aligned} \quad (1)$$

where

$$\text{AngleDiff}(\theta_y^a, \theta_x^b) = \frac{\hat{\Delta}(\theta_y^a, \theta_x^b)}{\pi}, \quad (2)$$

$$\text{LengthDiff}(L_y^a, L_x^b) = d(L_y^a, L_x^b), \quad (3)$$

$$\text{MassDiff}(M_y^a, M_x^b) = d(M_y^a, M_x^b), \quad (4)$$

$$\begin{aligned} \text{DensityDiff}(D_y^a, D_x^b) = & \frac{1}{4}(d(D_{y1}^a, D_{x1}^b) + \\ & d(D_{y2}^a, D_{x2}^b) + d(D_{y3}^a, D_{x3}^b) + d(D_{y4}^a, D_{x4}^b)) \end{aligned} \quad (5)$$

$$\text{TypeDiff}(T_y^a, T_x^b) = \begin{cases} 0 & \text{if } T_y^a = T_x^b, \\ 1 & \text{otherwise,} \end{cases} \quad (6)$$

where $\hat{\Delta}(\theta^a, \theta^b)$ is the minimum angle difference between θ^a and θ^b , and where d is defined as follows:

$$d(a, b) = 2 \left(1 - \frac{a + b}{2 \max(a, b)} \right). \quad (7)$$

If we are using the rotation invariant representation, $\text{AngleDiff}(\theta_y^a, \theta_x^b) / \pi$ should be replaced with $\text{DeltaDiff}(\Delta_y^a, \Delta_x^b) = d(\Delta_y^a, \Delta_x^b)$.

Values for weights k_1 to k_5 can be obtained through validation as explained in section 3.

2.4 Cyclic distance measure

Once a distance between N^a and N^b has been defined we must define it for the equivalence classes $[N^a]$ and $[N^b]$. The exact distance can be naively computed with cost $\mathcal{O}(m^2 n^2)$ as

$$CD([N^a], [N^b]) = \min_{\substack{0 \leq k < m \\ 0 \leq l < n}} D(\sigma^k(N^a), \sigma^l(N^b)) \quad (8)$$

and in $\mathcal{O}(m^2 n)$ [7] as

$$CD([N^a], [N^b]) = \min_{0 \leq k < m} (\min D(\sigma^k(N^a), N^b), D(\sigma^k(N^a) N_{k+1}^a, N^b)) \quad (9)$$

This last equation can also be solved in $\mathcal{O}(mn \log m)$ by means of a more elaborate divide & conquer approach [7].

Also, suboptimal solutions can be found in $\mathcal{O}(mn)$, e.g., calculating the standard DTW between the concatenations of N^a and N^b with themselves, i.e.,

$$\text{ApproxCD}([N^a], [N^b]) = D(N^a N^a, N^b N^b) \quad (10)$$

Measuring the distance with the cyclic DTW seems reasonable when using the rotation invariant descriptor $[\Delta N]$, but its usefulness when employing the angle-fixed $[N]$ is not that obvious. The reason why we use the cyclic DTW instead of the standard one is that on some descriptors there are many nodes with θ close to zero, above and below. Due to the cyclic nature of the angle difference, using the cyclic DTW we will allow some reasonable matches that were not possible with the standard DTW. This leads to better results even if the angles are fixed.

3 Experiments

To evaluate this cyclic representation of the layout we will test it with the Girona Archives database. The Girona database is a collection of documents from the Civil Government of Girona, in Spain, that contains documents related to people going through the Spanish-French border from 1940 up to 1976 such as safe-conducts, arrest reports, documents of prisoners transfers, medical reports, correspondence, etc. Even if it is a mostly-text database, in this case most of the pages also have images like stamps, signatures, etc in a non-manhattan disposition. We have used a subset of the database that contains 823 manually segmented images and is currently divided in 8 different categories. Some of these images are slightly skewed (see Fig. 1), but in most cases the skew is almost nonexistent.

For all of our experiments the testing strategy will be leaving-one-out cross-validation. Experiments will be done both for classification, using a simple nearest neighbour classifier, and retrieval. First, we will try to obtain the best values

for the weights k_1 to k_5 , used in cost function of the approximate cyclic DTW (10) both with $[N]$ and $[\Delta N]$ representations. Later, we will compare the results to another common layout distance measure, the minimum weight edge cover. Finally, we will apply some random rotations to the pages and test how this affects the different distance measures.

3.1 Validation of cost function weights

As we explained in section 2.3, appropriate values for factors k_1 to k_5 in the γ cost function must be found. We will try to find the weights that minimize the classification error and, also, the weights that maximize the average precision for retrieval. This objective maximization has been done by means of a genetic programming approach.

Table 1 shows the results for the three best classification error and retrieval rates. We can observe that the best weights for the classification are different than those optimized for retrieval. When classifying, we are only interested in the similarity with the nearest neighbour, while in retrieval we try to obtain a good similarity with as many documents of the class as possible. If we had used a different classifier that took into consideration not only the nearest neighbour but all the documents in the class, the weights would have most likely been similar to those obtained in retrieval. We can see that the type of the regions is very relevant, and so is the absolute angle when dealing with retrieval rates. However, in classification, angle has a much lower impact and mass is much more important.

3.2 Comparison with another layout distance

To test the results of this representation, we will compare our results against another common layout distance, the Minimum Weight Edge Cover (MWEC) [5]. This distance is based on the assignment problem and has an asymptotic cost of $\mathcal{O}(n^3)$. It has also been proven to provide better

	k_1	k_2	k_3	k_4	k_5	
Girona	(Angle)	(Length)	(Mass)	(Density)	(Type)	Score
Classification	0.109	0.043	0.348	0.065	0.435	1.742
Error	0.042	0.155	0.338	0.084	0.380	1.876
	0.049	0.147	0.328	0.098	0.377	1.876
Average	0.180	0.100	0.28	0.170	0.270	0.7749
Precision	0.165	0.137	0.257	0.154	0.156	0.7746
	0.180	0.132	0.241	0.168	0.277	0.7745

Table 1: Best classification and average precision rates for $[N]$ as a function of factors k_1 to k_5 .

results than other similar measures like the pure assignment problem or the earth’s mover distance [9].

Table 2 shows the best results obtained in classification and retrieval with the approximate cyclic DTW over $[N]$ compared to those obtained using the MWEC along with their average time¹. We can see that $[N]$ obtains better results than MWEC both in classification and retrieval.

	Error Rate	Av. P.	Av. Time
$[N]$	1.74	0.774	5.0s
MWEC	1.74	0.698	17.4s

Table 2: Classification and retrieval rates with different tuned distances.

It should be noted that a deep inspection of the results reveals that most of the $[N]$ errors are produced within two categories whose pages contain a very low number of zones. In this case, the translation and rotation invariance is inconvenient, as this is a critical information for the correct classification. Moreover, in pages with few zones, the center of mass is subject to a high variance, leading to inappropriate descriptions. On the other hand, the MWEC correctly classifies these regions most of the time. A combined classifier using MWEC when the page contains few zones and $[N]$ oth-

erwise would most likely outperform any of the methods separately.

3.3 Rotation

In this last experiment we will observe how the cyclic DTW over $[N]$ and $[\Delta N]$ compares to the MWEC when the pages can be rotated. To do so, we will first apply to each page an uniform random rotation in the range $[-\pi/2, \pi/2]$ radians and we will re-learn the appropriate weights of the γ function for $[N]$ and $[\Delta N]$ both for classification and retrieval. Then we will apply a second, different rotation to all the pages and we will check the results using $[N]$ and $[\Delta N]$ with the new weights, and MWEC.

Table 3 shows these results. As expected, the results obtained with $[\Delta N]$ on retrieval are better than those obtained either with $[N]$ or MWEC. Nonetheless, the results obtained with $[N]$ are quite close to those obtained with $[\Delta N]$ in retrieval and *slightly* better on classification. However, in both cases the angle weight is set to zero, so in this case both methods are equal and the difference is on the other weights selection, and suggests that $[\Delta N]$ weights were overfit. It should also be noted that the results of $[\Delta N]$ are not exactly equal to those obtained previously. This difference is due to the fact that we are using an approximate cyclic DTW and not an exact one.

¹Time to compare each of the 823 layouts against each other

	Error Rate	Average Precision
$[N]$	3.21	0.743
$[\Delta N]$	3.61	0.752
MWEC	5.89	0.41

Table 3: Classification and retrieval rates over a set of randomly rotated pages.

4 Conclusions

In this paper we have presented a descriptor for page layouts and a distance measure between page descriptions that can be computed in $\mathcal{O}(n^2)$. This descriptor is translation invariant and can be modified to be scale and rotation invariant. Experiments with the non-manhattan Girona Archives database prove these methods to perform better than the common Minimum Weight Edge Cover both in the classification and retrieval and in speed. Most of the classification errors are produced on documents with few zones, so the use of this descriptor is not encouraged in such cases and an hybrid method is suggested.

When the documents are rotated, the rotation invariant descriptor clearly outperforms the MWEC. The rotation invariant descriptor performs only slightly better than the unmodified one when its angle weight is set to a low value, which leads to think that the exact angle is not as important as the overall structure. Finally, using the exact cyclic DTW would yield better results both in classification and retrieval using $[\Delta N]$ over rotated images, but the computation cost would increase.

References

- [1] A D Bagdanov and M Worring. First order gaussian graphs for efficient structure classification. *Pattern Recognition.*, 36(6):1311–1324, 2003.
- [2] F Cesarini, M Lastri, S Marinai, and G Soda. Encoding of modified x-y trees for document classification. *Proceedings. Sixth International Conference on Document Analysis and Recognition*, pages 1131–1136, 2001.
- [3] N Chen and D. Blostein. A survey of document image classification: problem statement, classifier architecture and performance evaluation. *International Journal on Document Analysis and Recognition.*, pages 1–16, 2007.
- [4] Floriana Esposito, Donato Malerba, and Francesca A List. Machine learning for intelligent processing of printed documents. *Journal of Intelligent Information Systems.*, 14(2):175–198, 2000.
- [5] D. Keysers, T. Deselaers, and H. Ney. Pixel-to-pixel matching for image recognition using hungarian graph matching. In *DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, pages 154–162, 2004.
- [6] Jian Liang, D. Doermann, M. Ma, and J.K. Guo. Page classification through logical labelling. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 3:477–480 vol.3, 2002.
- [7] Andrés Marzal and Vicente Palazón. Dynamic time warping of cyclic strings for shape matching. *Pattern Recognition and Image Analysis*, pages 644–652, 2005.
- [8] Y Rubner, L Guibas, and C Tomassi. The earth mover’s distance, multi-dimensional scaling , and color-based image retrieval. *Proceedings of the ARPA Image Understanding Workshop*, pages 661–668, May 1997.
- [9] J. van Beusekom, D. Keysers, F. Shafait, and T.M. Breuel. Distance measures for layout-based document image retrieval. *Document Image Analysis for Libraries, 2006. DIAL ’06. Second International Conference on*, pages 11 pp.–242, April 2006.

Comparison of Seal Detection by Different Character Shape Features

Partha Pratim Roy* Umapada Pal⁺ and Josep Lladós*

* *Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain*
E-mail:(partha, josep)@cvc.uab.es

⁺ *Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata - 108, India*
E-mail:umapada@isical.ac.in

Abstract

Due to noise, overlapped text/signature and multi-oriented nature, seal (stamp) object detection involves a difficult challenge. This paper deals with automatic detection of seal from documents with cluttered background. A seal object is characterized by scale and rotation invariant spatial feature descriptors (distance and angular position) computed from recognition result of individual connected components (characters). Recognition of multi-scale and multi-oriented component is done using Support Vector Machine classifier. Generalized Hough Transform (GHT) is used to detect the seal and a voting is casted for finding possible location of the seal object in a document based on these spatial feature descriptor of components pairs. The peak of votes in GHT accumulator validates the hypothesis to locate the seal object in a document. Experimental results show that, the method is efficient to locate seal instance of arbitrary shape and orientation in documents.

1 Introduction

Content-based Image Retrieval (CBIR) consists of retrieving visually similar images to a given query image from a database of images. In machine readable documents searching, the query consists of a

collection of words and it retrieves documents according to user query. Moreover graphical information can be very relevant. The searching can be generalized by object terms such as logo, seal (stamp) etc. Indexing of documents having certain seals can be done based on seal and information obtained from seals could be used for efficient storage and retrieval of the documents. Undoubtedly, automatic seal detection and recognition is an important stage to follow this approach. It also allows to identify the document sources reliably. Detection and recognition of seal is a challenging task. Because, seals are generally unstable and sometimes contain unpredictable patterns due to imperfect ink condition, uneven surface contact, noise etc.[13]. Overlapping of seal with text/signatures and missing part of seal are typical problems and add difficulty in seal detection. Seals generally bear some constant character strings to convey the information about the owner-organization and its locality. Besides, in many instances seal contains some variable fields [7] for e.g., date. A seal instance may be affixed on any position within the document, which requires detection to be carried out on the entire document. Also a seal may be placed in arbitrary orientation. See Fig.1, where a part of an archive document containing a seal is shown. The seal is overlapped with part of sig-

nature and text regions. As a result some information of the seal is missing/illegible. A prior



Figure 1: A part of document showing a seal overlapped with signature

knowledge of seal-shape structure [13] is helpful to localize them in documents. But, it is difficult to detect the seal if text information within seal is different though the shape is similar. Due to the fact that, in many instances seals and text are in different colors, some researchers [4, 12] have studied the detection process based on color analysis. Sometimes seal has been treated as a symbol and methodology like segmentation by Delaunay tessellation [3] are applied for seal recognition. Text information matching [6, 7] has also been exploited to recognize textual objects. Detection of seal objects in documents can be treated as localizing objects in different pose. In object detection methodology, spatial organization of local key-point descriptors or appearance codebook have been used [1, 5]. These methods are promising in terms of accuracy, time and scalability. Document objects (symbol, seal, logos) are synthetic entities consisting of uniform regions which are highly structured [10, 11]. These facts make geometric relationships between primitives a discriminative cue to spot symbols. Given a single instance of a symbol queried by the user, the system has to return a ranked list of segmented locations where the queried symbol is probable to be found.

When imprinted, seal produces a fixed set of text characters with a logo (sometimes) by its nature to describe its identity. It contains text characters in different rotations and scale. Moreover, as the text document image where seal object appears also contains text characters and it is difficult to detect seal by matching only text characters. Since, the spatial arrangement among text characters in seal objects are structured in nature, using features

computed from spatial information will allow us to detect the seal. Hough transform provides a way of dealing with the complexity issue of searching and has been used for various pose estimation problems including shape detection [2]. By detecting the local parts to vote for possible transformations (translation, scale and rotation) of the seal object, we can use the peaks of voting space for locating seal objects.

We label the local connected components of a seal as alpha-numeric text characters. To obtain this, we employ multi-scale and multi-oriented text character recognition system. The recognized text characters are used as high level descriptor in our approach. For each component (text character) we find its n -nearest neighbors. Next for each pair of components, we encode their relative spatial information using their distance and angular position. Given a query seal we compute the relative spatial organization for pair-wise text characters within seal. These information are stored into a spatial feature bank. In an image, for each component pair, we use this spatial feature bank to retrieve query seal hypothesis. A vote is casted for possible location of the seal based on the matching of components pairs to ensure detection of partial seal objects. The peak of votes in GHT accumulator validates the hypothesis to locate the seal object.

The main contribution of this paper is to use of recognized local components as high level descriptors and to generate hypothesis of the seal object location based on spatial arrangement of these descriptors. This approach is robust to detect seal in noisy, cluttered document. The rest of the paper is organized as follows. In Section 2, we describe the representation of the query seal object and their detection process. The experimental results are presented in Section 3. Finally conclusion is given in Section 4.

2 Seal Detection using GHT

In documents, the seals are affected mainly by rotation, occlusion and overlapping. To take care of these problem, our approach is based on partial

matching which is inspired from the Generalized Hough Transform (*GHT*) [2]. In *GHT*, the edge information of a shape is used to define a mapping from the orientation of an edge point to a hypothetical reference point of the shape and detect the shape based on the accumulation of these reference points. The purpose is to find similar instances of objects by a voting procedure. The voting procedure is carried out in a parametric space, from which object candidates are localized as local maxima in voting space. In our approach, extracted text characters in a seal are considered as high level descriptors. The spatial information of these descriptors are used to vote for seal object detection under a certain pose. Here, we describe the architecture of our method with three key-parts namely, spatial information encoding, seal representation and hypothesis voting. The former two allow to represent model shapes in a compact way in hashing structures. The latter is the main step when a query is formulated into the image database.

2.1 Spatial Information Encoding

Our representation model for a seal is based on the connected components (recognized as text characters) within seal region. Again, text characters are too local and very generic to be used for matching primitives. We characterize the seal information based on spatial organization of these local text character shapes and the attributes used in our approach are described below.

Reference point of seal: The center of the minimum enclosing circle (MEC) of the seal is considered as the reference point (R). Note that, R will be fixed for each class of seal shapes in size and rotation invariant way.

List of component pair: To represent the spatial arrangement of the components present in the seal a list of component pair is built. These component pairs are selected based on their proximity [10]. Each component is associated to its n -nearest components from its boundary. The neighbor components are decided using region growing algorithm [8] from the boundary of the correspond-

ing component.

Formally, A proximity graph $G = (V, E)$ is created to store these component pair. If seal contains n components, so, $\|V\| = n$. An edge, $e_{ij} \in E$ is formed by component pair (C_i, C_j) if, components C_i and C_j are neighbor. See Fig.2(a), where the n -neighbor ($n = 5$) components of ‘U’ (according to distance) are shown by joining line, which are ‘L’, ‘L’, ‘F’, ‘A’ and ‘T’.

From each neighbor component pair, the location of reference point (R) of seal object is learned. The spatial information of each component pair with respect to R is done by simple basic features: distance and relative angle discussed as follows.

Distance: We compute the CG (center of gravity) of each component for this purpose. The distance (see, “ AD ” in Fig.2(b)) of component pair is defined by d_{ij} for component pair C_i and C_j as given by Eq.(1). Here, CG_c is CG of component C and mean component size S_{mean} is the average size of component pair.

$$d_{ij} = Euclidean(CG_{ci}, CG_{cj}) / S_{mean} \quad (1)$$

Relative Angle: Two angles are considered for each component pair to identify reference point (R). For each pair of components we may consider a triangle by their CG_{ci} , CG_{cj} with R (see Fig.2(b)). Angle α_i is formed at CG_{ci} by CG_{cj} ($j \neq i$), and R . These angles are denoted by α_i and α_j . In Fig.2(b), these angles are “ $\angle RAD$ ”, “ $\angle ADR$ ” respectively.

Spatial feature vector: Finally, we encode the spatial information for each component pair C_i and C_j in a 3-tuple feature vector f_{ij} considering distance and relative angle by Eq.(2). Given a component pair, the feature vector f is used to represent their spatial organization.

$$f_{ij} = \{d_{ij}, \alpha_i, \alpha_j\} \quad (2)$$

2.2 Seal Representation

We selected a model seal for each class of seal images to be representative of that class. Before obtaining all local features of this model seal automatically, first we manually removed all non text

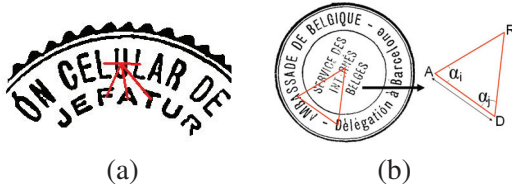


Figure 2: a) A character and its n -nearest neighbor
b) A character pair is shown in a seal. Here, R is the reference point.

regions or graphical portions (seal boundary, large strokes of signature etc.) and variable field (like date information). Next, the remaining components are recognized by a text character classifier to have text character labels. For each character component, we find the n -nearest neighbors and enlist all character pairs of the seal by using proximity graph (G_q). Then for each pair of text characters we compute the spatial descriptor f_i (from Eq. (2)) with respect to reference point. This 3-tuple feature vector describes the local spatial arrangement. Thus each of the model seal is represented by a list of high level spatial feature descriptors $F_q = \{f_1, f_2, \dots, f_p\}$, where, p is the total number of character pairs found. The feature vectors are entered into a Spatial Feature Bank (S) using a hash table.

Hash table is used to index the spatial information of component pair of a model seal. A hash function H projects a value from a set of members to a value from a set with a fixed number of (fewer) members. Here, a hash function is created to obtain the index key to store spatial feature vector for pair-wise components. Determination of index key (k) for a component pair is described as follows.

Hash key generation: A Look Up Table (LUT) has been used as hash function to create the index key. It is to find a suitable code (k) (Eq.(3)) for each pair of text characters, where we record the spatial feature vector of corresponding character pair. Here, L_i and L_j denote the text character label of component C_i and C_j respectively.

$$k = H(L_i, L_j) \quad (3)$$

If the number of distinct character class is x , then

the length of LUT is $(x \times x)/2$. The spatial information of character pair is entered in spatial feature bank (S) using this key (k). If there are m (more than one) spatial feature vector having same key, the collision is resolved by chaining. Thus using Eq.(2, 3), the storage of spatial information in S by index key (k) can be written as Eq.(4).

$$S[k] = \{f_1, f_2, \dots, f_m\} \quad (4)$$

2.3 Hypothesis Voting

Given a document image, we used a text separation method [9] to extract text character from the document. Next, a text character recognition process is employed to classify these text component. For each character component, we find n -nearest neighbor and list all character pair of the document.

For a given component pair C_x and C_y , we can make hypothesis of the location of seal reference point R_{xy} based on the spatial feature vector of component pair. Accordingly, we cast a vote in that location. By accumulating hypothesis voting, the similar seal image to the query seal is detected. To achieve it, for each pair of text character a hashing key k_{xy} is generated (see Eq.(3). Each hashing key allows us to obtain a set of spatial feature vector from Spatial Feature Bank (S) using

$$S[k_{xy}] \Rightarrow \{f_i\}$$

From each of these feature vector, we can simulate the hypothetical reference point location from the angle information provided the distance parameter matches. We compute the distance d_{xy} between C_x and C_y by Eq.(1). The matching is performed as given in Eq.(5), where d_i is obtained from f_i and D_{th} is set empirically to 0.05.

$$|d_{xy} - d_i|/d_{xy} \leq D_{th}. \quad (5)$$

This voting mechanism accumulates evidences in the locations where we find the same component pair having similar spatial arrangement. The presence of a seal object in a document provokes a peak in the voting space.

As the seal is formed by several text characters, high voting to the accumulation array points out the presence of many similar high level character-pair descriptors of that query seal object. The more we find the matching of character pair descriptor, the more accumulation of voting will be there. Finally, we detect the existence of the query seal using number of votes in local peaks of accumulation space.

3 Experimental Results

For the experiment, documents containing seal were collected from mainly historical documents. We constructed the database with 370 documents containing seals of English text characters to test the seal detection approach. The documents were digitized by a flatbed scanner at 200 dpi and these are in binary form. We checked that in our database there are mainly 3 types of shapes, eg. circular, elliptical and rectangular. There were total 19 classes of seals composed from these 3 different shapes. The database sometimes contains document images in up side down way. The seals are posted in different orientation and in different locations. There were missing of seal information many times by noise or overlapped signature. Sometimes, due to posting on text document, additional text characters of document are also placed inside the seal region. We have included additional 160 documents in this dataset which do not have the seals. Thus, our seal dataset consisted of 530 documents having different complexity.

Extracted text components in a document are recognized by a Support Vector Machine (SVM) classifier. The multi-size and multi-rotation feature used in the character recognition experiment are Angle based features [8], ART, HU, Zernike moments and Fourier Mellin. Their performance among 3 different fixed dataset are shown in Fig.3. It is noted that Angle based features and HU moments are best and worst features to classify text characters. We have done two experiments using these two different features for seal detection system. It is to understand how seal detection system

gets affected with different character recognition approach.

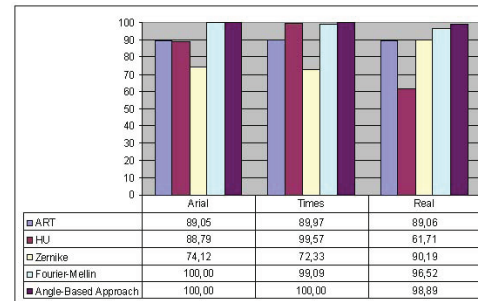


Figure 3: Detection of seals in documents

To evaluate the performance of the system with a query seal against a collection of documents, we use precision (P) and recall (R) for evaluation of retrieval of documents. The retrieved documents, the system results are ranked by the number of votes. Each result image is considered as relevant or not depending on the groundtruth of the data. In Fig.4 we show the average precision and recall plot of querying seal models in the whole database. We also compare the performance of seal detection system using two different shape features in Fig.5.

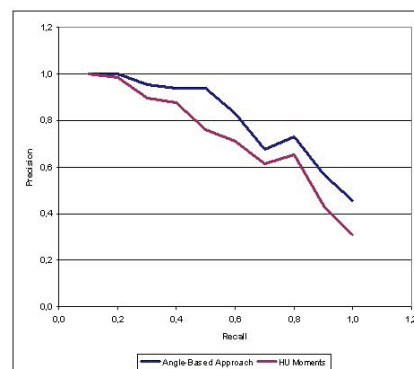


Figure 4: Precision-recall curve using different character recognition features

Given a query seal model, a ranked list of retrieved documents are shown in Fig.6. The ranking is done based on voting in accumulation array. It is understandable from the ranking of these



Figure 6: Ranking of different documents of a seal model

Shape Descriptor	Precision	Recall	F-Ratio
Angle-based Approach	91.66%	91.91%	91.78%
Hu's Descriptor	85.35%	83.97%	84.65%

Figure 5: Performance of seal detection system

images, how noise, overlapping and occlusion do affect the ranking process. Another advantage of our method is that since it works considering only labeled text characters, we can use this to detect arbitrary shape of seal. The detection accuracy is affected mainly due to broken text characters in the seal and the presence of long graphical lines over text regions (see Fig.7). If broken components of a character could not be joined through preprocessing, the character recognition algorithm does not perform well, and hence the seal detection task.



Figure 7: Documents having very low vote for the seal model

We show the detection of seal in documents using our approach in Fig.8. We used Angle based features and SVM to classify text characters. The location of a query seal is detected in these documents by our GHT approach. Also, we show here four different shapes of seal detection without any prior knowledge of the seal shape but character component information within seal. To visualize, we draw a circle of radius of MEC of the query

seal in that particular location. It is appreciable to see that our seal recognition performs well when there exists signature or other text overlapping with seal region. Here, Doc#1 and Doc#2 have two different seals though they share similar shape structure (circle). Using our approach, we correctly distinguish them and label the associate seal model with it. In Doc#3, it is shown that the query seal is detected in the document in spite of 180 degree rotation of the document. In Doc#4, we see our approach works instead of having variable date field in the seal location.

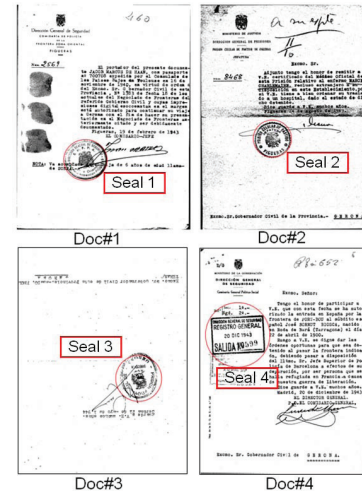


Figure 8: Detection of seals in documents

4 Conclusion

We presented a seal detection approach based on spatial arrangement of seal content. The contribution in this paper is two-fold. We have used multi-oriented and multi-scale character recogni-

tion method to generate high level local feature to take care of complex multi-oriented seal information. Recognition result of these text components within seal region are used to generate local spatial information to classify the seal. Relative positional information of text string characters are used for this purpose and hypothesis were generated based on that.

We tested our method in a database of seal documents containing noise, occlusion and overlapping. In retrieval of document image from database, all components of a document pass through a recognition process which is time consuming. So, the improvement of the performance in terms of time could be achieved if a pre-processing method (run-length smoothing) is applied to remove non-seal information from the document. The result shows that, our approach is efficient even if we do not extract all the text characters in a seal. At present we are working with 19 classes of seals and in future we want to increase the number of classes. There is not much effort in this area, so we hope it will be helpful for researchers in their future work.

5 Acknowledgements

This work has been partially supported by the Spanish projects CONSOLIDER-INGENIO 2010 (CSD2007-00018), TIN2008-04998 and TIN2009-14633-C03-03.

References

- [1] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, volume 4, pages 113–130, Denmark, 2002. Springer-Verlag.
- [2] D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. In *Pattern Recognition*, volume 13, pages 111–122, 1981.
- [3] J. Y. Chiang and R. Wang. Seal identification using the Delaunay tessellation. In *Proc. Natl. Sci. Counc. ROC (A)*, volume 22, pages 751–757, 1998.
- [4] A. S. Frisch. The fuzzy integral for color seal segmentation on document images. In *Int. Conf. on Image Processing*, volume 1, pages 157–160, 2003.
- [5] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Proceedings of the Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic, May 2004.
- [6] J. Mao, R. A. Lorie, and K. M. Mohiuddin. A system for automatically reading IATA flight coupons. In *ICDAR*, pages 153–157, USA, 1997.
- [7] F. Nourbakhsh, P. B. Pati, and A. G. Ramakrishnan. Automatic seal information reader. In *Int. Conf. on Computing: Theory and Applications*, pages 502–505, 2007.
- [8] P. P. Roy, U. Pal, J. Lladós, and M. Delalandre. Multi-oriented and multi-sized touching character segmentation using dynamic programming. In *ICDAR*, Barcelona, Spain, 2009.
- [9] P. P. Roy, U. Pal, and J. Llads. Touching text character localization in graphical documents using sift. pages 271–279, France, 2009.
- [10] M. Rusinol and J. Lladós. Word and symbol spotting using spatial organization of local descriptors. In *Proceedings of Document Analysis Systems*, pages 489–496, 2008.
- [11] K. Tombre and B. Lamiroy. Graphics recognition - from re-engineering to retrieval. In *ICDAR*, pages 148–155, 2003.
- [12] K. Ueda, T. Mutoh, and K. Matsuo. Automatic verification system for seal imprints on japanese bankchecks. In *ICPR*, pages 629–632, 1998.
- [13] G. Zhu, S. Jaeger, and D. Doermann. A robust stamp detection framework on degraded documents. In *SPIE Conference on Document Recognition and Retrieval*, 2006.

Perceptual Criteria on JPEG2000 Quantization

Jaime Moreno, Xavier Otazu and Maria Vanrell

*Computer Science Department, Computer Vision Center, Autonomous University of Barcelona
08193, Bellaterra, Cerdanyola del Vallès, Barcelona, Spain
E-mail: jmoreno@cvc.uab.es*

Abstract

The aim of this work is to explain how the Brightness Induction Wavelet Model used as a perceptual quantizer can be useful for improving a lossy compression and to introduce preliminary results. In fact, this approach consists in quantizing wavelet transform coefficients using the human visual system behavior properties. When compressing images, noise is fatal to compression performance, it can be both annoying for the observer and consuming excessive amounts of bandwidth when the imagery is transmitted. The perceptual quantization based on a chromatic induction reduces unperceivable details and thus improve both visual impression and transmission properties. The comparison between JPEG2000 without and with perceptual quantization shows that the latter is not favorable in PSNR, but the recovered image is more compressed at the same or even better visual quality measured with a weighted PSNR.

Keywords: Human Visual System, Contrast Sensitivity Function, Perceived Images, Wavelet Transform, Compression Algorithms, Bandwidth Reduction, Peak Signal-to-Noise Ratio.

1 Introduction

Digital image compression has been a topic of research for many years and a number of image compression standards has been created for different applications. The JPEG2000 [2] is intended to provide rate-distortion and subjective image quality performance superior to existing standards, as well as to supply functionality. However JPEG2000 do not provide the most relevant characteristics of the human visual system, since for removing information in order to compress the image only the information theory cri-

teria are applied. This information removal introduces artifacts to the image that are visible at high compression rates, since the compression is based in a data loss from a numerical threshold, because of the discard of many pixels with high perceptual significance.

Hence it is necessary an advanced model that removes information from its perceptual content, which preserves the pixels with high perceptual relevance regardless of the numerical information. The Brightness Induction Wavelet Model (BIWaM) is suitable for it. Both BIWaM and JPEG2000 use wavelet transform, but BIWaM uses it in order to generate an approximation to how every pixel is perceived from a certain distance taking into account the value of its neighboring pixels. By contrast, JPEG2000 applies a perceptual criteria for all coefficients in a certain spatial frequency independently of the values of its surrounding ones.

BIWaM attenuates the details that the human visual system is not able to perceive, enhances those that are perceptually relevant and produces an approximation to the image that the brain detects. At long distances, as Figure 3(d) depicts, the lack of information does not produce the well-known compression artifacts, rather it is presented as a softened version, where the details with high perceptual value remain.

The paper is organized as follows: Section 2 specifies quantization and dequantization model used by JPEG2000 for encoding and reconstruction of wavelet coefficients, thereby is described the Dead-zone Uniform Scalar Quantizer and the Visual Frequency Weighting. Section 3 describes the Brightness Induction Wavelet Model. In Section 4 the proposed method of quantization will be discussed. Experimental results applied for some test images are given in section 5. Ultimately, section 6 is where the conclu-

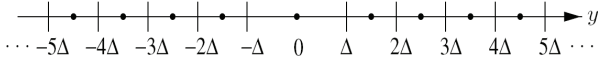


Figure 1: Dead-zone uniform scalar quantizer with step size Δ .

sions and future work will be exposed.

2 JPEG2000 Quantization Review

2.1 Dead-zone Uniform Scalar Quantizer

A uniform scalar quantizer is a function that maps each element in a subset of the real line to a particular value, which ensures that more zeros result [3]. In this way all thresholds are uniformly spaced by step size Δ , except for the interval containing zero, which is called the dead-zone and extends from $-\Delta$ to $+\Delta$, thus a dead-zone means that the quantization range about 0 is 2Δ .

For each spatial frequency s , a basic quantizer step size Δ_s is used to quantize all the coefficients in that spatial frequency according to Equation 1.

$$q = \text{sign}(y) \left\lfloor \frac{|y|}{\Delta_s} \right\rfloor \quad (1)$$

where y is the input to the quantizer or original wavelet coefficient value, $\text{sign}(y)$ denotes the sign of y and q is the resulting quantizer index. Figure 1 illustrates such a quantizer with step size Δ , where vertical lines indicate the endpoints of the quantization intervals and heavy dots represent reconstruction values.

The inverse quantizer or the reconstructed \hat{y} is given by the Equation 2, wherein δ is a parameter often set to place the reconstruction value at the centroid of the quantization interval and varies from 0 to 1.

$$\hat{y} = \begin{cases} (q + \delta)\Delta_s, & q > 0 \\ (q - \delta)\Delta_s, & q < 0 \\ 0, & q = 0 \end{cases} \quad (2)$$

The International Organization for Standardization recommends [2], the δ values are both 0.5 and 0.375, whereas Pearlman and Said suggest [6] $\delta = 0.38$, which places the reconstruction at the intervals midpoint. It is important to realize that when $-\Delta < y < \Delta$, the quantizer level and reconstruction value are both 0. For a spatial frequency, there may be many coefficients usually those of higher frequencies, that

s	HL	LH	HH
1	1	1	1
2	1	1	0.731 668
3	0.564 344	0.564 344	0.285 968
4	0.179 609	0.179 609	0.043 903
5	0.014 774	0.014 774	0.000 573

Table 1: Recommended frequency weighting for 400 dpi's

are set to 0. The array of quantizer levels q is further encoded losslessly.

2.2 Visual Frequency Weighting

In JPEG2000, only one set of CSF weights is chosen and applied according to a particular viewing condition (100, 200 or 400 dpi's) with fixed visual weighting. This viewing condition may be truncated depending on the stages of embedding, in other words at low bit rates, the quality of the compressed image is poor and the detailed features of the image are not available since at a relatively large distance the global features are more important. As more bits are received, the image quality improves, which is equivalent to decreasing the viewing distance.

The table 1 specifies a set of CSF weights which was designed for the luminance component based on the CSF value at the mid-frequency of each spatial frequency. The viewing distance is supposed to be 4000 pixels, corresponding to 10 inches for 400 dpi print or display. The table does not include the weight for LL , because it is always 1. Levels 1, 2, \dots , 5 denote the spatial frequency levels in low to high frequency order with three spatial orientations (HL , LH , HH).

3 Brightness Induction Wavelet Model

In order to explain the brightness assimilation/contrast phenomena as a unique perceptual process, Otazu et al. in [5] proposed a low-level brightness induction model, which combines three important stimulus properties: Spatial frequency, Spatial orientation and Surround contrast.

Thereby the achromatic input image \mathcal{I} is separated into different spatial frequency and orientation components from a multiresolution wavelet decomposi-

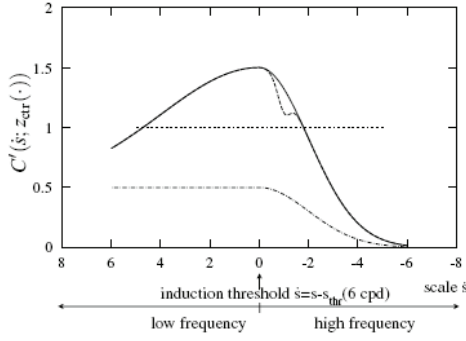


Figure 2: Contrast Sensitivity Function.

tion. Thus every single transformed coefficient is weighted using the response of the contrast sensitivity function (CSF, Figure 2) hence a perceptual brightness image \mathcal{I}_ρ is recovered. The CSF is modified considering both spatial surround information and observation distance, in this way the CSF value decreases when the surround contrast increases and vice versa.

\mathcal{I} can be decomposed a set of wavelet planes ω of different spatial frequencies, where each wavelet plane contains details at different resolutions of \mathcal{I} and it is described by:

$$\mathcal{I} = \sum_{s=1}^n \sum_{o=v,h,d} \omega_s^o + c_n \quad (3)$$

where n is the number of wavelet planes computed. The term c_n is the residual plane and the index o represents the spatial orientation either vertical, horizontal or diagonal at a certain spatial frequency.

The perceptual image \mathcal{I}_ρ recovered from the wavelet planes is defined by:

$$\mathcal{I}_\rho = \sum_{s=1}^n \sum_{o=v,h,d} C'(\dot{s}, z_{ctr}(s, o)) \cdot \omega_s^o + c_n \quad (4)$$

The term $C'(\dot{s}, z_{ctr}(s, o))$ is a weighting function, that tries to emulate some perceptual properties of human visual system, has a shape similar to the CSF and can be written as:

$$C'(\dot{s}, z_{ctr}(s, o)) = z_{ctr} \cdot C_d(\dot{s}) + C_{min}(\dot{s}) \quad (5)$$

where z_{ctr} is a non-linear function and an estimation of the central feature contrast relative to its surround

contrast. Its range oscillates from zero to one and is defined by:

$$z_{ctr} = \frac{\left[\frac{\sigma_{cen}}{\sigma_{sur}} \right]^2}{1 + \left[\frac{\sigma_{cen}}{\sigma_{sur}} \right]^2} \quad (6)$$

being σ_{cen} and σ_{sur} the standard deviation of the wavelet coefficients in two concentric rings, which represent a center-surround interaction around each coefficient.

The weighting function $C_d(\dot{s})$ is an approximation to the perceptual CSF [4] and to emulate some perceptual properties and is defined as a piecewise Gaussian function, such as:

$$C_d(\dot{s}) = \begin{cases} e^{-\frac{\dot{s}^2}{2\sigma_1^2}}, & \dot{s} = s - s_{thr} \leq 0, \\ e^{-\frac{\dot{s}^2}{2\sigma_2^2}}, & \dot{s} = s - s_{thr} > 0 \end{cases} \quad (7)$$

The term $C_{min}(\dot{s})$ avoids the $C'(\dot{s}, z_{ctr}(s, o))$ function to be zero and is defined by:

$$C_{min}(\dot{s}) = \begin{cases} \frac{1}{2} e^{-\frac{\dot{s}^2}{2\sigma_1^2}}, & \dot{s} = s - s_{thr} \leq 0, \\ \frac{1}{2}, & \dot{s} = s - s_{thr} > 0 \end{cases} \quad (8)$$

taking $\sigma_1 = 2$ and $\sigma_2 = 2\sigma_1$ so as to reproduce the approximate profile of the psychophysical functions. Both $C_{min}(\dot{s})$ and $C_d(\dot{s})$ depend on the factor s_{thr} , which is the scale associated to an induction threshold value equal to 4cpd when an image is observed from a distance d with a pixel size l_p and 1 visual degree, whose expression is defined by Equation 9.

$$s_{thr} = \log_2 \left(\frac{d \tan(1^\circ)}{4 l_p} \right) \quad (9)$$

Figure 3 shows three BIWaM images of *Lena*, which were calculated for a 19 inch monitor with 1280 pixels of horizontal resolution, at 30, 100 and 200 centimeters of distance.

4 Perceptual Method of Quantization

The block diagram of the JPEG2000 modification is illustrated in figure 4. To obtain transformed coefficients or \mathcal{I} a Forward Transformation with the 9/7 filter fast wavelet transform is first applied on the source


 Figure 3: BIWaM images of *Lena*.

image data. Then the perceptual quantized coefficients or Q with a known viewing distance are calculated by:

$$Q = \sum_{s=1}^n \sum_{o=v,h,d} \text{sign}(\omega_s^o) \left\lfloor \frac{|C'(\dot{s}, z_{ctr}(s, o)) \cdot \omega_s^o|}{\Delta_s} \right\rfloor + \text{sign}(c_n) \left\lfloor \frac{|c_n|}{\Delta_n} \right\rfloor \quad (10)$$

This expression is similar to Equation 1, but introduces a perceptual criteria. A normalized quantization step size Δ equal to $1/128$ is used, namely the range between the minimal and maximal values at \mathcal{I}_ρ is divided into 128 intervals. Finally, the perceptual quantized coefficients are entropy coded, before forming the output code stream or bitstream.

At the decoder, the code stream is first entropy decoded in order to reconstruct the perceptual quantized coefficients \hat{Q} . Second it is dequantized using 2 with a normalized quantization step size Δ equal to $1/128$ and δ equal to $3/8$. Finally, an inverse discrete transformed to recover $\hat{\mathcal{I}}_\rho$, thus providing the reconstructed perceived image data.

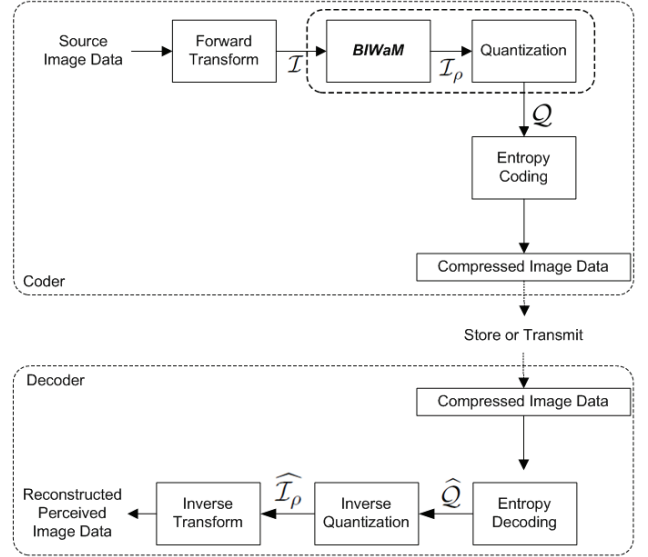


Figure 4: General block diagram of JPEG2000 compression.

5 Experimental Results

The Perceptual Criteria on JPEG2000 Quantization were tested on 44 images, but only results of the images *Peppers* and *Baboon* are reported, which are 256 gray-scale images and 512×512 of resolution (Figure 5). The BIWaM images were calculated for a 19 inch monitor with 1280 pixels of horizontal resolution at 50 centimeters of viewing distance.

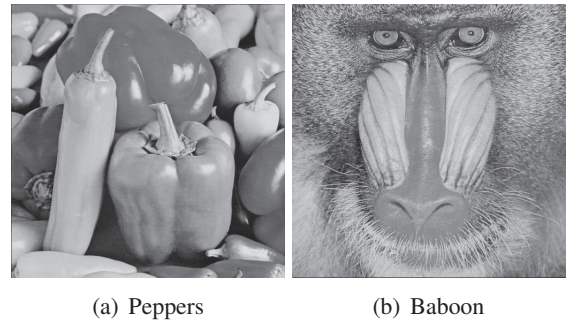


Figure 5: Tested Images.

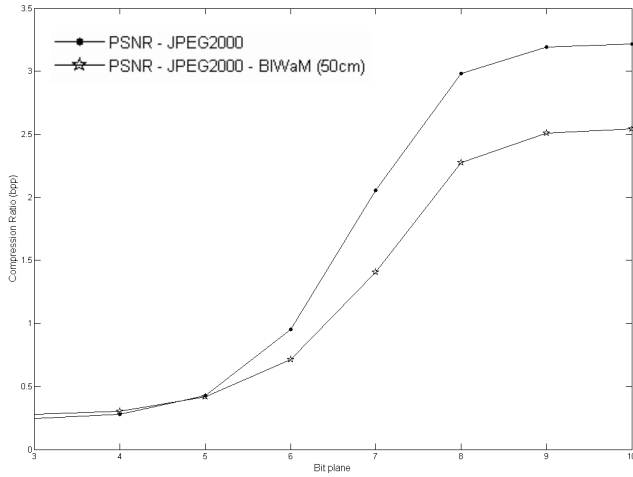
In order to measure the distortion between the original image $f(i, j)$ and the reconstructed image $\hat{f}(i, j)$ The Peak Signal to Noise Ratio was employed, however PSNR does not calculate perceptual quality measures. Therefore, it is necessary to weight each PSNR term by means of its local activity factor, taking into

account the local variance of the neighbors of the studied wavelet coefficients, thus defining a weighted PSNR or wPSNR [1]. The wPSNR increases with increasing variance and vice versa as:

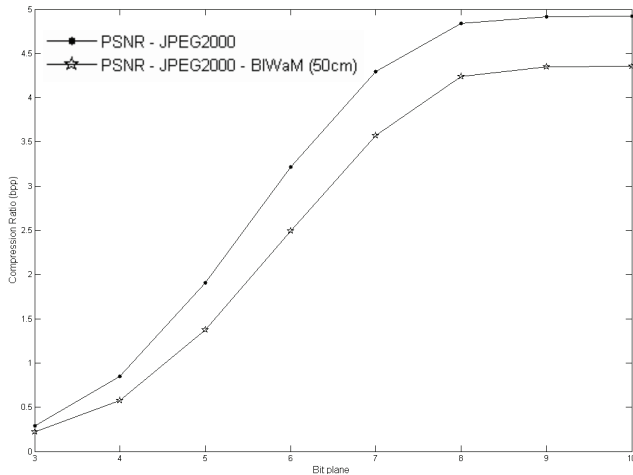
$$wPSNR = 10 \log_{10} \left(\frac{\mathcal{G}_{max}^2}{wMSE} \right) \quad (11)$$

where \mathcal{G}_{max} is the maximum possible intensity value in $f(i, j)$ ($M \times N$ size) and weighted MSE (wMSE) is defined as:

$$wMSE = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \left[\frac{f(i, j) - \hat{f}(i, j)}{1 + Var(i, j)} \right]^2 \quad (12)$$

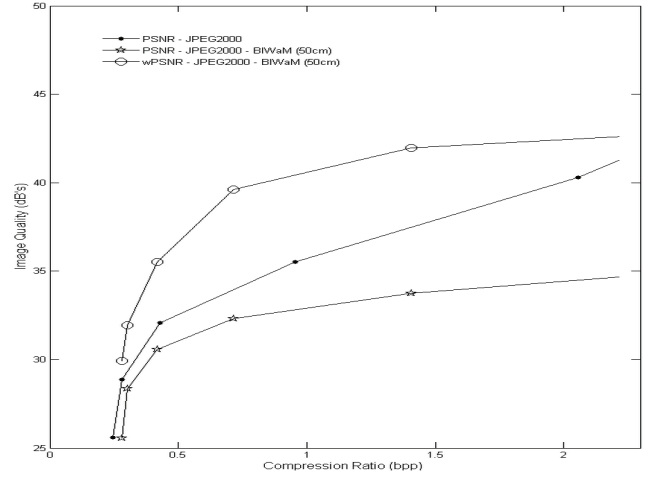


(a) Peppers.

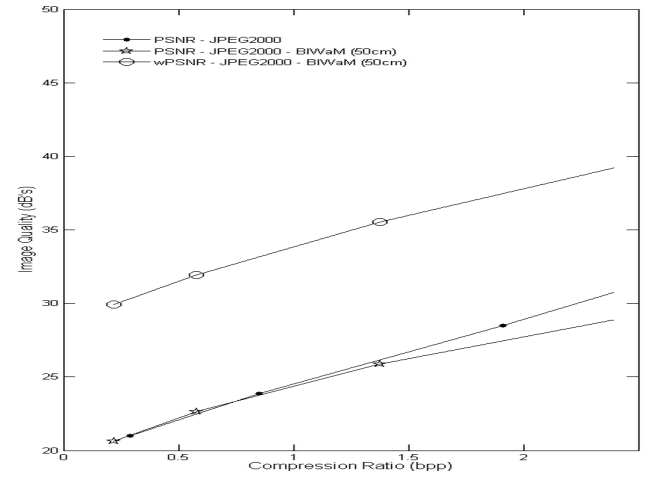


(b) Baboon.

Figure 6: Bit-plane compression ratio.



(a) Peppers.



(b) Baboon.

Figure 7: Comparison between compression ratio and image quality.

Figure 6 shows the assessment results of the compression performance at every bit-plane for a Dead-zone Uniform Scalar Quantizer and also for a BI-WaM Quantizer. In both Figure 6(a) and Figure 6(b) a BIWaM Quantizer achieves better compression ratios with the same threshold, that is because BIWaM reduces unperceivable coefficients. For example at the tenth bit-plane of *Peppers* a BIWaM Quantizer diminishes 21 percent less bits per pixel than a Scalar Quantizer, namely 22.3KB of information is perceptually irrelevant at 50 centimeters.

The comparison between compression ratio and image quality is depicted by the Figure 7, which shows that the reconstructed images quantized by BI-WaM has less PSNR but higher wPSNR than the ones

quantized by a scalar way, i.e. even if the reconstructed image has a lower objective quality, this image could have a higher perceptual quality.

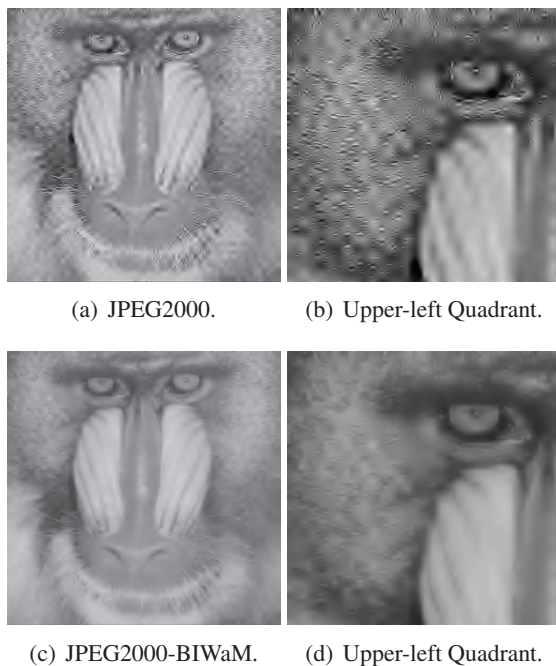


Figure 8: Reconstructed images compressed at 0.29 bpp.

The Figures 8(a) and 8(c) shows an example of reconstructed images compressed at 0.29 bits per pixel by means of JPEG2000 without and with perceptual Quantization, respectively. PSNR in 8(a) is 21.01 decibels and in 8(c) is 20.32 decibels but wPSNR is equal to 29.08 decibels, namely the reconstructed image quantized by BIWaM is perceptually better than the one quantized by a Scalar Quantizer, since the latter has more compression artifacts, as Figures 8(b) and 8(d) illustrate.

6 Conclusions and Future Work

This paper proposes an alternative of quantization for JPEG2000 using BIWaM. In order to measure the effectiveness of the perceptual quantization a performance analysis is done using the PSNR and wPSNR measured between reconstructed and original images. Unlike PSNR, wPSNR uses not only a single coefficient but also its neighbors as well as its psycho-visual properties. The experimental results show that a BI-WaM Quantization can help to improve the compres-

sion and image perceptual quality. One of the future tasks is the use of a threshold based on the CSF properties, namely a threshold based on perceptual importance that a coefficient has regardless of its numerical value.

Acknowledgements

This work was partially supported by The Spanish Ministry of Education and Science: Project TIN2007-64577 and The Mexican Science and Technology National Council: Financial Aid number 207950.

References

- [1] M. Bertini, R. Cucchiara, A. D. Bimbo, and A. Prati, "An integrated framework for semantic annotation and adaptation," *Multimedia Tools and Applications*, vol. 26, no. 3, pp. 345–363, August 2005.
- [2] M. Boliek, C. Christopoulos, and E. Majani, *Information Technology: JPEG2000 Image Coding System*, JPEG 2000 Part I final committee draft version 1.0 ed., ISO/IEC JTC1/SC29 WG1, JPEG 2000, April 2000.
- [3] M. W. Marcellin, M. A. Lepley, A. Bilgin, T. J. Flohr, T. T. Chinen, and J. H. Kasner, "An overview of quantization of JPEG2000," *Signal Processing: Image Communication*, vol. 17, no. 1, pp. 73–84, Jan. 2002.
- [4] K. T. Mullen, "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings," *The Journal of Physiology*, vol. 359, pp. 381–400, February 1985.
- [5] X. Otazu, M. Vanrell, and C. Parraga, "Multiresolution wavelet framework models brightness induction effects," *Vision Research*, vol. 48, pp. 733–751, 2007.
- [6] W. A. Pearlman and A. Said, "Image wavelet coding systems: Part II of set partition coding and image wavelet coding systems," *Foundations and Trends in Signal Processing*, vol. 2, no. 3, pp. 181–246, 2008.

Use of Filtered Back-projection Methods to Improve CT Image Reconstruction

Jorge Bernal and Javier Sánchez

Computer Science Department, Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra,, Barcelona, Spain
E-mail:jbernal@cvc.uab.es

Abstract

Actually, the use of CT scanners is not restricted to a medical environment and its use has been incorporated into industrial facilities. One of the biggest drawbacks related to the use of CT scanners in these environments is the cost (in memory and time) associated. There are many approaches to simulate the functioning of CT scanners with less expensive devices such as cameras. Out of all this approaches, a Filtered Back-projection method is presented in this article. Based on the Radon transform it outperforms the classical Euclidean Geometry Approach by offering a fast and accurate solution to the task of reconstructing images via projection methods.

Keywords: Computed Tomography, Projection, Back-projection, Radon Transform

1 Introduction

The main motivation of the project was being able to simulate by cameras the way CT scans work in order to avoid the high cost associated to this kind of devices. CT is the acronym of Computed Tomography, which consists on using X ray to generate 2D images of objects (i.e, human body). Images are taken by rotating the equipment 360 round the human body. The amount of radiation emitted

is measured by a ring of detectors placed in the gate-shape structure that the patient is introduced in. The image is created from this measures so the internal structure of the human body can be reconstructed from X ray projections. In the industrial environment that we are working on, there are several manufacturers that are applying CT techniques with several aims such as detecting failures in pieces or observing certain parts of an object in order to extract useful information. In Figure 1 one example of an industrial CT scan can be observed. In this case we do not have a ring of detectors but a flat moving area which will receive the rays emitted from the X-Ray tube. The problem is that these kind of devices, in order to get a perfect reconstruction of the object needs from 360 to 3600 images (depending on the angular offset) which leads to slow and high-memory consuming processes. The objective of this project is, by applying projection and back-projection techniques, to reconstruct accurately an image in a fast and feasible way, trying to find out how many different views are needed in order to reconstruct the image without losing too much accuracy. Different approaches to the solution have been tried. At first an Euclidean Geometry based approach was implemented but its performance still needs to improve in order to be considered as a solution. Taking this into account, a Filtered Back-projection solution has been developed. This method is based on the Radon transform and

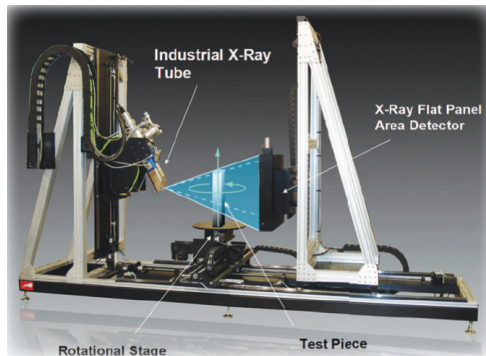


Figure 1: Industrial CT scanner

its relationship with the Fourier transform which will lead to faster processes.

Before presenting the method, in the next section we will do a fast review of the most recent articles related to our topic. Finally, before extracting the main conclusions of this article, in the results section we will analyze in depth the images obtained after doing several types of experiments in order to test the performance of the method.

2 Related Work

The majority of the bibliography can be enclosed in one of the two main research trends: Cone Beam Projection and Filtered Back-projection.

Related to Cone Beam Projection one interesting paper is 'Rectification for Cone Beam Projection and Back-projection' [1], where the authors try to find a technique that decreases computational time and increases the accuracy by providing a more efficient memory access. In July 2009 appeared the paper 'Directional View Interpolation for Compensation of Sparse Angular Sampling' [2], which gives a solution to increase the accuracy in the reconstruction by generating interpolated views. This approach is very interesting because it fits really well with our objective of reducing as much as possible the number of views without losing accuracy.

In the Filtered Back-projection field there are sev-

eral types of articles being the most frequent those related to improvements in the Radon transform or dedicated to explore the different types of interpolation or filters that could be used to improve the results. The 'Fourier-Based Forward and Back-Projectors in Iterative Fan-Beam Tomographic Image Reconstruction' [3] paper is focused on taking advantage of the relationship between the Radon and Fourier transforms with the objective of speeding the process along with some others techniques aimed to improve the accuracy results, such as blurring-avoidance. The effect of the distance between the source of the rays and the object is analysed in 'Direct Fan-Beam Reconstruction Algorithm via Filtered Back-projection for Differential Phase-Contrast Computed Tomography' [4] where the authors propose a method that avoids divergence effects. Focused on the filtering step, the article 'Filter design for Filtered Back-projection guided by the interpolation model' [5] introduces an approach that develops a filtering operator that has in mid the interpolation mode that has been applied to the sinogram, combining both operations into a single one. In this sense, but with a completely different objective, the article titled 'Combining Image Reconstruction And Image Analysis With An Application To Two-Dimensional Tomography' [6] combines the reconstruction of the image with the posterior analysis of it, developing application specific algorithms optimized for the purpose they are developed in mind.

As it can be seen there is not a main trend in research but several steps in the process chain than can be optimized, being the most relevant those dedicated to the improvement of the interpolation and filtering steps.

3 Method

In this section we present the basis (Radon transform) of our method along with details of its implementation.

3.1 Mathematical Foundations of the Method

Radon Transform

Bidimensional Radon Transform is an integral transformation of a function along a group of lines. As an example, if a line is represented by the equation $x\cos\theta + y\sin\theta = s$, where s is the minimum distance between the line and the origin and θ is the angle that the line makes with the x axis, the Radon transform is equal to:

$$R[f](\theta, s) = \int_{-\infty}^{+\infty} \delta(x\cos\theta + y\sin\theta - s) dx dy \quad (1)$$

In a n -dimensional space, the Radon transform is the integral of a function on hyperplanes. The integral of a function along a group of lines in the n -dimensional space is also known as X-ray transform. As it has been stated, Radon and Fourier transform are strongly related. Bi dimensional Fourier transform of $\mathbf{x} = (x, y)$ is:

$$\widehat{f}(\mathbf{w}) = \frac{1}{(2\pi)} \int f(\mathbf{x}) e^{-i\mathbf{x}\mathbf{w}} dx dy \quad (2)$$

We will use the next notation:

$$R_\theta[f](s) = R[f](s, \theta) \quad (3)$$

because we will do the Fourier transform of variable s . Projection-slice theorem is formulated as:

$$R_\theta[\widehat{f}](\sigma) = \sqrt{2\pi} \widehat{f}(\sigma \mathbf{n}(\theta)) \quad (4)$$

where $\mathbf{n}(\theta) = (\cos\theta, \sin\theta)$.

By using this property we have an explicit way to invert the Radon transform (and to study whether it is possible or not to invert it). But this method is way too complex and the operations involved can consume a lot of time.

Filtered Back-projection

A computational efficient algorithm in the bi-dimensional domain is the one that is used in this project, Filtered Back-projection. If we take the adjoint operator of R (Radon transform) the equation now takes this value:

$$R^*\widehat{g}(x) = \int_{\theta=0}^{2\pi} g(\theta, \mathbf{n}(\theta)x) d\theta \quad (5)$$

This operator is also known as 'backprojector' because it takes the projections along the different lines 'smearing' or projecting them back over the line to produce an image. This operator by no way is the inverse Radon transform.

Now we define the ramp filter h for one variable:

$$H[\widehat{h}](w) = [w] \widehat{h}(w) \quad (6)$$

If we apply now the Projection-slice theorem changing the integration variables, we can observe that for f , a two variables function, and $g = R[f]$

$$f = \frac{1}{4\pi} R^* H[g] \quad (7)$$

All this means that the original image f can be reconstructed from the sinogram g by applying a ramp filter (over s variable) and then backprojecting.

3.2 Implementation of the Filtered Back-projection algorithm

The solution implemented is based on the Filtered Back-projection method that has been presented and lets the user change dinamically the form of interpolation or the type of filter used, among other parameters which will be explained next:

Input Parameters

The parameters which values can be adjusted are the following ones:

- **Angles:** Slider that lets the user vary the number of angular positions (taking as starting point the 0) that act as rays' source points.
- **Steps:** It is a slider that lets the user adjust the value of a scalar in the range $[1, 20]$ that modifies the separation between the initial angles.
- **Interpolation:** Type of interpolation that will be used in the Back-projection process. The types that the methods admit are Nearest

Neighbour, Linear, Spline (where the interpolant is a special type of piecewise polynomial called spline, which tries to fit each division of the curve) or several types of cubic interpolation (from the simpler one to a shape-preserving pChip interpolation).

- **Filter:** The user can choose to use several types of filters (even none of them), from the most basic 'Ram-Lak' ramp filter to windowed versions of it (using sinc, cosine, Hamming or Hanning windows).

The frequency response of all the filters presented can be observed in Figure 2:

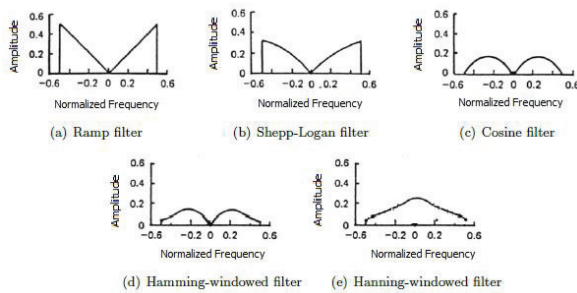


Figure 2: Frequency response of the filters

- **Frequency Compression:** It is a scalar in the range (0,1] that modifies the filter by rescaling its frequency axis.

4 Results

In this section results from several tests (each one of them studying the effect of a certain parameter in the reconstruction results) are shown. The original image to recover is shown in Figure 3

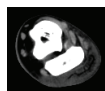


Figure 3: Original Image

Variation of the Angle Value

In Figure 4 it can be seen how the reconstruction is affected by the value of the maximum turn angle that is considered.

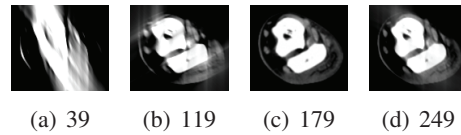


Figure 4: Results for different number of sources

As it can be observed, the reconstruction of the image is fair good for a number of initial angles equal or higher than 179. It is also true that if the number of angles is higher than 179 the system performance is not as good as with 179, because of having redundancy in the information (we are throwing projections from angles that are opposite to ones that we have used before).

Variation of the number of views

One of the objectives of the project was to find out if reducing the number of views has a drastic effect in the accuracy results. In Figure 5 we show how the reconstruction result is affected by the number of degrees of separation between angles (we have chosen 179 as the number of initial angles).

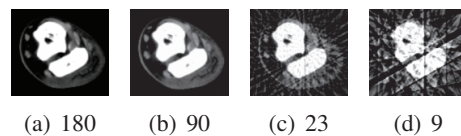


Figure 5: Results for different number of views

If we observe the figure above it is clear that there is difference in the reconstruction as we increase the separation between angles. Although we lose some precision in the reconstruction, if we take half the initial angles the loss of exact accuracy is less than 2 %, as it can be seen in Figure 6 and still the reconstruction can be considered as good.

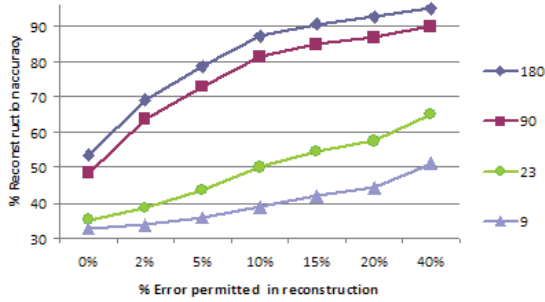


Figure 6: Reconstruction Accuracy Results

Variation of the Interpolation Form

In Figure 7 results obtained by varying the interpolation form (using an intermediate initial angles value (229)) are shown:

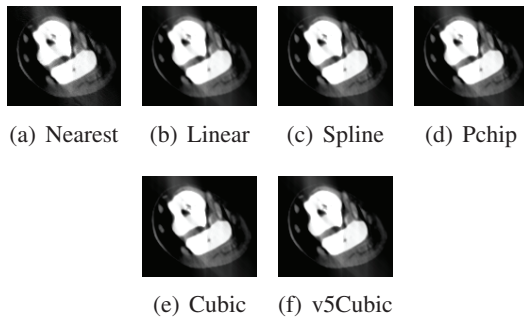


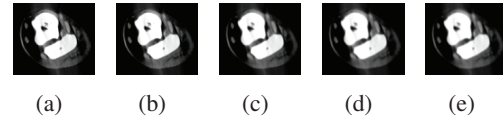
Figure 7: Results varying the form of interpolation

By observing the results it is difficult to see the difference in the reconstruction results whether we choose one form of interpolation or another. In terms of exact accuracy the form of interpolation that performs better is the spline interpolation (we get the higher accuracy percentage, 45.46) so it will be the option to choose if we want to achieve slightly better results.

Effect of the Variation of the Filter used

Results obtained by varying the type of filter applied (using a random number of initial angles (209) and spline interpolation) are shown in Fig-

ure 8:

Figure 8: Results obtained varying the type of filter: (a) *Ram-Lak* (b) *Sinc* (c) *Cosine* (d) *Hamming* (e) *Hanning*

The results vary a little between the different types of filters but, again, the variation is not high enough to discard any type of filter. In this case the better results are surprisingly obtained with the simplest Ram-Lak filter followed closely by the Shepp-Logan filter, which is the one that we will use in the tests because it decreases the effect of higher frequencies. It is important to mention that the tested images have no noise, so the effect of the filters cannot be seen in full.

Variation of the Frequency Compression

Results obtained by using several values of the frequency compression parameter can be observed in Figure 9

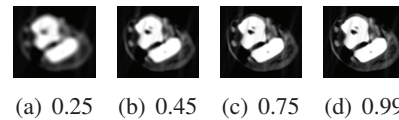


Figure 9: Results for several values of frequency compression

In this case there is indeed difference in the quality of the reconstruction related with the frequency compression chosen. The higher the value of the frequency compression is, the better the reconstruction because we are not eliminating necessary frequencial components.

Analysis of the Results

Taking a look at the results, it can be seen that the parameters that affect them most are the number of

source points and the distance between them (number of views). So, as a kind of summary, if we want to get the best reconstruction possible we would have to take 179 initial angles, spline interpolation, sinc windowed ramp filtered and no compression of the frequencial axis.

5 Conclusions and Future Work

In this paper it has been presented a Filtered Back-projection algorithm to reconstruct bidimensional Computed Tomography images with the aim of simulating the functioning of CT scans but in a more feasible way. The method is based on the Radon transform which performs a similar process to the Cone Beam projection methods but by taking advantage of the relationship between Radon and Fourier transforms (by the Projection-slice theorem of the Fourier transform) and the addition of a filter at the end of the process, improves the reconstruction results. The results obtained by the use of Filtered Back-projection are clearly better in terms of accuracy (more than 49 % of difference) and also helps to get rid of some drawbacks of the euclidean approach such as the excessive computing time and the use of memory. While it can be improved (maybe by the use of more complex filters or forms of interpolation) the results are fair good. Taking into account the environment we are working on, it has been proved that a slight reduction in the number of views can be perfectly considered because it does not lead to a great loss in accuracy and lets us save memory and time, two important constraints that needed to be overcome. The future work in this area could be adapting some of the most recent articles to our method (like the creation of artificial intermediate views) or combining the benefits of the Euclidean Approach (which gave better results than Unfiltered Back-projection) and the presented method in order to improve the results. The work that has been explained can be also thought as the first step in a whole analysis process, so another possible future line could be preparing

the results to a posterior analysis step.

Acknowledgements

This work has been done with the collaboration of the Universitat Autònoma de Barcelona.

References

- [1] Riddell, C., Troussel, Y, "Rectification for Cone-Beam Projection and Backprojection", *IEEE Transactions on Medical Imaging* 25(7):950-962, 2006.
- [2] Bertram, M. and Wiegert, J. and Schafer, D. and Aach, T. and Rose, G., "Directional View Interpolation for Compensation of Sparse Angular Sampling in Cone-Beam CT", *IEEE Transactions on Medical Imaging* 28(7):1011-1022, 2009.
- [3] Zhang O'Connor, Y. and Fessler, J.A., "Fourier-Based Forward and Back-Projectors in Iterative Fan-Beam Tomographic Image Reconstruction", *IEEE Transactions on Medical Imaging* 25(5):582-589, 2006.
- [4] Zhihua Qi, Guang-Hong Chen, "Direct Fan-Beam Reconstruction Algorithm via Filtered Backprojection for Differential Phase-Contrast Computed Tomography", *X-Ray Optics and Instrumentation*, 2008.
- [5] Horbelt, S. and Liebling, M. and Unser, M., "Filter Design for Filtered Back-Projection Guided by the Interpolation Model", *Proceedings of the SPIE International Symposium on Medical Imaging: Image Processing (MI'02)*, San Diego CA, USA, Volume 4684, Part II, 806-813, 2002.
- [6] Louis, A.K., "Combining Image Reconstruction And Image Analysis With An Application To Two-Dimensional Tomography", *SI-IMS*, 2(1):188-208, 2008.

Towards Detection of Measurable Contractions using WCE

Michał Drożdżał^{*#}, Petia Radeva^{*#}, Santi Seguí^{*#}, Fernando Vilariño^{*^}, Carolina Malagelada⁺,
Fernando Azpiroz⁺ and Jordi Vitrià^{*#}

** Computer Vision Centre, Bellaterra, Spain*

E-mail:michal@cvc.uab.es

Dept. de Matemàtica Aplicada i Anàlisi, Facultat de Matemàtiques, Universitat de Barcelona, Barcelona, Spain

^ Computer Science Department, Univ. Autònoma de Barcelona, Barcelona, Spain

+ Digestive System Research Unit, Hospital General Vall d'Hebron, Barcelona, Spain.

Abstract Features extraction of endoscopic images of small intestine provide various information about various intestinal motility malfunctions. Considering Wireless Capsule Endoscopy videos clinical experts concluded that a main feature to analyze intestinal motility is contained in manifested contractions. The ability of robust detection, extraction and measuring of contractions is an important step which will allow to understand better the intestinal activity. In this study, we propose a three-step procedure to recognize measurable contractions based on: (a) lumen size evaluation, (b) model definition of the measurable contraction and (c) search of predefined pattern of measurable contraction in the lumen size time series. The algorithm results suggest that our algorithm is able to extract feasible data for further clinical, diagnostic and therapeutic applications.

Keywords: Wireless Capsule Endoscopy, Small Intestine Motility, Contraction Analysis, Dynamic Time Warping.

1 Introduction

According to the statistics one out of three persons in the developed countries suffers from serious digestive problems. This motivates the necessity of studies in the area of digestive system. The small intestine is the longest part of the gut and until now it was not studied well because of the lack of diagnostic tools.

Wireless Capsule Endoscopy (WCE) is a very recent imaging technique that was introduced in [1] as a non-invasive examination method of Gastrointestinal (GI) tract. The capsule that is designed as up to 26mm device is able to acquire

video of the GI tract in particular: of the esophagus, the stomach, the small intestine and the large intestine. The advantage of WCE is that it is non-invasive and does not need patient hospitalization. To compare, the manometry that is the basic alternative for GI tract analysis consists of thin, pressure-sensitive tube that is passed through the patient mouth or nose and into its stomach or intestines and thus is highly invasive.

The WCE capsule (11 x 26 mm, 3,7 g [2]) consists of: illumination, camera, battery and radio transmitter. While passing through GI tract WCE is capturing images at rate 2 fps and sending them wirelessly to external hard disk which is placed on the body of the patient. The working time of capsule is approximately 7-8 hours that implicates about 57.000 images in total to analyse for each study. The clinical experience shows that a specialist needs at minimum from 4 to 8 hours of screening per video per patient in search of digestive system pathologies.

The need of profound study of the small intestine motility is motivated by various hypotheses that motility dysfunctions manifest various intestinal malfunctions. Weak and disorganized contractions are associated with bacterial overgrowth, intestinal obstructions or paralytic ileus, while dysfunctions in, or absence of, contractions over a long period can present functional dyspepsia [4].

The small intestine motility is mainly manifested by the appearance of contractions. A contraction is a movement of the intestine which is seen in the WCE video by an open-closed-open pattern of the lumen. Most of the time, during the occlusion when the intestinal wall is closed, a star-wise

pattern is present [5]. Generally, we can divide contractions into different groups depending on (1) duration and (2) level of occlusion.

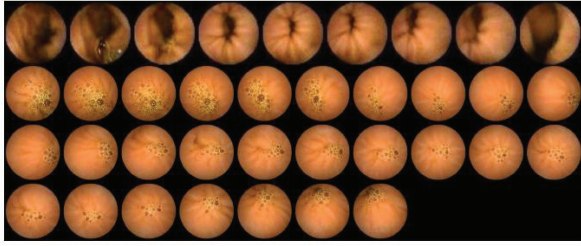


Figure 1: Examples of: phasic contraction (first row) and tonic contraction (2-4 rows).

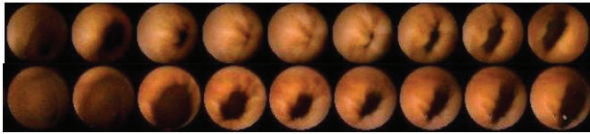


Figure 2: Examples of: occlusive contraction (first row) and non-occlusive contraction (second row).

Therefore, we can introduce the following types of contractions:

- Depending on the duration (Fig. 1):
 - *Short, phasic contractions*, that are characterized by sudden closing of the intestinal lumen, followed by a posterior opening (duration 4-5 s.);
 - *Long, tonic contractions* (sustained), that are characterized by closed intestinal lumen for an undefined period of time.
- Depending on the occlusion level (Fig. 2):
 - *Occlusive contractions*, where closing of the intestinal lumen is complete;
 - *Semi-occlusive contractions*, where closing of the intestinal lumen is incomplete.

Analysis of the WCE videos shows that contraction can last from 5 frames (phasic) to more than 25 frames (tonic). Hence, straightforward methods for contractions detection and characterization should be developed that do not depend on the duration of the contraction event.

On the other hand, there exists a hypothesis that intestinal motility is manifested also in parts where there are no contractions. In this case, the motility should be defined by detecting the motion between frames. That information will be used in

recovering of the capsule movement. However, these studies are out of the scope of this paper.

Taking into account that the capsule moves freely in the GI tract, and thus can partly observe a contraction, we introduce the term of measurable contractions as contractions where intestinal movement, in particular changes of lumen size, can be followed during the whole duration of the phenomenon. This will allow not only to detect and count the number of contractions during the patient video but also to perform measurements like: symmetry/asymmetry, level of occlusion, duration of contraction and duration of occlusion. The complexity of the problem of finding measurable contractions is not only hidden in variety of contractions but also in the free movement of capsule through intestine, as a result the camera is not always acquiring the centre of the lumen. According to [6] we can distinguish three types of scenarios of partially acquired contractions:

- *Incomplete contractions*: The central frame shows the intestinal lumen but it is not centered in the image. Part of the lumen lies out of plane.
- *Lateral contractions*: The first or the last part of the contraction sequence is missed, but the central frame is present.
- *Out-of-plane contractions*: The central frame of the contraction is completely out of plane. Nevertheless, the contraction event is deduced by the remaining part of the sequence.

In order to detect the measurable contractions, in this paper we propose a three step procedure based on the following steps: (a) lumen size evaluation, (b) model definition of the measurable contraction and (c) search of predefined pattern of measurable contraction in the lumen size time series. Note that up to our knowledge there is no work published in the bibliography that addresses the problem of extracting measurable contractions.

The paper is organized as follows: in *section 2* the problem of measurable contractions detection is defined, *section 3* provides a short discussion about pattern search in variable environment, *section 4* presents the three-step procedure description and in *section 5* the validation results

are presented. The paper finishes with discussions and conclusions.

2 Problem definition

In order to extract clinical measurements of contractions, it is necessary to find the beginning and the end of them. The parts of video which manifest contractile activity are calculated by a machine learning cascade system described in [7]. The contractions extracted by the cascade system are used as a point of reference in search of measurable contractions. In order to measure contractions, we need to pass through the following steps:

- Divide the results of the cascade system into two clusters: (1) measurable contractions, (2) non-measurable contractions;
- For the cluster of measurable contractions, mark: (i) the beginning, (ii) the end, and (iii) the centre of contraction.

Based on WCE video analysis, performed with gastroenterologists, the conclusion that the lumen (blob) size delivers majority of the information about contraction, was drawn. The centre of contraction was defined as a frame with maximum lumen occlusion, the beginning of contraction as the frame where the lumen starts decreasing. Analogously, the end of the contraction was defined as the image of the sequence where the lumen stops increasing. The goal of this work is to develop an algorithm to distinguish measurable contractions from the rest and present validation results of the applied methodology.

3 Finding patterns in variable environment

As we can see from the description presented in section 1, the intestines can be seen as variable environment. The contractions as an evidence of intestinal motility appear in a very diverse manner. The lumen size during the contraction can vary not only in level of occlusion but also in dynamics. Euclidean distance is not good enough to measure differences between contractions. The limitation of Euclidean distance stems from the fact that it is very sensitive to distortions in the time axis. This motivates the need for robust algorithms which would be resistant to contractile dynamics. Our proposition is to use a novel

technique to match sequences with different length and without the strict one-to-one frame correspondence, taking into account that according to the capsule movement, contractions can be acquired in different ways. To this purpose, we base our approach on the Dynamic Time Warping (DTW) method together with Uniform Scaling (US) to analyze and extract contractile sequences.

3.1 Dynamic Time Warping

DTW is a method frequently used as a basis of recognition algorithms. A profound approach description can be found in various papers, for example [8], [9] or [11]. DTW is an algorithm for measuring similarity between two sequences which may vary in time or speed in different ways. The cleverness of DTW lies in the computation of the distance between input stream and template. Rather than comparing the value of the input stream at time t to the template stream at time t , the algorithm is used to search optimal mappings results from the input stream to the template stream, so that the total distance between corresponding frames is minimized. In other words, DTW aligns the time axes allowing many-to-one and one-to-many frame matching before calculating the Euclidean distance (Fig. 3). Definition 1 formally defines the DTW distance [9].

Definition 1 (Time Warping Distance (DTW)). *Given two sequences $X = X_1, X_2, \dots, X_n$ and $Y = Y_1, Y_2, \dots, Y_m$, the time warping distance DTW is defined as follows:*

$$\begin{aligned} DTW(\phi, \phi) &= 0 \\ DTW(X, \phi) &= DTW(\phi, Y) = \infty \\ DTW(X, Y) &= D_{base}(First(X), First(Y)) + \dots \\ &\quad \min \left\{ \begin{array}{l} DTW(X, Rest(Y)) \\ DTW(Rest(X), Y) \\ DTW(Rest(X), Rest(Y)) \end{array} \right\} \end{aligned}$$

where, ϕ is the empty sequence, $First(X) = X_1$, $Rest(X) = X_2, X_3, \dots, X_n$, and D_{base} denotes the distance between two entries.

3.2 Uniform Scaling

Although DTW solves the problem of small local misalignments, it fails while working with real-live data [9] or [10]. The necessity of using Uniform Scaling is well understood in human

motion analysis, peoples movements not only vary locally, but “(people can) perform faster or slower than usual” [11]. The solution proposed in [9] or [10] is to use US together with DTW. US stretches or compresses vectors before calculating DTW distance, as a result the input and the template stream may be of equal length.

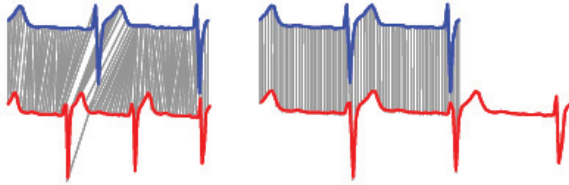


Figure 3: Aligning of two time series using DTW (left) and US (right) [10].

4 Methodology

We propose a three-step methodology (Fig. 4), based on the blob size information to detect the measurable contractions: (1) lumen size evaluation, (2) model definition of the measurable contraction in terms of the lumen size, (3) search

of predefined pattern of measurable contraction in the lumen size time series.

(Step 1). The goal of this step is to construct, from the WCE video, a function f describing the change of lumen size in every frame. The lumen size is defined as the sum of pixels in which the blob is visible.

(Step 2). In order to find measurable contraction model, an expert marked a set of “ideal” contractions (Fig. 5). The model m (Fig. 6) was obtained by calculating median value from all marked contractions (which were previously normalized to values from 0 to 1).

(Step 3). The goal of this step is to divide sequences of frames with contractile activity into two groups: (a) sequences which belong to the measurable contractions, (b) other sequences. In order to do so, a sequence (contractile and surrounding frames) is compared with the predefined model. The description of the pseudo-algorithm is presented in table 1.

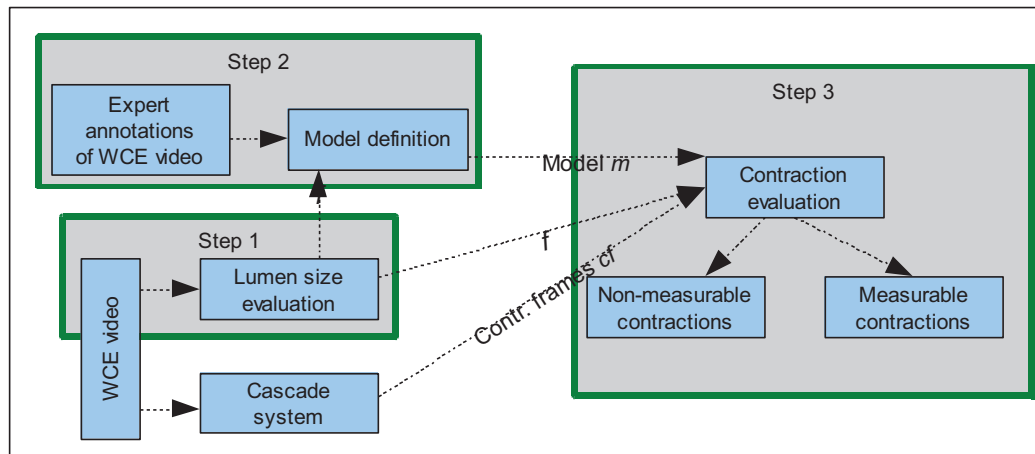


Figure 4: The three-step procedure draft.

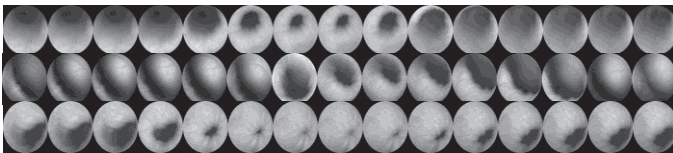


Figure 5: Example of expert annotations of measurable contractions (“ideal contractions”).

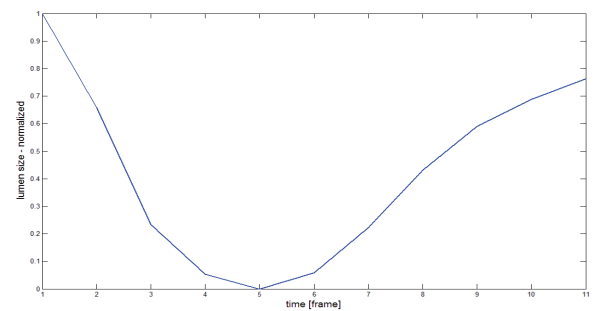


Figure 6: Obtained model of lumen size change of measurable contraction.

Table 1: The contraction evaluation algorithm.

```

Function find_measurable_contractions(f, m, cf)
  For every index I of contractile frame cf:
    Prepare sequences sj from function f, of length from 5 to 25 with a centre in I, (from
    [f(i-2):f(i+2)] to [f(i-12):f(i+12)]).
    For every sequence sj:
      Apply uniform scaling to sequence sj, and predefined model m to receive vectors
      of equal length sj' and m' (length(sj')>sj && length(m')>m).
      Calculate distance d=DTW(sj', m')/length(sj').
      Select the sequence sjbest with the best DTW fitting dbest.
    If (dbest<threshold) then sjbest is a measurable contraction.
end

```

An important tuning parameter is the *threshold* parameter. In our implementation, we defined distance *d* as a mean value of differences between points in the model and the compared sequence. This allows better understanding of the results of DTW fitting. For example, *d*=0.08 means that the mean difference between compared points is equal to 8%, so the sequence is deformed by 8% with reference to the model. Thus, changing the *threshold*, different levels of distortion can be accepted.

5 Validation

The experimental results were obtained using four capsule studies from four different healthy volunteers. The studies were conducted at the Digestive Diseases Dept. of the General Hospital de la "Vall D'Hebron", in Barcelona. For all four videos, the frames with contractile activity were calculated using cascade system described in [7]. The ground truth data (number of measurable contractions) was obtained by manual examination of the cascade system exit by an expert. Table 2 presents the data used in the experiments.

Table 2: The data for experiment.

Video number	Number of frames	Contractile frames	Measurable contractions
1	27321	2621	207
2	20623	2138	191
3	28431	6001	705
4	37656	7034	446

The testing procedure of the algorithm was implemented in Matlab. To evaluate the performance of the method the following information was calculated for every video:

- The number of correctly detected measurable contractions TP.
- The number of wrongly detected measurable contractions FP.
- The number of not detected measurable contractions FN.
- The number of correctly rejected non-measurable contractions TN.

In order to evaluate data the following criteria was calculated:

- Sensitivity = TP/(TP+FN).
- Specificity = TN/(TN+FP).
- False Alarm Rate = FP/(TP+FP).

The results were obtained using three different values of threshold. As it can be seen (table 3) the proposed algorithm performs quite well. Increasing the *threshold* parameter the loss in specificity is compensated by significant improvement of sensitivity. Analyzing all three criteria, the *threshold*=0.06 was chosen by the experts.

From the visual validation of the obtained results, we can draw the following conclusions:

- The wrongly detected measurable contractions (FP) appear because of: (1) the cascade system detected as contractile sequence frames with a presence of intestinal content (especially, frames with bubbles), (2) during the occlusion the centre of contraction lies out of plane, (3) the *threshold* parameter value is too large.
- The system does not detect measurable contractions (FN) when: (1) an error

during the lumen size evaluation has occurred, the blob size was wrongly calculated for one of the frames in which the contraction is visible, (2) value of the *threshold* parameter is too small.

6 Conclusions

This paper for the first time presented a method for recognition of measurable contractions which can be used for characterization of small intestine motility. The experimental results show that the

algorithm is able to detect more than 80% of measurable contractions, with 30% FAR. It is worth reminding that the algorithm outcome quality highly depends on performance of the cascade system used as a pre-filter of video frames. Our future work concerns extracting clinical information from the measurable contractions like speed and acceleration, duration of the contraction and other clinical measurements.

Table 3: The results of detection of measurable contractions with DTW and US.

Thresh.	Video	Positives	TP	FP	FN	TN	Sens	Spec	FAR
0,04	1	198	142	56	65	2358	0,69	0,98	0,28
	2	135	94	41	97	1906	0,49	0,98	0,3
	3	558	428	130	277	5166	0,61	0,98	0,23
	4	443	297	146	149	6442	0,67	0,98	0,33
	Average						0,62	0,98	0,29
0,06	1	265	185	80	22	2334	0,89	0,97	0,3
	2	184	150	34	41	1913	0,79	0,98	0,18
	3	780	538	242	167	5054	0,76	0,95	0,31
	4	603	370	233	76	6355	0,83	0,96	0,39
	Average						0,82	0,97	0,3
0,08	1	326	193	133	14	2281	0,93	0,94	0,41
	2	229	156	73	35	1874	0,82	0,96	0,32
	3	955	604	351	101	4945	0,86	0,93	0,37
	4	742	399	343	47	6245	0,89	0,95	0,46
	Average						0,88	0,95	0,39

Acknowledgments: This work is partially supported by a research grant from projects TIN2009-14404-C02, TIN2006-15308-C02, FIS-PI061290 and CONSOLIDER-INGENIO 2010 (CSD2007-00018), MI 1509/2005.

7 References

- [1] G. Iddan, G. Meron, et al., *Wireless capsule endoscope*, Nature 405 (2000) 417.
- [2] American Society for Gastrointestinal Endoscopy ASGE, *Technology status evaluation report wireless capsule endoscopy*, Gastrointest, Endoscopy 56 (5) (2002) 1866 – 1875.
- [3] H. Vu, T. Echigo, et al., *Detection of contractions in adaptive transit time of the small bowel from wireless capsule endoscopy videos*, Computers in Biology and Medicine 39 (2009) 16-26.
- [4] P. Spyridonos, F. Vilarino, et al., *Anisotropic feature extraction from endoluminal images for detection of intestinal contractions*, in: Proc. of the MICCAI, Lecture Notes in Computer Science, vol. 4225, Springer, Berlin 2006
- [5] F. Vilarino, *A machine learning approach for intestinal motility assessment*, Ph.D. thesis, Universidad Autonoma de Barcelona (June 2006).
- [6] F. Vilarino, P. Spyridonos, et al., *Intestinal Motility Assessment with Video Capsule Endoscopy: Automatic Annotation of Phasic Intestinal Contractions*, IEEE Trans. Medical Imaging (in press).
- [7] D. Sankoff, J. B. Kruskal: Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley, Reading, MA, 1983.
- [8] A. W. Fu, E. Keogh, et al., *Scaling and Time Warping in Time Series Querying*, Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005
- [9] E. Keogh, T. Palpanas, et al., *Indexing Large Human-Motion Databases*, Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004
- [10] N. Kosugi, Y. Sakurai, M. Morimoto, SoundCompass: A Practical Query-by-Humming System. Proceedings of the 2004 ACM SIGMOD Int. Conf. on Management of Data, 2004.
- [11] M. Waleed Kadous, A General Architecture for Supervised Classification of Multivariate Time Series. URL: <http://www.cse.unsw.edu.au/~waleed>.

Reactive object tracking with single uncalibrated PTZ camera

Murad al Haj*, Andrew D. Bagdanov* and Jordi González*

* *Computer Vision Center, UAB Campus, Bellaterra, Barcelona, Spain*

E-mail: malhaj@cvc.uab.es

Abstract

In this paper we describe a novel approach to reactive tracking of moving targets with an arbitrary pan-tilt-zoom camera. Our technique is based on estimating the position and velocity of the tracked object in the 3D environment and the pan, tilt and focal length of the active camera tracking it. The approach uses an extended Kalman filter to jointly track object position in the real world, its velocity in 3D, and the camera intrinsics and their rates of change. The filter outputs are used as inputs to a PID controller which continuously adjusts the camera motion in order to reactively track the object at a constant image velocity while simultaneously maintaining a constant object scale in the image plane. We provide experimental results on simulated and real tracking sequences to show how our tracker is able to accurately estimate both 3D object position and camera intrinsics with very high precision over a wide range of focal lengths.

Keywords: PTZ Camera Control, Active Tracking, Reactive Zoom.

1 Introduction

Many applications in the computer vision benefit from high resolution imagery, for example license-plate or face identification where a minimum size of the target is required [2]. For other applications,

such as identifying people in surveillance videos, having highly zoomed images is a must [9]. The problem with zoom control is that two opposing concepts are desirable: the first is obtaining a maximum resolution of the tracked target while the second is minimizing the risk of losing it. So, zoom control can be thought of as a trade-off between the effective resolution per target and the desired coverage of the area of surveillance. With a finite number of fixed sensors, there is a fundamental limit on the total area that can be observed, and maximizing both the area of coverage and the resolution of each observed target requires an increase in the number of cameras. However, such an increase is highly costly in terms of installation and processing, especially that cross-calibration and fusion of data would be necessary. Therefore, a system utilizing a single Pan-Tilt-Zoom (PTZ) camera can be much more efficient if it is properly designed to overcome the obvious drawback of having less information about the target.

Towards this end, different works have investigated the use of PTZ cameras to address this problem of *actively* surveying a large area in an attempt to obtain high-quality imagery while maintaining coverage of the region [4, 10]. However, this problem of how to control the PTZ camera in order to obtain the highest quality imagery of a moving target in the area of surveillance is still not solved. Most current approaches use multiple, calibrated cameras to provide the best possible depth information for active tracking.

Accurate reactive tracking of moving objects is a problem of both control and estimation. The speed at which the camera is adjusted must be a joint function of current camera position in pan, tilt and focal length, and the position of the tracked object in the 3D environment. In this paper we formulate the problem of jointly estimating the camera state and 3D object position as a Bayesian estimation problem. This formulation of the problem lends itself well to analysing the non-linearity introduced in the measurement model through the process of projection. We estimate the joint state with an extended Kalman filter. Our technique is designed to work on a single, uncalibrated active camera. No camera or scene calibration is required. In the next section, we review the relevant literature on active camera control. In section 3 we describe the joint model of camera and 3D world providing the stage for estimation, which is formulated in section 4. In section 5 we report on experiments conducted on both simulated data and live cameras. We conclude in section 6 with a summary and indications of future research directions.

2 Related work

Starting two decades ago, the area of active vision has been gaining much attention, mainly to improve the acquired visual data by keeping a certain object at a constant scale. Up to the current day, scale variations and its effect are still under investigation, an example of recent work can be found in [7]. Early on, Aloimonos et. al, [1], introduced the first general framework for active vision in order to improve the quality of perceptual results. Fayman et. al, [6], presented a method for zoom tracking showing two uses of it: recovery of depth information and improving the performance of scale-invariant algorithms. However, unlike us, they don't incorporate in their framework the uncertainty in the localization of the object and they assume that the zoom lens is calibrated.

Another work for active control is presented in

[5], where the authors model the motion of the tracked object using a Kalman filter then try to select the camera focal length that minimizes the uncertainty in state estimation with respect to the observation. However, the authors use a dual-camera, stereo set-up which significantly simplifies the problem. A newer approach is presented in Tordoff et al., [11], where the authors tune a constant velocity Kalman filter in order to insure proper tracking while the focal length is varying. Their approach is to correlate all the parameters of the filter with the focal length. However, they don't concentrate on the overall estimation problem, so their filter does not take into account any information from the real world making it irrelevant to any depth estimation problem and as stated in [6], the depth estimation is important for many applications. The focus of both of these works is primarily on zoom-control, and not on total object and camera position and control.

The most recent work we could find is that of Neslon et. al, [8], where an optical zoom system which relies on information from two cameras is introduced. They noted that in the event of a lost fixation, a single-camera system is unstable. Therefore, they chose to introduce a second camera with fixed focal length in order to solve this problem. Despite the fact that our system is a single-camera, we are able to avoid lost fixation by adjusting our controller based on the output of the extended Kalman filter, insuring a smooth transition without fixation problems.

3 The camera and world model

We use a pinhole camera model as shown in figure 1. The camera center is located at the origin of the world coordinate system, the principal point is at the origin of the plane of projection, and at zero pan and tilt, the axis of projection is aligned with the z -axis and the center of the object.

The object being tracked is assumed to be a rigid rectangular patch perpendicular to the axis of pro-

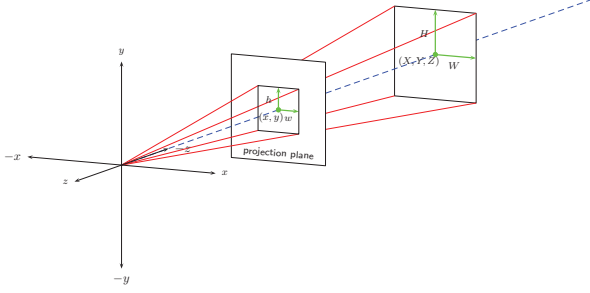


Figure 1: The pinhole camera model. The axis of projection at zero pan and tilt is aligned with the z -axis and the object being tracked is assumed to be a rigid rectangular patch perpendicular to the axis of projection.

jection regardless of camera orientation. It is located at world position $(X, Y, Z)^\top$ with known width W and height H . Changes in camera orientation due to panning and tilting are modeled as pure rotations of the coordinate system:

$$R(T, P) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos T & -\sin T \\ 0 & \sin T & \cos T \end{bmatrix} \begin{bmatrix} \cos P & 0 & -\sin P \\ 0 & 1 & 0 \\ \sin P & 0 & \cos P \end{bmatrix},$$

where P and T represent the pan and tilt angles, respectively. In our model, positive P corresponds to *clockwise* rotation of the camera about the x -axis and positive T *counterclockwise* rotation of the camera about the y -axis. This corresponds to the most common rotational orientation of commercial pan-tilt-zoom cameras.

We assume that the camera projection is reasonably approximated using equal scaling in the x and y directions in the plane of projection (i.e. square pixels). The center of projection is also assumed to be at the origin of the world coordinate system. The camera matrix is then parametrized by a single focal length parameter F :

$$K(F) = \begin{bmatrix} F & 0 & 0 \\ 0 & F & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The projection of the object at position $\mathbf{X} = (X, Y, Z)^\top$ onto the plane of projection can now

be written in terms of object position and the camera intrinsics by applying the rotation followed by scaling by focal length:

$$\begin{aligned} \mathbf{f}(P, T, F, \mathbf{X}) &= K(F)R(P, T)\mathbf{X} \\ &\equiv \begin{bmatrix} \frac{X'}{Z'} \\ \frac{Y'}{Z'} \end{bmatrix}, \end{aligned}$$

where X' , Y' and Z' are obtained from the initial, linear transformation:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = K(F)R(P, T)\mathbf{X}$$

This camera model relates the geometry and position of the tracked object in the 3D world to the internal camera parameters. In the next section we describe how the estimation problem can be formulated.

4 Estimation

In this section we formulate the problem of jointly estimating the camera and world parameters in a recursive Bayesian filter framework [3].

4.1 The estimation problem

At time k , the state configuration of the joint camera/object model is represented as by the spatial coordinates of the tracked object in the real world, the camera intrinsics and the velocities corresponding to the object position and camera intrinsics:

$$\mathbf{x}_k = [\mathbf{O}_k \mid \mathbf{C}_k \mid \dot{\mathbf{O}}_k \mid \dot{\mathbf{C}}_k]^\top, \quad (1)$$

where each component is defined as:

$$\begin{aligned} \mathbf{O}_k &= [X_k, Y_k, Z_k] \\ \mathbf{C}_k &= [P_k, T_k, F_k] \\ \dot{\mathbf{O}}_k &= [\dot{X}_k, \dot{Y}_k, \dot{Z}_k] \\ \dot{\mathbf{C}}_k &= [\dot{P}_k, \dot{T}_k, \dot{F}_k]. \end{aligned}$$

$[X_k, Y_k, Z_k]$ is the position of the planar patch in world coordinates at time k , $[P_k, T_k, F_k]$ represents the camera pan angle, tilt angle and focal length, respectively. The remaining elements $[\dot{X}_k, \dot{Y}_k, \dot{Z}_k, \dot{P}_k, \dot{T}_k, \dot{F}_k]$ represent the velocities of the previously mentioned components.

From time $k - 1$ to time k , the state is updated by the linear function U :

$$\mathbf{x}_k = U\mathbf{x}_{k-1} + \mathbf{v}_{k-1}, \quad (2)$$

where U is defined as:

$$U = \begin{bmatrix} \mathbf{I}_6 & \mathbf{I}_6 \\ \mathbf{0}_6 & \mathbf{I}_6 \end{bmatrix}$$

where \mathbf{I}_n and $\mathbf{0}_n$ are the $n \times n$ identity and zero matrices, respectively. The term \mathbf{v}_{k-1} in equation (2) is considered to be a zero-mean, Gaussian random variable adding noise to the system update.

At each time k , a measurement \mathbf{z}_k of the unknown system \mathbf{x}_k is made. Like the system state defined in equation (1), our measurement is composed of object and camera measurements:

$$\mathbf{z}_k = [\mathbf{o}_k \mid \mathbf{c}_k]^\top + [\mathbf{n}_k^o \mid \mathbf{n}_k^c]^\top, \quad (3)$$

where \mathbf{n}_k^o and \mathbf{n}_k^c are a zero-mean Gaussian processes on the object and camera measurements, respectively, and \mathbf{o}_k and \mathbf{c}_k are defined as:

$$\begin{aligned} \mathbf{o}_k &= [\mathbf{f}(P_k, T_k, F_k, \mathbf{O}_k) \mid (K(F_k)[W, H, 0]^\top)^\top \\ &\quad - \mathbf{f}(P_k, T_k, F_k, \mathbf{X}_k)^\top]^\top \\ \mathbf{c}_k &= [p, t, f]. \end{aligned}$$

The object measurement \mathbf{o}_k consists of the projection of the object position \mathbf{O}_k and the known object size $W \times H$ into the coordinate system of the plane of projection. Camera measurements arrive in the form of the camera's imprecise internal measurements p, t, f of the pan angle, tilt angle and focal length.

Given the system update and measurement processes defined in equations (2) and (3), the Bayesian estimation problem is to find an estimate of the unknown state \mathbf{x}_k that maximizes the likelihood $p(\mathbf{x}_k | \mathbf{z}_{1:k})$.

4.2 Estimation by extended Kalman filter

The estimation of the likelihood is complicated by the non-linearity introduced in the measurement process through the projection \mathbf{o}_k . Because of this, Gaussian uncertainty in the state space is non-Gaussian in the projective measurement space. The standard Kalman filter cannot be directly applied, as it assumes a linear measurement model [12].

The extended Kalman filter (EKF) uses a local linearization of the measurement function to approximate the measurement process as a linear function of the system state \mathbf{x}_k . Defining \hat{H} as the Jacobian of the measurement function:

$$\hat{H} = \left. \frac{\partial \mathbf{h}(\mathbf{x}_k)}{\partial \mathbf{x}_k} \right|_{\mathbf{x}_k = m_{k-1}},$$

the EKF recursively approximates the likelihood as the Gaussian:

$$p(\mathbf{x}_k | \mathbf{z}_{1:k}) \approx \mathcal{N}(\mathbf{x}_k; m_k, P_k),$$

where

$$\begin{aligned} m_k &= Um_{k-1} + K_k(\mathbf{z}_k - \mathbf{h}(m_{k-1})) \\ P_{k|k-1} &= Q + UP_{k-1}U^\top \\ P_k &= P_{k|k-1} - K_k\hat{H}P_{k|k-1} \\ S_k &= \hat{H}P_{k|k-1}\hat{H}^\top + R \\ K_k &= P_{k|k-1}\hat{H}^\top S_k^{-1}. \end{aligned}$$

Q and R represent the (assumed known) Gaussian covariance processes in system update and measurement, respectively.

The extended Kalman filter recursively estimates the parameters of a Gaussian approximation of the uncertainty in prediction of the true state \mathbf{x}_k .

5 Experimental analysis

In this section, the robustness of our method is demonstrated. Experiments were done using both

simulated scenarios and live scenes of a PTZ camera. The simulated scenario consisted of a random object motion and the error was averaged over many runs. The camera used in the live scenes was an Axis 214 PTZ network camera. A PID controller was later used in order to move the camera based on the output of the extended Kalman filter, with a proportional gain (K_p) set to 1 and a derivative gain (K_d) set to 0.2 in order to eliminate overshoot. The integral gain (K_i) was found unnecessary since the output of the filter is accurate enough not to have an error at the steady state, i.e. the state where the object is centered with maximum zoom. In both the simulated scenarios and the live scenes, our method shows an excellent performance.

5.1 Results on simulated data

To validate our model, we performed a number of simulations. The error metric we used in all model parameters estimation is:

$$\text{RMSD}(\phi_i) = \sqrt{E((\bar{\phi}_i - \phi_i)^2)},$$

where ϕ_i is one of the model parameters defined in equation 2, $\bar{\phi}_i$ is the estimated model parameter and the expectation is taken over the entire sequence. The RMSD is measured for several runs of the simulation (we used 100 runs in our experiments), and the average RMSD is used as a measure of estimation performance.

Figure 2 shows a box-and-whisker summary of the RMSD for a simulation where a moving object is tracked by a moving camera. In these experiments we simulate the motion the camera would execute due to corrections coming from the PID controller described above. Camera motion is choreographed through dynamic adjustment of PID gains so that the camera intercepts the target right as it is about to randomly change direction. In these simulations, some noise is introduced in the different state parameters. To investigate sensitivity to varying measurement noise, this value is scaled by a constant $a \in \{1, 5, 10\}$. Similar results

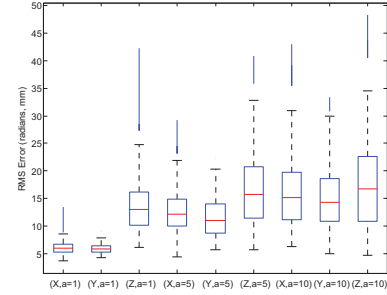


Figure 2: Error in 3D position parameters (X, Y, Z) estimated for a camera actively tracking a randomly moving object. All measurements are in millimeters.

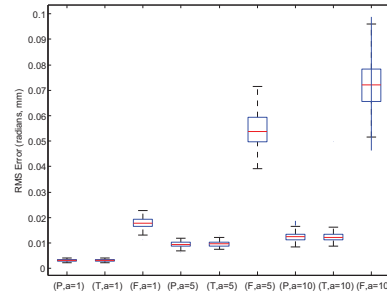


Figure 3: Error in PTF. Angular measurements are in radians, focal length in millimeters.

can be seen in figure 3 for camera parameters estimation. From these figures, one can conclude that the estimation is robust and the outliers are more frequent with increased measurement noise.

5.2 Results on live cameras

We also tested our method by actively tracking two objects: a dark blue cup and a face. We used simple detectors since the detection problem is not the aim of this paper.

The results on tracking a moving cup are shown shown figure 4. The two red dots shown on the cup represent the center of the cup and the upper left corner, both are outputs of the detection process. The green circles represent the projection of the estimates of the center and the bounding box position. Similar results for tracking a moving face



Figure 4: Reactive tracking of a moving object. This figure is best viewed in color.

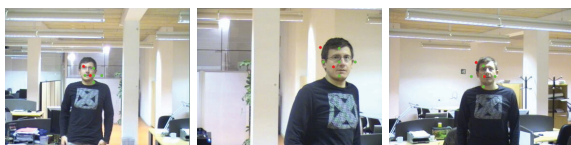


Figure 5: Reactive face tracking. This figure is best viewed in color.

are shown in figure 5. The tracker was able to successfully follow the face making correct decisions when to zoom in and when to zoom out.

6 Conclusions and future work

In this paper we described an approach to active camera tracking, using a single camera without calibration, that jointly estimates the orientation and focal length of a pan-tilt-zoom camera and the position of the tracked object relative to the camera center in the 3D environment. Experiments show that object position estimates are robust in the presence of camera motion and increased measurement noise. We are currently investigating alternate methods of measuring uncertainty projected from Gaussian processes into the image plane.

Acknowledgements

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDIVideo project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 MIPRCV CSD2007-00018. The first author also thanks the Generalitat de Catalunya for financial support through an FI scholarship.

References

- [1] J. Aloimonos, I. Weiss, and Amit Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988.
- [2] C. K. Anagnostopoulos, I. E. Anagnostopoulos, I. D. Psoroulas, , and E. Kayafas. License plate recognition from still images and video sequences: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 9(3):377–391, September 2008.
- [3] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *IEEE Transactions of Signal Processing*, 50(2):174–188, February 2002.
- [4] Andrew D. Bagdanov, Alberto del Bimbo, Walter Nunziati, and Federico Pernici. A reinforcement learning approach to active camera foveation. In *VSSN '06: Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, pages 179–186, New York, NY, USA, 2006. ACM.
- [5] J. Denzler, M. Zobel, and H. Niemann. Information theoretic focal length selection for real-time active 3-d object tracking. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, pages 400–407. IEEE Computer Society Press, 2003.
- [6] J.A. Fayman, O. Sudarsky, E. Rivlin, and M. Rudzsky. Zoom tracking and its applications. *Machine Vision and Applications*, 13(1):25–37, August 2001.
- [7] A. Madabhushi, J. Udupa, and A. Souza. Generalized scale: Theory, algorithms, and application to image inhomogeneity correction. *Computer Vision and Image Understanding*, 101:100–121, 2006.
- [8] E.D. Nelson and J.C. Cockburn. Dual camera zoom control: A study of zoom tracking stability. In *in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE Computer Society Press, April 2007.
- [9] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. Frvt 2006 and ice 2006 large-scale results. Technical Report NISTIR 7408, National Institute of Standards and Technology, 2007.
- [10] E. Sommerlade and I. Reid. Information-theoretic active scene exploration. In *Proceedings of: Computer Vision and Pattern Recognition, 2008. CVPR 2008*, June 2008.
- [11] B. J. Tordoff and D. W. Murray. A method of reactive zoom control from uncertainty in tracking. *Computer Vision and Image Understanding*, 105(2):131–144, 2007.
- [12] Greg Welch and Gary Bishop. An introduction to the kalman filter. Technical report, (TR 95-041), University of North Carolina at Chapel Hill, 2004.

Robust Background Subtraction Approach based on Chromaticity and Intensity Patterns

Ariel Amato *, Mikhail Mozerov † and Jordi Gonzàlez ‡

* *Universitat Autònoma de Barcelona, Centre de Visió per Computador, Bellaterra, Spain*

E-mail: aamato@cvc.uab.es

† *Universitat Autònoma de Barcelona, Centre de Visió per Computador, Bellaterra, Spain*

E-mail: mozerov@cvc.uab.es

‡ *Universitat Autònoma de Barcelona, Centre de Visió per Computador, Bellaterra, Spain*

E-mail: poal@cvc.uab.es

Abstract

An efficient Real-Time method for detecting moving objects using a background subtraction technique is presented. A sophisticated background model that combines spatial and temporal information based on similarity measure in angles and intensity between two color vectors is introduced. The comparison is done in RGB color space. A new feature based on chromaticity and intensity pattern is extracted in order to improve the accuracy in the ambiguity region where there is a strong similarity between background and foreground and to cope with cast shadows. The effectiveness of the proposed method is demonstrated in the experimental results and comparison with others approaches is also shown.

Keywords: Background Subtraction, Motion Detection, Shadows Removal, Color and Texture.

1 Introduction

Moving object detection plays an important role in the field of Computer Vision [1], [2], and is implemented in different applications such as tracking, video compression, video surveillance etc. One

of the most common and effective approach to localize moving objects is the background subtraction, i.e. the differencing between the current image frame and the background model. This technique has been actively investigated and applied for many researchers during the last years [3] [4] [5] [6] [7]. To obtain a precise segmentation, is necessary to overcome several difficulties like illuminations changes, moving shadows, camouflages etc. Consequently, a robust and accurate algorithm to segment moving object in different scenarios (indoor, outdoor) has to be developed. In this paper, we present several new modifications of similarity measurements that are used in background subtraction methods. These modifications concern with the chromaticity similarity and the neighborhood similarity. Instead of measuring the chromaticity similarity in the non-linear HSV color space we propose a simple measurement based on an angle between two color vectors. This modification helps us to solve the shadow suppression problem. The information provided by a single pixel, sometimes is not enough to distinguish between background and foreground due to the strong similarity that can exist between regions. In this case, we propose to measure similarity not between two color vectors, but between two

sets of color vectors that form a neighborhood patterns. Such modification helps us to improve the ability of our background subtraction method to overcome the camouflage problem. This paper is organized as follows. Section 2 introduces a brief of more classical Real-Time Background Subtraction approaches. Section 3 presents our method. In Section 4 experimental results are discussed. Concluding remarks are available in Section 5.

2 Related work

Haritaoglu et al. in W4 [6] use a model of background subtraction built from order statistics of background values during a training period. The background scene is modeled by representing each pixel by three values: its minimum and maximum intensity values and the maximum intensity difference between consecutive frames observed during this training period. Pixels are classified as foreground if the difference between the current value and the minimum and maximum values are greater than the values of the maximal interframe difference. However, this approach is rather sensitive to shadows and lighting changes, since the only illumination intensity cue is used. Horprasert et al. [7] implement a statistical color background algorithm, which use color chrominance and brightness distortion. The background model is built using four values: the mean, the standard deviation, the variation of the brightness and chrominance distortion. However, the chromaticity noise model in this paper is not correct. Indeed, according to the logic of the thresholds definition the chromaticity noise does not depend on the change of illumination, but such an assumption is wrong in general. Kyungnam Kim et. al [5] use a similar approach, but they obtain more robust motion segmentation in the presents of the illumination and scene changes using background model with codebooks. The codebooks idea gives the possibility to learn more about the model in the training period The authors propose to cope with the unstable information of

the dark pixels, but still they have some problems in the low and the high intensity regions. Stauffer and Grimson [8] address the low and the high intensity regions problem by using a mixture of Gaussians to build a background color model for every pixel. Pixels from the current frame are checked against the background model by comparing them with every Gaussian in the model until a matching Gaussian is found. If so, the mean and variance of the matched Gaussian is updated, otherwise a new Gaussian with the mean equal to the current pixel color and some initial variance is introduced into the mixture. Cucchiara et al. [3] use a model in Hue-Saturation-Value (HSV) and stress their approach in shadow suppression. The idea is that shadows change the hue component slightly and decrease the saturation component significantly. In the HSV color space a more realistic noise model can be done. However, this approach also has drawbacks. The similarity measured in the non-linear HSV color space, as a result the noise value seems to depend on the chromaticity value of the background color and such an assumption is not generally correct. Also, the color space transformation from RGB to HSV increases complexity of the approach and decreases the calculation error.

3 Proposed Algorithm

3.1 Similarity Measurements

To compare a background image with a current frame we use four similarity measurements. Those are:

- **Angular similarity measurement $\Delta\theta$** between two color vectors in the RGB color space \mathbf{p}_1 and \mathbf{p}_2 which is defined as follows.

$$\Delta\theta(\mathbf{p}_1, \mathbf{p}_2) = \cos^{-1} \left(\frac{\mathbf{p}_1 \mathbf{p}_2}{|\mathbf{p}_1| |\mathbf{p}_2|} \right) \quad (1)$$

- **Intensity similarity measurement ΔI** between two color vectors in the RGB color space \mathbf{p}_1 and \mathbf{p}_2 .

$$\Delta I(\mathbf{p}_1, \mathbf{p}_2) = |\mathbf{p}_1 - \mathbf{p}_2| \quad (2)$$

With each of the described similarity measurements we associate a threshold function:

$$\begin{aligned} T\theta(\Delta\theta) &= \begin{cases} 1 & \text{if } \Delta\theta > T^\theta \\ 0 & \text{else} \end{cases}, \\ TI(\Delta I) &= \begin{cases} 1 & \text{if } |\Delta I| > T^I \\ 0 & \text{else} \end{cases} \end{aligned} \quad (3)$$

where T^θ and T^I are intrinsic parameters of the threshold functions of the similarity measurements. To describe a neighbourhood similarity measurement let us first characterize the index vector $\mathbf{x} = (\mathbf{n}, \mathbf{m})^t \in \Omega = \{0, 1, n, N; 0, 1, m, M\}$, which define the position of a pixel in the image. Also we need to name the neighbourhood radius vector $\mathbf{w} = (\mathbf{i}, \mathbf{j})^t \in \mathbf{W} = \{-W, 0, 1, i, W; -W, 0, 1, j, W\}$, which define the positions of pixels that belong to the neighbourhood relative any current pixel. Indeed, the domain \mathbf{W} is just a square window around a chosen pixel.

- **Angular neighborhood similarity measurement** $\eta\theta$ between two sets of color vectors in the RGB color space $\mathbf{p}_{1+\mathbf{w}}$ and $\mathbf{p}_{2+\mathbf{w}}$ ($\mathbf{w} \in \mathbf{W}$) can be written as

$$\eta\theta(\vartheta) = \sum_{\mathbf{w} \in \mathbf{W}} T\theta(\Delta\theta(\vartheta)) \quad (4)$$

where the functions $T\theta$ and $\Delta\theta$ are defined in Eq. (3) and Eq. (1) respectively and ϑ is $(\mathbf{p}_{\mathbf{x}+\mathbf{w}}^1, \mathbf{p}_{\mathbf{x}+\mathbf{w}}^2)$.

- **Intensity neighborhood similarity measurement** $\mu\theta$ between two sets of color vectors in the RGB color space $\mathbf{p}_{1+\mathbf{w}}$ and $\mathbf{p}_{2+\mathbf{w}}$ ($\mathbf{w} \in \mathbf{W}$) can be written as

$$\mu\theta(\vartheta) = \sum_{\mathbf{w} \in \mathbf{W}} TI(\Delta I(\vartheta)) \quad (5)$$

where the functions TI and ΔI are defined in Eq. (3) and Eq. (2) respectively. With each of the neighbourhood similarity measurements we associate a threshold function

$$\begin{aligned} T\eta\theta(\eta\theta) &= \begin{cases} 1 & \text{if } \Delta\theta > T^{\eta\theta} \\ 0 & \text{else} \end{cases}, \\ T\mu I(\mu I) &= \begin{cases} 1 & \text{if } \Delta I > T^{\mu I} \\ 0 & \text{else} \end{cases} \end{aligned} \quad (6)$$

where $T^{\eta\theta}$ and $T^{\mu I}$ are intrinsic parameters of the threshold functions of the neighborhood similarity measurements.

3.2 Background Modelings

Our background model (BG) will be represented with two classes of components one we call running components (RC) and another we call training components (TC). The RC is a color vector in RGB space and only this component can be updated in running process. The TC is a set of fixed thresholds values obtained during the training. The background model is represented by

$$BG_{\mathbf{x}} = \left\{ \{\mathbf{p}_{\mathbf{x}}\}, \{T_{\mathbf{x}}^\theta, T_{\mathbf{x}}^I, T_{\mathbf{x}}^{\eta\theta}, T_{\mathbf{x}}^{\mu I}, W\} \right\} \quad (7)$$

where $\mathbf{T}_{\mathbf{x}}^\theta$ is maxima of the chromaticity variation (temporal-base); $\mathbf{T}_{\mathbf{x}}^I$ is maxima of the intensity variation (temporal-base); $\mathbf{T}_{\mathbf{x}}^{\eta\theta}$ is the chromaticity pattern threshold (spatial-base); $\mathbf{T}_{\mathbf{x}}^{\mu I}$ is the intensity pattern threshold (spatial-base); W is the half size of the neighbourhood window. To obtain the background parameters in the definition of Eq. (7) the training process has to be performed. This first step consists of estimating the value of the RC and TC during the training period. To initialize our BG we put the $RC = \{p_x^0\}$ as the initial frame and

these values are supported during all training process. To estimate T_x^θ and T_x^I during the training period, we compute the intensity and chromaticity difference between a background image pixel and the related pixel in the current frame belonging to the training process

$$\begin{aligned} T_x^\theta &= \max_{f \in \{1,2,..F\}} \left\{ \Delta\theta \left(\mathbf{p}_x^0, \mathbf{p}_x^f \right) \right\}, \\ T_x^I &= \max_{f \in \{1,2,..F\}} \left\{ \Delta I \left(\mathbf{p}_x^0, \mathbf{p}_x^f \right) \right\}, \end{aligned} \quad (8)$$

where F is the number of frames in the training period. In this paper we consider the simplified version of our algorithm. In this case the spatial-base thresholds $T_x^{\eta\theta}$ and $T_x^{\mu I}$ we put as 1 for the neighborhood radius equaling 1 (the 3x3 square windows), and for all pixels in the frame. However, locally adaptive version of our approach allows estimating this parameter in each pixels like the temporal-based thresholds. After the initialization has been done the following equations show how to obtain the TC values. These values are estimated in the Running and Foreground Classification Process. Our classification rules are enunciated in two steps:

Step One: This step is concentrate on the pixels that have strong chromaticity and intensity differences with the background, it means that the following rule expression have to be TRUE or 1

$$\begin{aligned} T\theta \left(\Delta\theta \left(\mathbf{p}_x^{bg}, \mathbf{p}_x^f \right) | \gamma T^\theta \right) \cap \\ T I \left(\Delta I \left(\mathbf{p}_x^{bg}, \mathbf{p}_x^f \right) | \gamma T^I \right) = 1; \end{aligned} \quad (9)$$

where γ is an experimental scale factor for the training set thresholds and the value of this factor greater than 1. In this case the tested pixel is directly classified like foreground. Otherwise the classification will be done in the next step.

Step Two: Due to the angular similarity measurement is inaccurate for low intensity color vectors, we define a confidence threshold T^{cf} , where if the vectors are greater than T^{cf} the classification will be done according to the rule of Eq. (10)



(a)



(b)

Figure 1: (a) Original Image, (b) Foreground Detection (classification done based on angular patterns in red, classification done based on intensity patterns in yellows)

otherwise according to the rule in Eq. (11). An example of this segmentation process is shown in Fig. 1.

$$\begin{aligned} T\theta \left(\Delta\theta \left(\mathbf{p}_x^{bg}, \mathbf{p}_x^f \right) \right) \cup \\ T\eta\theta \left(\eta\theta \left(\mathbf{p}_{x+w}^{bg}, \mathbf{p}_{x+w}^f \right) \right) = 1 \end{aligned} \quad (10)$$

$$\begin{aligned} T I \left(\Delta I \left(\mathbf{p}_x^{bg}, \mathbf{p}_x^f \right) \right) \cap \left(\Delta I \left(\mathbf{p}_x^{bg}, \mathbf{p}_x^f \right) > 0 \right) \\ \cup T\mu I \left(\mu I \left(\mathbf{p}_{x+w}^{bg}, \mathbf{p}_{x+w}^f \right) \right) = 1 \end{aligned} \quad (11)$$

3.3 Updating Models

The update of the background follows the next rule: put the current pixel value to the BG if this pixel is not classified as foreground.

4 Experimental Results

In this section we present the performance of our approach in term of quantitative and qualitative results.

• Quantitative Results

We applied the proposed algorithm in several videos scenes, under different conditions (indoor/outdoor). In order to evaluate the performance of the proposed technique we generated a ground-truth segmentation masks by manual segmentation. Due to the numbers of frames tested, from different sequences was considerably high, we choose 20 frames per scene which these were sampled in such way to maintain the same inter-frame distance from the beginning to the end of the sequence and four different sequence has been evaluated. Using the same sequences we implements algorithms from others authors [5] [6] [7] [8] to obtain a general comparison of our method. Two metrics were utilized to evaluate segmentation process, False Positive Error (FPE) and False Negative Error (FNE). Eq (12). The FPE means that the background pixel was set as a Foreground and FNE the foreground pixels that were set as a Background. In Fig. 2 a comparison, in term of accuracy is shown.

$$Error(\%) = \frac{N. \text{ of misclassification pixels}}{N. \text{ of correct foreground pixels}} 100 \quad (12)$$

• Qualitative Results

Fig. 3 shows the comparison between our techniques and some well known methods. It can be seen that our method performs better in terms of segmenting camouflage areas and suppressing strong shadows.

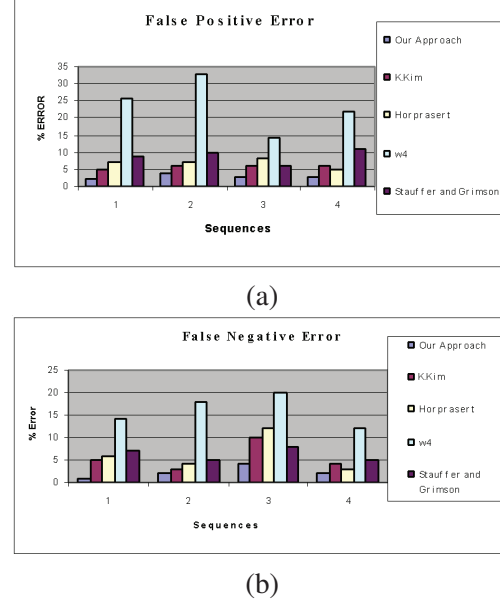


Figure 2: Segmentation errors. (a) FPE and (b) FNE

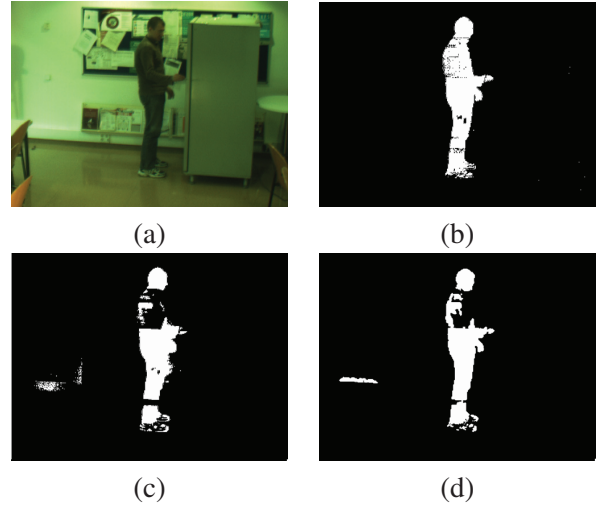


Figure 3: (a) Original Image, (b) Our method, (c) Stauffer method and (d) K.Kim method

5 Conclusions

We proposed a Background Subtraction algorithm whose effectiveness was demonstrated in the comparison with other methods. The background model combines spatial and temporal information based on similarity measure in angle and intensity. Also, a feature based on chromaticity and intensity pattern is extracted in order to resolve the ambiguity that exists between similar regions and to cope with moving shadows.

Acknowledgments

This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 MIPRCV CSD2007-00018.

References

- [1] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice Hall, August 2002.
- [2] G. Obinata and A. Dutta, *Vision Systems: Segmentation and Pattern Recognition*. I-Tech Education and Publishing, 2007.
- [3] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti, "Improving shadow suppression in moving object detection with hsv color information," in *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, 2001, pp. 334–339.
- [4] M. K. Lutz, L. Goldmann, D. Yu, and T. Sikora, "Comparison of static background segmentation methods," *Visual Communications and Image Processing*, vol. 5960, pp. 2140–2151.
- [5] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. S. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.
- [6] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 809–830.
- [7] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *ICCV Frame-Rate WS*. IEEE, 1999.
- [8] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, 2000.

Interest Point based Human Action Recognition

Bhaskar Chakraborty and Andrew D. Bagdanov and Jordi González

Computer Vision Center, UAB Campus, Bellaterra, Barcelona, Spain

E-mail: bhaskar@cvc.uab.es

Abstract

The combination of spatial interest points, and local space-time features on those points, capture local events in video and can be adapted to the size, the frequency and the velocity of moving patterns. In this paper we demonstrate how such features from the detected spatial interest points in each frame can be used for recognizing human actions. Vocabularies of quantized spatio-temporal descriptors are used to construct vocabularies for use in a Bag of Words (BoW) representation. Histograms of spatio-temporal words are then used to classify human actions using an ensemble of SVM classifiers. We provide experimental results on a standard action recognition database, and the preliminary analysis of these results is encouraging. In particular, we show how the potential of a vocabulary for classifying results can be computed independently of the SVM classifier.

Keywords: action recognition, spatio-temporal interest points, bag of words.

1 Introduction

The recognition of the human actions in videos is considered an important problem in the field of computer vision. This is due, in part, to the large number of potential applications of action recognition in areas such as surveillance, video retrieval, human-computer interaction and sports video analysis. Action classification is greatly

complicated by the variation in recording conditions under which action sequences are recorded. Scenes with cluttered, moving backgrounds, non-stationary cameras, scale variations, variations in appearance and clothes of people, changes in light and view point, etc., are all common variations that must be handled by robust action recognition techniques. These problems have been addressed in various ways in the literature on human action recognition [5, 9, 13].

Several methods for learning and recognizing human actions directly from image measurements have been proposed [3, 2, 7, 1]. Image measurements such as optical flow or spatio-temporal gradients depend on the recording conditions such as position of the pattern in the frame, spatial resolution and relative motion with respect to the camera. Moreover, global image measurements can be influenced by motions of multiple objects and variations in the background. These kind of dependencies are always considered as a handicap for real applications of action recognition. Whereas these problems can be solved in principle by external mechanisms for spatial segmentation and/or camera stabilization, such techniques might be unstable in complex situations. This motivates the need for alternative action recognition representations that are stable with respect to changes in recording conditions.

The work of Laptev et. al [11, 8] demonstrates that action recognition can be achieved using local measurement such as spatio-temporal interest points. Such features capture local motion events in video and can be adapted to the size, fre-

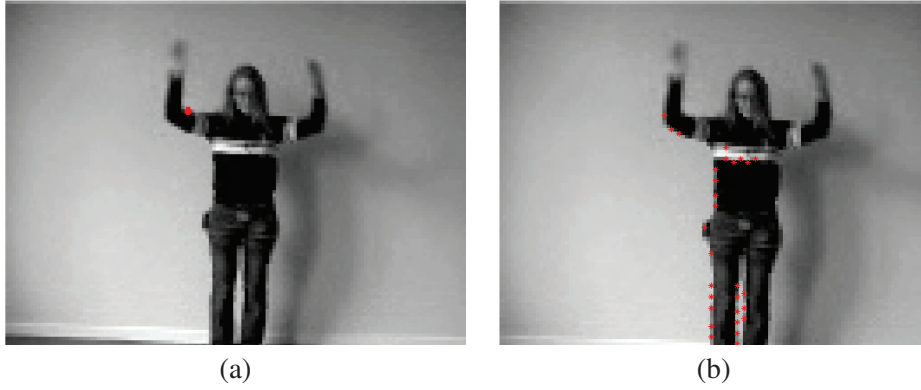


Figure 1: Interest point detection. (a) Spatio-temporal interest point and (b) spatial interest point in a waving action. This illustrates how spatio-temporal interest points can fail to capture both static and dynamic component of the action. These figures are best viewed in color.

quency and velocity of moving patterns. This results in a stable video representation with respect to different recording conditions. Apart from local feature based action recognition approach, there is another influential model, the Bag of Words (BoW), that has recently been used by many researchers [12, 10] for action representation and recognition. This model represents each human action as collection of codewords in a predefined vocabulary generated from the training data.

However, features from spatio-temporal interest points can fail to capture the *static* component of action since they are defined as points that are both spatially *and* temporally salient. Human action is often best characterized by combination of the static and dynamic state of different body parts, and thus it is important to consider the spatio-temporal features of both the moving and non-moving parts (see figure 1). In order to extract such features in this work we instead first apply Harris interest point detection [6] in each frame and then represent each feature point using a the spatio-temporal descriptor. The Harris detector captures interesting points, regardless of their temporal characteristics, while the spatio-temporal descriptor captures the motion features whether they are static or not. We then compute a vocabulary over these descriptors for all training samples and

use the BoW model to classify actions using an ensemble of SVM classifiers.

In the next section, feature detection and representation are discussed. Section 3 describes the Bag of Words model for action recognition. Experimental results along with the analysis of the obtained classification rates are reviewed in Section 4. Finally we conclude with some discussion and indications of future research in Section 5.

2 Feature Detection and Representation

To represent the motion pattern of an action we use local spatio-temporal features from detected spatial interest points on the agent in each frames.

2.1 Feature Point Detection

For detecting the interest points in each frame, we have use the scalespace Harris corner point detector. The corner points are locations in an image L where the image values have significant variation in both directions. For a given scale of observation σ_l^2 , such interest points can be found by computing μ at the scale σ^2 :

$$\mu = g(\cdot; \sigma^2) \times \begin{pmatrix} (L_x)^2 & L_x L_y \\ L_x L_y & (L_y)^2 \end{pmatrix}$$

where L_x and L_y are Gaussian derivatives of the image L defined as:

$$\begin{aligned} L_x(\cdot; \sigma^2) &= \partial_x(g(\cdot; \sigma^2) * L) \\ L_y(\cdot; \sigma^2) &= \partial_y(g(\cdot; \sigma^2) * L) \end{aligned}$$

and $g(\cdot; \sigma^2)$ is the Gaussian kernel:

$$g(x, y; \sigma^2) = \frac{1}{2\pi\sigma^2} \exp(-(x^2 + y^2)/2\sigma^2)$$

The eigenvalues λ_1, λ_2 where $(\lambda_1 < \lambda_2)$ of μ represent characteristic variations of L in both image directions. Two significant values λ_1, λ_2 indicate the presence of an interest point. To detect such points, Harris and Stephens [6] propose to detect positive maximum of the corner function:

$$\begin{aligned} H &= \det(\mu) - k * \text{trace}(x)^2(\mu) \\ &= \lambda_1\lambda_2 - k(\lambda_1 + \lambda_2)^2 \end{aligned}$$

as positive indications of salient points in the image.

2.2 Feature Point Representation

The feature points described in the previous section are detected at every frame in an image sequence. The features we use are spatio-temporal descriptors of the local space-time structure of the image sequence. To compute the local features in the image sequence $L(x, y, t)$, we construct its spatio-temporal scale-space representation:

$$L(\cdot, \sigma^2, \tau^2) = L * g(\cdot, \sigma^2, \tau^2)$$

where $g(\cdot; \sigma^2, \tau^2)$ is a spatio-temporal Gaussian kernel defined a spatial scale σ^2 and temporal scale τ^2 . The spatio-temporal neighborhood of a feature point in space and time characterizes the local space-time structure of a salient point. They contain information about the motion and the spatial appearance of events in image sequences. To capture this information, we compute spatio-temporal n-jets:

$$l^j = (L, L_x, L_y, L_t, L_{xx}, L_{yy}, L_{tt}, L_{xy}, L_{xt}, L_{yt})^j$$

where j indicates a set of interest points extracted from a particular frame.

3 A BoW Model for Action Recognition

After extracting the local space-time features from the interest points of action sequences we use a BoW model for classification. We first compute a vocabulary over the extracted spatio-temporal descriptors, then build an ensemble of SVM classifiers to recognize each action class.

3.1 The BoW model

Let N be the number of action classes to recognize. The feature set is defined as:

$$S = \{l_i^j\},$$

where $i \in \{1, \dots, N\}$ and j ranges over each interest point descriptor for action i . We have apply the Fuzzy C-means clustering algorithm S to obtain k “words” that constitute the vocabulary of descriptors in the BoW model. These words represent are the primitive events within each action. We then quantize every descriptor in S to the closest word in the vocabulary. Finally, the histogram of word occurrences in each action sequences is the final feature representation for every action sequence. These features are used to train the SVM classifiers.

3.2 SVM design

Support Vector Machines (SVMs) are state-of-the-art large margin classifiers which have recently gained popularity within visual pattern recognition [15]. In this section we provide a brief review of the theory behind this type of algorithm; for more details we refer the reader to [4, 14].

Consider the problem of separating the set of training data $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ into two classes, where $x_i \in \mathbb{R}^N$ is a feature vector and $y_i \in \{-1, +1\}$ its class label. If we assume that the two classes can be separated by a hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ in some space \mathbf{H} , and that we have no prior knowledge about the data distribution, then the optimal hyperplane is the one which maximizes the margin [14]. The optimal values for \mathbf{w} and b can

be found by solving a constrained problem, using Lagrange multipliers $\alpha_i (i = 1, 2, \dots, m)$.

$$f(x) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}) \right)$$

where α_i and b are found by using and SVC learning algorithm [14]. Those x_i with non-zero α_i are the "support vectors". For $K(\mathbf{x}, \mathbf{y}) = x \cdot y$, this corresponds to constructing an optimal separating hyperplane in the input space \mathbb{R}^N .

In our method we have designed main *six* SVMs f^i , where $i \in A = \{\text{walking, running, jogging, boxing, waving, clapping}\}$. Furthermore, each f^i is a combination of SVMs f_j^i for $i \neq j$ and $i, j \in A$. So in this way for each action class $i \in A$ we have *five* SVMs. For a test sequence ω we get the final score ρ as

$$\rho = \max_{i \in A} \left(\sum_{j \in A, i \neq j} f_j^i(\omega) \right)$$

Based on this ρ value the class of ω is defined. We have used χ -square kernel for all the SVMs.

4 Experimental Results

SVM classification combined with motion descriptors in terms of local features (LF) from the spatial interest points is used for action recognition. In this section we evaluate the performance of our approach to human action recognition.

4.1 Datasets

For the evaluation we have used KTH dataset [8]¹. It is a well known video database containing six different types of human actions like walking, jogging, running, boxing, hand waving and hand clapping. All these actions are performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variance, outdoor with different cloth and indoor. In this way there are 100

¹<http://www.nada.kth.se/cvap/actions/>

sequences per action class. The frames are having spatial resolution of 160 x 120.

First, we randomly split the dataset into training (50 sequences per class), validation (20 sequences per class), and testing (20 sequences per class) sets. All parameter tuning was performed on the independent validation set, and reported test results are on the independent test set.

4.2 Results

Table 1 shows the confusion matrix of the our approach. The first thing to note about these results is that overall they are well below state-of-the-art performance for action recognition. However, many individual discriminations between action classes are at or beyond the state-of-the-art. Note how the confusions tend to be clustered within similar action classes: walking, running and jogging forming one group, and boxing, hand waving and clapping another.

	W	R	J	B	HW	C
W	0.46	0.1	0.2	0.02	0.14	0.08
R	0.12	0.34	0.48	0	0.04	0.02
J	0.2	0.28	0.42	0	0.04	0.02
B	0.04	0	0.06	0.42	0.32	0.16
HW	0.08	0	0	0.18	0.46	0.28
C	0.06	0.04	0.04	0.14	0.28	0.44

Table 1: Confusion matrix of action recognition using proposed method. Walking (W), running (R), jogging (J), boxing (B), waving (HW) and clapping (C) actions are taken into account.

4.3 Discussion and Analysis

The confusion matrix (Table 1) shows that our method can clearly distinguish between "leg" and "arm" classes of actions. Running, walking and jogging, for example, are mostly characterized by leg motions, while boxing, hand waving and clapping are primarily characterized by arm motions. The confusions in 1 are predominantly clustered into then action groups, while very few "cross-group" confusions are generated.

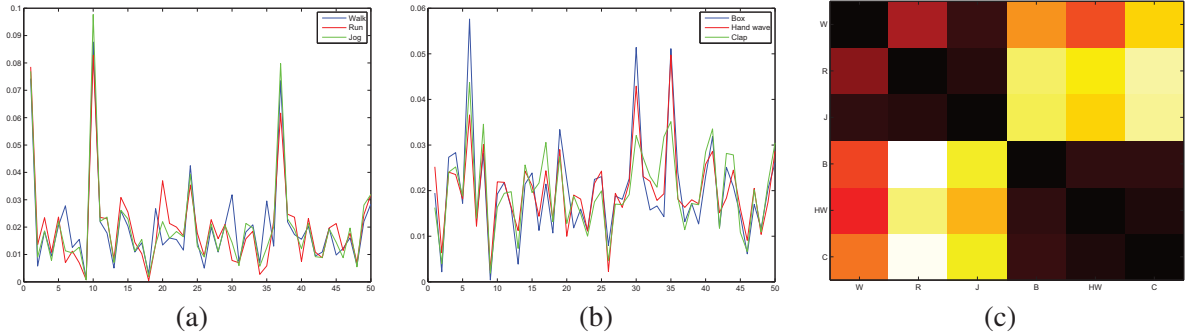


Figure 2: Analysis of the mean histograms of each action class. (a) The plot of the mean histogram of “leg” actions. (b) Mean histograms of the “arm” actions. (c) A matrix of KL-divergence between all class means. These figures are best viewed in colour.

By analyzing the histograms of individual action classes we can gain some insight into why this is happening. Figure 2(a) and (b) show plots of the mean histograms of all action classes grouped according to action type: leg actions in figure 2(a) and arm actions in figure 2(b). Inspection of these two groups of mean histograms shows how the mean histograms within these groups are highly correlated. The histograms defined over our vocabulary are simply not discriminative enough to make distinctions within these groups.

More insight can be obtained by considering the histograms of word occurrences as probability distributions and inspecting the KL-divergence between the mean histograms of all action classes. Let \bar{H}_a designate the mean histogram for action class a , so that $H_a(w_i)$ indicates the mean frequency of occurrence of word w_i in action class a . The KL-divergence between two mean action histograms for actions a and b is defined as:

$$d(H_a||H_b) = \sum_i H_a(w_i) \log \frac{H_a(w_i)}{H_b(w_i)}.$$

The KL-divergence between two mean histograms is a measure of the similarity of those two histograms.

Figure 2(c) shows a matrix indicating the KL-divergence between all pairs of action classes we consider. In this figure, the darker cells indicate low KL-divergence between the two histograms, while brighter ones indicate high KL-divergence.

Intuitively, we want high KL-divergence between all mean histograms as this indicates good distinction between two distributions. Instead, in figure 2(c) we see the same grouping affect between similar actions. The “leg” and “arm” groups are clearly distinguishable. This KL-divergence matrix is, in fact, strongly negatively correlated with the confusion matrix in 1.

5 Conclusions and Future work

We have demonstrated how local spatio-temporal features can be used for representing and recognizing motion patterns such as human actions. Extraction of local spatio-temporal feature from spatial corner points is a novel approach to model both the static and dynamic components of human action.

We hypothesize that the poor performance of the classifier is mainly due to deficiencies in vocabulary construction. The analysis given in the previous section indicates that the mean histograms of similar actions are tightly clustered in feature space. Future work will investigate alternative methods for vocabulary construction as well as the possibility of using the KL-divergence matrix as a guide to vocabulary selection.

Representations of motion patterns using local features have advantages of being robust to variations in the scale, the frequency and the velocity of the pattern. Future work is to investigate more on

the different vocabulary sizes and the different set of features that can reduce the confusions present inside the hand and leg actions. This work can be extended to the action recognition in scenes with complex non-stationary backgrounds. Finally, using the locality of features, recognition of multiple actions in the same scene can be addressed.

References

- [1] Saad Ali and Mubarak Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2008.
- [2] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.
- [3] Olivier Chomat, Olivier Chomat, and James L. Crowley. Probabilistic recognition of activity using local appearance. In *IEEE Conference on Computer Vision and Pattern Recognition, Fort Collins*, pages 104–109, 1999.
- [4] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 1 edition, March 2000.
- [5] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98, 1999.
- [6] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.
- [7] Jesse Hoey and James J. Little. Representation and recognition of complex human motion. In *In Proc. of the Conference on Computer Vision and Pattern Recognition*, pages 752–759, 2000.
- [8] Ivan Laptev and Tony Lindeberg. Space-time interest points. *Computer Vision, IEEE International Conference on*, 1:432, 2003.
- [9] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, November 2006.
- [10] Juan Carlos Niebles and Fei-Fei Li 0002. A hierarchical model of shape and appearance for human action classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, Minneapolis, Minnesota, USA, June 2007.
- [11] Christian Schödl, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local svm approach. *Pattern Recognition, International Conference on*, 3:32–36, 2004.
- [12] Paul Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th International Conference on Multimedia*, pages 357–360, 2007.
- [13] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, pages 1473–1488, 2008.
- [14] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998.
- [15] Christian Wallraven, Barbara Caputo, and Arnulf B. A. Graf. Recognition with local features: the kernel recipe. In *ICCV*, pages 257–264. IEEE Computer Society, 2003.

Adaptive Dynamic Space Time Warping for Real Time Sign Language Recognition

Sergio Escalera, Alberto Escudero, Petia Radeva, and Jordi Vitrià

Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Spain UB

Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain

Abstract

The problem of automatic action recognition in uncontrolled environments becomes a hard because due to the high changes in action appearance because of illumination changes, frame resolution, occlusions, background moving objects, etc. In this paper, we propose a general framework for automatic action classification applied to the sign language recognition problem. The system is based on skin blob detection and tracking. Using a coordinate system estimated from a face detection procedure, the final sign is recognized by means of a new adaptive method of Dynamic Space Time Warping. Results over a Sign Language database show high performance improvement classifying more than 20 signs.

Keywords: Sign Language Recognition, Dynamic Space Time Warping.

1 Introduction

Automatic Action and gesture recognition is a challenging task in the fields of social signal processing, affective computing, communication or psychology, between others. In the case of sign language, recognition is a hard task because of the high changes of gestures in motion and appearance. Recent works try to deal with this problem by means of tracking blobs mainly corresponding to hands. Afterwards, temporal knowledge is used to perform sign classification.

Concerning the segmentation and feature extrac-

tion steps, several works use special clothes or cumbersome devices such as colored markers or gloves [1]. Common approaches for hand location base on skin detection, motion detection, edges, or background subtraction [2, 3]. After localizing the regions of interest, motion information or local descriptors, such as SIFT [4] or HOG [5], are frequently used to describe the region content.

In our work, subjects can appear either with short-sleeved or long-sleeved clothes. In this domain, working with multiple region hypotheses is recommended. In order to work with multiple candidates, Sato and Kobayashi [6] extended the Viterbi algorithm in the Hidden Markov Model (HMM) accommodating multiple hypothesis at each query frame. Dynamic Space Time Warping (DSTW) [7] was defined as an extension of Dynamic Time Warping (DTW) [8] in order to deal with a fixed number of candidates by frame. Recently, Conditional Random Fields (CRF) have been also applied to sign language recognition in order to learn an adaptive threshold able to distinguish between vocabulary and non-sign patterns [9].

In this paper, we suppose that the number of candidates should vary based on the size of the segmented body region. This allows a problem-dependent adaptation that reduces time complexity while preserving (or even improving) the performance of the sign language recognition system. The scale and translation invariant approach, based on [10], defines a bottom-up procedure where skin

regions are segmented, described, and temporally recognized as signs of the vocabulary using the new Adaptive Dynamic Space Time Warping (A-DSTW) procedure.

The rest of the paper is organized as follows: Section 2 describes the different steps of the bottom-up sign language recognition system, including the A-DSTW algorithm. Section 3 presents the evaluation of the methodology, and finally, Section 4 concludes the paper.

2 Sign language recognition

The process for sign language recognition is shown in Figure 1. First steps of the procedure focus on segmentation of arm-hand blocks, tracking, and description of the object content, meanwhile final step uses temporal knowledge to perform sign classification.

2.1 Segmentation

In this work, we use image sequences from uncontrolled environments. In order to avoid false arm-hand detections, first, a face detection procedure based on Viola & Jones detector is applied [11]. Using the content of the detected face, a skin color model is defined [12]. This step reduces false positive detection at the same time that robustly segments arm-hand regions. Size and position of the face region are used to define a coordinate system centered on the face and normalized using the face area. The face resolution is also used to define the size of the candidate regions. This step makes the procedure invariant to scale and translation. Arm-hand regions are segmented just by capturing the highest density blobs at the expected locations. An example of this procedure is shown in Figure 2(a)-(c). First, the face region and skin candidates are shown. Next, some candidate regions over the highest density blobs are captured. Finally, an example of region tracking is shown over the input image sequence. An example of an ideal and obtained tracked sign trajectory considering both hands are shown in Figure 3.

2.2 Feature extraction

In order to describe the content of the candidate regions, we take advantage of the state-of-the-art region descriptors. In [10, 13], the authors define the feature vector $Q_{jk} = \{x_{jk}, y_{jk}, u_{jk}, v_{jk}\}$ for arm-hand candidate k at the j th frame, tracking just one arm-hand sign. x and y correspond to the spatial coordinates and u and v to the components of the movement vector. In our case, working with two arm-hand signs, the feature vector becomes $Q_{jk} = \{x_{jk}^1, y_{jk}^1, u_{jk}^1, v_{jk}^1, F_{jk}^1, x_{jk}^2, y_{jk}^2, u_{jk}^2, v_{jk}^2, F_{jk}^2\}$, where the two super-index correspond to the left-right candidate arm-hand, and F is the HOG feature vector of the candidate region [5].

2.3 Adaptive DSTW classification

The original DTW algorithm [8] was defined to match temporal distortions between two models. Among all the defined variants of this method, in [7], the authors defined a spatio-temporal Dynamic Time Warping in order to work with a fixed number of multiple candidates. This approach has been later applied in [10, 13], where the authors defined an one hand sign recognition system. Given the high computational complexity of the DSTW approach for a high number of fixed candidates, the authors of [13] introduced the pruning of classifiers in order to reduce cost. In this section, we present an Adaptive Dynamic Space Time Warping (A-DSTW), where the number of candidates is adapted based on the arm-hand tracking process. Next, we overview the basis of the Dynamic Time Warping approaches and the A-DSTW proposal.

2.4 A-DSTW

The goal of DTW is to find an alignment warping path between two time series $Q = \{q_1, \dots, q_n\}$ and $C = \{c_1, \dots, c_m\}$. In order to align these two sequences, a $n \times m$ matrix is designed, where the position (i, j) of the matrix contains the distance between q_i and c_j . The Euclidean distance is the most frequently applied. Then, a warping path

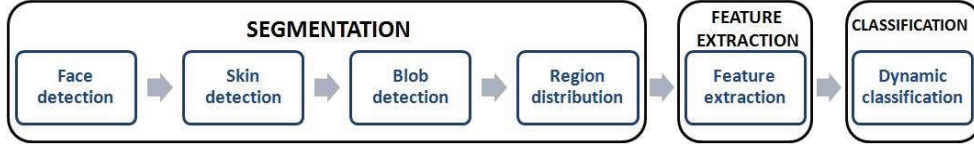


Figure 1: System scheme.

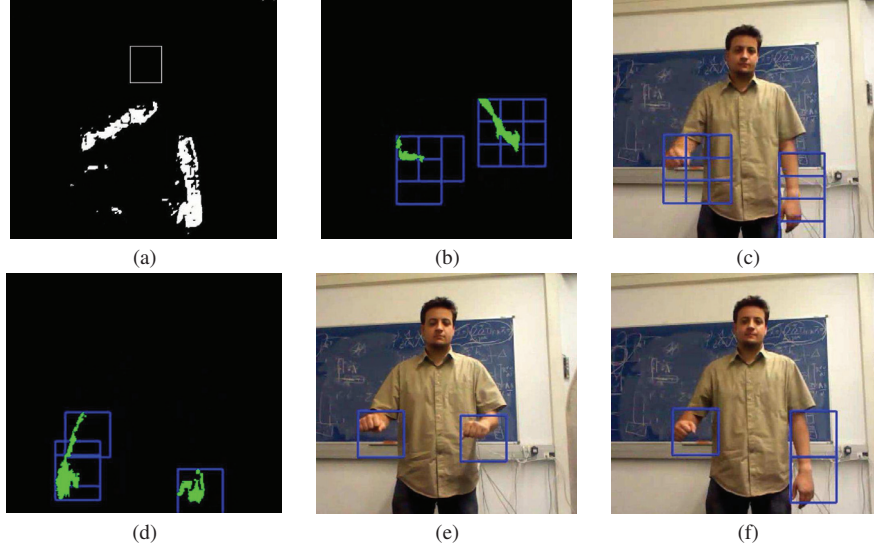


Figure 2: (a) Skin color segmentation based on face color model, (b)(c) Region distribution of DSTW for a fix number of regions over highest density blobs, and (d)(e)(f) Region distribution of A-DSTW for a variable number of regions over highest density blobs.

$W = \{w_1, \dots, w_T\}$, $\max(m, n) \leq T < m + n + 1$ is defined as a set of "contiguous" matrix elements that defines a mapping between Q and C . This warping path is typically subjected to several constraints:

Boundary conditions: $w_1 = (1, 1)$ and $w_T = (m, n)$.

Continuity: Given $w_{t-1} = (a', b')$, then $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$.

Monotonicity: Given $w_{t-1} = (a', b')$, $w_t = (a, b)$, $a - a' \leq 1$ and $b - b' \leq 1$, this forces the points in W to be monotonically spaced in time.

We are generally interested in the final warping path that satisfying these conditions minimizes the warping cost:

$$DTW(Q, C) = \min \left\{ \frac{1}{T} \sqrt{\sum_{t=1}^T w_t} \right\} \quad (1)$$

where T compensates the different lengths of

the warping paths. This path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i, j)$ as the distance $d(i, j)$ found in the current cell and the minimum of the cumulative distance of the adjacent elements:¹

$$\gamma(i, j) = d(i, j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\} \quad (2)$$

The first image in Figure 4 shows an exemple of a warping path for a two time series matched in a DTW matrix.

In the case of the DSTW of [7], the two-dimensional matrix is extended into a three-

¹Note that though different adjacency elements can be considered varying the warping normalization factor T , here we follow the present adjacency rule as the most extended one.

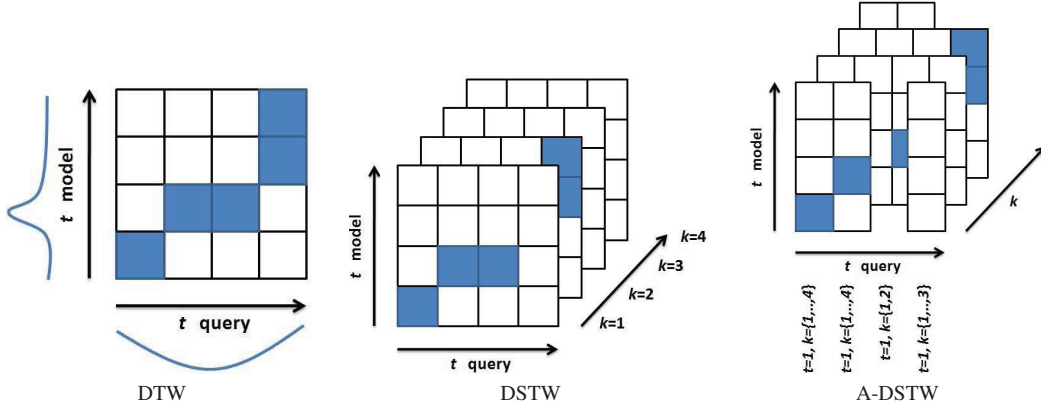


Figure 4: Examples of dynamic matching by DTW variants.



Figure 3: Left: ideal sign hand trajectories. Right: tracked hand trajectories.

dimensional one in order to match K multiple candidates of the third dimension that appear at each instant of time. An example of this procedure is shown in the second image of Figure 4. In this case, the warping path constraints are re-defined in a three-value space as follows:

Boundary conditions: $w_1 = (1, 1, k)$ and $w_T = (m, n, k')$, $k, k' \in [1, \dots, K]$.

Continuity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1$, $k, k' \in [1, \dots, K]$.

Monotonicity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1$, $k, k' \in [1, \dots, K]$, this forces the points in W to be monotonically spaced in time.

Now, continuity and monotonicity are required only in the temporal dimensions. No such restrictions are needed for the spatial dimension; the warping path can "jump" from any spatial candidate k to any k' . In this case, the cumulative distance of the adjacent elements is re-defined as follows:

$$\gamma(i, j, k) = d(i, j, k) + \min_{(3)} \{ \gamma((i-1, j-1), (i-1, j), (i, j-1) \times \{1, \dots, K\}) \}$$

Concerning our A-DSTW, we follow the DSTW

rules, but, instead of using a predefined number of candidate cases, we adapt the number of candidates based on the length of the segmented blobs. This results in a variable number of candidates per instant of time (frames in our case).

Given a sequence of model feature vectors $M_i, 1 \leq i \leq m$, and a sequence of sets of query feature vectors $Q_j = \{Q_{j1}, \dots, Q_{jK}\}, 1 \leq j \leq n$, where K varies among different j , now the A-DSTW warping constraints are defined as follows:

Boundary conditions: $w_1 = (1, 1, k)$ and $w_T = (m, n, k')$, $k, k' \in [1, \dots, \max(\text{length}(Q))]$.

Continuity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1$, $k, k' \in [1, \dots, \max(\text{length}(Q))]$.

Monotonicity: Given $w_{t-1} = (a', b', k')$, then $w_t = (a, b, k)$, $a - a' \leq 1$ and $b - b' \leq 1$, $k, k' \in [1, \dots, \max(\text{length}(Q))]$, this forces the points in W to be monotonically spaced in time.

The A-DSTW algorithm is shown in Algorithm 1. $N(w) = N(i, j, k)$ defines the set of all possible values of w_{t-1} that satisfy the warping path constraints:

$$N(i, j, k) = \{(i-1, j), (i, j-1), (i-1, j-1)\} \times \{1, \dots, K\} \quad (4)$$

taking into account that the value of K varies over j . An example of an A-DSTW (i, j, k) space is shown in the last image of Figure 4. Note that the number of query candidates changes across time instants. In Figure 2(d)-(f) an example of candidates distribution over segmented blob regions and original images is shown. Note that different num-

ber of candidates are distributed over the skin region based on the segmented areas.

Table 1: The A-DSTW algorithm

```

Input: A sequence of model feature vectors  $M_i, 1 \leq i \leq m$ ,
and a sequence of sets of query feature vectors  $Q_j = \{Q_{j1}, \dots, Q_{jK}\}, 1 \leq j \leq n$ , where  $K$  varies among different  $j$ .
Output: A global matching cost  $D^*$  and an optimal warping path  $W^* = (w_1^*, \dots, w_T^*)$ .
 $j = 0$  // Initialization
for  $i = 0 : m$  do
  for  $k = 1 : \max(\text{length}(Q))$  do
     $D(i, j, k) = \infty$ 
  end
end
 $D(0, 0, 1) = 0$ 
for  $j = 1 : n$  do
  for  $i = 1 : m$  do
    for  $k = 1 : \text{length}(Q_j)$  do
      if  $i = 0$  then
         $D(i, j, k) = \infty$ 
      end
      else
         $w = (i, j, k)$ 
         $D(w) = d(w) + \min_{w' \in N(w)} D(w')$ 
         $b(w) = \text{argmin}_{w' \in N(w)} C(w', w)$ 
      end
    end
  end
end
 $k^* = \text{argmin}_k \{D(m, n, k)\}$  // Termination
 $D^* = D(m, n, k^*)$ 
 $w_T^* = (m, n, k^*)$ 
 $w_{t-1}^* = b(w_t^*)$  // Backtrack

```

3 Results

Before the presentation of the results, first, we discuss the data, methods and parameters, and validation protocol of the experiments.

Data: The data used in our experiments consists of 200 video sequences corresponding to 20 signs from the Spanish sign language dictionary from 10 different subjects. Half of the sequences are captured using short-sleeved clothes meanwhile the remaining half of the data is recorded using long-sleeved clothes. The resolution of the video sequences is 640×480 and 15 FPS. Some samples of the captured signs are shown in Figure 5.



Figure 5: Frame samples of the sign language database.

Methods and parameters: We use the DSTW with 15 fixed candidates per frame for comparison [10]. Concerning the A-DSTW approach, the number of candidates is adapted based on the number of regions that uniformly fall in the length of the detected blobs [14] with size 75% of the face area with an overlapping of 50% among regions. For both dynamic methods, spatial coordinates, movement vectors and HOG descriptors are computed per candidate. The weight of the four first features is of 0.5 and the same weight is assigned for the normalized HOG descriptor in the Euclidean computation in the dynamic matching.

Validation measurements: We apply stratified ten-fold cross-validation and test for the confidence interval with a two-tailed t-test. The ground truth is obtained using the samples from the ten-fold iteration containing short-sleeved clothes and tracking just one candidate region per hand. The remaining data is used for testing. Each test sample is categorized applying 3-KNN over the first three retrieved sign models from the database after computing the warping path.

3.1 Sign recognition

Applying stratified ten-fold evaluation as commented before over the sign language database for both DSTW and A-DSTW, we obtained the results shown in the top row of Table 2. A-DSTW obtains near 13% more of performance, corresponding to a relative performance improvement near

20%. This result is due to the main drawback of the DSTW algorithm. Using a fixed number of candidates, when a small arm-hand region is segmented, redundant information may be computed meanwhile when the segmented skin region is large, we may compute few candidate regions, reducing efficiency. On the other hand, simply adapting the number of candidates does not only increase the final performance, we can also save time. This can be seen by the mean number of candidate regions shown in the middle row of Table 2. In comparison to the mean of 15 fixed regions of DSTW, the A-DSTW only required a global mean of 9.75 regions, allowing a real time computation, as shown in the bottom row of Table 2.

Table 2: Sign language recognition results.

	DSTW	A-DSTW
Performance	79.27±3.15	92.18±2.12
Mean candidate regions	15	9.75
Computed frames/second	18	26

4 Conclusion

In this paper, we proposed a general framework for real time action classification applied to the sign language recognition problem. The system is based on skin blob detection and tracking. Using a coordinate system estimated from a face detection procedure, the final sign is recognized by means of a new adaptive method of Dynamic Space Time Warping. The A-DSTW uses a variable number of segmented region candidates to match temporal series, yielding a better performance while reducing the computational complexity of the classification task.

5 Acknowledgement

This work has been partially supported by the projects TIN2009-14404-C0 and CONSOLIDER-INGENIO 2010 (CSD2007-00018). We are specially grateful to Noelia H. for her support during the development of this work.

References

- [1] J. Triesch, C. von der Malsburg, Robotic gesture recognition, *Gesture Workshop* (1997) 233–244.
- [2] F. Chen, C. Fu, C. Huang, Hand gesture recognition using a real-time tracking method and hidden markov models, *Image and Video Computing* 21 (8) (2003) 745–758.
- [3] J. Martin, V. Devin, J. Crowley, Active hand tracking, *Automatic Face and Gesture Recognition* (1998) 573–578.
- [4] D. G. Lowe, Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision* 60 (2004) 91–110.
- [5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: *International Conference on Computer Vision & Pattern Recognition*, Vol. 2, 2005, pp. 886–893.
- [6] Y. Sato, T. Kobayashi, Extension of hidden markov models to deal with multiple candidates of observations and its application to mobile-robot-oriented gesture recognition, *ICPR* 2 (2002) 515–519.
- [7] J. Alon, V. Athistos, Q. Yuan, S. Sclaroff, Simultaneous localization and recognition of dynamic hand gestyres, *IEEE Motion Workshop* (2005) 254–260.
- [8] J. Kruskall, M. Liberman, The symmetric time warping algorithm: From continuous to discrete, *Time Warps*. Addison-Wesley.
- [9] H.-D. Yang, S. Sclaroff, S.-W. Lee, Sign language spotting with a threshold model based on conditional random fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (7) (2009) 1264–1277.
- [10] A. Stefan, V. Athistos, J. Alon, S. Sclaroff, Translation and scale-invariant gesture recognition in complex scenes, in: *PE-TRA: Proceedings of the 1st international conference on Pervasive Technologies Related to Assistive Environments*, 2008, pp. 1–8.
- [11] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, 2001, pp. 511–518.
- [12] M. Jones, J. Rehg, Statistical color models with application to skin detection, in: *International Journal of Computer Vision*, Vol. 46, 2002, pp. 81–96.
- [13] J. Alon, V. Athistos, Q. Yuan, S. Sclaroff, A unified framework for gesture recognition and spatiotemporal gesture segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (9) (2009) 1685–1699.
- [14] O. B. segmentation software, <http://opencv.willowgarage.com/wiki/cvblobslib>.

An Analysis of Theoretical and Practical Aspects of Spatio-Temporal Regular Flow (SPREF)

Juan Diego Gomez*, Carlo Gatta* and Petia Radeva*

* *Computer Vision Center, Barcelona, Spain*
E-mail: jdgomez@cvc.uab.es

Abstract

Describing movement direction is a key operation in video analysis and is often used for many applications, e.g. tracking. The directions in which a video (or at least a part of it) is regular, can be extracted and described by means of SPREF (SPatio-temporal REGularity Flow). SPREF is a novel method that models the directions of regular movement along the axis x , y and t , with a 3D vector field. The method estimates the vector field by minimizing an energy/cost function. Such minimization is performed solving a linear system. The contribution of this paper is twofold: (1) we perform a theoretical study on the previous related work. This analysis gave rise to a new mathematical formulation able to solve the same minimization problem with a smaller linear system. As a consequence, the linear system does not depend on video length. (2) We performed a statistical analysis on a SPREF parameter (standard deviation of Gaussian filtering) hardly mentioned in previous work and whose importance has been overlooked. The experiments show that an empirical optimal value for the standard deviation can be obtained independently of image size and movement magnitude.

Keywords: SPREF, motion estimation.

1 Introduction

Movement vector flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene. In the literature, there are many techniques to describe and measure motion, such as, Phase correlation [3], Block-based methods [4], Partial derivatives-based methods (e.g. LucasKanade [5], HornSchunck [6], BlackJepson [7]) and Discrete optimization methods [8].

SPREF is a lately proposed method, described in [1], to model and estimate the movement directions in a video. A video is determinate to be regular along the directions in which the pixels intensity change the least. These directions depend on the type of motion and spatial structure of the scene. SPREF, as introduced in [1], is a method designed to calculate these directions, only if the motion and spatial structure are composed according to a translation or affine transformation. The regularity directions of a video is a very useful feature often used in image and video processing for tasks such as, motion analysis, compression, and video inpainting. The work presented in [1], proposes a mathematical approach to find spatiotemporal regular features, independently of edge detection. Hence its success does not depend on the presence of strong edges in the scene. Instead it analyzes the whole regular region, and tries to find the best directions that model the overall regularity of the region. The movement directions of regular-

ity in the video are modeled with a 3-D vector field, called the spatiotemporal regularity flow (SPREF) field.

In [1], the authors introduce two SPREF models; the translational SPREF (T-SPREF) and the affine SPREF (A-SPREF). T-SPREF is a simplified computational method that provides a good estimation of spatiotemporal regularity flow, when the direction of regularity in the video is a function only of the flow-propagation axis (translational flow model). Nevertheless, the T-SPREF precision decreases when the directions of regularity depend on multiple axes. In this case, A-SPREF model, is more robust since its components still propagate along one major axis. However, each component is also a function of the other axes. Hence, A-SPREF can estimate spatiotemporal regularity flows produced by any affine transformation. The major strength of the method is treating a video not as 2-D independent sequence of images but as a 3-D volume, processing all of its information simultaneously. This provides general information about the motion along the video, instead of describing local variations between subsequent frames. In [1], the regular regions of a video are found by decomposing the video by applying and oct-tree method.

In this paper, we consider translational movements as a particular case of affine transformations. Therefore, we only focus our analysis on the A-SPREF, which will be called SPREF, from now on.

2 Previous Work

At the best of our knowledge, the work in [1] is pioneer of SPREF concept and together with [2] summarize the whole previous work related to this topic. The work presented in [1] defines \mathcal{F} , as a 3-D vector field that shows the directions, along which intensity in a spatiotemporal region is regular. In [1], the authors state that: "the condition that the intensity should vary regularly in the flow direction is perceived as a requirement to follow the

directions, in which the sum of directional gradients is minimum". Therefore, they poses a general flow energy function, for \mathcal{F} , as:

$$E(\mathcal{F}) = \int_{\Omega} \left| \frac{\partial (I * H)(x, y, t)}{\partial \mathcal{F}(x, y, t)} \right|^2 dx dy dt \quad (1)$$

Where H is a regularizing filter, e.g a Gaussian, and x, y , are the coordinates of the images. The flow in (2), is approximated by block translations orthogonal to the directions of flow propagation. This results in planar (cross-sectional) parallelism in the SPREF, which is defined as all the vectors on a plane being uniform. A cross-sectional parallel flow field consists of the following three components: xy-parallel (\mathcal{F}_t), xt-parallel (\mathcal{F}_y), and yt-parallel (\mathcal{F}_x). In this paper the flow-propagation axis is always t , and we model the regularity that depends on the motion in a spatiotemporal regular region Ω . Defining the vector field \mathcal{F} according to the affine model, $\mathcal{F}_t = (c'_1(x, y, t), c'_2(x, y, t))$. The general flow energy function (1) is expanded and discretized accordingly to (2); where,

$$\begin{bmatrix} c'_1(x, y, t) \\ c'_2(x, y, t) \end{bmatrix} = \begin{bmatrix} a_{11}(t) & a_{12}(t) & a_{13}(t) \\ a_{21}(t) & a_{22}(t) & a_{23}(t) \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

In order to achieve a global solution that uses all the information in the spatio-temporal region Ω . In [1] the parameters $a_{ij}(t)$ are approximate by translated box splines function of the first degree, $b(u)$ in (3), with $\alpha_n (n = 1, \dots, 2^l)$ as the n^{th} spline coefficient, u is one of the volume axis x, y, t indicating the flow-propagation axes; and $l = 1, \dots, k$. Here k is a scale factor, and 2^k is the width of Ω . Then, these parameters are expanded as:

$$a_{ij}(t) = \sum_n \alpha_n^{ij} b(2^{-l}t - n) = \sum_n \alpha_n^{ij} b_n^l(t) \quad (3)$$

The SPREF parameters are estimated by minimizing the flow energy function (1) by quadratic minimization over the spline parameters. One single spline could be fit to all parameters separately. However, since this will increase the search space,

in [1] and [2] is assigned one spline per flow component. The spline parameters for a temporal flow propagation axis are:

$$A = [\alpha_1^{11} \dots \alpha_{l_1}^{11}, \alpha_1^{12} \dots \alpha_{l_1}^{12}, \alpha_1^{13} \dots \alpha_{l_1}^{13}, \alpha_1^{21} \dots \alpha_{l_2}^{21}, \alpha_1^{22} \dots \alpha_{l_2}^{22}, \alpha_1^{23} \dots \alpha_{l_2}^{23}] \quad (4)$$

Finally, they minimize and rearrange (2), according to these variables, to take the form:

$$A_{3[l_1+l_2]} (B_{3[l_1+l_2] \times 6N} M_{6N \times 3[l_1+l_2]} + C_{3[l_1+l_2]} = 0 \quad (5)$$

Where N is the maximum value of t , and l_1, l_2 are given from the spline theory. Further details on this derivation can be found in [2].

3 Alternative derivation of the SPREF solution equation

Analysing in detail the derivation of equation (5) as presented in [2], we found an alternative derivation that leads to a more compact linear system. Moreover, the linear system solution formulated in (6) depends on both the number of spline parameters and the length of the video (N); as it can be verified by inspecting the size of its matrices. The number of spline parameters (usually small), is a variable we can establish according to our needs of accuracy and computational cost. On the contrary, t is an independent variable that typically assumes large values. We propose a linear system solution for the SPREF that depends only on the number of spline parameters, but not on the temporal length (frames number) and in addition, this system is always square. As a consequence the computational cost to solve the SPREF minimization, is reduced. The following subsection outlines the proposed derivation.

3.1 Formulation of the linear system

Being

$$e = \sum_{x,y,t} [f_x(a_1(t)x + a_2(t)y + a_3(t)) + f_y(a_4(t)x + a_5(t)y + a_6(t)) + f_t] \quad (6)$$

where f_x, f_y, f_t are the derivatives in (2), and a_i is described in (7). Expanding the summation in (7) on n , we obtain,

$$a_i(t) = [\alpha_1^i b_1(1) + \dots + \alpha_N^i b_N(1), \alpha_1^i b_1(2) + \dots + \alpha_N^i b_N(2), \dots, \alpha_1^i b_1(T) + \dots + \alpha_N^i b_N(T)] \quad (8)$$

then, deriving over the parameters α_n^i

$$\frac{\partial a_i(t)}{\partial \alpha_n^i} = [b_n(1), b_n(2), \dots, b_n(T)] = b_n(t). \quad (9)$$

Now, using the chain rule, we can write

$$\frac{\partial e}{\partial \alpha_n^i} = \frac{\partial e}{\partial a_i(t)} \frac{\partial a_i(t)}{\partial \alpha_n^i} = \frac{\partial e}{\partial a_i(t)} b_n(t) \quad (10)$$

and then,

$$\begin{aligned} \frac{\partial e}{\partial \alpha_n^1} &= x f_x b_n(t), \quad \frac{\partial e}{\partial \alpha_n^2} = y f_x b_n(t), \\ \frac{\partial e}{\partial \alpha_n^3} &= f_x b_n(t), \quad \frac{\partial e}{\partial \alpha_n^4} = x f_y b_n(t), \\ \frac{\partial e}{\partial \alpha_n^5} &= y f_y b_n(t), \quad \frac{\partial e}{\partial \alpha_n^6} = f_y b_n(t). \end{aligned} \quad (11)$$

Being B a ($N \times N$) matrix defined for a particular t , as follow:

$$B = \begin{bmatrix} b_1(t) & \dots & b_N(t) \\ \vdots & \ddots & \vdots \\ b_1(t) & \dots & b_N(t) \end{bmatrix} \begin{bmatrix} b_1(t) & \dots & b_N(t) \\ \vdots & \ddots & \vdots \\ b_1(t) & \dots & b_N(t) \end{bmatrix} \quad (12)$$

$$E(\mathcal{F}_t) = \sum_{\Omega} \left| I * \frac{\partial H}{\partial x} c'_1(x, y, t) + I * \frac{\partial H}{\partial y} c'_2(x, y, t) + I * \frac{\partial H}{\partial t} \right|^2 \quad (2)$$

$$a_i(t) = \sum_n \alpha_n^i b_n(t) = \left[\sum_n \alpha_n^i b_n(1), \sum_n \alpha_n^i b_n(2), \dots, \sum_n \alpha_n^i b_n(T) \right] \quad (7)$$

$$M = \sum_t \begin{bmatrix} B \sum_{x,y} x^2 f_x^2 & B \sum_{x,y} xy f_x^2 & B \sum_{x,y} x f_x^2 & B \sum_{x,y} x^2 f_x f_y & B \sum_{x,y} xy f_x f_y & B \sum_{x,y} x f_x f_y \\ B \sum_{x,y} xy f_x^2 & B \sum_{x,y} y^2 f_x^2 & B \sum_{x,y} y f_x^2 & B \sum_{x,y} xy f_x f_y & B \sum_{x,y} y^2 f_x f_y & B \sum_{x,y} y f_x f_y \\ B \sum_{x,y} x f_x^2 & B \sum_{x,y} y f_x^2 & B \sum_{x,y} f_x^2 & B \sum_{x,y} x f_x f_y & B \sum_{x,y} y f_x f_y & B \sum_{x,y} f_x f_y \\ B \sum_{x,y} x^2 f_x f_y & B \sum_{x,y} xy f_x f_y & B \sum_{x,y} x f_x f_y & B \sum_{x,y} x^2 f_y^2 & B \sum_{x,y} xy f_y^2 & B \sum_{x,y} x f_y^2 \\ B \sum_{x,y} xy f_x f_y & B \sum_{x,y} y^2 f_x f_y & B \sum_{x,y} y f_x f_y & B \sum_{x,y} xy f_y^2 & B \sum_{x,y} y^2 f_y^2 & B \sum_{x,y} y f_y^2 \\ B \sum_{x,y} x f_x f_y & B \sum_{x,y} y f_x f_y & B \sum_{x,y} f_x f_y & B \sum_{x,y} x f_y^2 & B \sum_{x,y} y f_y^2 & B \sum_{x,y} f_y^2 \end{bmatrix} \quad (13)$$

Notice that for all t , B can be seen as a $(N \times N \times T)$ volume. Finally, M is the $(6N \times 6N)$ matrix defined in (14).

Every sub-matrix B into the M , is multiplied by a scalar given from a particular t .

4 On the regularizing filter H

In this work we want to highlight the benefices of applying a preprocessing filter to solve the SPREF equation. This is because in [1] and [2], the authors over look the importance of this topic and the filter is hardly mentioned in the formulation of (2). Pre-process filtering is an essential step to extend (2) from (1), since I in (1), must be continue and derivable over all its points, and in this case it is no true, due to the digital image nature. Given the importance of filtering to the SPREF solution, we want to propose a method to estimate the optimal Gaussian filter to pre-process a digital video.

4.1 Using a Gaussian Filter as H

In this section we mainly want to answer two questions about the optimal Gaussian filter for a digital video: (1) Does the optima σ depend on the video size? and (2) Does the optima σ (Gaussian standard deviation) depend on the magnitude of the unknown flow? For our propose, we created 3 different groups of 100 synthetic videos as testing data. For each video of every class, we applied seven different and random affine transformations and, for

each transformation we finally applied 11 growing sizes of Standard deviation of Gaussian filter. Then, a video with 7 different transforms can be seen as a 7 particular test videos. We processed a total of 2100 ($7 \times 100 \times 3$) videos with 11 different Gaussian filters. Thus, the relative error produced by a certain standard deviation is described as follow:

$$\varepsilon_r = \frac{\|E - R\|}{\|R\|},$$

where E is the computed transformation by SPREF, R is the real transformation and the size of the videos vary for each class. The first class of videos is compose by videos of 100 frames with size 50×50 (pixels). Second class is 100×100 pixels per frame with 200 frames. And finally, the third class is $200 \times 200 \times 50$ frames. All the videos were created from a random image data base of 500 different natural pictures. Figures 1, 2 and 3, show the mean of the error curves for each class. Each mean curve is composed by 700 curves, obtained from 7 transformation applied to 100 videos of the same size, according to the class. Test results are summarized as follow:

As it can be seen in figure 1, 2 and 3, the minimum of the error is clearly achieved at $\sigma = 1.5$. This value provides the best behavior of SPREF independently of the size and transformation of the video. This fact is given due to the digital nature of the images used. Since the method is based on the derivatives of the images, this kind of image commonly present hardlim edges which have in the continue domain, an infinitive derivative value.

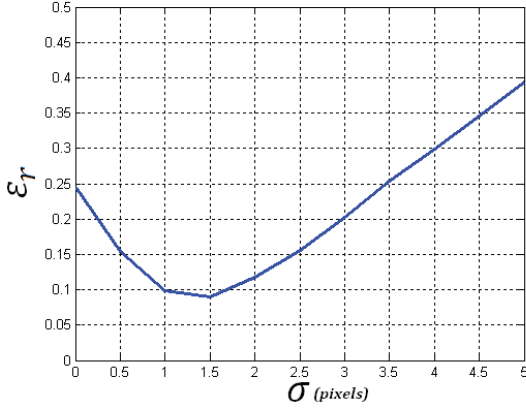


Figure 1: Error mean in translation (Spref-real) vs Std Filter - for 100 videos of 50x50 pixels

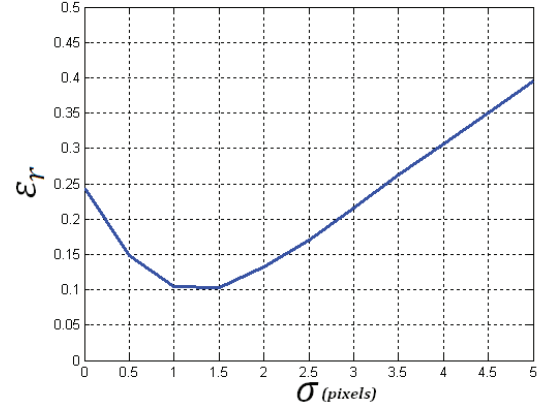


Figure 3: Error mean in translation (Spref-real) vs Std Filter - for 100 videos of 200x200 pixels

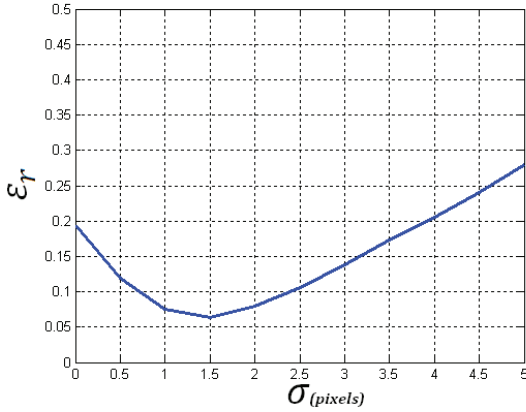


Figure 2: Error mean in translation (Spref-real) vs Std Filter - for 100 videos of 100x100 pixels

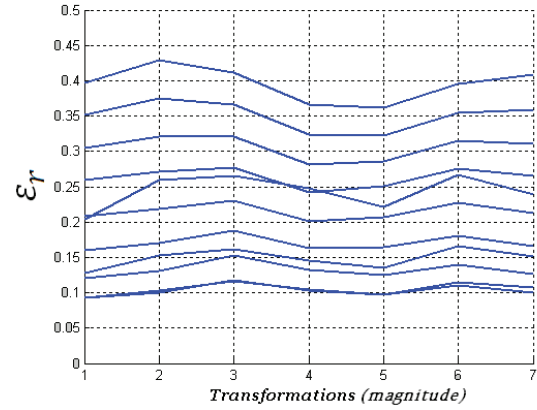


Figure 4: Error mean vs Transformations, for each Standard Deviation

Growing Gaussian filters σ , reduces this limitation by smoothing these edges. However, after a value of 1.5, the edges undergo an over smoothing phenomena. Thus, the method loses important information and the accuracy of the solution begins to decrease. Summarising and as expected, the optimal σ is independent on the transformation magnitude since, as it is observed in figure 4, the 11 error curves (one per each σ) vary as minimum (almost flat) and the curve with lowest errors belong to the $\sigma=1.5$. Moreover, the optimal σ is also inde-

pendent of the video size, since the minimum was achieved in the same value, for all the three class of video sizes. This optimal σ is only for natural images. Then, if it is required to find the optimal σ for another kind of images, we only have to test a set of videos with the same size, and just one transformation.

5 Conclusions and Future Works

In this paper we performed a theoretical study on the previous related work on SPREF. As a consequence, we presented a linear system more compact than the one presented in [2] and independent on video length. We also performed an analysis on standard deviation of Gaussian filtering, hardly mentioned in previous work. We obtained a optimal value of 1.5 for the standard deviation of the Gaussian filter. This value is independent of the magnitude of the transformation that the regular video presents, and also of the size of the video.

As future work we want to improve the algorithm efficiency and reduce the computational time. We also want to extend the integral image concept to 3D volumes and finally, to explore alternatives models for the flow \mathcal{F} .

References

- [1] Orkun Alatas, Pingkun Yan: SpatioTemporal Regularity Flow (SPREF): Its Estimation and Applications, *IEEE Transactions on circuits and systems for video technology*, Vol 17, No 5, May 2007
- [2] Orkun Alatas, Pingkun Yan: Technical Report. Regularity Flow (SPREF): Its Estimation and Applications,
- [3] Liu, J. G. and Yan, H., 2006. Robust phase correlation methods for sub-pixel feature matching. *Proceeding of 1st Annual Conference of Systems Engineering for Autonomous Systems, Defence Technology Centre, A13, Edinburgh, UK*.
- [4] Hongshi Yan, Jian Guo Liu: Robust Phase Correlation based feature matching for image co-registration and demgeneration *Int. J. Comp. Vision*. vol. 29, No. 1, pp. 59-77.
- [5] S. Baker and I. Matthews, Lucas-kanade 20 years on: A unifying framework, *IJCV*, vol. 56, no. 3, p. 221255, 2004
- [6] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *Int. J. of Computer Vision*, 2:283310, 1989.
- [7] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141158, April 2006.
- [8] Ian T. Young, Lucas J. van Vliet: Recursive implementation of the Gaussian filter. *IPattern Recognition Group, Faculty of Applied Physics, Lorentzweg 1, Delft University of Technology, NL-2628 CJ, Delft, Netherlands*
- [9] Viola Paul, Jones Michael: Robust Real-time Object Detection. *Second international Workshop on statical and Computational Theories of Vision - Modeling, learning, computing, and sampling*
- [10] W. Pratt, *Digital Image Processing*. New York: Wiley, 1978.
- [11] Tomasi, C. Manduchi, R.: Bilateral filtering for gray and color images. *Computer Vision*, 1998. *Sixth International Conference on Publication Date: 4-7 Jan 1998 On page(s): 839-846*

3D Human Action Recognition using Key Poses

Wenjuan Gong, Andrew D. Bagdanov and Jordi Gonzàlez

Centre de Visió per Computador, Universitat Autònoma de Barcelona, Barcelona, Spain

E-mail: wenjuan@cvc.uab.es, bagdanov@cvc.uab.es and jordi.gonzalez@cvc.uab.es

Abstract

This paper presents a novel approach to recognition of 3D human actions using key poses. After representing each motion sequence using a vector of direction cosines, we decrease the dimensions by applying Principle Component Analysis and projecting these representations into decreased dimension space. Introducing the idea of bag of words, all the performances are represented as histograms of key poses extracted from the mean performance of each action. We use a Support Vector Machine in training and labelling the test performance. Experimental results show that our method is effective and is near state-of-the-art.

Keywords: Action recognition, direction cosine, key poses, bag of words.

1 Introduction

Action recognition is an important problem in computer vision. Applications are mainly in human computer interaction, in which the recognition of human actions is important to understand the human behavior so that the computer can react accordingly.

There are many different ways to solve this problem. [3] use a multi-class AdaBoost method to classify human actions. Seven types of features like 3D positions of a joint, or 3D positions of each

non-root joint and so on are used. For each feature and each action, one Hidden Markov Model (HMM) is learned. Each HMM contains three hidden states and each state is modelled by a three component Mixture of Gaussians (MOG). The parameters of each HMM are learned by the Baum-Welch algorithm. Then, they use the AdaBoost algorithm to combine the results of weak classifiers.

Other methods interpret actions as multi-dimensional time series that captures the deformations of the body during activity [6]. Dictionaries of temporal scale, mean, and action primitives are created. Action primitives are defined as a time series subsequence that encodes a single dimension of a commonly occurring deformation.

Action types that we consider are walking, running, boxing and jumping. We use a 3D stick figure to model the human body and direction cosines to represent the body poses. After representing each motion sequence using a vector of direction cosines, we decrease the dimensionality by applying Principle Component Analysis (PCA) and project the motions into decreased dimension space. Introducing the idea of bag of words, all the performances are represented as histograms of the vocabulary. The vocabulary is obtained by concatenating key poses from different actions. Given the motion models in decreased dimension space, we explored five methods to extract key poses: poses with local maximum and local minimum energies, poses with local maximum distances, poses corresponding to equally spaced frames, poses cor-

responding to randomly spaced frames and the center poses of the clusters. According to the experimental results, we choose the first method. We use a Support Vector Machine (SVM) in training and labelling the test performance. We show our action classification method is effective and near state-of-the-art.

Our main contribution to solving these problems are: To choose an effective method to extract key poses and to propose an effective method to classify different actions. We experiment on five different methods for key poses extraction and choose the most representative one. We introduce the idea of bag of words and represent performance sequences as a collection of representative poses, but discard the ordering between these poses.

This paper is organized as follows: section 2 introduces our representation of human posture and human motion, section 3 explains our classification method, we show the experimental results in section 4 and we conclude our method and discuss about our future work in section 5.

2 Human Posture and Human Motion Models

In this section, we explain how to represent the human posture as a vector of direction cosines, and how to model human motion in a decreased dimension space of these vectors.

2.1 Human Posture Representation

The body model employed in our work is composed of twelve rigid body parts (hip, torso, shoulder, neck, two thighs, two legs, two arms and two forearms) connected by a total of ten inner joints, see figure 1(a). Body segments are structured in a hierarchical manner, constituting a kinematic tree rooted at the hip, which determines the global rotation of the whole body.

The 3D data we used in our method are from

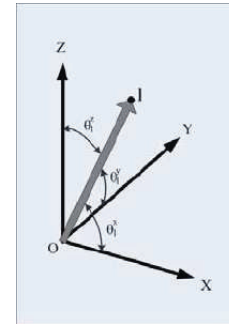


Figure 2: Angles $(\theta_l^x, \theta_l^y, \theta_l^z)$ between the limb l and the axes [8]

CMU motion capture dataset¹ which include 23 categories (running, walking, jumping, varied and so on). The data are joint positions in world coordinate captured using the marker set. The correspondence between the marker set and the joints in our human model is shown in figure 1(b).

Limb orientation is modelled using three parameters, without modelling self rotation of limbs around its axes, as shown in figure 2. As a result, each posture represented using a vector of direction cosines from 12 limbs in each frame, which results in a 36-dimensional representation: $\psi = \{\cos(\theta_1^x), \cos(\theta_1^y), \cos(\theta_1^z), \dots, \cos(\theta_{12}^x), \cos(\theta_{12}^y), \cos(\theta_{12}^z)\}$, where $\theta_l^x, \theta_l^y, \theta_l^z$ are the angles between the limb l and the axes as shown in figure 2. Directional cosines constitute a good representation method for modelling a limb's orientation since it does not lead to discontinuities, in contrast to other methods such as Euler angles or spherical coordinates. Additionally, unlike quaternions, they have a direct geometric interpretation [10].

2.2 The Projected Parameter Space: *aSpace*

The natural constraints of the human body motion lead to highly correlated data in the original space [9]. Therefore, to find a more compact representation, we consider a set of performances corre-

¹<http://mocap.cs.cmu.edu/>

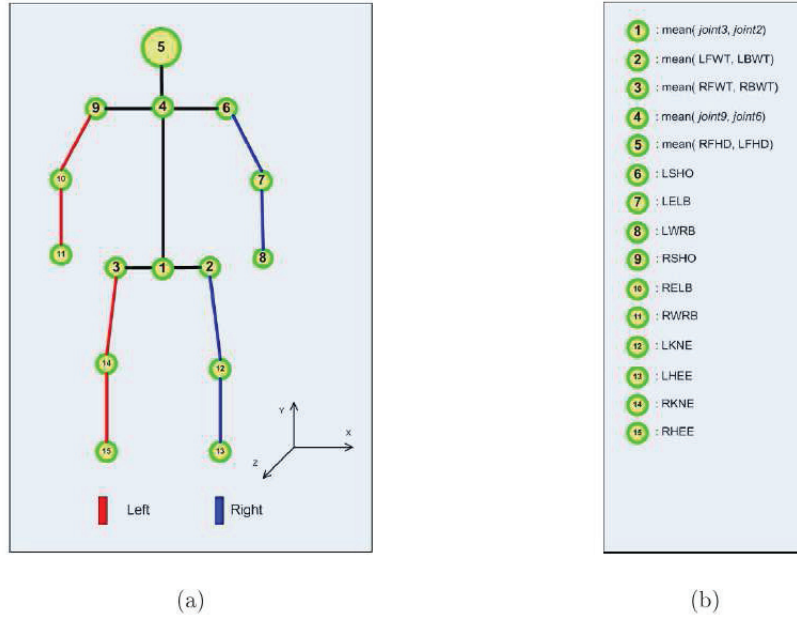


Figure 1: (a) Details of the human body model used, (b) The correspondence between the joint number and the markers [7]

sponding to a particular action A_k , and perform PCA on all the postures that belong to that action. Then we project all the postures to the PCA space:

$$\tilde{\psi} = [\mathbf{e}_1, \dots, \mathbf{e}_b]^T (\psi - \bar{\psi}), \quad (1)$$

where ψ refers to the original posture, $\tilde{\psi}$ denotes the lower-dimensional version of the posture represented in the PCA space, $[\mathbf{e}_1, \dots, \mathbf{e}_b]$ correspond to the first b selected eigenvectors, and $\bar{\psi}$ is the mean of all the postures.

As a result, we obtain a lower-dimensional representation of human postures which is more suitable to describe human motion. Choosing different values for b lead to models of different complexities in terms of their dimensionality. Hence, while the major motion is explained by the eigenvectors corresponding to the bigger eigenvalues, subtle motions require that more eigenvectors be considered.

Figure 3 illustrates the main modes of variation found from PCA and reprojected to the original dimension space. The first (a), the second (b) and

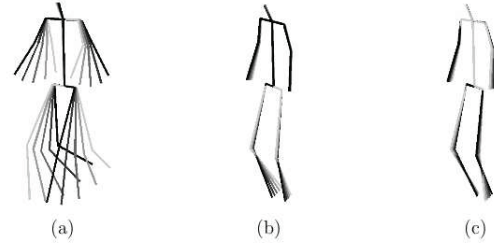


Figure 3: First (a), second (b), and third (c) principal components of the *aWalk* space [8]

the third (c) components of the mean posture are varied from -3 to 3 times the standard deviation found. We observe that the main motion present is related to arms and legs, as the first dimension accounts for the coupled motion between them. Complementarily, the second and the third components encode more subtle motions of legs and arms.

By applying PCA to each action separately (figure 4 shows motion samples for each action), we have *aSpaces* for all actions, walking, running,

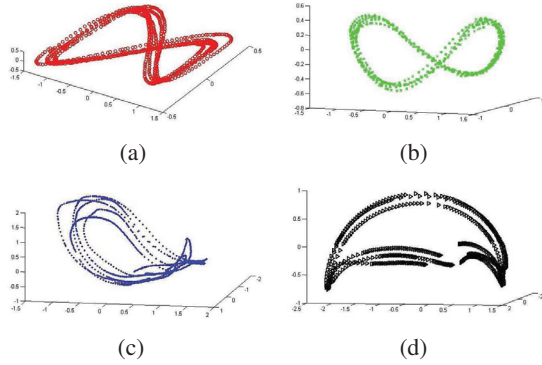


Figure 4: *aSpaces*: (a) *aWalk*; (b) *aRun*; (c) *aBoxing*; (d) *aJump* with motion samples.

boxing and jumping: *aWalk*, *aRun*, *aJump* and *aBox*. Figure 4 shows the mean representation for each action in the *aSpace*. The axes correspond to the first three selected eigenvectors $[e_1, \dots, e_b]$ where $b = 3$ (from equation (1)) of the applied PCA.

The training sequences are acquired under very different conditions. As a result, it is difficult to perform useful statistical analysis to the raw training set since we cannot put in correspondence postures from different cycles of the same action. Therefore, we use the method from [7] to synchronize the training set and establish a mapping between postures from different cycles. We synchronize the training set for each action separately in their own *aSpaces* (see figure 4).

3 Classification

We use a bag of words approach to classify between different actions. In this section, we describe how a vocabulary of key poses is extracted from training sequences and how this is used to classify actions.

The vocabulary consists of the extracted key poses which are representative among all classes. Instead of using only k-means to extract the vocabulary, we explore four other methods. First, we extract key poses in each action, then we concate-

nate the key poses from the four actions to obtain the vocabulary. In this subsection, we explain how to compute key poses.

3.1 Poses Corresponding to Randomly Spaced Frames

We explore a method of extracting key poses from randomly spaced frames of the mean posture in an action [1]. That is, key poses $p^r = \{p_1, p_2, \dots, p_k\}$, where each p_j is randomly selected.

3.2 Poses Corresponding to Equally Spaced Frames

We also explore a method of extracting key poses from equally spaced frames of the mean posture in an action [5]. The are key poses $p^e = \{p_1, p_2, \dots, p_k\}$, where each p_j and p_{j+1} are equally spaced.

3.3 Poses with Local Maximum Distances

Key poses may also be extracted with the maximum distances from the mean of the synchronized training postures of an action [2]. Key poses are defined from a probabilistic point of view: characteristic postures are the least likely body postures of each mean performance.

3.4 The Center Poses of the Clusters

Key poses are extracted as the cluster centers of all the training postures in an action, that is, key poses $p^r = \{p_1, p_2, \dots, p_k\}$, where p_j is the center of the cluster C_j calculated using the k-means methods. In our experiments, the number of the clusters k is also equal to the number of the key poses extracted using the first method: poses with local maximum and local minimum energies.

3.5 Poses with Local Maximum and Local Minimum Energies

In this method, we extract key poses with the local maximum and local minimum motion energies. Given the mean sequence of an action, $\bar{\Psi}_i = \{\bar{\psi}_1^i, \dots, \bar{\psi}_{F_i}^i\}$, the motion energy at the i -th frame is defined as:

$$E_i = |\bar{\Psi}_i - \bar{\Psi}_{i-1}|^2, \quad (2)$$

where $|\cdot|$ denotes Euclidean distance. The energy function is very noisy.

Instead of sliding window method [4], we introduce the scale-space representation to first smooth the energy function globally and then extract the local maximum and the local minimum values. With this multi-scale representation, we calculate the gradient of the energy functions with different scale and the local maximum and local minimum values are the zero crossings of the gradient functions. According to the experiment results, we choose the scale to be 5.5.

After representing the performances using histograms, we use a SVM in training and labelling the test performance.

4 Experimental Results

Here, we test the previously described five classification methods and show the classification accuracy for each method with different scale of noise. To simulate the noisy data extracted from video or images, we add a zero-mean Gaussian noise $\varepsilon(\kappa \cdot \sigma)$ to the performance data with covariance equals $\kappa \cdot \sigma$, where σ is the covariance between the mean performance $\bar{\Psi}$ and all the training performance in an action, and κ is a scale factor controlling the noisy degree. We test our classification method using leave-one-out validation on the unsynchronized performance data.

The pseudo-code of the leave-one-out validation method is as follows:

Algorithm 1 leave-one-out validation

```

for  $in = 0$  to  $50$  do
  for  $ip = 1$  to  $N$  do
    Add noise  $in$  to the  $ip$ -th performance;
    Learn SVM using histograms of all performance data except  $ip$ -th performance;
    Label the test performance;
     $ip \leftarrow ip + 1$ 
  end for
   $in \leftarrow in + 5$ 
end for

```

We test the five methods with κ in the range of $[0, 50]$ with step size of 5. N is 210, the number of all the performances. We set the number of the clusters in the fourth method to be 10. K-means method is a time consuming method in calculating the vocabulary, and more clusters need more time. We set the number of key poses in the first and the second methods to be $N_a/40$, where N_a is the number of frames in action a , and $a \in \{walking, running, boxing, jumping\}$.

The results are shown in Table 1. The first row represents different scales of the added noise, and the second to the sixth rows are the accuracy results. From Table 1, we can see the method using energy have a better performance, so we choose this method to extract key poses.

5 Conclusions and Future Work

We use a 3D stick figure to model the human body and direction cosines to represent the body poses. After representing each motion sequence using a vector of direction cosines, we decrease the dimensionality by applying PCA. Introducing the idea of bag of words, all the performances are represented as histograms of the vocabulary. The vocabulary is obtained by concatenating key poses from different actions. We use a SVM in training and labelling the test performance. We show our action classification method is effective and near state-of-the-art.

Instead of images or videos, we currently only

Table 1: Classification accuracy for with scale κ from 0 to 50 with step size of 5.

Scale	0	5	10	15	20	25	30	35	40	45	50
Random	1.000	0.971	0.943	0.900	0.924	0.943	0.924	0.905	0.881	0.919	0.910
Equal	0.995	0.995	0.967	0.962	0.919	0.957	0.943	0.919	0.948	0.929	0.948
Distance	0.995	0.962	0.924	0.876	0.895	0.914	0.843	0.810	0.829	0.848	0.810
Kmeans	0.976	0.857	0.771	0.757	0.757	0.771	0.733	0.752	0.738	0.767	0.748
Energy	0.995	0.995	0.981	0.957	0.910	0.952	0.962	0.962	0.948	0.933	0.962

consider 3D data. Also, this method works better if we do not need to distinguish between the sequence of the motion. For example, if we consider sit-down and stand-up, this method might not be able to classify between them.

To address these problems and further improve our method, our future work is: First, to define a likelihood function that copes with the images and videos by directly estimating pose; Second, introduce some techniques to classify actions to which the sequence of executing the poses matters.

References

- [1] Jezekiel Ben-Arie, Zhiqian Wang, Purvin Pandit, and Shyamsundar Rajaram. Human activity recognition using multidimensional indexing. volume 24, pages 1091–1104, Washington, DC, USA, 2002. IEEE Computer Society.
- [2] Jordi González. *Human Sequence Evaluation: the Key-frame Approach*. PhD in Informatics, Universitat Autnoma de Barcelona, Barcelona, Spain, 2004.
- [3] Fengjun Lv and Ramakant Nevatia. Recognition and segmentation of 3d human action using hmm and multi-class adaboost. In *9th European Conference on Computer Vision*, volume 4, pages 359–372, 2006.
- [4] Fengjun Lv and Ramakant Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2007.
- [5] Osama Masoud and Nikos Papanikolopoulos. A method for human action recognition. In *Image and Vision Computing*, volume 21, pages 729–743, 2003.
- [6] Kamil Wnuk Michalis Raptis and Stefano Soatto. Flexible dictionaries for action classification. In *In Proceedings of the Workshop on Machine Learning for Visual Motion Analysis*, 2008.
- [7] Ignasi Rius, Jordi González, Mikhail Mozerov, and F. Xavier Roca. Automatic learning of 3d pose variability in walking performances for gait analysis. volume 1, pages 33–43, 2007.
- [8] Ignasi Rius, Jordi González, Javier Varona, and F. Xavier Roca. Action-specific motion prior for efficient bayesian 3d human body tracking. volume 42, pages 2907–2921, New York, NY, USA, 2009. Elsevier Science Inc.
- [9] Vladimir M. Zatsiorsky. *Kinematics of Human Motion*. 1998.
- [10] Vladimir M. Zatsiorsky. *Kinetics of Human Motion*. 2002.

Advances in Variational Optical Flow

Naveen Onkarappa and Angel D. Sappa

Computer Vision Centre, Autonomous University of Barcelona, Bellaterra, Barcelona, Spain

E-mail: naveen@cvc.uab.es and asappa@cvc.uab.es

Abstract Optical flow is the pattern of apparent motion in a visual scene caused by the relative motion between an observer and the scene. Even though the use of optical flow dates back to decades, considering its importance in different applications, it is still one of the present research topics in computer vision. The literature in this topic suggests that variational techniques are producing accurate motion estimates being recently able to perform in realtime. Hence the research in this field is focussed more towards variational techniques. This paper presents a review of development in variational optical flow techniques since its first proposal by Horn and Schunck [1] in 1981. This paper also presents an initial formulation and state of the art in variational optical flow.

Keywords: Motion estimation, Variational optical flow.

1 Introduction

Perception of motion is an important function of the human visual system as well as the artificial vision systems. The estimation of motion information, which is the prominent source of temporal variations in image sequences, is one of the key problems in computer vision. Motion in image sequences acquired by a video camera is induced by movements of objects in a three-dimensional scene and by camera motion. Horn and Schunck [1] define motion field as pure geometric concept, without any ambiguity, is the projection of three-dimensional motion vectors onto a two-dimensional image. On the other hand, optical flow is defined as a velocity field in the image, which transforms one image into the next image in a sequence. The motion estimation, in particular the optical flow estimation

techniques are being investigated and used for many applications such as human action recognition, structure from motion, autonomous navigation, just to mention a few. In addition to the extensions of optical flow towards different applications, a continuous effort to improve the accuracy can be observed in the literature.

Even though many attempts have been made to estimate motion since a long time, concrete formulations were proposed by Horn and Schunck [1], and Lucas and Kanade [2] in 1981. An overview of the developments upto the times can be found in [3, 4]. Further Fleet et al. [5] detail the design and use of linear parametrised models for optical flow. Despite the volumes of research on the topic of optical flow, few attempts have been made to empirically evaluate the performance of optical flow algorithms on complex image sequences. A major obstacle in performing any empirical study in computer vision is obtaining ground truth data. The major previous works in the area of quantitative optical flow evaluation is the work of Barron *et al.* [3] and Otte and Nagel [6]. Barron *et al.* conduct an empirical and qualitative analysis of nine optical flow algorithms. The algorithms are tested on five synthetic image sequences where ground truth motion fields were available and four real image sequences for which no ground truth motion field were available. Otte and Nagel [6] present a good evaluation of their optical flow algorithm by comparing it against others using one of the synthetic sequences from Barron *et al.* Their paper is significant because they actually measured the ground truth motion field for their real sequence and have made the sequence and the motion field publicly available. Galvin et al. [7] evaluate eight different optical flow algorithms among those, two approaches are other than those evaluated by Barron et al. Presently the vision group at Middlebury [8] presents the evaluation results of almost all optical flow techniques including those to be published yet. The different

This work has been supported by projects TRA2007/62526/AUT and CTP/2008ITT00001 and research programme Consolider-Ingenio 2010: MIPRCV (CSD2007/00018).

approaches are rank ordered separately based on different error measures by experimenting on various datasets with ground truth.

Recently McCane et al. [9] propose benchmarking suite of image sequences and tools for the purpose of evaluating optical flow algorithms. They provide a comprehensive set of complex synthetic scenes, a simple interactive method for extracting ground truth motion from real polyhedral scenes, and also three new ground truth motion data sets from such scenes. Baker et al. [10] has proposed few sequences with ground-truth and evaluation methodology recently. Most commonly used error measures to compare optical flow techniques are average angular error (AAE) [3] and average end-point error (or average magnitude of difference) [9]. Some other error measures are interpolation error, normalised interpolation error [8].

Generally, the optical flow techniques are classified [3] as gradient based techniques, correlation based techniques, energy-based techniques and phase-based techniques. Gradient based methods can be classified as local methods and global methods. Local methods optimise local energy functions while global approaches, also referred as variational approaches, attempt to minimise a global energy functional.

There is no recent survey of optical flow methods in the literature. Since there is lot of interest in this topic recently, it is of our interest to review the recent developments in this area. We focus on the class of variational optic flow approaches that give currently the best results in the literature. The paper is organised as follows: the next section details about advantages of variational techniques, its formulation, further developments on initial formulation and very recent advances. Finally concluding remarks are given.

2 Variational Optical Flow

The variational optical flow method is first proposed by Horn and Schunck [1]. This is the first introduction of variational techniques to the machine vision field. There are many other optical flow techniques in literature such as: local differential methods that are based on the same constancy assumptions as variational techniques but minimise local energy-like expressions [2, 11]; feature-based techniques that seek correspondences for sparse but characteristic

image features such as edges or corners [12, 13]; area-based approaches that rely on matching complete image patches by aggregating local information [14, 15]; and phase-based approaches that make use of velocity-tuned filters in the Fourier domain [16, 17]. The variational methods are global differential methods. The main advantage of variational techniques is it generates dense flow field and as demonstrated in recent literature on optical flow [18, 19, 20, 21, 22, 23], the variational techniques offer high accuracy.

2.1 Formulation

The classical variational method of Horn and Schunck [1] is based on two assumptions: the *brightness constancy assumption* (BCA), which is also called as *optical flow constraint*, assumes the grey value of objects remains constant over time and *homogeneous regularization*, which assumes that the resulting flow field varies smoothly all over the image. The BCA can be formulated as:

$$I_1(x + u(x)) - I_0(x) = 0, \quad (1)$$

where I_0 and I_1 is the image pair, $\mathbf{x} = (x_1, x_2)^T$ is the pixel location within a rectangular image domain $\Omega \subset R^2$; $u = (u_1(x), u_2(x))^T$ is the two-dimensional displacement field. Linearising above equation using first-order Taylor expansion we get *optical flow constraint* as:

$$I_{x_1} u_1 + I_{x_2} u_2 + I_t = 0, \quad (2)$$

where subscripts denotes the partial derivatives.

Since optical flow is a highly ill-posed inverse problem, using only local intensity constraints does not provide enough information to infer meaningful flow fields. In particular, optical flow computation suffers from two problems: first, no information is available in un-textured regions. Second, one can only compute the normal flow, i.e. the motion perpendicular to the edges. This problem is generally known as the *aperture problem*. In order to solve this problem it is clear that some kind of regularization is needed. The Horn and Schunck [1] method overcomes this by assuming the resulting flow field is globally smooth all over the image. This can be formulated as penalizing large spatial flow gradients $\nabla_2 u_1$ and $\nabla_2 u_2$. Combining BCA and smoothness assumptions in a single variational framework and squaring both constraints in order to penalize negative and positive derivations in the same way, the following energy functional is obtained.

$$E(u) = \int_{\Omega} \left((I_{x_1} u_1 + I_{x_2} u_2 + I_t)^2 + \alpha (|\nabla_2 u_1|^2 + |\nabla_2 u_2|^2) \right) dx, \quad (3)$$

where α is a regularization parameter.

2.2 Improvements on initial formulation

Since the proposal of first variational method in 1981 [1], a lot of research has been carried out to improve the performance of such techniques. As the variational models consist of data term and smoothness term, the developments in corresponding concepts hereinafter are discussed separately. Also the minimization and solving approaches are discussed.

2.2.1 Data term:

Regarding the data term in the variational approaches, there are two fields of research in the literature, that are *robust data terms* that use non-quadratic penaliser functions to improve the performance in the presence of outliers in the image data [25, 26, 18, 27] and *modified constraints* that allow for a more accurate estimation in the case of varying illumination, large displacements and noise [28, 29, 30, 31, 19].

Robust data terms: Black and Anandan [25, 26] suggested the use of M-estimators from robust statistics. As compared to quadratic penaliser proposed in [1], these functions penalize outliers less severely and thus reduce the influence of corrupted data on the result. The usefulness of M-estimators is also explored by Memin and Perez [18], but solved resulting non-convex optimization problem by an iteratively weighted least square method. Hinterberger et al. [27] investigated similar non-quadratic growth functions for continuous quasi-convex energy functional. Aaubert et al. [32] proposed a robust L_1 norm in the data fidelity term and Zack et al. [21] used dual formulation to solve energy function with L_1 data term. Recently Govindu [33] presented a probabilistic formulation for brightness constraints and demonstrated its superiority over previous methods.

Modified constraints: Many ideas have been proposed in the literature regarding modifying the constraints in the data term. Schnorr [30] proposed suitable constraints with respect to a changing image brightness. Nagel [34] proposed the approximation via a second order Taylor expansion for more accurate estimation of small displacements. Nagel and Enkelmen [29] and Alvarez et al. [31] proposed to use non-linear

constancy assumption for estimating large displacements. Bruhn et al. [19] proposed the integration of local least square fit to improve the performance with respect to noise. The use of blended texture component of the image is proposed by Wedel et al. [22] to get rid of violations in the optical flow constraints due to illumination changes. Papenberg et al. [24, 20] proposed higher-order derivatives based constancy assumption addressing additive or multiplicative illumination changes. The concept of separate robustification, which employs a robust penaliser function to each of the constancy assumptions, has been proposed [19]. Zang [35] proposed dense optical flow estimation based on monogenic curvature tensor where the intensity constraint equation is replaced by the local-phase vector. Color information has also been used. The work in [23] used HSV color space for multiplicative illumination changes and in particular under shadow and shading.

2.2.2 Smoothness Term:

Regarding smoothness term the research in literature is focused on *preserving the motion discontinuities* [28, 29, 36, 25, 37, 38, 39, 40, 41, 32, 42] and the *integration of temporal information* [43, 25, 44].

Discontinuity-Preserving smoothness terms: The first approach to adapt regulariser is proposed by Nagel [28, 29] that orients smoothness constraints inhibiting the filling-in-effect across image discontinuities. Shulman and Herve [36], Heitz and Bouthemy [28], as well as Nesi [39] proposed alternative approaches based on the use of robust statistics in the context of discrete variational approaches. These methods penalize the smoothness term less severely as compared to Horn and Schunck, thus allowing discontinuities in the flow field. Evolving flow-field driven non-quadratic regularisers have been proposed by Wickert and Schnorr [42] in the context of continuous formulation of variational techniques. Their work classifies continuous smoothness terms based on their induced diffusion process such as isotropy, anisotropy in combination with image features and flow features. Similar flow driven discontinuity-preserving smoothness terms have been proposed by Cohen [37] and Aubert et al. [32]. Total variation based regularization is also presented in [21, 22], where they use a dual formulation to solve it. Recently a geometric framework, using Beltrami paradigm based

regularisers, is proposed [45]. A structure and motion adaptive regularization for high accuracy optical flow is also proposed by Wedel et al. [46].

Spatiotemporal regularisers: Murray and Buxton [47] first proposed spatio-temporal smoothness term that use a discrete optical flow method based on spatio-temporal Markov random fields in order to estimate multiple flow fields simultaneously. Nagel also extended his previous work toward spatio-temporal smoothness term [43]. Black and Anandan [25] proposed to use computed flow estimates as prior knowledge to obtain temporally piecewise homogeneous motion field. Later, spatiotemporal concept is adopted to discontinuity-preserving regularisers by Weickert and Schnorr [44] and in [20]. These smoothness terms preserve both spatial and temporal discontinuities in the unknown flow field.

2.3 Minimisation / Solvers

Variational optical flow energy functions can be minimized in a number of ways. The most used way is to express and solve the set of Euler-Lagrange equations of the energy model. The thesis [4] presents various numerical linear and non-linear equation systems solvers such as basic Gauss-Seidel method, its variants, advanced methods such as successive overrelaxation (SOR) technique, unidirectional multigrid methods in the form of coarse-to-fine strategies and bidirectional multigrid methods. Brox et al. [24] proposed to solve the energy function keeping the data term as nonlinear and solving the Euler-Lagrange equations using nested fixed point iterative methods and SOR. Bruhn et al. [48, 49] experiments various solvers on many variational techniques and demonstrates the real-time performance of multigrid methods. Recently Zach et al. [21] proposed a dual formulation based on iterative alternating steps to solve TV-L1 optical flow energy model. Further they also proposed a numerical scheme to implement it on graphic processing units with real-time performance.

2.4 Recent advances:

Apart from the discussions in the previous section about efforts to improve data terms and smoothness terms separately, here we discuss few very recent attempts, which involve either of both data and smoothness terms, or attempts to adapt to a particular application or of a different

interpretation or it makes use of a different concept into variational framework. A variational optical flow based technique for long-range motion estimation using point trajectories has been proposed by Sand and Teller [50]. Wedel [22] improved TV-L1 optical flow proposed in [21] by using blended version of the image texture component and also used a median filter to reject outliers. A geometric framework and a new alignment criterion for optical flow modelling is proposed by Ben-Ari and Sochen [45]. Zimmer et al. [23] proposed an optical flow technique which is the rank one in [8] as on today with regard to angular error and based on the concept of complementarity between data term and smoothness term. Here the data term incorporates HSV color channels with higher-order constancy assumptions, separate robust penalization with constraint normalization, while the anisotropic smoothness term reduces smoothing in the data constraint direction instead of image edge direction, thus enforcing a filling-in effect. Structure and motion adaptive regularization [46] proposed by Wedel et al. is at rank 2 according to AAE, but rank 1 according to end-point error.

A segmentation based variational model for accurate optical flow estimation is proposed by Xu [51]; each segment in the image is constrained by an affine motion and a confidence map is used for global minimization. Brox et al. [52] proposed hierarchical region-mapping based variational optical flow techniques for large displacements. A variational method to segment the image motion in parallel to optical flow computation is proposed by Brox et al. [53]. This method uses level set framework following the idea of motion competition. Sun [54] proposes to learn statistics from both ground-truth optical flow and brightness constancy to formulate a fully learned probabilistic model for optical flow estimation.

3 Conclusion

In this paper, a thorough review of the developments in the variational optical flow techniques since its first proposal is presented. The related work is classified into major subsections such as constancy constraints, smoothness terms and the minimization of the variational model. Further few recent developments with their ranking according to [8] are also mentioned. The literature suggests variational techniques give dense estimates with

more accuracy as compared to other approaches. The recent advances in this field hint that optical flow techniques can be incorporated into many applications with efficiency and effectiveness in real life.

References

- [1] B. Horn, and B. Schunck, "Determining optical flow", *Artificial Intelligence*, 17:185–203, 1981.
- [2] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision". *Proc. IJCAI*, pp. 674–679, 1981.
- [3] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques", *IJCV*, 12(1):43–77, 1994.
- [4] A. Bruhn, "Variational Optic Flow Computation- Accurate Modelling and Efficient Numerics", *PhD Thesis*, 2006.
- [5] D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson, "Design and Use of Linear Models for Image Motion Analysis", *IJCV*, 36(3), 171–193, 2000.
- [6] M. Otte and H-H. Nagel. "Estimation of optical flow based on higher-order spatiotemporal derivatives in interlaced and non-interlaced image sequences", *Artificial Intelligence*, 78:5–43, 1995.
- [7] B. Galvin, B. McCane, K. Novins, D. Mason, and S. Mills, "Recovering motion fields-An evaluation of eight optical flow algorithms", *Proc. BMVC*, pp. 195-204, 1998.
- [8] <http://vision.middlebury.edu/flow/>
- [9] B. McCane, K. Novins, D. Crannitch and B. Galvin, "On Benchmarking Optical Flow", *CVIU*, 84, 126–143, 2001.
- [10] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. "A database and evaluation methodology for optical flow", *Proc. ICCV*, 2007.
- [11] J. Bigun, G. H. Granlund, and J. Wiklund, "Multidimensional orientation estimation with applications to texture analysis and optical flow", *IEEE Trans. PAMI*, 13(8):775–790, August 1991.
- [12] B. F. Buxton and H. Buxton. "Computation of optic flow from the motion of edges in image sequences", *Image and Vision Computing*, 2:59–75, 1984.
- [13] 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.
- [14] P. Anandan. "A computational framework and an algorithm for the measurement of visual motion", *IJCV*, 2:283–310, 1989.
- [15] A. Singh. "An estimation-theoretic framework for image-flow computation", *Proc. ICCV*, pp. 168–177, 1990.
- [16] D. J. Fleet and A. D. Jepson. "Computation of component image velocity from local phase information". *IJCV*, 5(1):77–104, 1990.
- [17] M. Felsberg, "Optical flow estimation from monogenic phase", *1st Intl. Workshop on Complex Motion*, LNCS 3417, 2004.
- [18] E. Memin and P. Perez. "Dense estimation and object-based segmentation of the optical flow with robust techniques", *IEEE Trans. Image Proc.*, 7(5):703–719, 1998.
- [19] A. Bruhn, J. Weickert, and C. Schnorr, "Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods", *IJCV*, 61(3):211–231, 2005.
- [20] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert, "Highly accurate optic flow computation with theoretically justified warping", *IJCV*, 67(2):141–158, 2006.
- [21] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow", *Proc. DAGM*, pp. 214-223, 2007.
- [22] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L1 optical flow computation", *Proc. Dagstuhl Visual Motion Analysis Workshop*, 2008.
- [23] H. Zimmer, A. Bruhn, J. Weickert, L. Valgaerts, A. Salgado, B. Rosenhahn, and H.-P. Seidel. "Complementary optic flow", *Proc. EMMCVPR*, 2009.
- [24] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping", *Proc. ECCV*, pp. 25-36, 2004.
- [25] M. J. Black and P. Anandan. "Robust dynamic motion estimation over time", *Proc. CVPR*, pp. 292–302, 1991.
- [26] M. J. Black and P. Anandan. "The robust estimation of multiple motions: parametric and piecewise smooth flow fields", *CVIU*, 63(1):75–104, 1996.
- [27] W. Hinterberger, O. Scherzer, C. Schnorr, and J. Weickert, "Analysis of optical flow models

- in the framework of calculus of variations”, *Numerical Functional Analysis and Optimization*, 23(1/2):69–89, 2002.
- [28] H.-H. Nagel. “Constraints for the estimation of displacement vector fields from image sequences”, *IJCAI*, vol. 2, pp. 945–951, 1983.
- [29] H.-H. Nagel and W. Enkelmann. “An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences”. *IEEE Trans. PAMI*, 8:565–593, 1986.
- [30] C. Schnorr. “On functionals with greyvalue - controlled smoothness terms for determining optical flow”, *IEEE Trans. PAMI*, 15:1074–1079, 1993.
- [31] L. Alvarez, J. Weickert, and J. Sanchez. “Reliable estimation of dense optical flow fields with large displacements”, *IJCV*, 39(1):41–56, 2000.
- [32] G. Aubert, R. Deriche, and P. Kornprobst, “Computing optical flow via variational techniques”, *SIAM J. Appl. Math.*, 60(1):156–182, 1999.
- [33] V. M. Govindu, “Revisiting the Brightness Constraint: Probabilistic Formulation and Algorithms”, *Proc. ECCV*, Part III, LNCS 3953, pp. 177–188, 2006.
- [34] H.-H. Nagel, “Displacements vectors derived from second-order intensity variations in image sequences”, *CVGIP*, 21:85–117, 1983.
- [35] D. Zang, L. Wietzke, C. Schmaltz, and G. Sommer, “Dense Optical Flow Estimation from the Monogenic Curvature Tensor”, *SSVMCV*, LNCS 4485, pp. 239 – 250, 2007.
- [36] D. Shulman and J. Herve. “Regularization of discontinuous flow fields”, *Proc. Work. on Visual Motion*, pp. 81–90, 1989.
- [37] I. Cohen. “Nonlinear variational method for optical flow computation”, *Proc. SCIA*, vol. 1, pp. 523–530, 1993.
- [38] F. Heitz and P. Boutheymy. “Multimodal estimation of discontinuous optical flow using Markov random fields”, *IEEE Trans. PAMI*, 15(12):1217–1232, 1993.
- [39] P. Nesi. “Variational approach to optical flow estimation managing discontinuities”, *IVC*, 11(7):419–439, 1993.
- [40] M. Proesmans, L. V. Gool, E. Pauwels, and A. Oosterlinck, “Determination of optical flow and its discontinuities using non-linear diffusion”, *Proc. ECCV*, LNCS 801, pp. 295–304, 1994.
- [41] A. Kumar, A. R. Tannenbaum, and G. J. Balas. “Optic flow: a curve evolution approach”, *IEEE Trans. Image Processing*, 5(4):598–610, April 1996.
- [42] J. Weickert and C. Schnorr. “A theoretical framework for convex regularizers in PDE-based computation of image motion”, *IJCV*, 45(3):245–264, 2001.
- [43] H.-H. Nagel. “Extending the ‘oriented smoothness constraint’ into the temporal domain and the estimation of derivatives of optical flow”, *Proc. ECCV*, LNCS. 427, pp. 139–148, 1990.
- [44] J. Weickert and C. Schnorr. “Variational optic flow computation with a spatio-temporal smoothness constraint”, *JMIV*, 14(3):245–255, 2001.
- [45] R. Ben-Ari and N. Sochen, “A Geometric Framework and a New Criterion in Optical Flow Modeling”, *JMIV*, 33:178–194, 2009.
- [46] A. Wedel, D. Cremers, T. Pock, and H. Bischof, “Structure- and Motion-adaptive Regularization for High Accuracy Optic Flow”, *Proc. ICCV*, 2009.
- [47] D. W. Murray and B. F. Buxton. “Scene segmentation from visual motion using global optimization”, *IEEE Trans. PAMI*, 9(2):220–228, 1987.
- [48] A. Bruhn, J. Weickert, C. Feddern, T. Kohlberger, and C. Schnorr, “Variational optical flow computation in real-time”, *IEEE T. IP*, 14(5):608–615, May 2005.
- [49] A. Bruhn, J. Weickert, T. Kohlberger, and C. Schnorr. “A multigrid platform for real-time motion computation with discontinuity - preserving variational methods”, *IJCV*, 70(3), 2006.
- [50] P. Sand and S. Teller, “Particle Video: Long-Range Motion Estimation Using Point Trajectories”, *IJCV*, 80: 72–91, 2008.
- [51] L. Xu, J. Chen, J. Jia, “A Segmentation Based Variational Model for Accurate Optical Flow Estimation”, *Proc. ECCV*, Part I, LNCS 5302, pp. 671–684, 2008.
- [52] T. Brox, C. Bregler and J. Malik, “Large Displacement Optical Flow”, *CVPR*, 2009.
- [53] T. Brox, A. Bruhn, J. Weickert, “Variational Motion Segmentation with Level Sets”, *Proc. ECCV*, LNCS 3951, pp. 471–483, 2006.
- [54] D. Sun, S. Roth, J. P. Lewis, and M. J. Black, “Learning Optical Flow”, *Proc. ECCV*, Part III, LNCS 5304, pp. 83–97, 2008.

Image Description using Local Binary Patterns: Application to Scene Classification

Noha Elfiky* and Jordi González*

* *Computer Vision Center, UAB Campus, Bellaterra, Barcelona, Spain*

E-mail:noha@cvc.uab.es

Abstract

This work presents a novel and efficient scene image representation based on local binary pattern (LBP) texture features. First, we introduce a new image descriptor that represents local image texture based on local binary pattern (LBP) texture features and their spatial layout, together with a spatial pyramid kernel. Second, we generalize the spatial pyramid kernel, and learn its level weighting parameters on a validation set, thus improving classification performance. Finally, we investigate how different cues of image information can be used together. We show that texture, shape and appearance kernels can be combined and that fusing additional information cues increase classification performance. The performance of the proposed method is assessed in the scene classification problem on the Vogel and Schiele dataset under different challenges, and we show that the class specific optimization that we investigate exceeds the state-of-the-art performance by more than 2%.

Keywords: Texture features, Local Binary Patterns, Spatial Pyramid Kernel, Parameter Optimization, Scene classification.

1 Introduction

Classifying images into semantic categories (e.g. coast, mountains, and forest) is a challenging problem of great interest in the computer vision research community. We consider the problem of scene classification where our main goal is to explore the benefit of the spatial distribution of texture.

We introduce a new descriptor which has the advantages of both: it captures the spatial distribution of textons, but is formulated as a vector representation. Similarity on descriptor vectors between two images (for example measured using histogram intersection or ²) then measures the similarity of their spatial distribution of textons. Our descriptor is mainly inspired by two sources: (i) the image pyramid representation of Lazebnik et al. [12], (ii) the Histogram of Local Binary Patterns (LBP) of [1], and (iii) the parameter optimization as in [5]. Basically, we wish to assess how well an exemplar image matches (the texture of) another image. To this end, the structure of the paper is as follows: In section 2 we review existing work of the bag-of-words representation, which form the basis for this work. In section 3 we introduce a new texture descriptor that captures the spatial distribution of the LBP texture features, termed PLBP (for Pyramid of histograms of Local Binary Patterns). section 4 presents the experimental setup. Section 5 discusses the experiments and the experimental results.. Finally, section 6 concludes the paper.

2 Related Work

In recent years we can find in the literature on image classification, an increasing number of proposals which make use of the bag-of-words model representing the images using histograms of quantized appearances of local patches[2]. The first works using the bag-of-words representation can be found in the literature related to texture classification. The goal of these works

is to recognize textures captured from different camera viewpoints, and under varying illumination. [14] quantized responses of a filter bank applied densely over an entire image. These quantizations of appearance descriptors are called texons. And textures are represented by distributions of texons. [11] modified this approach by quantizing small image patches rather than filter responses. [10] address texture classification using quantized affine covariant regions. Ordinary bag-of-words techniques, as the described above, do not take the spatial information into account. However, in complex natural images, image classification systems can be further improved by using contextual knowledge like common spatial relationships between the absolute position of objects in certain scenes [13]. While the above methods have shown to be effective, their neglect of spatial structure ignores valuable information which could be useful to achieve better results for image classification. [12] proposed a method which is based on the spatial pyramid matching of [3].

3 Spatial Texture Descriptor- PLBP

In our work, we use the texture operator called Local Binary Patterns (LBP) which has been successfully applied in the literature for various computer vision problems, such as face recognition and background subtraction. LBP [8] has been shown to be one of the best performing texture descriptors and it has been widely used in various applications, yielding in outstanding results as shown in [7]. It has been proven to be highly discriminative and its key advantages, namely its invariance to monotonic gray-scale changes, it's shown to discriminate a large range of rotated textures efficiently, Moreover its computational simplicity and efficiency as the operator can be realized with a few operations in a small neighborhood and a lookup table, make it suitable for demanding image analysis tasks.

Our objective is to represent an image by its local texture and the spatial layout of the texture. The descriptor consists of a histogram

of local binary patterns over each image sub-region at each resolution level a Pyramid of Local Binary Patterns (PLBP). The distance between two PLBP image descriptors then reflects the extent to which the images contain similar textures and the extent to which the textures correspond in their spatial layout. The following sub-sections will describe these two aspects (local texture and spatial layout correspondence) in more detail.

3.1 Local Texture

The LBP descriptor and its variants use short binary strings to encode properties of the local micro-texture around each pixel. In this subsection we will review the most recent and significant works in the texture literature for a family of related LBP descriptors namely: LBP, $B_{P,R}^{riu2}$, CSLBP, TPLBP and FPLBP.

The LBP operator was originally designed for texture description. The most simple form of LBP is created at a particular pixel location by threshold the 3×3 neighborhood surrounding the pixel with the central pixel's intensity value, and treating the subsequent pattern of 8 bits as a binary number. A histogram of these binary numbers in a predefined region is then used to encode the texture of that region. To be able to deal with textures at different scales, the LBP operator was later extended to use neighborhood of different sizes [8]. $B_{P,R}^{riu2}$ is the definition of so called "uniform" patterns, which are fundamental properties of local image textures and their occurrence histograms are proven to be a very powerful texture feature [8]. These "uniform" patterns, provide a vast majority, sometimes over 90%, of the 3×3 texture patterns in examined surface textures. In the Center-Symmetric Local Binary Patterns, termed (CSLBP) as in [9], the technique encodes at each pixel the gradient signs at the pixel at four different angles. [1] introduce *Three-Patch LBP Codes* (TPLBP), Four-Patch LBP (FPLBP) descriptors, to encode additional types of local texture information by using different ways of bit strings to encode the similarities between neighboring patches of pixels, thus capturing information which is complementary to pixel based ones. In TPLBP codes are produced

by comparing the values of three patches to produce a single bit value in the code assigned to each pixel. In *Four-Patch LBP Codes (FPLBP)* for every pixel in the image, two rings of radii r_1 and r_2 centered on the pixel are examined, and S patches of size spread out evenly on each ring. To produce the Four-Patch LBP (FPLBP) codes two center symmetric patches in the inner ring with two center symmetric patches in the outer ring are compared. One bit in each pixel's code is set according to which of the two pairs being compared is more similar.

3.1.1 Spatial Layout

In order to introduce spatial information to our texture descriptor, termed PLBP, we follow the scheme proposed by [12]. The technique works by partitioning the image into increasingly fine sub-regions and computing histograms of local features found inside each sub-region. The resulting spatial pyramid is a simple and computationally efficient extension of an orderless bag-of-features image representation.

In our case, the PLBP representation of a given image is computed for each grid cell at each pyramid resolution level. The concatenation of all these histograms constitutes the image's signature. This spatially enhanced histogram encodes both the appearance and the spatial relations of the scene regions.

In the spatially enhanced histogram (PLBP), we effectively have a description of the scene on three different levels of locality: The LBP labels for the histograms contain information about the patterns on a pixel level, the labels are summed over regions to produce information on a regional level, and the regional histograms are concatenated to build a global description of the scene. Similarity between a pair of PLBPs is computed using a distance function, with appropriate weightings for each level of the pyramid.

4 Experimental setup

In this section we describe the kernels that will be used in the SVM classifiers for classifying images according to their class. To evaluate the classification performance of our implementa-

tion we use the Mean Average Precision (MAP) score, which is an ideal measure of the quality of classification. High average precision implies that most actual positives are classified as true positives for earlier (more conservative) thresholds.

4.1 Kernel Definition

If images I and J are represented by the PLBP feature vectors S_I and S_J , then the similarity between the images is defined as

$$(S_I, S_J) = \sum_{l=0}^L w_l d_l(S_I, S_J) \quad (1)$$

where w_l is the weight at level l and d_l is the distance between S_I and S_J at pyramid level l . We use the χ^2 on the normalized PLBP descriptors to compute it, as it is demonstrated to be a good distance for histogram comparison [5]. This defines the kernel for PLBP similarity. We then investigate two methods for learning the pyramid level weights as in [5]:

Global Level-Weights (GLW). Instead of giving a fixed weight to each pyramid level as in [12], we learn the weights w_l which give the best classification performance over all categories.

Class specific Level-Weights (CLW). Instead of learning weights common across all classes, the weights w_l are learnt for each class separately by optimizing classification performance for that class using one vs the rest classification.

4.2 Merging Features

$$(x, y) = w_{App} \cdot App + w_{Shp} \cdot Shp + w_{Tex} \cdot Tex \quad (2)$$

Where w_{App} , w_{Shp} and w_{Tex} are the weights for the appearance, shape and texture kernel in Equation (2) respectively. It has the capacity to give higher weights to the more discriminative features during learning. Moreover it also has the capability to ignore features which do not match well as in [5].

4.3 Dataset

For our experiments we have used the Vogel & Schiele (VS) dataset, which includes 7 natural

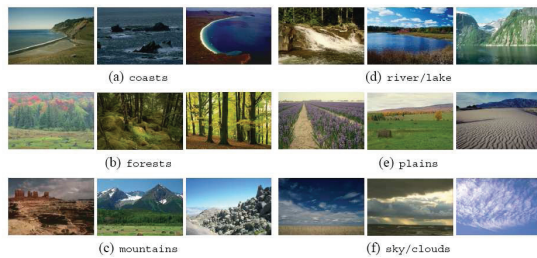


Figure 1: Example Images for Each Category. Figure taken from [13].

scenes consisting of 6 categories: 142 coasts, 111 rivers/lakes, 103 forests, 131 plains, 179 mountains and 34 sky/clouds¹. The size of the images is (720×480) (landscape format) or (480×720) (portrait format). Every scene category is characterized by a high degree of diversity and potential ambiguities since it depends strongly on the subjective perception of the viewer as shown in [13]. Example images for each category are displayed in Figure 1.

5 Experimental Results

This section explains in detail: First, the creation of the newly proposed texture descriptor that incorporates the spatial information showing how the spatial information of texture can help in the scene classification problem. Second, investigate the methodology used to combine the vocabularies of texture, shape and appearance in the kernel level. Experiment 1 examines a family of LBP operators that were discussed so far. Experiment 2 concerns the optimization of the pyramid levels. Experiment 3 explores the effect of different distance measures in SVM. In experiment 4 investigates the effect of shape information. Experiment 5 investigates the effect of appearance information. Finally, experiment 6 explores the effect combining the texture, shape and appearances showing these aspects have complementary information that will help in distinguishing one scene from another, and overcoming the ambiguities that exist between scenes categories.

¹url: <http://www.cs.ubc.ca/~vogel/private/images.zip>

Experiment 1: Examining Different LBP Operators

In this experiment we investigate the performance score using different LBP operators namely, riu^2 , TPLBP and FPLBP to encode the image and preserving the spatial information using a spatial pyramid. The best results are obtained using TPLBP operator and consequently this texture operator will be used from now on to represent our Pyramid of texture (PLBP) over pyramid level $L=2$, see table 1.

Table 1: Mean Average Precision using Different LBP Patch-Based Operators with Histogram Intersection Kernel, $L=2$ and $R=1$.

Method	Parameters	MAP%
B_{riu^2}	$= 8$	73.4
TPLBP	$= 8 \quad = 2$	80.9
FPLBP	$= 6 \quad 1 = 4 \quad 2 = 5$	71.1

Experiment 2: Optimizing PLBP Levels

This experiment shows that the representation of images with just one LBP histogram is not very distinctive between classes(78.5%), while the performance increases(80.9%) when introducing spatial information proposed in [4] at $L=2$, however, it begin to decrease(79.9%) at $L=3$, so the PIBP becomes more discriminative at the pyramid level($L=2$). Consequently, we will represent our Pyramid of texture with TPLBP descriptor over pyramid level $L=2$. Moreover, using CLW then the score increases as far as (80.9% vs. 83%).

Experiment 3: Distance Measures

Fig.2 explore two distance measures: histogram intersection and the χ^2 . The best results are obtained using the χ^2 and the flat fusion (1×3) Spatial pyramid over $L=2$ and consequently this distance is used for the rest of this section.

Experiment 4: Shape Features

We refer to the PHOG descriptor [5] as shape_gray. Then we examined combining a robust color space. We compute the PHOG for each HSV component, we termed this as shape_colour. For VS dataset shape_colour

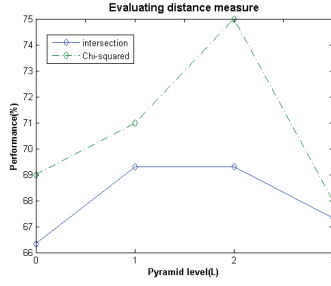


Figure 2: Performance for Different Distance Measures over Pyramid using Texture Separately (TPLBP Patch-Based Operator) over Pyramid Levels L=0 to L=3.

works better than Shape_Gray(77.1% vs 84.5%). If we use CLW then the score increases to 86.0% for shape_color as shown in table(2).

Table 2: Investigating the Effect of Shape.

Method	MAP%
shape_gray	77.1
shape_color	84.5
shape_color_clw	86.0

Experiment 5: Appearance Features

A dense SIFT representation is computed (termed AppGray) as described in [6]. The image can be represented using colour (termed AppColour). For AppColour the SIFT descriptors are computed for each Opponent color space componen. For VS dataset AppColour performance works better than AppGray performance (73.4% vs 79.4%). When we use CLW then the score increases to 85% for AppColour_clw as shown in table (3).

Table 3: Investigating the Effect of Appearance.

Method	MAP%
AppGray	73.4
AppColour	79.4
AppColour_clw	85.0

Experiment 6: Merging Features

By combining different aspects in a flexible manner one could expect to achieve improved performance. We use the kernel in (2) with CLW. We have AppColour as an appearance cue, shape_colour as a shape cue and PLBP as a texture cue. When merging all the kernels representing these cues using CLW, our best performance is achieved when we merge all the cues using CLW and CFW then we obtain the best performance overall: 92% as shown in table(4).

Table 4: Comparison with other methods results using V.S. database [13]. The pLSA, SP and SP-pLSA results are shown in [6].

#cat.	#train	#test	VS	pLSA	SP	SP-pLSA	Ours
6	600	100	74.1	88.8	89.5	89.7	92

6 Conclusion and Future Work

PLBP, which flexibly represents the spatial layout of the local image texture, improves the overall performance significantly. The spatial pyramid shows significantly improved performance on challenging scene categorization tasks. Learning the specific level weights for each class; which we refer to it as CLW enhances the results significantly. We also studied the influence of various descriptors and we have shown that using color features are good for scenes so most of them can easily be recognized by its appearance. The combination of the PLBP, PHOG and PHOW descriptors achieve a significant improvement using the global optimization of the feature kernel; which demonstrates that texture, shape and appearance descriptors are complementary. We then conclude that complementary spatial pyramid based descriptors, together with class-specific optimization of pyramid weights and class-specific kernel selection for merging are all important for good performance. Furthermore, We will study the effect of other classifiers like the random forests (and random ferns) classifier, the advantage of such classifiers (over multi-way SVM for example) is the ease of training and testing. Moreover we will investigate the usage of smart

dictionaries which will allow us to make use of larger vocabulary sizes with trade-off between accuracy and computational time. Finally, we will investigate the usage of motion features as we are working towards image and video retrieval applications.

References

- [1] L. Wolf, T. Hassner, Y. Taigman, "Descriptor Based Methods in the Wild", *Real-Life Images workshop at the European Conference on Computer Vision*, 2008.
- [2] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study", *International Journal of Computer Vision*, 73(2), 213–238, 2007.
- [3] K. Grauman, T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features.", *International Conference on Computer Vision*, 2, 1458–1465, 2005.
- [4] M. Marszalek, C. Schmid, H. Harzallahand, J. van de Weijer, "Learning Object Representations for Visual Object Class Recognition", *Visual Recognition Workshop in conjunction with ICCV 2007*, 2007.
- [5] A. Bosch, A. Zisserman, X. Muñoz, "Representing shape with a spatial pyramid kernel", *Proceedings of the 6th ACM international conference on Image and video retrieval*, New York, NY, USA, 401–408, 2007.
- [6] A. Bosch, A. Zisserman, X. Muñoz, "Scene classification via pLSA", *Proc. at the European Conference on Computer Vision, ECCV*, 517–530, 2006.
- [7] T. Ahonen, A. Hadid, M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition", *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 2037–2041, 2006.
- [8] T. Ojala, M. Pietikäinen, T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7), 971–987, 2002.
- [9] Heikkilä, M. and Pietikäinen, M. and Schmid, C., "Description of interest regions with local binary patterns", *Pattern Recognition*, 42(3), 425–436, 2009.
- [10] S. Lazebnik, C. Schmid, J. Ponce, "A sparse texture representation using local affine regions.", *Pattern Analysis and Machine Intelligence, IEEE*, 27, 1265–1278, 2005.
- [11] M. Varma, A. Zisserman, "Texture classification: are filter banks necessary?", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, 2, 691–698, 2003.
- [12] S. Lazebnik, C. Schmid, J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2, 2169–2178, 2006.
- [13] J. Vogel, B. Schiele, "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval", *International Journal of Computer Vision*, 72(2), 133–157, 2007.
- [14] T. Leung, J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons", *International Journal of Computer Vision*, 43(1), 29–44, 2001.
- [15] D. Lowe, "Distinctive image features from scale invariant keypoints", *International Journal on Computer Vision*, 60(2), 91–110, 2004.

Graph-based representations for Object Recognition

Jaume Gibert and Ernest Valveny

*Computer Science Department, Computer Vision Centre, Campus UAB,
08193 Bellaterra (Barcelona), Spain
E-mail: {jgibert, ernest}@cvc.uab.es*

Abstract

Structural-based pattern representations offer certain advantages over the use of feature vectors in terms of representation. In this work we explore the use of graphs for object recognition proposing three different ways of representing objects using graphs. Two different techniques for graph comparison are adopted: one in the graph domain, the graph edit distance and another one in the vector domain converting graphs into vectors using graph embedding. A comparison between a reference statistical-based technique, the Bag-of-Words, and the proposed graph-based representations constitute the experiments of this work.

Keywords: Object Recognition, Structural Representation, Graph Matching.

1 Introduction

Pattern recognition systems usually make use of feature vectors in order to represent patterns. Vector spaces make possible the use of several analysis tools, mathematically efficient and smart machine learning techniques. However, in some cases, one might be interested in representing structural relations of patterns such as, for instance, relations between parts of an object. This situation is not possible when using feature vectors. Moreover, the

use of vectors for representing patterns is restricted to the representation of patterns using always the same number of features. Graphs-based pattern representations overcome such situation. They offer the possibility of expressing binary relations between parts of patterns and they are not restricted to a predefined size.

With the properties of graphs in mind, in this work we formulate the hypothesis of whether the introduction of structural information in the representation of objects increases the accuracy of a recognition system. In order to prove this hypothesis, we will have to classify the same objects using both statistical-based and graph-based representations. The Bag-of-Words technique will be used as a reference approach for the statistical-based representation. In the case of graphs we propose different ways to represent objects and their classification will be done in two different ways: using a k NN classifier with a similarity measure between graphs, the so-called Graph Edit Distance, and using k NN and SVM classifiers after embedding graphs in a vector space.

In Section 2 we first describe the graph-based representations for the objects. Then in Section 3 the experiments are presented as well as how the classification is done. Also some discussions are done so a better understanding of the representations can be extracted. Finally, some conclusions are outlined in Section 4.

2 Structural Representations of Objects

2.2 Graph-of-Words

In this work, we explore two kinds of graph representations for the problem of object recognition. The first one aims to provide hard structural relations between salient points in a given object while the second one, using feature descriptors, tries to extend the bag-of-words approach codifying the relations between salient points corresponding to each of the words. This section is devoted to the explanation of both representations.

2.1 Harris-based representations

Given an object, a good representation keeping structural relations between points of the image is to join interesting points as nodes of a graph and relate those points with edges under certain conditions. In our case, we pick a set of salient points $H = \{p_i\}$ using the Harris corner detector algorithm. This set, which we will assume to have different cardinals for each object, constitutes the set of nodes V of the graph representing the object. We label such nodes with the (x, y) pixel coordinates of the corresponding points.

Once the nodes are settled, we build the set of edges E of the graph by using a Delaunay triangulation over the points (this representation can already be found in the literature in [1, 2]). The edges will be labelled with the Euclidean distance between the pixel coordinates of the points they are linking. We will refer to this kind of representation as **HR** (Harris Representation).

In order to obtain a richer description of the object and to go one step further with respect to the previous representation, we can also label the nodes by using a feature descriptor of their corresponding salient points. In our case, the fact that we want to recognize objects made us decide to use an appearance descriptor such as SIFT, which is quite a very used feature descriptor for the problem of object recognition. This representation will be called **HSR** (Harris-SIFT Representation).

The backbone of this work is built under the assumption that the addition of structural information to the statistical approaches for object recognition may increase the accuracy of the classifiers used. This is, in principle the system should be able to learn more properties of the classes it deals with. To this end, in this section we propose a new graph representation -closely related to the statistical approach called the Bag-of-Words (**BOW**)- which tries to codify the same information one is keeping in the histogram of visual words in the Bag-of-Words technique, but adding some structural properties. Let us explain this representation more deeply.

The **BOW** technique represents an image by a histogram of visual words [3]. Such words are representatives of the set of feature descriptors that are computed from interesting points of all the images in the train part of the dataset. The closest word in the vocabulary is assigned to each salient point in the image. Finally counting how many times each word appears in the image a histogram of words is built.

Following this idea we create the new graph representation for our objects, the **GOW** (Graph-of-Words) representation. Assume we have our **HSR** representation for an image, we can assign to each descriptor at each point (to each label of each node) the closest word from a previously built vocabulary. Then we can create a graph by taking the appearing words as nodes of the graphs and label them with the frequency of apparition. With this procedure, we already obtain exactly the same information as the histogram of words is providing: we know which words appear in the image and how often they do it. In Figure 1(a) we show an example of the way to build the nodes of the graph.

We still need to define relations between these nodes, we still need to define the edges of the graphs, which will allow us to describe how these words in the input image are related. The proce-

dure is as follows: in the new representation, two nodes (appearing words) are connected when there exists an edge in the **HSR** representation connecting points whose labels are the words in the new representation. For instance, taking Figure 1(b) as an example, if there exists a point with label green connected to a point with label yellow, the green and yellow words will be linked in the new representation. We will label this new edge with the frequency this fact occurs.

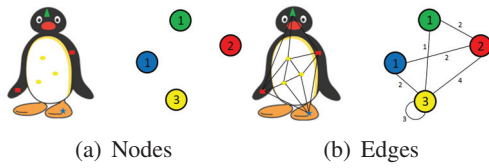


Figure 1: An illustration of the **GOW** representation. For instance, the point on the head (and only this) is assigned to one word. This word appears in the graph representation as a node and its label is 1. The same for the two points on the arms. In the **HSR** representation, such points are linked by the Delaunay triangulation (using two different edges), then, in the **GOW** representation, there exists an edge with label 2 between the words corresponding to the head and the arms.

3 Experiments

This section is devoted to a deepest study of the nature of the representations presented in this work by means of classification rates on a well known database called the COIL-100 database. First, we describe the database. Then we explain how graphs are compared and also a description of the evaluation methods is given. Finally, classification results and discussions are outlined.

3.1 COIL-100 Database

Columbia Object Image Library [4], also known as COIL-100, is a database of 7200 color images of 100 different objects (72 images per object). Objects have a wide variety of complex geometric and illumination characteristics. Images of an

object are taken every 5 degrees of rotation from the whole 360 degrees view ($72 \times 5 = 360$).

Due to the fact that graph matching paradigms are computationally expensive, instead of using the whole dataset we just perform our validation experiments on 8 classes from the existing 100. The selection of the 8 objects we work with has been done using the same criteria as the spirit of the whole database: differences on geometry of the objects and illumination characteristics. A final experiment using the 100 classes is run using the obtained parameters¹ for the 8 selected classes.

For the experiments, we also need to create the training, the validation and the test sets. In this aspect, we have followed the same protocol as in [1]. From the 72 images of every class, 24 constitute the training set (one image every 15 degrees of rotation). From the 48 remaining images per class, 5 are randomly selected for the validation set and 10 for the test set. This leads to a training set of 192 images, a validation set of 40 images and a test set of 80 images.

Finally, let us talk about the processing we have performed to the images and which is common to all the graph-based representations and to the Bag-of-Words approach. The set of Harris corner points is first extracted, then SIFT descriptors are computed around each of such points. The vocabulary of visual words is calculated by clustering all the descriptors of all the points in the training images using the *k*-Means algorithm. Finally, words are assigned to points by taking the nearest neighbour (1NN) among all the words in the vocabulary.

3.2 Graph matching

Graph matching is the process of evaluating the structural similarity of two graphs. To do so, we have used two different approaches: Graph Edit Distance and Graph Embedding in vector spaces. *Graph Edit Distance* defines the similarity between two graphs by the minimum amount of distortion

¹Graph Edit Distance has parameters related to the costs of edit operations on graphs.

that is needed to transform one graph into another. It is quite an interesting approach because its flexibility and adaptability to different kind of problems and graph representations.

Graph Embeddings associate a vector to each graph in order to embed graph in a vector space so all the existing original vector-based techniques become available for the graph domain, but trying to keep, at the same time, the properties of the structural representation. We have used an embedding based on Graph Edit Distance [5] by which each graph is embedded into a vector space, by first, selecting a set of prototypes and then computing the distance between each graph to each of the prototypes. All these distances are though as components in a n -dimensional vector space, where n is the number of selected prototypes.

3.3 Performance evaluation

The classification of graph-based representations is not a trivial issue due to their complex structure. In the graph domain, we are restricted to the classification of graphs by means of the k -Nearest Neighbour classifier (k NN) which consists in assigning to a query graph the class that most frequently occurs among the k nearest neighbours from a training labelled set of graphs (in our case, the neighbouring relations are based on the graph edit distance paradigm).

Once the query graphs will be labelled, we will evaluate the power of our graphs-based representations by counting how many of the objects have been correctly classified out of the whole set of testing objects. This measure is called the *Accuracy rate*.

For the sake of integrity of the work, another classifier will be tested. Support Vector Machines (SVM) is a really well known statistical technique that has been gaining popularity during the last years due to its success in real world applications [6]. Its main idea is to separate pattern classes in \mathbb{R}^n by means of hyperplanes. Using such statistical based technique forces us to embed graphs in

a vector domain. There, a classification using a SVM but also a k NN classifier will be performed; their evaluation will also be done by means of accuracy rates.

3.4 Working in the graph domain

After tuning parameters for the edit cost functions and the k NN classifier using the validation set, we can run the experiments on the test set and see which are the results we are able to obtain up to this point. As we work in the graph domain (no embedding is yet performed) the only classifier that can be used is, as we have said, the k NN technique. Table 1 shows which are the obtained accuracy rates.

<i>k</i> NN Classifier	
HR	93.75%
HSR	100%
GOW	97.5%
BOW	87.5%

Table 1: Accuracy rates for the graph representations working in the graph domain. Results also for the reference system **BOW**.

As we can see, all the graph-based representations significantly outperform the accuracy rate of the Bag-of-Words using the k NN classifier. This is something that confirms our main hypothesis in this work, which was that the addition of structural information would help us in the problem of object recognition, where in principle relations about part of the objects are an important issue to take into account in the representation of the objects themselves.

3.5 Working in the vector domain

The next step forward is to use a different classifier (such as SVM) for both the vector and the graph-based representations. As we have explained in Section 3.2 we have embedded graphs into a vector space using the embedding defined in [5]. We

have have used two different sets of prototypes for the embedding².

The first set of prototypes is exactly the training set of graphs (corresponding to 192 images), which leads to embed graphs into a 192 dimensional euclidean space, \mathbb{R}^{192} . The second set of prototypes is to take half of the graphs in the train set, in particular, images every 30 degrees of rotation for each object. In this case, the graphs are embedded into \mathbb{R}^{96} . After tuning parameters for the classifiers using the validation set, results are as follows. Table 2 shows results after embedding graphs into \mathbb{R}^{192} while Table 3 shows results for \mathbb{R}^{96} .

<i>k</i> NN Classifier		SVM Classifier	
HR	83.75%	HR	86.25%
HSR	98.75%	HSR	100%
GOW	83.75%	GOW	77.5%
BOW	87.5%	BOW	92.5%

Table 2: Accuracy rates for the graph representations working in \mathbb{R}^{192} . *Linear* kernels have been used for the SVM classifier in the case of the graph representations. *Radial basis function* kernel for the case of the reference technique **BOW**.

<i>k</i> NN Classifier		SVM Classifier	
HR	72.5%	HR	88.75%
HSR	87.5%	HSR	100%
GOW	77.5%	GOW	78.75%
BOW	87.5%	BOW	92.5%

Table 3: Accuracy rates for the graph representations working in \mathbb{R}^{96} . *Linear* kernels have been used for the SVM classifier in the case of the graph representations. *Radial basis function* kernel for the case of the reference technique **BOW**.

Several things can be said about these results we have just shown. The most important one to be remarked seems to be that the **HSR** representation

²The sets of prototypes we selected are as simple as possible. We leave for future research the selection of more complex sets as in [5].

is able to classify correctly all the query objects in the test set when using a SVM classifier, and all but one when using the *k*NN. This thing also happened when working in the graph domain. This is telling us about the strength of this representation, which, however, is the one demanding more computation time. It is also worth noticing the fact that we are always above the **BOW** technique using this representation. Even though the **GOW** representation presents lower results, it is, as we already said, less time consuming than the other representations. This is a strong point to take into account.

From results in the graph domain (Table 1), in general, the accuracy of the classifiers decreases. This makes total sense if we consider the fact that when embedding graphs into a vectorial space, some information is necessary missed in the transition. More proper selections of the prototypes for the embedding and even different embedding procedures might fix this problem.

3.6 Using the whole database

Up to this point, we have just used 8 classes out of the 100 existing ones in the COIL-100 database. We have justified this fact saying that graph matching paradigms, Graph Edit Distance in our case, are computationally expensive and, therefore, slow. Due to time issues, optimizing parameters for the whole database is something which, unfortunately, has to remain beyond the scope of this work. However, we can give it a last chance and check how everything works using the parameters we have got for the 8 selected classes.

So, using exactly the same protocol for the training and the test sets (no validation set is used any more because no parameter will be optimized), right now we deal with 2400 train images and 1000 test images. Leading to, by just working in the graph domain, the results shown in Table 4.

As we can see, the results for the reference system **BOW** and for the **HR** representation keep being more or less similar (still the parameters optimization is needed so the results showed in [1] can

<i>k</i> NN Classifier	
HR	75.7%
HSR	98.8%
GOW	41.3%
BOW	76.1%

Table 4: Accuracy rates for the graph representations working in the graph domain and using 100 classes. Results also for the reference system **BOW**.

be reached). The results for the **GOW** representation are really low, which make us think on the real need to tune the parameters for this representation as well as to define the cost function in a different manner.

On the other hand, we can see how the results for the **HSR** representation are still awesomely high. This is telling us that, besides the fact that probably the parameters used are already the optimum ones, the introduction of an appearance descriptor like SIFT gives important information with respect to the **HR** representation where this kind of features are not considered.

4 Conclusions

We started this paper by considering whether the introduction of structural information in the representation of objects may increase the accuracy of a recognition system. For that three different graph-based representations have been introduced. A classification of these representations as well as a statistical reference approach for the objects have been done. The conclusions are summarized as follows.

In the graph domain when using 8 classes all the representations overcome the **BOW** technique which confirms our main hypothesis. The fact that the **GOW** representation is better than the **BOW** is saying that not only the frequency of appearing words is important but also how these words are spatially related. When embedding graphs in the

vector domain, apart from losing information, results of the SVM classifier are shown to be better than the ones for the *k*NN, so it is worth investigating embedding techniques, and even other techniques than the one used here which still demands lot of computational time.

The good results for the whole dataset in the **HSR** representation reinforce our hypothesis but still some work needs to be done. Also other problems need to be investigated using more challenging datasets and apply the **GOW** representations to different problems such as scene classification.

References

- [1] K. Riesen, H. Bunke, "IAM Graph Database Repository for Graph Based Pattern Recognition and Machine Learning". In *SSPR & SPR '08*. Berlin, Heidelberg, pages: 287-297, 2008. Springer-Verlag.
- [2] B. Luo, R.C. Wilson, E.R. Hancock, "Spectral embedding of graphs". *Pattern Recognition*, 40(3):1042-1056, 2007.
- [3] C. Dance, J. Willamowski, L. Fan, C. Bray, G. Csarka, "Visual categorization with bags of keypoints". In *ECCV International Workshop on Statistical Learning in Computer Vision*, pages 1-22, 2004.
- [4] Nayar and H. Murase, "Columbia Object Image Library: COIL-100". *Technical Report CUCS-006-96*, Dept. of Computer Science, Columbia University. February 1996.
- [5] K. Riesen, H. Bunke, "Recent Developments in Graph Classification and Clustering using Graph Embedding Kernels". In *Pattern Recognition in Information Systems*, pages: 3-13, 2008. INSTICC PRESS.
- [6] J. Shaer-Taylor, N. Cristianini, "Kernel Methods for Pattern Analysis". Cambridge University Press, June 2004.

Semantic Segmentation of Images Using Random Ferns

Josep M^a Gonfaus, Jordi González, Theo Gevers

Centre de Visió per Computador, UAB, Barcelona

E-mail: gonfaus@cvc.uab.cat, poal@cvc.uab.cat, gevers@cvc.uab.es

Abstract

Object-based segmentation is a challenging topic. Significant advances have been recently made by combining local, global and contextual features. Although most recent work merges this information using Conditional Random Field (CRF), in this approach we investigate a faster classification method called Random Ferns. The image segmentation is performed in 3 steps; first, we learn the local appearance of objects from the patches around each pixel; second, based on result of appearance extraction, we learn contextual features that constrain the recognition; and third, using the image as a whole, an image prior is computed. Finally, all steps are merged in a Bayesian framework.

Keywords: Random Ferns, Random Forests, Semantic Segmentation of Images

1 Introduction

An ongoing topic in computer vision is image understanding, where the aim is to know what is happening in an image. This reasoning can be split into two different steps; first, what appears in the image must be detected (objects, animals, humans or even regions), and then by using a high-level reasoning method, infer what these instances are doing, or even more difficult, predict why they are doing it. However, since the first step is still not well solved, the second step becomes much

harder. Hence, we focus on the recognition problem. Since there are many difficulties to overcome in the recognition step, as much information as possible must be used.

We can approach the recognition problem from different points of view. For example, from a wide overview of the image, the *classification* task tries to decide whether an image contains at least one instance of a specific object or not. The Bag-Of-Words technique has emerged as the most suitable for this task. Although it has achieved good results, this framework cannot give us much information about what is happening in the scene. The next step is usually to determine the position of object instances in the image; this task is called *detection*. Generally, the candidate detections are represented by means of bounding boxes, which fit a rectangle around the detected objects. However, region classes such as sky, road or grass are not well characterized by rectangles, due to the unconstrained nature of their shape. Hence, another kind of representation is required. An alternative to detection is pixel-wise labeling of images, commonly known as *segmentation*. This approach has been the focus of interest of many recent publications [4, 10, 11, 12, 13, 15, 16], and is becoming one of the “hot” topics in the field.

This work faces the recognition problem by classifying each pixel into its semantic label. That is, assigning to each pixel a label describing to which particular class it belongs. Given an unseen image, a new semantic image is created according

to the predicted labels. We restrict our investigation to a small number of predefined classes like grass, trees, water, buildings and sky (which could be understood as regions), and a few object and animal instances like cows, dogs, cars, bicycles and chairs.

2 Related Work

Although there is a large literature on image segmentation, dating back over 30 years, great improvements have been recently achieved. During the initial stages, the work was focused in splitting the images into regions that shares some features [3, 6]. However, nowadays the effort is concentrated in divide the contents of an image into semantic regions [4, 10, 12, 15, 16].

In this work we evaluate our method with the MSRC-21 dataset, which has been widely used in the literature [13, 15, 10, 16]. The latest improvements come from the fusion of multiple local cues, which are able to encode different aspects of the objects. Moreover, a further enhancement is due the use of more contextual features which adds a semantic relation between the regions [12, 4].

Most of the techniques are based on the probabilistic frameworks of the Conditional Random Fields (CRF). In that way, it is easy to force coherence within the regions. However, the use of this techniques in a pixel level is computationally expensive and the use of some unsupervised pre-processing techniques becomes mandatory [3, 14].

Recently, Random Forests is becoming a popular classification method. Its main strengths are its efficiency (both in training and testing time) and that it allows to incorporate multiple cues easily for the classification step. It has been lately used in many applications: keypoint recognition [5], clustering method [7], image classification [1] and semantic segmentation [12]. It shows state-of-the-art results within a very short time of computation.

Based on the idea of Random Forests, in [9] extends the ensemble to what they call Random

Ferns. This ensemble is faster to compute and simpler to code while obtaining similar accuracy. The Random Ferns method was first used in the same way as [5] for keypoint recognition, so combining the work of [12] and [9], for image segmentation, is a logical next step.

3 The Use of Random Ferns

One of the main advantages of Random Ferns is the required time for training and testing, which is drastically reduced with respect to other state-of-the-art methods, such as SVM.

3.1 Random Forests

The use of Random Forests (RF) was motivated by the observation that Decision Trees can cause overfitting on the training data [2]. Instead of using only one tree, an ensemble of T trees is used. The final output is calculated by combining the information of the output of each singular Tree.

The use of multiple trees gives the name of “Forest” to the ensemble. However, the “Random” part of the name comes from the randomized way to choose the questions in the splitting nodes. Contrary to Boosting, where all the possible questions (features) have to be pre-calculated before choosing the best ones, in RF the questions are chosen and evaluated on-the-fly during training. Usually, each node creates a set of random questions and the best one is selected. The fact that not all the possible questions have to be pre-calculated reduces the computational time drastically, as well as the memory requirements. Moreover, using this scheme the size of the feature space is not directly related to the effort needed to train the classifier.

One of the requirements when using ensemble methods (such as Random Forests or Boosting) is that each classifier be accurate, but also different from the rest. If any of these requirements is not satisfied, the ensemble becomes useless. There are some techniques to force this diversity. Boosting

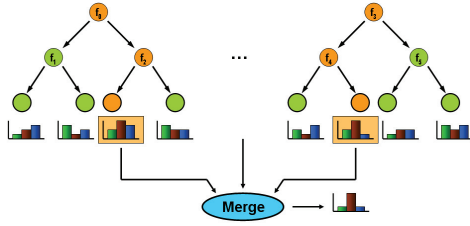


Figure 1: **Random Forest Framework.** A query is passed down in each tree. Once each tree reaches a leaf, the final output is obtained by merging their stored information.

weights misclassified examples by putting more emphasis on them. Bagging uses a random subset of the data for each classifier, training each one of them with different data, making them more diverse and discriminative as a whole. In the case of Random Forests, this diversity comes implicitly since the features used are randomly chosen in each step. However, some works also use the bagging technique with Random Forests in order to be even more divergent [12].

3.2 Random Ferns

Based on the idea of Random Forests, [9] extended the ensemble, making it faster to compute and simpler to code while obtaining similar results. The main contribution comes from the idea that when the questions are randomly chosen, it does not matter which ones are selected. Following this assumption, Random Ferns use the same question for each level of the tree (now called ferns). As can be seen in Fig. 2, this allows us to convert the tree structure into a look-up table based on the answers to each question. This implies that for each fern there are as many questions as the tree is deep, but also that the questions are independent of the answers at previous levels. Considering a certain order of questions, it allows the use of a binary code that indexes a certain leaf (or bin).

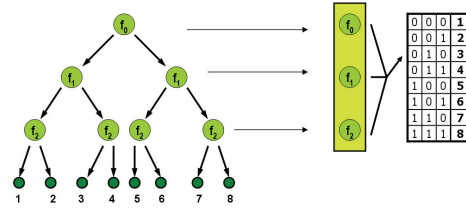


Figure 2: **Random Ferns structure.** Since the questions are randomly chosen, a tree can be transformed into a fern by constraining the tree to systematically perform the same test across all the nodes at the same level. Therefore, the hierarchical structure can be converted into a look-up table.

4 Learning the Model

This section describes our image segmentation approach. As described above, Random Ferns are the core of the method, but as important as taking advantage of them is the way to use them. If accuracy and speed are necessary, it is also required to work with speedy splitting nodes. In Fig. 3 the global schema is presented. The method is based on merging local, contextual and global information. This is done by three independent steps; the first looks for the local appearance, and the second merges the contextual information, looking for common patterns. For example, the cow is usually on the grass. Finally, a third step takes care about the whole image and computes a prior that emphasize the likely categories and discourage the others.

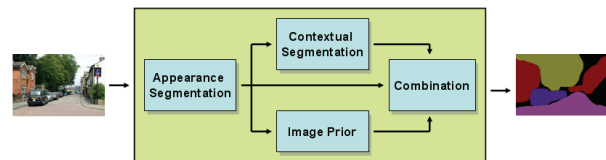


Figure 3: **Full model of segmentation algorithm.** First of all, a patch-based appearance method gives a probability to each pixel to belong to each class. Afterwards, this information is used by the contextual segmentation which enforces a spatial coherence between the classes. Finally, an estimated image prior is fused with the appearance and the contextual information to obtain the segmented image.

4.1 Appearance Segmentation

Many methods have been developed to describe patches, whether characterizing the color, the shape or the texture. However, since describing each pixel with its surrounding patch is computationally expensive (due to the many times that the process is repeated), the method must be as fast as possible. Considering this assumption, common techniques such as SIFT, HOG or LBP are not suitable for a pixel-wise framework. Therefore, the solution requires quick features, being the simplest ones, the pixels themselves, where no preprocessing is required. The appearance method is based on basic operations such as differences of pixels, or thresholds over the pixels values, and by using integral images, the method can be extended to regions. Moreover, the operations can be done in any of the three channels, so the Ferns are able to learn the shape and also color aspects.

During the training step, the method chooses the operation to perform over the patch randomly. After evaluating all the tests, each Fern has learnt different cues, such as having the upper part brighter than the lower part, or having more blue components than green. Despite the simplicity of the questions, by combining these trivial tests the method is able to distinguish between smoothed and sharper patches, or colourful and flat patches. In Fig. 4 a toy example shows how the method works. It can be seen that patches that fall in the same leaf share similar appearance.

During the test step, each patch around each pixel is passed across the test nodes of each Fern. Since there is not any dependency between the tests, the evaluation can be done independently for all the patches, but also for the tests of the same Fern. In this way, the required time is significantly reduced. Once the tests have been evaluated, each binary output leads to a probability distribution for each Fern. Merging this information using the Semi-Naive Bayesian approach, each pixel obtains a probability of belonging to each class. Finally, the class with highest is assigned.

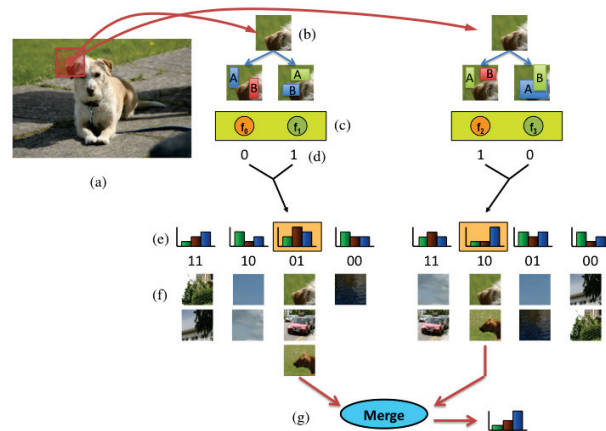


Figure 4: **Appearance patch segmentation example.** The original image (a) is split in several patches (b). In this case, the test nodes (c) look if the region A is greater than B. The binary output leads to the corresponding leaf (d). Each leaf has its learned probability distribution (e), which has been calculated during the training step from the patches (f). Finally, the output is computed by merging the individual output of each fern (g).

4.2 Contextual Segmentation

Since the local appearance of pixels is not enough to correctly classify all the pixels, more information is required. In fact, this information has not been taken into account in the appearance segmentation. Moreover, it is also interesting to note that humans also uses this behavior, as presented in [8]. In our case, by looking in the neighborhood of an object, the method is able to learn what is reasonable to find around it, but also what is not. For example, a boat surrounded by *sky* or *grass* (at least in the general case) is not usual, so the boat class decreases its likelihood to be found in that location.

The learning method for this second step consists of another bunch of Ferns. The main difference is in the splitting nodes, which are now based on semantic information. Thus, instead of looking for pixel values, the method looks for the predicted probability of each class provided by the appearance segmentation. The training procedure follows the same schema as the previous level. During training time, each splitting node chooses a region to be tested relative to the pixel. However, in this case, in spite of computing differences of pixels

and regions, the splitting nodes look for the probability to find a certain class in a region that surrounds the current pixel. This is done by choosing 4 random parameters: the relative location to the pixel, the size of the region, the class to be evaluated and a certain threshold. Then, the output of a node is computed by evaluating if the mean probability inside the region of a certain class is greater than the threshold.

4.3 Combining appearance and context

Since the output of the contextual segmentation is quite smooth at the borders of objects, the accuracy is not as good as it could be. Knowing this drawback and extending [12], our method merges both information sources. Given that we are working in a probabilistic framework, for each pixel we use these two probability distributions (from the appearance and the context). Then, the final output is obtained by weighting both probabilities.

4.4 Image Prior

During the previous stages, most of the information is extracted from the local appearance of a pixel. Though contextual segmentation is looking beyond the local patch, it cannot be considered global information. From a glance at an image, humans are able to get an idea about what is expected to be in this image. Our goal is to also exploit this information in a similar manner. From a global view of the image, we can use all the features as a whole, and obtain a naive idea about what we can expect to find. This information can be understood as an *image prior*. We compute this information with a Bag-of-Words schema, similar to the existing *classification methods*.

5 Experiments

After evaluating our method, we have compared it with the latest segmentation methods. In Table 1 we can see a fairer comparison, both in accuracy

	Global	Average	Time
[4]	77 %	64 %	32 s.
[12]	72 %	67 %	0.6 s.
[13]	71 %	58 %	180 s.
[16]	75 %	62 %	60 s.
Appearance	44 %	30 %	0.2 s
Context	54 %	46 %	0.5 s
App + Ctx	59 %	50 %	0.5 s
Full Model	63 %	54 %	0.6 s

Table 1: **MSRC-21 segmentation results.** Comparison with the latest state-of-the-art methods.

and time. Regarding to the accuracy results, we can notice that each step in the method improves the results. However, with this simple schema, it is not enough to reach the same results of the other methods. One of the main reasons is that our method does not take into account the whole shape of the object, which is quite important for some of them. Despite do not overcome the other methods in accuracy, the required time to process an image is much lower, even though we are using an un-optimized Matlab code. Finally, we can see some results of the method in Fig. 5.

6 Conclusions

We have investigated the use of Random Ferns for semantic segmentation of images, which has not been previously investigated. This classification method is based on splitting the feature space using random tests. Given the nature of the randomness, each test is independent of the others, so all of them can be evaluated in parallel. However, we show that since the difficulties of the problem are so hard, the simple random implementation is not enough to achieve acceptable results. Though we are close to the latest state-of-the-art results, we think that a significant improvement can come if we learn the ferns in a class-specific way.

We also note, that each step of the framework improves the semantic overall segmentation.

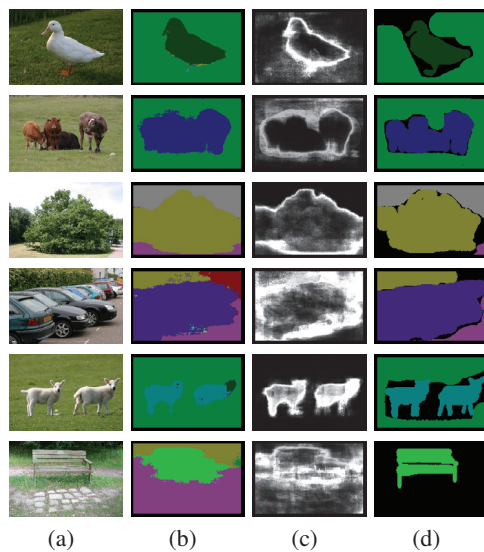


Figure 5: **Visual results of the full framework.** The original image (a) and our segmentation. We can also appreciate the confidence of the segmentation in (c) and what it is labeled in the ground truth in (d).

Therefore, by adding a new step in the segmentation framework which takes the shape of the objects into account would also improve the results. Similar to the image prior step, which can be obtained with another *categorization* method, this shape-based step is related to a *detection* method. Hence, the use of a more localized prior will help in the recognition of classes like objects, where shape is a very important cue.

References

- [1] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Image classification using random forests and ferns. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007.
- [2] Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.
- [3] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002.
- [4] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. Multi-class segmentation with relative location prior. *Int. J. Comput. Vision*, 2008.
- [5] V. Lepetit, P. Lagger, and P. Fua. Randomized trees for real-time keypoint recognition. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005.
- [6] D. R. Martin, C. C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2004.
- [7] Frank Moosmann, Bill Triggs, and Frédéric Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Advances in Neural Information Processing Systems 19*, 2006.
- [8] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 2007.
- [9] Mustafa Ozuysal, Michael Calonder, Vincent Lepetit, and Pascal Fua. Fast keypoint recognition using random ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009.
- [10] Caroline Pantofaru, Cordelia Schmid, and Martial Hebert. Object recognition by integrating multiple image segmentations. In *ECCV '08: Proceedings of the 10th European Conference on Computer Vision*, 2008.
- [11] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Object class segmentation using random forests. In *British Machine Vision Conference*, 2008.
- [12] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [13] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 2009.
- [14] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.
- [15] J. Verbeek and B. Triggs. Region classification with markov field aspect models. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007.
- [16] Lin Yang, Peter Meer, and David J. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

Ranking Error-Correcting Output Codes for Class Retrieval

Mehdi Mirza-Mohammadi, Francesco Ciompi, Sergio Escalera, Oriol Pujol, and Petia Radeva

Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Spain UB

Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain

Abstract

Error-Correcting Output Codes (ECOC) is a general framework for combining binary classification in order to address the multi-class categorization problem. In this paper, we include contextual and semantic information in the decoding process of the ECOC framework, defining an ECOC-rank methodology. Altering the ECOC output values by means of the adjacency of classes based on features and class relations based on ontology, we defined a new methodology for class retrieval problems. Results over public data show performance improvement when using the new ECOC-rank in the retrieval process.

Keywords: Retrieval, Ranking, Error-Correcting Output Codes.

1 Introduction

Information Retrieval deals with uncertainty and vagueness in information systems (IR Specialist Group of German Informatics Society, 1991). This information could be in forms such as text, audio, or image. The science field which deals with information retrieval in images is called Content-based image retrieval (CBIR). CBIR corresponds to any technology that helps to organize digital picture archives by visual content. In this sense, any system ranging from an image similarity function to a robust image annotation engine falls under the purview of CBIR [1].

In last decade, many work has been performed to describe color, shape, and texture features, with-

out considering image semantics. Most of these works are based on the retrieving of samples from the same category. However, in our case we address the class retrieval problem. Suppose we have an image from a cat animal category. In that case, we want to retrieve similar categories and not just samples from the same animal (e.g. tiger could be a possible solution). In order to deal with this problem, we can use on the output of multi-class classifiers to rank classes and perform class retrieval. In particular, we focus on the Error-Correcting output codes framework, which combines binary classifiers to address the multi-class problem.

Up to now, the ECOC framework has been just applied to the multi-class object recognition problem, where just one output label was required. Based on the ECOC framework, we extend this methodology to address class retrieval problems. Altering the ECOC output values by means of class adjacency matrix based on features and class relations within an ontology matrix, we alter the ECOC output ranking. This new ranking is used to look at the first retrieved classes to perform class retrieval. The results of the new ECOC-Rank approach show that performance improvements are obtained when including contextual and semantic information in the ranking process. The rest of the paper is organized as follow. In section 2 we define the ECOC rank and the proposed alteration methods. Section 3 presents experimental results. Finally, section 4 concludes the paper.

2 ECOC Rank

Retrieval systems retrieve huge amount of data for each query. Thus, sorting the results from most to less relevant cases is required. Based on the framework and application, there exists different ways for ranking the results based on the associated criteria.

In the decoding process of the ECOC framework [2], a "distance" associated to each class is computed. This "distance" can be then interpreted as a ranking measure. However, this ranking is the most trivial way for sorting the results. Moreover, the output of the ECOC system does not take into account any semantic relationship among classes, which may be beneficial for retrieval applications. As an example of an image retrieval system, suppose the query of "Dog". In the feature space, it is possible that there exists high similarity between "Dog" and "Bike", so based on features, the ranking will be higher for "Bike" than for some other class which can be semantically more similar to "Dog", such as "Cat". On the other hand, it is easy to see that similarity based on features also is important, and thus, a trade-off between appearance and semantics is required. In order to embed class semantic and contextual information in the ranking process, we define two matrices that will be used to vote the ranking process: one based on adjacency and another one based on ontology. These matrices are $n \times n$ matrices for n number of classes, where each entry represents the similarity between two classes. By multiplying the ranking vector of the ECOC output by these matrices, we alter the output ranking and improve retrieval results. The rest of this section describes the design of the class adjacency matrix, ontology matrix, and their use to modify the output ECOC rank.

2.1 Adjacency Matrix M_A

There are different approaches in literature for measuring the similarity between two classes. Support Vector Machines margin and the distance between cluster centroid are two common ap-

proaches. Here, we follow a method similar to the second approach. However, just considering the cluster centroid would not be an accurate criteria for non-gaussian data distributions. Instead, we re-cluster each class data into a few number of clusters and measure the mean distance of centroid of the new set of representant.

Since the objective is to alter the ranking, the defined adjacency matrix should be converted to a measure of likelihood, which means that the more two classes are similar, the more the new measure among them should be higher. Thus, we compute the inverse of the distance for each element and normalize each column of the matrix to one to give the same relevance to each of the classes similarities. The details of this procedure are described in algorithm 1.

Table 1: Adjacency Matrix M_A computation.

Given the class set $c = \{c_1, c_2, \dots, c_n\}$ and their associated data $W = \{W_{c_1}, \dots, W_{c_n}\}$ for n classes
For each c_i
1) Run k -means on W_{c_i} set and compute the cluster centroids for class c_i as $m_i = \{m_{i1}, \dots, m_{ik}\}$
Construct distance matrix M_D as follows:
For each pair of classes c_p and c_q
1) $M_D(p, q) = \frac{\sum_{i=1}^k \sum_{j=1}^k \delta(m_{pi}, m_{qj})}{\frac{k(k-1)}{2}}$, being δ a similarity function
Convert distance matrix M_D to adjacency matrix M_A as follows:
For each pair of classes c_p and c_q
1) $M_A(p, q) = \frac{1}{M_D(p, q)}$
Normalize each column p of M_A as follows:
1) $M_A(p, q) = \frac{M_A(p, q)}{\sum_{i=1}^n M_A(i, p)}$

Look at the toy problem of Figure 1. In the example, three representant are computed for each class using k -means. Then, the distance among all pairs of representant are computed for a pair of classes, obtaining an adjacency distance for that two classes as $M_D(1, 3) = \frac{8+10+9+7+9+8+7.5+9.5+8.5}{9} = 8.5$. After that, the remaining positions of M_D are obtained in the same way, defining the following distance matrix

M_D :

$$M_D = \begin{pmatrix} 1 & 4 & 8.5 \\ 4 & 1 & 10 \\ 8.5 & 10 & 1 \end{pmatrix} \quad (1)$$

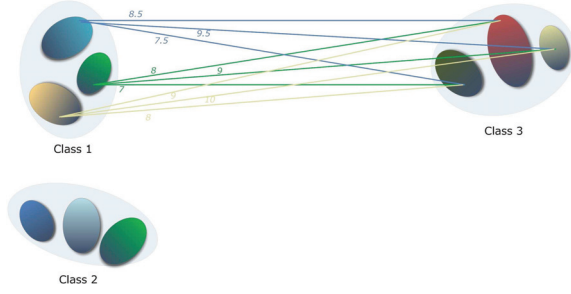


Figure 1: Toy problem for a 3-class classification task. For each class, three representant are computed using k -means. Then, the distance among all pairs of representant are computed for a pair of classes.

Finally, the adjacency matrix is computed changing distances to a likelihood values and normalizing each column of the matrix to unit. In this sense, the final adjacency matrix M_A for the toy problem of 1 is as follows:

$$M_A^{\text{Likelihood}} = \begin{pmatrix} 1 & 0.25 & 0.12 \\ 0.25 & 1 & 0.1 \\ 0.12 & 0.1 & 1 \end{pmatrix} \quad (2)$$

$$M_A = \begin{pmatrix} 0.73 & 0.18 & 0.08 \\ 0.19 & 0.74 & 0.07 \\ 0.09 & 0.08 & 0.81 \end{pmatrix} \quad (3)$$

2.2 Ontology Matrix M_O

The process up to here considered the relationship between classes by means of computational methods. However, some times no matter how good the system is, it can benefit of human knowledge. Here, we try to "inject" human knowledge of semantic similarity between classes into the system.

Taxonomy based on ontology is a tree or hierarchical classification which is organized by subtype-supertype relations. For example, Dog is a subtype of Animal. The authors of Caltech 256 data set compiled a taxonomy for all the categories included in their data set. Based on this taxonomy, we also defined a similar one for the MSRCORID

data set, which will be used to validate our methodology in the results section. The taxonomy of the Caltech data set can be found in [3]. The taxonomy tree defined for MSRCORID is shown in Figure 2.

Here we try to construct a similarity matrix like we did for the adjacency matrix, but now the similarity of classes is computed by means of the taxonomy tree.

In order to compute the distance among classes based on taxonomy, we look for common ancestor of nodes within the tree. Each category is represented as a leaf, and the non-leaf vertices correspond to abstract objects or super-categories. The less distance of the two leafs to their common ancestor, the less is their ontology distance. We construct the similarity matrix by crawling the tree from a leaf and rank all other leaves based on their distance. When we start from each leaf and crawl up the tree, at each step the current node is being explored based on depth-first search algorithm. In this search the less depth leaves get higher rank.

Finally, like in the case of the adjacency matrix, we need to convert distances into a measure of likelihood by inverting the values, and normalizing each column of the ontology matrix M_O to give the same importance for the taxonomy of all the classes. The whole process of computing the taxonomy distance and the ontology matrix is explained in algorithm 2. Figure 3 shows a possible ontology distance computation for the toy problem of Figure 1.

The final ontology matrix M_O obtained after computing all ranks from ontology distance and likelihood computation are the followings:

$$M_O^{\text{Ranking}} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 4 & 5 & 6 \\ 4 & 3 & 1 & 2 & 5 & 6 \\ 4 & 3 & 2 & 1 & 5 & 6 \\ 5 & 2 & 3 & 4 & 1 & 6 \\ 6 & 3 & 4 & 5 & 2 & 1 \end{pmatrix} \quad (4)$$

$$M_O^L = \begin{pmatrix} 1.0000 & 0.5000 & 0.3333 & 0.2500 & 0.2000 & 0.1667 \\ 0.5000 & 1.0000 & 0.3333 & 0.2500 & 0.2000 & 0.1667 \\ 0.2500 & 0.3333 & 1.0000 & 0.5000 & 0.2000 & 0.1667 \\ 0.2500 & 0.3333 & 0.5000 & 1.0000 & 0.2000 & 0.1667 \\ 0.2000 & 0.5000 & 0.3333 & 0.2500 & 1.0000 & 0.1667 \\ 0.1667 & 0.3333 & 0.2500 & 0.2000 & 0.5000 & 1.0000 \end{pmatrix} \quad (5)$$

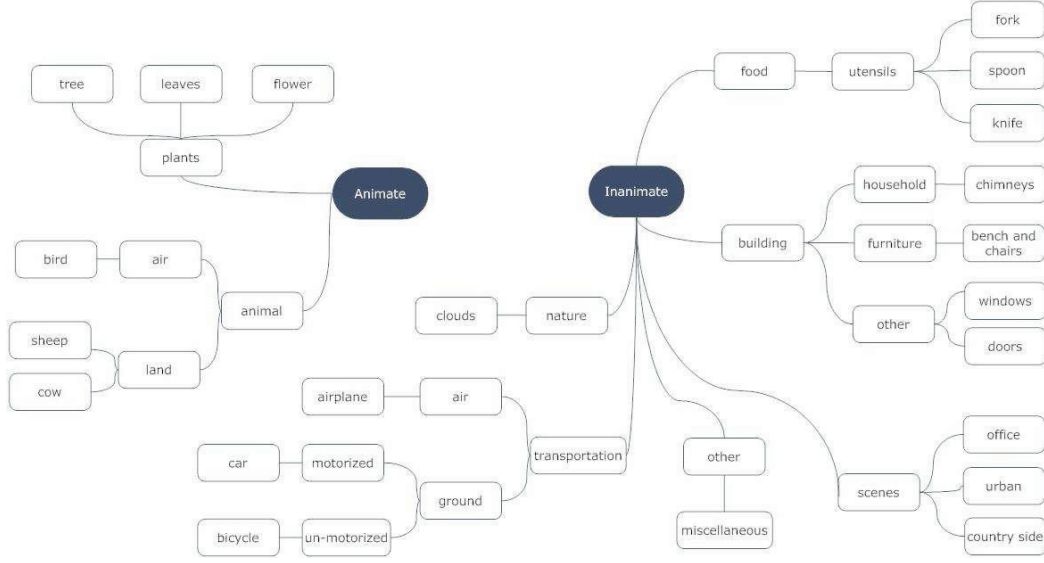
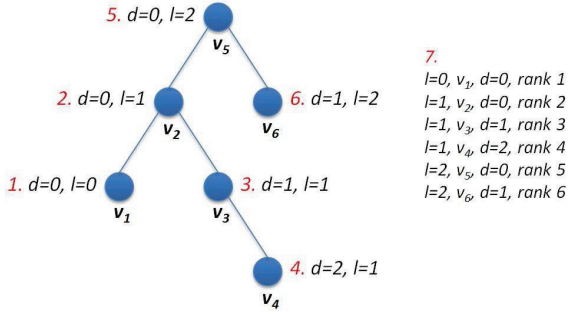


Figure 2: Taxonomy of object categories of the MSRCORID data set.

Figure 3: Example of the ontology distance computation of vertex v_1 to the rest of vertices. The steps of the distance computation are sorted. The final ranking is shown in the last step of the distance computation. This final ranking is then normalized and used as an ontology likelihood.

$$M_O = \begin{pmatrix} 0.4082 & 0.2041 & 0.1361 & 0.1020 & 0.0816 & 0.0680 \\ 0.2041 & 0.4082 & 0.1361 & 0.1020 & 0.0816 & 0.0680 \\ 0.1020 & 0.1361 & 0.4082 & 0.2041 & 0.0816 & 0.0680 \\ 0.1020 & 0.1361 & 0.2041 & 0.4082 & 0.0816 & 0.0680 \\ 0.0816 & 0.2041 & 0.1361 & 0.1020 & 0.4082 & 0.0680 \\ 0.0680 & 0.1361 & 0.1020 & 0.0816 & 0.2041 & 0.4082 \end{pmatrix} \quad (6)$$

2.3 Altering ECOC output rank using M_A and M_O

Given the output vector $D = \{d_1, \dots, d_n\}$ of the ECOC design, where d_i represents the distance of a test sample to codeword i of the coding ma-

trix, first, we convert the vector D to a measure of likelihood by inverting each position of D as $D^L = \{\frac{1}{d_1}, \dots, \frac{1}{d_n}\}$, and normalizing the new vector so that $\sum_{i=1}^n D_i^L = 1$. Then, using the previous M_A and M_O matrices, the new altered rank R is obtained by means of a simple matrix multiplication, as follows:

$$R = D^L \cdot M_A \cdot M_O \quad (7)$$

3 Results

Before the presentation of the results, first, we discuss the data, methods and parameters, and validation protocol of the experiments.

Data: The data used in our experiments consists on two public data sets: Caltech 256 [4] and 'Microsoft Research Cambridge Object Recognition Image data set' [5].

Methods and parameters: We use the classical Bag-Of-Visual-Words model (BOVW) [6] of 50 visual words to describe the data sets using the Harris-Affine detector and SIFT descriptor. For the ECOC classification, One-versus-one method with Gentle Adaboost with 50 decision stumps and RBF

Table 2: Ontology Matrix M_O computation.

<p>Given the class set $c = \{c_1, c_2, \dots, c_n\}$ and the taxonomy graph G</p> <p>For each leaf vertex v_i in G, $i \in [1, \dots, n]$, where n is the number of classes</p> <p>1) Visiting vertex $v_j = v_i$, Up Level $l = 0$, Depth $d = 0$</p> <p>Position list for each vertex v_p: $M_P(v_p) = [L_{v_p}, D_{v_p}]$ where L_{v_p} is the level of v_p and D_{v_p} is the depth of v_p</p> <p>2) Do while there are unvisited vertices</p> <p>1) $VisitVertex(v_j)$</p> <p>Function VisitVertex(v_p): If v_p is not visited $visitChild(v_p)$ if $\exists parent(v_p)$ $l = l + 1$ $M(v_p) = [l, d]$ $VisitVertex(parent(v_p))$</p>	<p>Function VisitChild(v_p): for each child v_p^c of v_p: if v_p^c has not been visited: if $child(v_p^c) \neq \emptyset$ $VisitChild(v_p)$ else $d = d + 1$ $M(v_p) = [l, d]$</p> <p>3) Filling the ranks $r = 0$ for $\nu = [1, \dots, \max(l)]$ for $\omega = [1, \dots, \max(d)]$ if $v_q M_P(v_q) = [\nu, \omega]$ is a leaf vertex of G $M_O(i, q) = r$ $r = r + 1$</p> <p>Convert distance matrix M_D to ontology matrix M_O as follows:</p> <p>For each pair of classes c_p and c_q 1) $M_O(p, q) = \frac{1}{M_O(p, q)}$</p> <p>Normalize each column p of M_O as follows: 1) $M_O(p, q) = \frac{M_O(p, q)}{\sum_{i=1}^n M_O(i, p)}$</p>
---	---

Support Vector Machines with parameters $C = 1$ and $\sigma = 0.5$ have been used. We use the Linear Loss-weighted decoding to obtain the class label [7]. For the adjacency matrix construction, the k parameter of k -means has been experimentally set to 3. For ranking the hist count we looked for one to seven matches at the first 15 positions using vector ontology and semantic distances of 0.001 and 0.0001.

Validation measurements: In order to analyze the retrieval efficiency, we defined an ontology distance based on taxonomy trees to look for the retrieved classes at the first positions of the ranking process. As explained in the previous section, the ranking result R is a sorted set of classes, where the first items have the highest rank. Then, we define an ontology distance m based on the taxonomy tree and adjacency matrices. Each c_i in R is accepted if its ontology distance d_i compared to the true label

class is less than m . The accepted results in the end of the list R are not desired, so another parameter k is used to analyze the results of the first positions of the ranking. If there are more than N accepted classes based on the value of m at the first positions defined by k , then we achieve a test hit. In order to perform a realistic analysis, we included this validation procedure in a stratified 10-fold evaluation procedure. The algorithm that summarizes the retrieval validation is shown in table 3.

3.1 Caltech 256 retrieval evaluation

In this case, we have defined an ontology distance of 0.001 and 0.0001 for Adaboost ECOC base classifier based on the taxonomy tree and the ontology distance defined in previous sections. For both distances we computed the BOVW features for this data set with different values of k first positions and number of hits. Some obtained performance surfaces are shown in Figure 4. The

performances are also shown in Table 4 estimated as the mean performance surface for each experiment. Note that we compared the classical ECOC output (Raw) with the ranking alteration using the adjacency matrix, ontology matrix, and both. In this case, the best results are obtained just altering the ECOC output by the ontology matrix.

Table 3: ECOC-Rank evaluation.

Given the sorted list of classes based on their rank $R = \{r_1, \dots, r_n\}$ For each item r_i in the top k positions of R $acceptedCount = 0$ 1) $d = \text{OntologyDistance}(r_i, \text{TrueLabel})$ 2) if $d > m$ then $acceptedCount + 1$ 1) If $acceptedCount > N$ then <i>Hit</i>

Table 4: Performances of Caltech 256 data set for different methods and parameters using Gentle Adaboost ECOC base classifier and ontology distance evaluation.

Problem	Adjacency	Ontology	Adj & Ont	Raw
m=0.001	0.4394	0.6901	0.4389	0.5530
m=0.0001	0.0718	0.1479	0.0719	0.0785

3.2 Microsoft Research Cambridge Object Recognition Image data set

In this case, we have defined an ontology distance of 0.001 and 0.0001 for Adaboost and RBF SVM ECOC base classifiers based on the taxonomy tree and the ontology distance defined in previous sections. For both distances we computed the BOVW features for this data set with different values of k first positions and number of hits. A sample of results are shown in the performance surfaces of Figure 5 and Figure 6 for Adaboost and SVM, respectively. The performances are also shown in Table 5 estimated as the mean performance surface for each experiment. In this experiment, though most of the experiments improve the classical ECOC rank, the adjacency matrix is selected as the first choice.

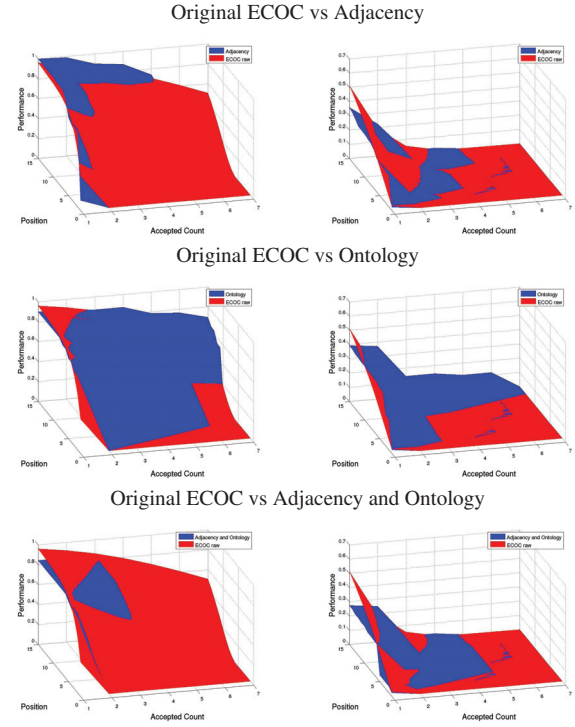


Figure 4: Results on Caltech 256 data set for Gentle Adaboost ECOC base classifier. Left column using ontology distance $m=0.001$ and right column using $m=0.0001$.

Table 5: Performances of Microsoft Research Cambridge Object Recognition Image data set for different methods and parameters using Gentle Adaboost ECOC base classifier and ontology distance evaluation.

Problem	Adjacency	Ontology	Adj & Ont	Raw
ADA m=0.001	0.3154	0.1744	0.2996	0.1568
ADA m=0.0001	0.1777	0.0659	0.1576	0.0667
SVM m=0.001	0.3714	0.1798	0.3001	0.2038
SVM m=0.0001	0.2511	0.0676	0.1577	0.0950

4 Conclusion

In this paper we altered the decoding process of the ECOC framework to define a new measure of semantic ranking that is applied on class retrieval problems. In order to include contextual and semantic information, we defined two matrices that mutates the ECOC output. An adjacency matrix is defined based on the feature space, and an ontology matrix is designed based on taxonomy trees. Results over public data show performance improvement when using the new ECOC-rank in the re-

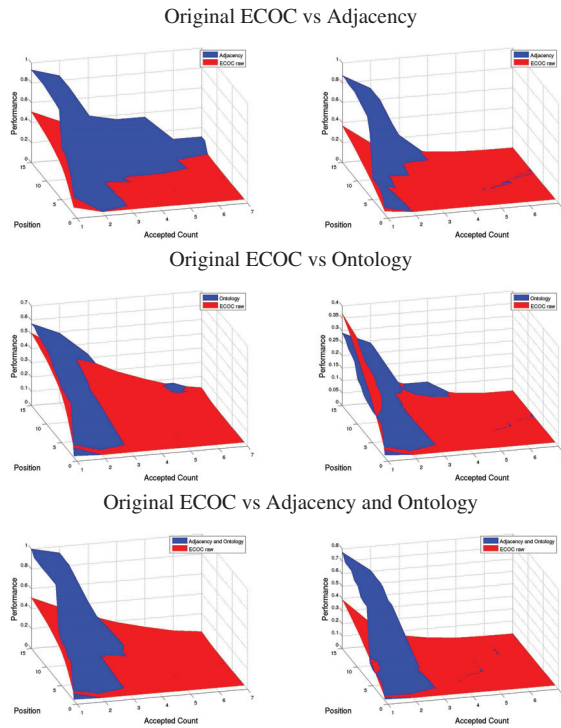


Figure 5: Results on Microsoft Research Cambridge Object Recognition Image data set for Gentle Adaboost ECOC base classifier. Left column using ontology distance $m=0.001$ and right column using $m=0.0001$.

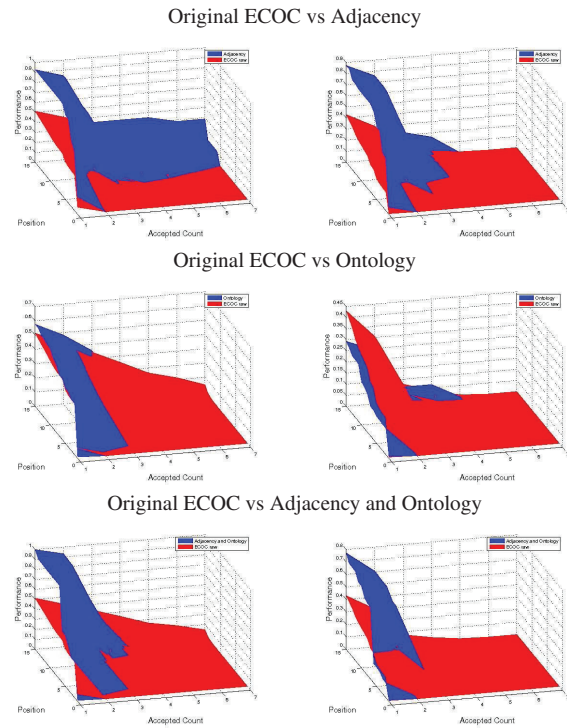


Figure 6: Results on Microsoft Research Cambridge Object Recognition Image data set for RBF SVM ECOC base classifier. Left column using ontology distance $m=0.001$ and right column using $m=0.0001$.

trieval process.

5 Acknowledgement

This work has been partially supported by the projects TIN2006-15694-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

- [1] R. Datta, D. Joshi, J. Li, J. Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, *ACM Comput. Surv.* 40 (2) (2008) 1–60.
- [2] T. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes 2 (1995) 263–282.
- [3] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Technical Report, California Institute of Technology.
- [4] G. Griffin, A. Holub, P. Perona, Caltech-256 object category dataset, Tech. Rep. 7694, California Institute of Technology (2007).
URL authors.library.caltech.edu/
- [5] Microsoft research cambridge object recognition image database.
URL <http://research.microsoft.com/>
- [6] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: *ECCV*, 2004, pp. 1–22.
- [7] S. Escalera, O. Pujol, P. Radeva, On the decoding process in ternary error-correcting output codes, in: *PAMI*, 2009.

Colour Logo Recognition

Farshad Nourbakhsh, Dimosthenis Karatzas and Ernest Valveny

Computer Vision Center, Edifici O - Campus UAB, Bellaterra, Spain

E-mail: {farshad,dimos,ernest}@cvc.uab.es

Abstract

In this paper, we propose a novel rotation and scale invariant method for colour logo retrieval and classification, which involves performing a simple colour segmentation and subsequently describing each of the resultant colour components based on a set of topological and colour features. A polar representation is used to represent the logo and the subsequent logo matching is based on Cyclic Dynamic Time Warping (CDTW). Finally, the proposed method is tested on a set of 100 different colour logo images and a classification result of 90.2% is obtained. The proposed method shows that a combination of topological and colour features is capable of extending the typical structural representations. We show also the dependency of method on the number of connected components.

Keywords: Colour Logos Recognition, Retrieval, Colour Segmentation, Topological Features.

1 Introduction

A logo is a graphical element designed for easy and definitive recognition, and is typically used to identify a company or organisation [1]. Although logos can be found in any style, they are bound by certain design restrictions. They have to follow certain rules depending on the type of document. Logos can be found in many kinds of paper documents, web images, scene images, signs

and banners. They can be rendered in either colour or grey scale, and their recognition and extraction are demanding tasks in computer vision and document analysis. Finding a robust and accurate algorithm which can handle different types of logos has many advantages for document retrieval, including enabling better compression techniques, lower environment requirements, and easier manipulation and management. As such, logo recognition has been a focus of research for the last several years and it remains an open problem.

Numerous approaches to shape analysis, with emphasis on logo recognition, have been reported in the domain of grey or two-level scales but the work on colour logo recognition is limited in the literature. Phan et al. [2] and [4] have presented logo and trademark detection based on edge gradient information (the colour edge gradient co-occurrence histogram (CEGCH)) which is an extension of a colour edge co-occurrence histogram (CECH [3]) in unconstrained colour images. In their work, the edge map is calculated first based on colour edge detection. To accomplish this, the image is quantized by a HSV colour quantization method and the information is extracted by CEGCH. In the next step, multiple sizes of the input queries are considered and the above methods are applied to each of them. Finally, overlapping search windows are used to search for the input query. The CEGCHs of every search window at a scale factor are computed and compared to the input query CEGCH and the logo and trademark are localized. Their proposed algorithm has some

drawbacks namely, the algorithm is not able to find multiple logos and trademarks, it is not able to detect a logo or trademark which is very deformed and the colour edge detection is very sensitive to noise. A. Hesson and D. Androutsos [5] have proposed a method which is based on their previous methods explained above, but this time they apply the Haar transform on the grey scale version of the quantized down sample image with co-occurrence of colour in each pixel. Furthermore, according to their report their accuracy is not yet promising. Z. Ahmed and H. Fella [6] have proposed a method for colour logo extraction based on chromatic densities, after removing areas with a small probability of having the logo with basic image processing techniques. The weak point of their presented method is that there are many assumptions so it is an application dependent method and the program is very sensitive to noise.

In this paper we present a rotation and scale invariant method for colour logo retrieval and classification based on colour segmentation of the logo into connected components and a polar representation of them using a set of topological and colour features, which is later flattened into a cyclic sequence, obtaining a vector representation of the logo. The cyclic dynamic time warping (CDTW) is used for comparison between the descriptors which are not of equal length. Finally, the proposed method is tested on a set of 100 different colour logo images.

This paper is organized as follows: In section 2 we explain the colour segmentation method used for transforming the logo images to different connected components. Section 3 proposes a component filtering followed by a background separation method. In section 4, we introduce a structural method for logo representation based on a selection of geometrical and colour features. Section 5 makes use of Dynamic Time Warping for logo matching to calculate the similarity between unequal length representations. Section 6 deals with the experimentation, and finally, in section 6 we summarize the obtained results.

2 Colour Segmentation

A. Clavelli and D. Karatzas [7] have implemented a method for text extraction in colour documents based on colour segmentation. To do this, they have applied a relatively simple colour segmentation method. The segmentation process used is a one-pass algorithm which creates 8-connected components based on colour similarity. The algorithm processes the image in a left-to-right, top-to-bottom fashion, and for each pixel calculates its colour similarity with the four of its neighbours that are already assigned to a connected component. The pixel is then either assigned to one of the existing components if their colour difference is below a set threshold, or it is used as the seed for a new component. Additional checks are performed in case a pixel is similar to more than one existing component, in which case components might be merged. The algorithm is further described in [8]. Colour similarity is assessed in the RGB colour space. After several tests with the images of our dataset, we set the threshold to 130 which permits to get components that are robust to rotation, scale and illumination.

3 Component Filtering and Background Separation

After segmenting the image into a set of connected components. We apply a filtering step with the aim of getting rid of small components which do not carry useful information for logo description and can hinder the matching process. Another type of noise that can appear is due to anti-aliasing artefacts in the original image. Typically anti-aliasing will result into components with a width of 1 pixel. These components do not carry any useful information, therefore they are discarded. The next step after component filtering is to extract the foreground components. Hence, the components that touch the border of the logo image are selected as potential background components. If the number

of touching pixels in these selected components is larger than half of the height or the width of the logo, the colour of the selected component will be picked as a background colour. In the next step, the colour of each connected component will be compared to the background components and if similar will be labelled as background. Figure 1(a) and 1(b) show one of the logo images in the dataset and its result after colour segmentation where each colour shows a different component. Figure 1(c) and 1(d) depict the result of component filtering and background separation respectively.

4 Logo Representation

After obtaining foreground connected components which are separated from each other, we have to extract a set of features for each of the components and define a representation which can convey the topological information between these components. In detail, we first extract a set of features for each connected component, then investigate different ways to represent the logo as a whole, based on the individual components. Moreover, The polar fashion (clockwise) is used to produce a rotation invariant representation where all components are linked to a reference point (Figure 1(e) and Figure 1(f)). The most important issue, after foreground components extraction and logo description is selecting a reference point to order our components in polar format. Some possible approaches are to select the biggest connected component or the center of mass of all components.

1- Biggest Connected Component: The center of the biggest connected component, the one with the highest number of pixels between foreground components, is selected as the reference point. The main reason for choosing the biggest connected component as a reference point is that this component is robust to different variations.

2- Center of Mass: Some earlier work also reported the usefulness of center of mass [9], which lead us to choose a different reference point com-

puted over all the components, which might be assumed to be more robust.

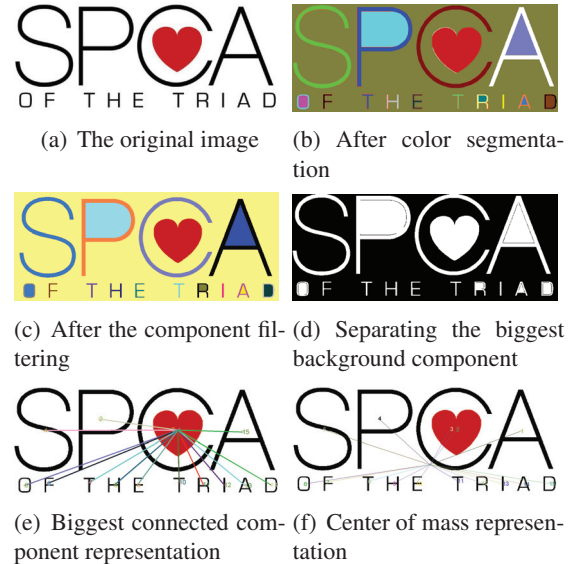


Figure 1: (a) to (d) show the different colour segmentation and component filtering steps; (e) and (f) show representations of foreground components.

4.1 Feature Selection

Each connected component in the logo description is represented by a feature vector. A number of features were examined that express both the connected component's own properties (i.e. its size and colours) as well as the spatial relationships between connected components. The features assessed are listed below:

1. Forward Angle: This is the angle between each connected component and the reference point which are ordered in clockwise format. This vector is normalized by dividing by 360 in each logo.
2. Backward Angle: This is like the forward angle but this is the angle with respect to the previous connected component. Like the forward angle, it is normalized.

3. Distance between other components to reference component: This feature vector should be normalized by dividing by the maximum distance value.
4. Average colour value of each connected component: Each connected component has a RGB colour value which is the average RGB colour value of all pixels in that component. This value will be normalized when the colour distance between two logos are calculated. For normalization, the colour distances will be divided by $255\sqrt{3}$.
5. Total Number of pixels in each connected component: This feature like distance gets normalized by dividing by maximum value for each logo image.

5 Logo Matching

Having obtained a logo representation, the next step is logo matching. The main problem here is that the length of the description is not the same for different logos. A possible approach for calculating similarity between unequal sequences is Dynamic Time Warping (DTW) which defines a dissimilarity measure based on an optimal alignment of two (non-cyclic) strings and has been successfully applied to speech recognition, on-line handwritten text recognition and time series alignment. For example, Gordo and Valveny [9] have presented a rotation invariant page layout descriptor based on cyclic dynamic time warping (CDTW). This work is a good example of using DTW for comparing descriptions which do not have the same length.

To calculate the difference between connected components (points in the DTW sequences) we are using a weighted sum of the difference between each feature.

$$D(a, b) = W_1 \cdot d(f_{a_1}, f_{b_1}) + W_2 \cdot d(f_{a_2}, f_{b_2}) + \dots + W_n \cdot d(f_{a_n}, f_{b_n})$$

where $W_i = 1/n$ for all $i \in \{1, \dots, n\}$. In the above equation f_{a_i} is a feature vector and W is the weight and $D(a, b)$ is the distance between components a and b that is used in CDTW.

6 Evaluation

To validate the system, we first describe the data and experiments for two application scenarios: recognition and retrieval.

6.1 Data Set

Because the amount of work on logo recognition in the colour domain is very limited colour logo datasets are sparse. Therefore, we decided to create our own data set. To do so we used Google picture search with the key word "Colour Logo". The first 100 logos returned were included the dataset. Additionally, we were not generally able to find different rotations and scales of the chosen logos on the web. Hence we created different rotation and scale settings for each logo in our dataset as below:

1. Rotation: An image of each class is rotated (bilinear interpolation) by the following angles: 9, 36, 45, 90 and 150 degrees (500 instances).
2. Scale: An image of each class is scaled (nearest neighbour interpolation) by the following factors: 165%, 150%, 135%, 125% and 85% (500 instances).
3. Rotation-Scale: An image of each class were scaled and then rotated by the above factors (2500 instances).

6.2 Evaluation of Different Feature Sets

For this experiment, the biggest connected component based logo representation is assessed on the first dataset (rotation only, 500 instances) to investigate the performance in a classification scenario

when employing different feature sets and their combinations. Specifically, we have different combinations of features to describe connected components (CC). In each of the cases the individual features are considered with equal weights. The combinations tried are A) Only the angle (forward and backward angle as a vector), B) only the distance of the component from the reference point, C) only the number of pixels of the component, D) only the colour of the components, E) the combination of the above four, and F) the combination of the above, but considering two separate features for the forward and the backward angle. Table 1 shows the result of this experiment. It is clear, after using all 5 features the accuracy increased from 79.6% to 90.2%.

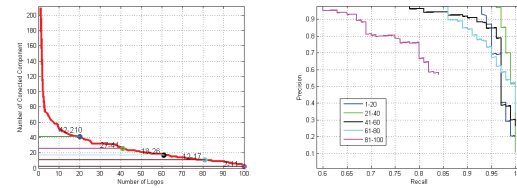
A	B	C	D	E	F
66%	57.2%	43.2%	64%	79.6%	90.2%

Table 1: Accuracy obtained with respect to the different feature sets.

6.3 Effect of Logo Complexity

As in the previous section, we have used the first dataset and the biggest connected component based representation, but now, for the retrieval scenario. The aim of this experiment is to investigate the performance in relation to the complexity (number of connected components) of a logo. We expect that the more components available in a logo, the richer the logo description is so the result of logo matching would be better. The 100 original logo images are sorted based on the number of their connected components. The maximum one was 210 and the minimum one was 2 (Figure 2(a)). Then the set of images is divided into 5 classes (each class contains 20 images). Next, a similarity distance between the original logo and its instances is used as a threshold for calculating the precision and recall. Figure 2(b) depicts the dependency of the proposed method on the number of connected

components. In detail when the number of components is too less, we do not have enough information to describe the logo and in contrast when the number of components is too high the performance is not acceptable because of the confusion that small components can create.

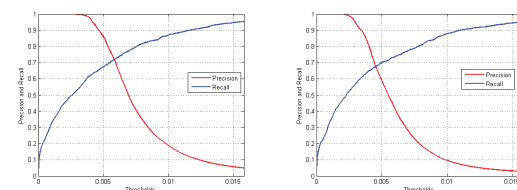


(a) the sorted images with (b) the resultant Precision-respect to their number of Recall plot of each divided connected components class

Figure 2: Comparison of results for different classes based on the number of CCs

6.4 Comparison of Different Logo Representations

The rotation-scaled (2500 instances) dataset and precision and recall strategy are used to investigate the performance of our method in the retrieval scenario for biggest connected component and center of mass. Figure 3 shows the result of the precision and recall with respect to the biggest cc and center of mass as a reference point in retrieval scenario respectively.



(a) Precision and recall for Biggest CC (b) Precision and recall for center of mass

Figure 3: Precision and recall with respect to different reference points

Figure 4 shows the precision vs recall plot for the above experiment. It is obvious the result of

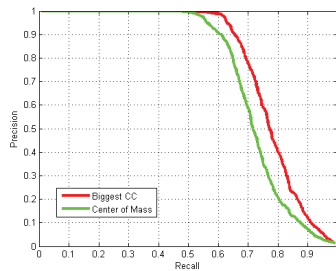


Figure 4: Precision and recall comparison for different reference points

the proposed method with respect to the biggest connected component as a reference point is better than the other one based on the value of the average precision.

7 Conclusion

In this paper we developed and tested a novel method for colour logo retrieval and classification. The method involves first performing a simple colour segmentation and then describing each of the resultant colour components based on a set of topological and colour features. A polar representation is used to represent the logo and logo matching is subsequently performed, using CDTW. Two flavours of the method are proposed, differing in the selection of the reference point, and a comparison between them is made. We concluded that combining the proposed features gives a better result than using individual features alone. An accuracy 90.2% was achieved with 5 combined features. We observed a dependency of the method on the number of connected components in the logos. The optimal result can be obtained when the number of connected components is not too high or too low. Logo representation based on the biggest connected component gives a better result compared to the other proposed method. This difference occurs because a bigger connected component is more robust to rotation and scale compared to smaller components.

References

- [1] Alina Wheeler, *Designing Brand Identity*, John Wiley and Sons, 2006.
- [2] Raymond Phan and John Chia and Dimitrios Androutsos, "Colour logo and trademark detection in unconstrained images using colour edge gradient co-occurrence histograms", *Canadian Conference on Electrical and Computer Engineering, CCECE*, 531-534, 2008.
- [3] J. Luo and D. Crandall, "colour object detection using spatial colour joint probability functions", *IEEE Transaction On Image Processing*, 1443-1453, 2006.
- [4] Raymond Phan and Dimitrios Androutsos, "Logo and Trademark Retrieval in General Images Database Using Colour Edge Gradient Co-Occurrence Histograms", *4th International Conference on Image Analysis and Recognition, ICIAR*, 674-685, 2007.
- [5] Ali Hesson and Dimitrios Androutsos, "Logo and Trademark Detection in Image Using Wavelet Co-Occurrence Histograms", *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 1233-1236, 2008.
- [6] Zeggari Ahmed and Hachouf Fella, "Logos extraction on picture documents using shape and colour density", *IEEE International Symposium on Industrial Electronics, ISIE*, 2492-2496, 2008.
- [7] Antonio Clavelli and Dimosthenis Karatzas, "Text Segmentation in Colour Posters from the Spanish Civil War Era", *In Proceedings of the Tenth International Conference on Document Analysis and Recognition, ICDAR09*, 2009.
- [8] Dimosthenis Karatzas, "Text Segmentation in Web Images Using Colour Perception and Topological Features", *University of Liverpool, UK*, 2002.
- [9] Albert Gordo and Ernest Valveny, "A rotation invariant page layout descriptor for document classification and Retrieval", *10th International Conference on Document Analysis and Recognition*, 2009.

Object Detection using Coarse-to-Fine relocalization

Marco Pedersoli*, Jordi González*, Andrew Bagdanov* and Juan José Villanueva*

* *Computer Vision Center, Campus UAB Edifici O, 08193 Bellaterra (Cerdanyola), Spain*

E-mail:marcopede@cvc.uab.es

Abstract

A common problem of object detection is the poor learning of complex objects like humans, animals, cars, etc., where many different aspects and shapes of the object must be modeled. In most cases, the poor performance of the detector is said to be due to the low discriminative capability of the features. In this paper we prove that this is not always the case. We show that, while using a linear SVM classifier with HOG features, we can improve detection results by using an iterative method that better localizes the position of the object in the training images. Results are presented on one of the most challenging and complete databases of objects: the Pascal Challenge 2007, which consist of 20 different classes and more than 10,000 images.

Keywords: object recognition, pattern recognition, object detection.

1 Introduction

In recent years, there has been an increasing interest in object localization (or object detection), which is the task of finding the precise location of a certain object class in an image. The possible applications are many, from video-surveillance to driver assistance and scene understanding. Although many improvements and enhancements have been developed, the state of the art for de-

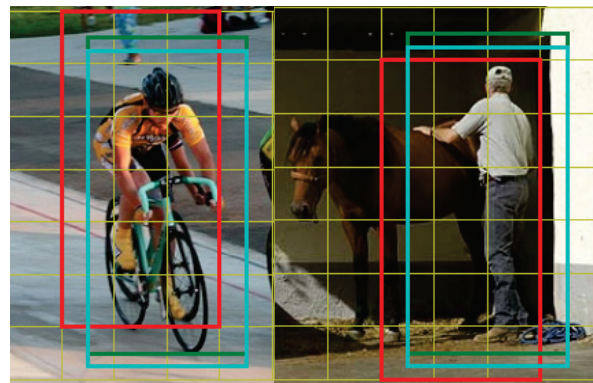


Figure 1: Bounding Boxes of the three detector levels in two training examples. Red is the coarse feature resolution, green is the middle feature resolution and cyan is the finer feature resolution. Increasing the feature resolution the object is better localized.

tection is still far from the level necessary for real applications.

The main reasons that object detection is still not ready for real applications are: (i) the computational complexity of the problem, because the object has to be searched for in all possible positions, locations and sizes in the image; (ii) the low accuracy of object localization in the training data. While the first problem has been tackled by many authors and with different solutions like cascades of detectors [6, 2, 1] and sparse search [7, 4], the second has been considered only partially and it is the focus of this paper.

The general way to localize an object in the training data is to provide the coordinates of the minimum bounding box that contains the object. This is generally considered a good compromise between the cost of manually annotating the database and the localization accuracy obtained.

However, in many situations and specifically for deformable objects, the bounding box is far from being a good localization of the object. Some examples of bounding box misalignment are shown in figure 1.

In this paper we show that the bad alignment of objects in training images results in a reduction in detection accuracy, and that fixing this problem can improve the accuracy of state of the art methods.

The rest of the paper is organized as follows: section 2 explains the HOG pyramid and its advantages, sections 3 and 4 detail the training and detection procedures, section 5 shows experiments and results about the performances of the method varying the most relevant parameters. Finally, section 6 states some concluding remarks.

2 Multiresolution Pyramid of HOGs

In contrast to the standard HOG classifier where the feature model is a grid of HOGs, our model representation is a pyramid of HOGs at different resolution. To make the model easy to compute, we uses a dyadic pyramid, where the features at each level have size the double of the following level. An example of the HOG pyramid of a person detector is represented in figure 2.

The advantages of this representation compared to the standard one are many: (i) the multiple resolutions of the HOG can better represent the different characteristics of an object class: for non-deformable and static parts, a highly discriminative description can be given by the high resolution features, while for deformable and moving parts, a better description can be obtained by low resolution features (see figure 2); (ii) this representation

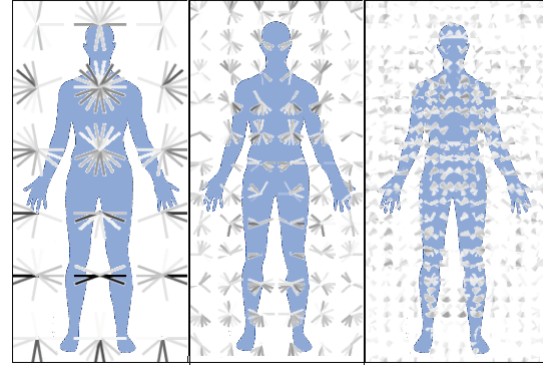


Figure 2: HOG pyramid model for the class person. The low resolution features are able to give a general coarse of the human silhouette, while the high resolution focus more in details.

does not need the computation of further features, because the same features are used for the multiresolution pyramid, but also for the multi-scale search of the sliding window approach; (iii) the detector can still work even if the higher resolution features are not available, which allows detecting both low and high resolution object without any further complication of the framework; (iv) localization is done from coarse-to-fine resolution, which permits scanning only interesting locations at high resolution; (v) the localization can be further sped up by using a multiresolution cascade similar to [5].

3 Training

Training Process

Selection of the positives examples T_p

Random Selection of the negative examples T_n

Repeat:

SVM training based on (T_p, T_n)

Refinement of T_p

Selection of T_n

Algorithm 1: Iterative training process

The complete training process is an iterative pro-

cedure which is summarized in the algorithm 1. In the following subsections a detailed explanation of every part is given.

Selection of positives examples

The first step of this process is the computation of the correct bounding box aspect ratio. This is computed from the mean bounding box of all training samples. Then, the initial selection of the positive examples is made. The selection is based on the minimum overlapping area between the bounding box of the example B_t and the bounding box of the model B_m using formula (1).

$$O = \frac{\text{area}(B_t \cup B_m)}{\text{area}(B_t \cap B_m)} \quad (1)$$

If the overlap O is greater than 0.5, the example is taken as a positive sample and it will be used for SVM training, otherwise it is discarded. Generally the overlapping is less than 0.5 when the object is very small or when the aspect ratio is very different from the one chosen for the detector model. The final number of positive samples depends on the number of examples provided by the database and the amount that have been discarded by the selection. Generally the number of positive samples is between 100 and 1,000.

Random selection of negative examples

The distribution of the negative examples in images not containing the object is drawn uniformly in scale and space. In practice we take 10 samples per image up to a total of 5,000 samples. An important detail that allows good performance is to select also samples that are partially outside the image.

SVM training

Positive T_p and negatives T_n samples are used to train a linear SVM using libSVM [9]. We notice that the C does not affect very much the final performance of the detector, thus we do not perform

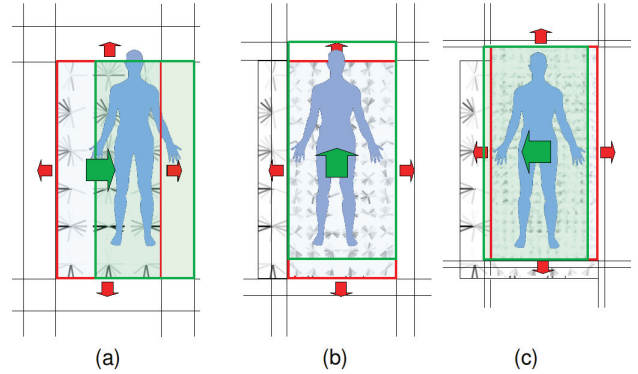


Figure 3: Coarse to fine refinement of an example. (a) low resolution refinement (b) middle resolution refinement (c) high resolution refinement

extensive cross-validation, but fix it to 0.01. Once the initial SVM classifier is built, this is used in an iterative process detailed in the following subsections.

Refinement of the positives examples

The localization of the bounding box of the positive examples is refined using the already computed classifier, in a way similar to [8]. We define the best location for the example as the place where the answer provided by the SVM classifier is maximum. This corrects problems due to the deformation of the object or to bad annotation selecting the location that best fit with the already computed classifier. We assume that the best bounding box location is relatively close to the current location, which makes possible to employ a local search instead of a more expensive global one, as in [8].

Furthermore, taking advantage of the pyramid structure, we compute the search in a coarse-to-fine manner as represented in figure 3. We first localize the best response using a 3x3 neighborhood of the bounding box at the lowest resolution level (figure 3(a)), which means big displacements at coarse resolution. Then, starting from the new location of the bounding box, the same process is

repeated at the next finer feature resolution. In this case the displacement will be half of the previous (figure 3(b) and (c)), and the resolution double, which provides better object distinctiveness. The process is repeated until reaching the finest feature resolution and consequently the smallest displacement stride.

Selection of negative examples

The space of all possible negative examples is huge, and to obtain a good representation of it a huge number of samples is required. This makes SVM training very costly in terms of time and memory.

A common way to reduce this number without losing performance is to select the training samples in a smart way. Considering that, in an SVM, only the support vectors are used in the final classifier, it is easy to see that only the samples close to the splitting hyperplane are needed to ensure good performance.

In our case, these samples can be obtained by looking the score value of the already computed classifier. If this value is close to 0 it is probable that the sample will be a support vector, and must be added to the negative training samples T_n . Otherwise, it will be discarded. Therefore the final process consists of applying the already computed SVM classifier to the negative images, and if the detection score d is close to 0 (exactly $-1 < d < 1$), this sample is added to T_n . We select a maximum of 10 negative samples per images to avoid oversampling similar errors.

4 Detection

The trained SVM classifier is used for the detection of object classes in images using a modified version of the common sliding window approach. The search of the object is computed as a convolution between the precomputed HOG pyramid of the image and the HOG pyramid of object model.

The search is accomplished from the lowest to the highest feature resolution of the object model. At the beginning, the HOG pyramid of the image is convolved with the lowest resolution model and the result is saved. After that, if we want to apply a filtering algorithm to select only the promising hypothesis, we can select a certain number of locations where the low resolution response has given a high score. In our case we use all the hypothesis because the focus of this work is not about speeding-up the scanning process (see [6]), but increasing the detection accuracy.

The second step is to sum to the partial result of the first convolution the result of the convolution of the HOG pyramid of the image with the corresponding HOG level of the object model. The correspondence can be easily found in the next higher resolution of the HOG pyramid of the image. The process is repeated for all the resolutions of the model until obtaining the final detection response which is the sum of all the detection levels.

Notice that for each detection position at the low feature resolution there are 9 possible detection positions at the next resolution (see Fig. 3). Considering that only one detection per position is possible, among the 9 candidates we select the one with highest partial score. This is done for all the resolution levels of the object model, so that the detector refines its search around the most probable location. In this way, without any filtering scheme, we fix the number of candidates to evaluate to the lowest resolution scan, which is quite small, being this one coarse.

5 Experiments and Results

We tested our method using the database provided by the Pascal Visual Object Recognition Challenge VOC 2007 [3] which is one of the most complete dataset for object detection. It contains 20 different classes of objects: vehicles (aeroplane, car, bus, bike, motorbike, train), person, animals (bird, cat, cow, dog, horse, sheep) and indoor objects (bottle,

Levels	Area	mAP 0	mAP 1	mAP 2	mAP 3	mAP 4	mAP best
1	100	10.7	12.3	13.6	14.0	13.3	15.2
2	31	9.8	12.7	13.9	14.4	14.1	15.7
3	15	9.4	14.6	15.3	15.7	15.2	17.0

Table 1: Comparison of different detectors configuration in the VOC 2007. *Levels* is the number of levels of resolutions in the HOG pyramid of the object model. *Area* is the number of HOG tiles in the lowest level features. X is the mean Average-precision computed on the complete database at the x^{th} reiteration. best is the mean Average-precision obtained selecting the iteration which gives the best Average -Precision for each class.

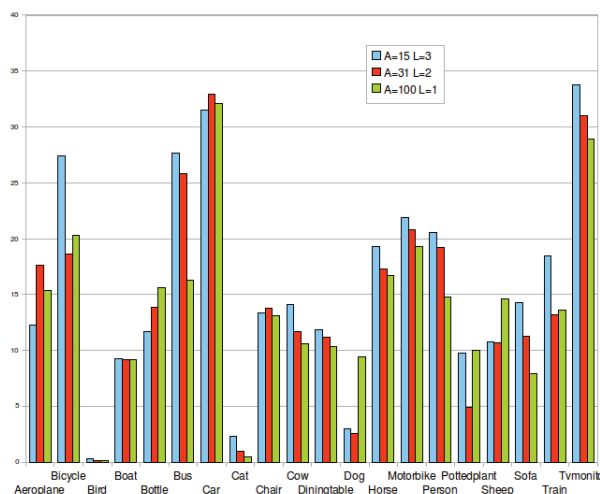


Figure 4: Average Precision for the 20 classes of the VOC 2007.

chair, dining table, potted plant, sofa, tv-monitor).

The first experiment shown on table 1 compares three different configurations of the HOG object model.

The first configuration (first row of table 1) uses a special case of pyramid with only one level. In this case no coarse-to-fine relocalization is possible and the model is reduced to the standard HOG detector. Performance is also similar to the standard linear HOG detector; the best retraining of our linear detector gives a mean Average-Precision of 14.0 while the best of [1] gives 14.6. We believe that the difference between the two scores is due to the fact that we did not tune the HOG features for every class.

In the second and third configuration (respectively second and third row of table 1) we really used the HOG pyramid. Also in this case the best scores are given in the third iteration with respectively 14.4 and 15.7 mean Average-Precision. The score with 3 levels shows a noticeable improvement of more than 1 point. It is also interesting to notice that, if we select the best iteration for every class instead that selecting the global best iteration, the mean Average-Precision can further increase up to 17.0 (see last column of table 1). Figure 4 shows the average precision for all object classes for the three detector configurations.

Finally, in table 2 we compare the best configuration of our detector for each class with the results of the standard HOG detector provided by [1].

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
Our Method	12.3	27.4	0.3	9.3	11.7	27.7	31.8	2.3	13.4	14.1	11.9	3	19.3	21.9	20.6	9.8	10.8	14.3	18.5	33.8	15.7
HOG from [1]	10.0	27.8	4.7	0.6	11.4	31.7	33.9	2.6	10.1	14.9	9.7	1.8	28.1	22.6	12.2	9.9	10.0	4.3	19.3	26.1	14.6

Table 2: Average-Precision Results for all classes of the Pascal 2007 database.

6 Conclusions

In this paper we have shown the importance of good localization in training examples for object detection. Specifically, we developed and tested an iterative method that uses a pyramid of HOG features to better align the training object images and thus obtain better detection accuracy.

References

- [1] H. Harzallah, F. Jurie, C. Schmid, "Combining efficient object localization and image classification", *IEEE Proc. of the Intl. Conf. on Computer Vision*, 2009.
- [2] A. Vedaldi, V. Gulshan, M. Varma, "Multiple Kernels for Object Detection", *IEEE Proc. of the Intl. Conf. on Computer Vision*, 2009.
- [3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*
- [4] M.B. Blaschko, C.H. Lampert, Learning to Localize Objects with Structured Output Regression, *IEEE Proc. of the European Conf. on Computer Vision*, 2008
- [5] M. Pedersoli, I. Ivn, J. Gonzlez, J.J Villanueva, Fast Human Detection using Multiresolution Cascade, *Current Challenges in Computer Vision. Proc. of the Third International Workshop*, 2008
- [6] M. Pedersoli J. Gonzlez and J.J. Villanueva, High-Speed Human Detection Using a Multiresolution Cascade of Histograms of Oriented Gradients, *4th Iberian Conf. on Pattern Recognition and Image Analysis*, 2009
- [7] K. Mikolajczyk, B. Leibe, and B. Schiele, Multiple object class detection with a generative model, *IEEE Proc. of the Conf. on Computer Vision and Pattern Recognition*, 2006.
- [8] P.F. Felzenszwalb, D.A. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, *IEEE Proc. of the Conf. on Computer Vision and Pattern Recognition*, 2008
- [9] C. Chang and C. Lin, LIBSVM: a library for support vector machines, 2001.

Calibration and Rectification of Multimodal Stereo Rigs

Fernando Barrera*, Felipe Lumbreras⁺, and Angel Sappa[‡]

Department, Computer Vision Center, Edifici O, Barcelona, Spain

*E-mail:jfbarrera@cvc.uab.es **

E-mail:felipe.lumbreras@cvc.uab.cat⁺

E-mail:angel.sappa@cvc.uab.cat[‡]

Abstract

This paper presents an extension of two classical image rectification algorithm to the multimodal field, it works for unconstrained multimodal stereo rigs (intensity and thermal infrared camera). A briefly analysis of geometry of multimodal stereo systems gives insight into the limitations and requirements that motives the formulation of the developed algorithm. Using calibration data of a stereo rig, we compute two transformations that determine the rotations of image planes, such that conjugates epipolar lines are collinear and parallel to image coordinate frame. Experiments demonstrate the utility of the method for multimodal rectification, a qualitative and quantitative evaluation measured its accuracy for align vertically features.

Keywords: Multimodal stereo vision, Intensity\Thermal infrared sensing, and Multisensor fusion.

1 Introduction

According to the reviewed literature, in most of cases, the rectification problem is avoided in multimodal stereo applying constrains on the scene and camera's position. For instance, assuming that the image planes are approximately parallel [4], or that a 2-D homogeneous transformation relates the im-

ages [10]. These restrictions on the geometry of the system are possible, but constrain the application domain, since that are not always valid in a real environment. The image rectification is an alternative way to tackle the multimodal stereo problem, considering a general scenario where these restrictions could be suppressed.

A stereo reconstruction algorithm could be decomposed into: image acquisition, calibration, *rectification*, stereo matching algorithm, and triangulation. Some authors consider that image rectification is an optional step. However, it had been shown that search correspondences in rectified images is more efficient than to use epipolar lines, specially in the case of a dense algorithm. The structure presented above suggests a sequential pipeline with high dependency of data, a wrong image rectification introduces errors to following stage, and so on. Therefore, the accuracy of calibration and rectification limits the performance of stereo matching algorithm.

The calibration is a well known problem, which has been largely studied for intensity images, and its contributions could be used for infrared images calibration; it is only is necessary to pay attention to few details (see section 2).

In a multi-camera setup, each sensor can be at a different position in the world and have different intrinsic parameters. But, in the case of multimodal stereo rig, the cameras are capturing sig-

nals related with two physical phenomena, reflection and emitted of energy, but seen at different spectral band, infrared and visible spectrum. Due to this, corresponding objects in each image may have different sizes, shapes, positions and pixel values. Therefore, the rectification process must take into account these factors, as well as camera effects as reduced field of view force wide baseline and significant rotation.

This paper presents a calibration procedure and a rectification algorithm for multimodal images; the rectification algorithm is based on [1] and [3]. But, our approach support multiples modalities, and considers the parameter of the infrared camera and their contribution to complexity in geometry of stereo rig. Other methods as [8] and [6] are based solely on the estimation of the fundamental matrix, which are susceptible to noise, and share the same problem assume that point matches have already been accurately determined.

The remainder of this paper is structured as follows. Section 2 presents a procedure to calibration of multimodal rigs. In Section 3, the proposed method of rectification and their mathematical foundations are introduced. Section 4 describes our proposed evaluation methodology, including the performed experiments. Finally, section 5 concludes with a discussion of results and future works.

2 Multimodal stereo rig calibration

Infrared cameras are thermal sensors, it means that two kinds of calibration could be performed. The first one is the *thermal calibration*, that relate the temperature measurements with the real data. In this case generally, the camera provider supply targets and software for this purpose. In the second kind of calibration, the intrinsic and extrinsic parameter are estimated.

The intrinsic parameters of an infrared camera are computed like an intensity camera, under the

cited conditions. The coefficients of camera matrix are found for each camera: focal lengths, principal points, skew, and distortion coefficients. Next, the extrinsic parameters as the rotation (R) and translation (T) are obtained. In this way the multimodal stereo rig is calibrated, by available calibration tools.

An elegant procedure for multimodal image calibration has been presented in [7], re-using existent calibration tools like [2] or [9], originally developed for intensity images. However, they could be extended to infrared images, introducing changes. The strategy consists in illuminating the scene with high intensity halogen bulbs placed behind the cameras, this warms the calibration pattern, making visible the chessboard in both modalities. We prefer to use the edges of the board, which is attached the calibration pattern. The difference of contrast between the background and the board produce edges in intensity images, as well as in the infrared images. Then, four lines are detected and intersected over each image. Notice that, lines correspond edges of board. These intersections are calibration points for the toolbox. Our proposal shows a detection stable and high repeatability. Specially, when the cameras have wide baseline.

3 Proposed multimodal image rectification

This section presents the proposed extensions from [1] and [3] to tackle the multimodal stereo rig rectification problem. Fig. 1(b) shows an illustration of a multimodal stereo rig, defined with an intensity and an infrared camera, used through this work.

3.1 Camera model

A projective camera model is assumed for infrared camera as well as intensity camera, see figure 1(a). Let C be optical center and \mathcal{R} the image plane, at $Z = f$. The projection of an arbitrary 3D point

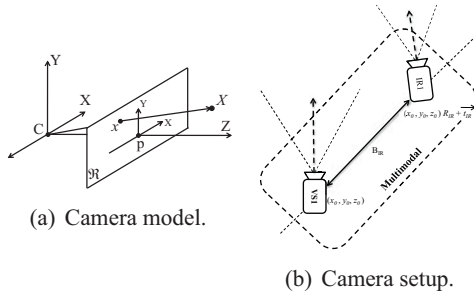


Figure 1: Multimodal stereo rig.

$X = [X \ Y \ Z]^T$ to image plane is given by the intersection of \mathcal{R} with the line containing C and X . The centre of projection C is called the *camera centre* or *optical centre*. The line perpendicular to the image plane passing through the camera center C is called the *principal axis* or *principal ray* of the camera, its intersection with \mathcal{R} is the *principal point* p . The distance between C and p is the focal length f [5].

If the world coordinates and image points are represented by homogeneous vectors, then central projection is simply expressed as a linear mapping between their homogeneous coordinates. Let $\tilde{X} = [x \ y \ z \ 1]^T$ and $\tilde{x} = [x \ y \ 1]^T$ be the homogeneous coordinates of X and x respectively, then:

$$\tilde{x} = P\tilde{X}. \quad (1)$$

The camera is therefore modeled by its *perspective projection matrix* P , which can be decomposed, using the QR factorization, into the product:

$$P = K[R \ | \ t]. \quad (2)$$

The matrix K or camera calibration matrix only depends on the intrinsic parameters, and has the following form:

$$K = \begin{bmatrix} \alpha_x & \gamma & x_0 \\ 0 & \alpha_y & y_0 \\ 0 & 0 & 1 \end{bmatrix}, \quad (3)$$

where: $\alpha_x = fm_x$ and $\alpha_y = fm_y$ represent the focal length of the camera in terms of pixel dimensions in the x and y directions respectively; f is

the focal length; m_x and m_y are the effective number of pixels per millimeter along the x and y axes; (x_0, y_0) are the coordinates of the principal point; finally γ is the skew factor.

In general, points in space will be expressed in terms of a different Euclidean coordinate frame, known as the world coordinate frame. Two coordinate frames are related via a rotation and a translation. If X is an inhomogeneous 3-vector representing the coordinates of a point in the world coordinate frame, and \tilde{x} represents the same point in the camera coordinate frame, then it may write $\tilde{x} = R(X - C)$, where C represents the coordinates of the camera centered in the world coordinate frame, and R is a 3×3 rotation matrix representing the orientation of the camera coordinate frame. This equation may be written in homogeneous coordinates as:

$$\tilde{x} = \begin{bmatrix} R & -RC \\ 0 & 1 \end{bmatrix} \tilde{X}. \quad (4)$$

Putting (4) together with (2) and (1) leads to the following equation:

$$\tilde{x} = KR[I] - C]X, \quad (5)$$

where x is now in a world coordinate frame. It is often convenient not to make the camera centre explicit, in this case the camera matrix is simply

$$\tilde{x} = K[R|t]X, \quad \text{where } t = -RC. \quad (6)$$

3.2 Rectification of camera matrices

At this point, it is assumed that the stereo rig is calibrated (sec. 2), and the perspectives projection matrix P_1 and P_2 are known. The rectification process defines two new projective matrices P'_1 and P'_2 obtained by rotating the old ones around their optical centers until focal planes becomes coplanar, thereby containing the baseline. This ensures that epipoles are at infinity; hence, epipolar lines are parallel. To have horizontal epipolar lines, the baseline must be parallel to the new \hat{x}' axis of both

cameras. In addition, in a proper rectification corresponding points must have the same vertical coordinate.

The calibration procedure returns the following values: $K_1, R_1, t_1, K_2, R_2, t_2$. Notice that, according to where the world reference is placed, either t_1 or t_2 is equal to $[000]^T$; the same happen for rotation results, one of them is equal to the identity matrix.

In general, the equation (2) can be written as follow, since a general projective camera may be decomposed into blocks.

$$P_n = \begin{bmatrix} q_{11} & q_{12} & q_{13} & q_{14} \\ q_{21} & q_{22} & q_{23} & q_{24} \\ q_{31} & q_{32} & q_{33} & q_{34} \end{bmatrix} = [M_n | q_n], \quad (7)$$

where M_n is a 3×3 matrix and q_n a column vector of camera n . The coordinates of optical centres $C_{1,2}$ are given by:

$$C_1 = -R_1 K_1^{-1} q_1, \quad (8)$$

$$C_2 = -R_2 K_2^{-1} q_2. \quad (9)$$

In order to rectify the images is necessary to compute the new principal axes \hat{x}', \hat{y}' , and \hat{z}' . A special constrain is applied for computing \hat{x}' , since theses axes must be parallel to the baseline, and to have horizontal epipolar lines, then:

$$\hat{x}' = \frac{(C_2 - C_1)}{\|C_2 - C_1\|}. \quad (10)$$

The new \hat{y}' axis is orthogonal to \hat{x}' and old \hat{z} :

$$\hat{y}' = \frac{(\hat{z} \times \hat{x}')}{\|\hat{z} \times \hat{x}'\|}, \quad (11)$$

where old \hat{z} axis is third vector (r_3^T) of rotation matrix R_1 and R_2 . The new \hat{z}' axis is orthogonal to \hat{x}' and \hat{y}' :

$$\hat{z}' = \frac{(\hat{x}' \times \hat{y}')}{\|\hat{x}' \times \hat{y}'\|}. \quad (12)$$

The previous procedure shows the steps for computing the new axes, but none image has been rectified. In order to do this image rectification, the

new projective matrices P'_1 and P'_2 should be expressed in terms of their factorization.

$$P'_1 = K_1[R'_1 | -R'_1 C_1] = [M'_1 | q'_1], \quad (13)$$

$$P'_2 = K_2[R'_2 | -R'_2 C_2] = [M'_2 | q'_2]. \quad (14)$$

At this point the problem can be seen in different ways: (i) it follows the original formulation of [3], or (ii) Compute a linear transformation T_1 such that $P_1 \xrightarrow{T_1} P'_1$ and $P_2 \xrightarrow{T_2} P'_2$ are true (our approach). This transformation corresponds to the matrix R'_1 and R'_2 , because the optical center of cameras C_1 and C_2 are not translated. Then, R'_n are bases, which spans the points of rectified images, and the vectors \hat{x}', \hat{y}' , and \hat{z}' their basis. They are linearly independent and can be written as the linear combination:

$$R'_n = \begin{bmatrix} \hat{x}'_n \\ \hat{y}'_n \\ \hat{z}'_n \end{bmatrix}. \quad (15)$$

Finally, the image transformations T_1 and T_2 are expressed as follow:

$$T_1 = M'_1(M_1)^{-1} \quad (16)$$

$$T_2 = M'_2(M_2)^{-1} \quad (17)$$

The transformations are applied to the original images in order to rectify them (see figure 2). It is not common that the value of rectified pixel, after the transformation, corresponds to its initial values. Therefore, the pixels values of the rectified image are computed by an interpolation.

4 Experimental results and validation

The figure 2 shows some results of the implementation of the proposed method. As it was mentioned, the baseline (B) is wide, and the entries of rotation matrix (R) show that the optical axes are not aligned. Hence, the rectified images (see fig. 2(c) and 2(d)) exhibit strong distortion. Notice

that, a mapped point of the scene on the new rectified image plane could be found, only searching over a horizontal line.

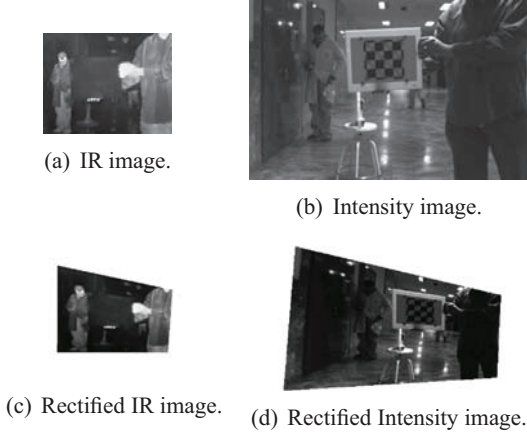


Figure 2: Rectification results.

We propose a evaluation different to one presented in [3] and [1]; instead of measure the errors introduced by the rectification on reconstruction, we measure the misalignment of epipolar lines in both modalities. If any pair of rectified images have epipolar lines parallel and co-linear, then whatever point of first image and its corresponding in the second one is contained in a single line. These points are extracted with a corner detector in order to build an initial point collection for each modality.

The multimodal correspondence problem was dealt, matching points by hand, due to unstable detection of interest point, the change of view and complexity that supposes a multimodal descriptor (still open research line). For these reason a supervised matching approach was followed.

The previous steps were applied to a set of multimodal images, in total 14 images (7 for each modality). In this evaluation frame the corners are an important feature, because images correspond to indoor environments, rich in edges and very differentiable in both modalities.

Let P_I and P_{IR} be a list of pair, that stores

the correspondences by each modality; $P_I(n)$ and $P_{IR}(n)$ are points of form $[x, y]$. The accuracy of rectification is measured by root-mean-squared error of the vertical positions of a pair of correspondences. It takes as reference the y coordinates at $P_I(n)$, and measuring its difference against to its corresponding y coordinate in P_{IR} . Formally, it defined as:

$$E_a = \left(\frac{1}{N} \sum_{(n)} |P_I(n) - P_{IR}(n)|^2 \right)^{\frac{1}{2}}, \quad (18)$$

where N is the total number of pairs. The error of alignment (E_a) is computed for every pair of conjugate points varying the interpolation method. Finally, the experiment is repeated 7 times, changing the couple of multimodal images. The results are shown in table 1, and values of error are the central tendency (mean), together with its respective standard deviation.

Four alternative interpolation methods have been tested, in order to find one that reduces the error of alignment. The evaluated interpolation methods are *linear*, *splines*, *cubic functions*, and *nearest neighbor*.

Table 1: Vertical alignment error.

Interpolation method	E_a (px)	Std. Dev. E_a
Linear	2.93	0.31
Spline	3.44	0.26
Cubic	3.46	0.28
Nearest neighbor	3.11	0.18

In general, the standard deviation of error in table 1 indicates that the measurements of E_a are clustered closely around the mean, confirming the consistency of results and precision level in the measurements.

Linear interpolation is the method with minimum error. However, its precision is wider than

other one. The best method of interpolation has an accuracy of 2.93 pixels, and a precision of ± 0.31 pixels. Therefore, the follow stage in the pipeline, will have an initial accumulated error similar to it. Matching algorithms with windows size smaller than 7 px could not find correct correspondences, if it is restricted to one line the space of search, or it is not coded a method with sub-pixel precision.

5 Concluding remarks and future works

A classical mathematical formulation for rectification, that was initially developed for one modality (intensity images), has been extend to two modalities (infrared and intensity images). Moreover, a comparative study reveals significant improvement in the accuracy using linear interpolation, and noise in the pipeline.

The proposed evaluation method shown an important source of error, at first sight introduced by image rectification. But, in [3] and [1] low error rates were reported using intensity images, their success rest in two facts, the quality of calibration, and that some experiments used synthetic data. However, our results were obtained from an unconstrained stereo rig, without controlled condition.

The equations (10), (11), and (12) are constraints mathematically valid for rectification, but impose strong transformations over the axis \hat{x}' , \hat{y}' , and \hat{z}' , which modified the thermal and intensity information. A future works is to include a new restriction that finds the minimum rotation where the image planes are aligned, and minimize the effect of image distortion.

6 Acknowledgement

This work has been supported by the projects TRA2007-62526/AUT and CTP-2008ITT00001 and research programme Consolider-Ingenio 2010:

MIPRCV (CSD2007-00018).

References

- [1] N. Ayache and C. Hansen. Rectification of images for binocular and trinocular stereovision. *ICPR*, 1988.
- [2] J. Bouguet. Matlab camera calibration toolbox. 2000.
- [3] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and App.*, 12(1):16–22, 2000.
- [4] J. Han and B. Bhanu. Fusion of color and infrared video for moving human detection. *Pattern Recognition*, 40(6):1771–1784, 2007.
- [5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. 2 edition, 2004.
- [6] R. I. Hartley. Theory and practice of projective rectification, 1998.
- [7] S.J. Krotosky and M.M. Trivedi. Mutual information based registration of multimodal stereo videos for person tracking. *CVIU*, 106(2-3):270–287, 2007.
- [8] J. Mallon and P.F. Whelan. Projective rectification from the fundamental matrix. 23(7), July 2005.
- [9] Z. Zhang. A flexible new technique for camera calibration. *PAMI*, 22(11), 2000.
- [10] L. Zheng and R. Laganier. Registration of ir and eo video sequences based on frame difference. *Canadian Conf. Computer and Robot Vision*, pages 459–464, 2007.

BeaStreamer-v0.1 : a new platform for Multi-Sensors Data Acquisition in Wearable Computing Applications.

Pierluigi Casale*, Oriol Pujol* and Petia Radeva*

* *Computer Vision Center, Campus UAB, Edifici O , Bellaterra, Spain*
Dept. of Applied Mathematics and Analysis, University of Barcelona, Barcelona, Spain
E-mail: pierluigi@cvc.uab.es

Abstract

In this paper, we present *BeaStreamer-v0.1*, a new wearable computing platform designed fusing the Beagleboard hardware platform and the GStreamer software platform. The device has been designed for monitoring a variety of day-to-day activities and to be used as a 24/24h digital personal assistant. *BeaStreamer-v0.1* can acquire data collected from multiple sensors in controlled and uncontrolled environments. The benefits of using *BeaStreamer-v0.1* are multiple. First, the small size of the Beagleboard allows to use a really portable computer device. In addition, Beagleboard ensures laptop-like performances despite its dimensions. Using GStreamer makes managing the parameters in the acquisition of many different media types simple and allows to joint the acquisition of different types of data under a unique and compact framework. We demonstrate how the acquisition of audio, video and motion data can be easily performed by *BeaStreamer-v0.1* and we point some highlights in the computational power of the system, some of them to be exploited as future lines of the work.

Keywords: Wearable Sensors, BeagleBoard, Social Sensors, Multimodal Data Fusion, Pattern Recognition Applications.

1 Introduction

In a recent interview to CNN, Gordon Bell (from *Microsoft Research*) tells how and why he had been recording every single event in his life over the last decade. He carried around video equipment, cameras and audio recorders to capture conversations, trips and any kind of experiences. Bell says that this huge amount of data (more than 350 GigaBytes, not including the streaming audio and video) is a replica of his biological memory. This digitized *eMemory* is never forgotten. *Microsoft* is working on a *SenseCam* [1], shown in Figure 1. *SenseCam* is a camera that can be worn around the person's neck and automatically captures every detail of daily life with photos.

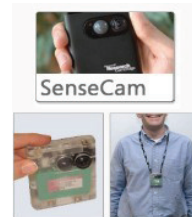


Figure 1: Images of the SenseCam, from [1].

In the same direction, *Intel* has been working on the “*Everyday Sensing and Perception*” project[2]. A team of twelve researchers has been working for three years on the *90/90 Challenge*, i.e. in build-

ing a real-time system for egocentric recognition of handled objects that is accurate at 90% over 90% of our days.

These examples clearly show an increasing interest in developing perception-based systems capable of monitoring a variety of day-to-day activities, both in the research community and in the industry as well. A system being aware of both context and activities during daily life not just would be able to give assistance in memory-retrieval tasks, but also for real-time assistance to not completely self-sufficient people.

In this paper, we present the first version of *BeaStream*, a wearable system for multi-sensors data acquisition and analysis that, in experiences similar to [1] and [2], could be successfully used as a 24/24h digital personal assistant. Using wearable devices such *BeaStream*-v0.1 opens the opportunity to define new use cases, such as healthcare monitoring in patients rehabilitation or studying of social behaviour of people.

The article is organized as follows. In Section 2, we will describe the system and its parts, both hardware and software. In Section 3, we will show some examples of data acquisition and analysis with *BeaStream*-v0.1. Finally, we will discuss conclusions and future works.

2 *BeaStream*-v0.1

BeaStream-v0.1 is a wearable system designed for real-time multi-sensors data acquisition. In this work we use it for acquiring audio, video and motion signals, but its capabilities are not restricted to these data types. Any kind of data flow might be acquired from the system and stored in memory. In Figure 2(a) we show the system disassembled on a table, showing all the components. In Figure 2(b) we show a tester wearing the system.

The system can be easily brought in one hand or in a little bag around the users waist. The audio and video dataflows are acquired using a standard low-cost webcam that can be hooked to the shirt

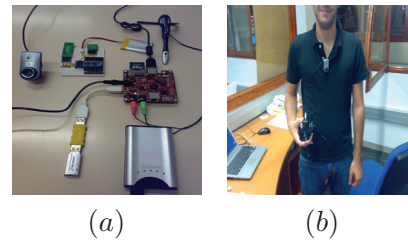


Figure 2: (a) The *BeaStream*-v0.1 system; (b) The *BeaStream*-v0.1 system worn by a tester.

just down the neck or at chest level. An Arduino-based bluetooth accelerometer, can be put in the pant pocket or in the shirt pocket. Audio and video data are acquired via GStreamer, motion data are acquired via bluetooth. Although at the moment, the main functionality of the system is data acquisition, the system has been designed also for data analysis.

The core of the system is based on *BeagleBoard*, an OMAP-based board with high computational power. The system is equipped on-board with a 4 Gigabytes SD-Card where both operating system and data acquired are stored. In the next sections, we describe the hardware components, the development environment and finally, the operating system and the application software running on the board.

2.1 The Hardware Core : BeagleBoard.

The *BeagleBoard* (BB)[3], shown in Figure 3, is a low-power, low-cost single-board computer produced by *Texas Instruments* (TI).

With open source development in mind, BB has been developed to demonstrate the potential of TI's *OMAP3530* system-on-chip, though not all OMAP functionalities are available on the board. The BB sizes approximately $80\text{mm} \times 80\text{mm}$ and it provides all the functionalities of a basic computer.

The *OMAP3530* system-on-chip includes an *ARM Cortex-A8* CPU at 600 MHz which can run Windows CE or Linux, a *TMS320C64x+ DSP* for

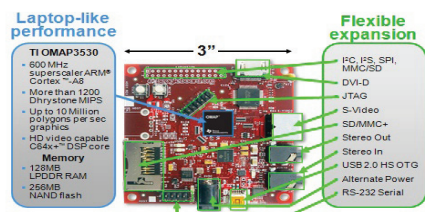


Figure 3: Beagleboard front view, from [3].

accelerated video and audio codecs, and an *Imagination Technologies PowerVR SGX530 GPU* to provide accelerated 2D and 3D rendering that supports OpenGL ES 2.0. Built-in storage and memory is provided through a *Package on Package* chip that includes 256MBytes of NAND flash memory and 256MBytes of RAM. The board carries a single SD/MMC connector, supporting a wide variety of device such as WiFi Cards, SD/MMC Memory Cards and SDIO Cards. One interesting feature of the OMAP3530 is the possibility of booting the processor from SD/MMC card.

Video output is provided through separate S-Video and HDMI connections. A 4-pin DIN connector is provided to access the S-Video output of the BeagleBoard. This is a separate output from the OMAP processor and can contain different video output data from what is found on the DVI-D output. The BB is equipped with a DVI-D connector that uses an HDMI connector. It does not support the full HDMI interface and it is used to provide the DVI-D interface only.

Two USB ports are present on the board. Both ports can be used as host ports, using High Speed USB devices conform to USB 2.0 protocol, using a maximum of 500 mA to power the host device. If additional power is needed or multiple devices as mouse, keyboard and USB mass storage devices must be used, one USB port can be used as OTG (On-The-Go) port to drive a self-powered USB hub. The USB OTG port can be also used to power the board from a standard external USB port. If both USB ports need to be used, there exists an

additional 5 mm power jack to power the board. DC supply must be a regulated and clean 5 Volts supply. The board uses up to 2 Watts of power.

Beagleboard presents on board a populated RS-232 serial connection where a serial terminal is present. Using the terminal, it is possible to set the boot parameters and the size of the video buffer. Furthermore, a 14-pins JTAG connection is present onboard to facilitate the software development and debugging on-board using various JTAG emulators. Two stereo 3.5mm jacks for audio input and output are provided. An option for a single 28 pin header is provided on the board to allow the connection of various expansion cards. Due to multiplexing, different signals can be provided on each pin providing more than 24 actual signal accesses. This header is not populated on the BB and, depending on the usage scenario, it can be populated as needed. Because of the efficient power consumption, the board requires no additional cooling.

Typical usage scenario for the BB are shown in Figure 4. BB might be considered a laptop substitute. There are many projects using BB in

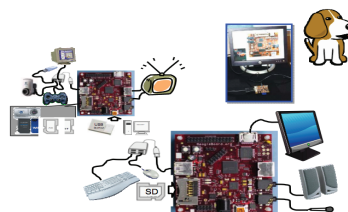


Figure 4: Typical Usage Scenarios for Beagle-board, from [3].

robotic applications ([4], [5]). Nevertheless, up to now, there is no literature using BB as a wearable device, despite of the low dimensions of the board. The major issue of using BB in wearable applications is the need of a portable power supply source. In our applications, we use an A.C. Ryan *MobiliT* external USB battery at 3400mAh, allowing 4 hours of autonomy for the system in complete functionality.

2.2 The Motion Sensor : Arduino

Arduino ([6]) is an open-source electronics prototyping platform based on flexible, easy-to-use hardware and software. *Arduino* can sense the environment by receiving input from a variety of sensors and can affect its surroundings by controlling lights, motors, and other actuators. The microcontroller on the board is programmed using the *Arduino* programming language and the *Arduino* development environment. The boards can be built by hand or purchased preassembled. The software can be downloaded for free. Although *Arduino* was built for artists and hobbyists, there are many people working on real electronic interactive projects, thanks to the rapid prototypization *Arduino* allows. We prototype a Bluetooth-based accelerometer using the *Arduino* board, an analogic *ADXL 345* accelerometer and a *BlueSMiRF Gold* Bluetooth modem.

2.3 The Development Side : OpenEmbedded + Ångström.

Openembedded (OE) offers a complete cross-compiler environment and allows developers to create complete Linux Distributions for embedded systems. OE offers different kernels for the BB. All kernels come with several patches and the support of the BB hardware is not perfect yet. Figure 5 summarizes the current status of hardware support, where it is possible to see that all the features of OMAP processors available in the BB are actually available for the use.

Boot	USB OTG	USB host	DVI	Audio out	Audio in	S-Video out	MMC / SD	RS232	Flash	DSP	SGX
2.6.27-r11	host only			Working on	NO	NO					
2.6.28-r12											

Figure 5: Current Status of Hardware Support for BB in OE, from [3].

The Linux Kernel 2.6.28r12 runs on our system. This particular Linux Kernel has *V4L2* (Video for Linux 2) drivers, allowing to plug in the system almost every Linux-compatible webcam. Furthermore, it contains *BlueZ*, the official Bluetooth protocol stack. Using this kernel version, several problems appear when using the DSP on the board. At the moment, all the problems related to DSP have been officially resolved in the Linux kernel version 2.6.29 but there exist concerning issues regarding the use of USB ports. For that reason, we use the “old” kernel release leaving aside, provisionally, the DSP functionalities.

The *Ångström Distribution* (AD) is the Linux Distribution running on the board. AD is a specific Linux distribution for embedded systems. A complete image of AD can be built using OE or with an online tool [9], where it is possible to choose the packages to be installed in the system. In the distribution we build, we include a toolchain for developing source codes on board. We install the *arm-gcc* compiler, *arm-g++* compiler and the *Python-Numpy* development environment. In addition, we build the *GStreamer* and the *OpenCV* packages, as we will explain in the next section.

2.4 The Software Side: GStreamer + OpenCV

GStreamer is a framework for creating streaming media applications. The *GStreamer* framework is designed to make easy writing applications handling audio/video streaming. Nevertheless, *Gstreamer* is not restricted to audio and video, and it can process any kind of data flow. One of the most obvious uses of *GStreamer* consist in using it to build a media player. *GStreamer* already includes components for building a media player that can support a very wide variety of formats, including MP3, Ogg/Vorbis, MPEG-1/2, AVI, and more.

The main advantages of *GStreamer* are that the software components, called plugins, can be mixed and matched into arbitrary pipelines so that it is

possible to write complete streaming data editing applications. Plugins can be linked and arranged in a pipeline. The GStreamer core function is to provide a framework for connecting plugins, for data flow management and for media type handling and negotiation. Using GStreamer, performing complex media manipulations becomes very easy and it integrates an extensive debugging and tracing mechanism. In BeaStreamer-v0.1, we use a pipeline for acquiring audio and video from webcam, with the possibility to encode the dataflow with the request quality and the resolution and the possibility to change the acquisition parameters at run time.

In addition, we compile on BeaStreamer-v0.1 the well-known *OpenCV* libraries and its Python bindings.

3 Experiments with BeaStreamer-v0.1

In this section, we show some experiment performed with BeaStreamer-v0.1 to demonstrate its capabilities. In Figure 6, we show six sequential photos taken wearing BeaStreamer-v0.1, walking in the street. Using GStreamer, we take photos with a framerate of one photo/second with a resolution of 320×240 pixels, compressed in *jpeg* format. At the same time, in a separate thread, we record a continuous audio flow from the webcam microphone, sampled at 44100 samples/s and compressed in *ogg* format. GStreamer allows setting online the parameters of acquisition making simple to change the resolution of photo and the encoding audio quality.

In order to get an estimate of the autonomy of the system, we record an audio/video stream compressed in *ogg* format and receive motion data from the bluetooth accelerometer. We are able to record up to *4 hours* of audio, video and motion data.

Using OpenCV, we setup a face detector running on photos acquired sequentially with GStreamer. The face detector can compute detections at a

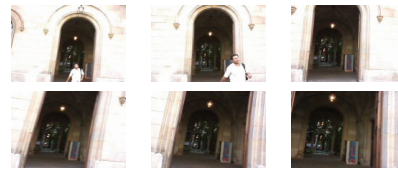


Figure 6: A sequence of photos taken with BeaStreamer-v0.1

framerate of *5-10* frames/second depending of the images resolution, without using DSP. An example of faces successfully detected is shown in Figure 7. Using images with resolution of 80×60 pixels, the face detector can scan the image in less than *100 ms* and detect faces in *200 ms*.

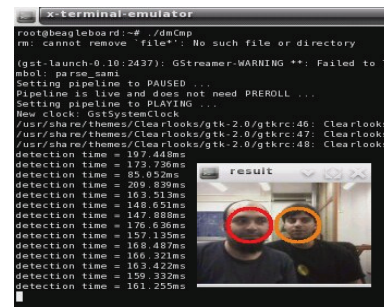


Figure 7: Face Detector running on BeaStreamer-v0.1.

Finally, in Figure 8 we show how BeaStreamer-v0.1 receives motion data. The acceleration analog values are converted in *10-bit* values from Arduino *ADC* (Analog-to-Digital Converter) at 40Hz , stored in a buffer and sent every second via Bluetooth as *UNICODE* characters with a label showing the axis. BeaStreamer-v0.1 receives the data and stored them in a text file.

4 Conclusions and Future Works

In this paper we presented BeaStreamer-v0.1, a new platform for multi-sensors data acquisition. BeaStreamer-v0.1 is small and easy to bring, al-


```

x-terminal-emulator
root@beagleboard:~# python btConnect.py
Connected with 00:08:60:01:4C:FE
522:400
x521y522:400
x521y522:401
x521y521:40
1
x521y522:401
x
x521y522:401
x5
x521y522:400
x521y522:401

```

Figure 8: Acquisition of motion data.

lowing its use in wearable computing applications, for controlled and uncontrolled environments. We showed that different types of data can be easily acquired joining the potentiality of the Beagleboard and GStreamer.

The Beagleboard allows to connect different types of sensors communicating via Bluetooth or via the principal types of communication protocols implemented in the OMAP processor. GStreamer provide a framework to manage different types of data flows using a single and coherent environment.

At the moment, just a basic face detector has been developed on the system to demonstrate its capabilities. Furthermore, the computational power of the system will increase as soon as the DSP side of the OMAP will be completely operative.

Finally, we consider that unifying Bluetooth and general sensors acquisition under GStreamer will provide a powerful and complete platform for general multi-sensors data acquisition and analysis.

Acknowledge

This work is partially supported by a research grant from projects TIN2006-15308-C02, TIN2009-14404-C02, FIS-PI061290 and CONSOLIDER-INGENIO 2010 (CSD2007-00018), MI 1509/2005.

Special thanks to the “Computer Vision Group/DSPLab” of *Istituto di Fisiologia Clinica*, Consiglio Nazionale di Ricerche (CNR), Pisa,

Italy for their help in the first stage of system development.

References

- [1] Microsoft Research, *Introduction to SenseCam*, <http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/>
- [2] X. Ren, M. Philipose, “Egocentric Recognition of Handled Objects: Benchmark and Analysis”, *Proc. of 1th Workshop on Egocentric Vision*, <http://www.seattle.intel-research.net/egovision09/>
- [3] Beagleboard.org, *Beagleboard Technical Documentation*, <http://beagleboard.org>
- [4] Home Brew Robotic Club, *A tutorial on setting up a system to do image processing with a BeagleBoard.*, <http://www.hbrobotics.org/wiki>
- [5] Beaglebot, *Beagle powered robot* <http://www.hervanta.com/stuff/Beaglebot>
- [6] Arduino.cc , *Arduino Technical Documentation*, <http://www.arduino.cc>
- [7] Openembedded.org, *Open-Embedded Official Manual*, <http://docs.openembedded.org/usermanual.html>
- [8] The Ångström Distribution, *Ångström Documentation*, <http://linuxtogo.org/gowiki/Angstrom>
- [9] The Ångström Distribution, *Online Ångström Building*, <http://www.angstrom-distribution.org/narcissus/>
- [10] GStreamer, *GStreamer Technical Documentation*, <http://gstreamer.freedesktop.org/documentation/>

Quadric Surface Fitting: Orthogonal versus Estimated Distances

Mohammad Rouhani and Angel Sappa

Computer Vision Center, 08193 Bellaterra, Spain

E-mail: {rouhani, asappa}@cvc.uab.es

Abstract

This paper presents a comparative study between two different approaches for fitting general quadric surfaces to 3D data. In both cases the same nonlinear minimization scheme—Levenberg-Marquardt algorithm—is used for finding the corresponding surface parameters. In the first case an iterative approach that find the shortest distance—orthogonal distance— between the data points and the fitted surface is considered. In the second case an approximation of that orthogonal distance is used for solving the minimization problem. The performance of both fitting schemes is compared with different real and synthetic data sets.

Keywords: Surface Fitting; Orthogonal Distance Fitting; PCA-based Distance Estimation.

1 Introduction

In general, surface fitting algorithms can be classified into two main categories, according to the definition of the metric to be minimized: *i)* algebraic and *ii)* geometric fitting [1]. In the *algebraic fitting* the model is described by means of an implicit equation $f(\mathbf{c}, \mathbf{X}) = 0$ with surface parameters \mathbf{c} and the error distances are defined with the deviations of functional values from the expected value (i.e., zero) at each given point. Although

this approximation is highly attractive because of its closed-form solution and direct computation, it should be noticed that algebraic fitting leads to a biased estimation for small surfaces with low curvature.

On the contrary, in the *geometric fitting*, also referred as *orthogonal distance fitting*, the error distance is defined as the shortest distance from the given point to the fitting surface. Since there is no closed formula to compute the shortest distance between a point p and a general quadric surface two solutions have been proposed in the literature: *a)* compute the orthogonal distance by means of an iterative approach; and *b)* compute an approximation to this distance and use it as the residual value of p . Different criteria have been proposed for estimating this residual value. In [2] the minimum distance between the given point p and the fitting surface along the x, y, z axes is used as an orthogonal distance approximation. A more evolved approach, based on finding a closest point to the surface using an estimation of the surface orientation, is proposed in [6].

Although different approaches have been proposed for quadric surface fitting, only a few comparisons of their advantages/disadvantages have been proposed in the past (e.g., [3], [4]). In the current work two recently published quadric surface fitting approaches are implemented under the same minimization framework to make a compar-

ative study. The rest of the paper is organized as follows. Section 2 describes the problem we are focusing on as well as some backgrounds. Section 3 details the two approaches implemented for comparisons: orthogonal distance based [1] and distance estimation based [6]. A brief description of the Levenberg-Marquardt algorithm is provided in section 4 since it has been used for the non-linear minimization. Section 5 gives experimental results and comparisons. Finally conclusions are provided in section 6.

2 Problem Formulation

Fitting problems aim at finding a curve or surface *close* to a given cloud of points $\mathbf{X} = \{p_i\}_{i=1}^n$. The current work is focussed on quadric surfaces, which have been widely used in computer vision for 3D object representation, due to their simplicity and compactness. A general quadric surface is described as zero set of quadratic polynomials:

$$f(\mathbf{c}, \mathbf{X}) = c_1x^2 + c_2y^2 + c_3z^2 + c_4xy + c_5xz + c_6yz + c_7x + c_8y + c_9z + 1 = 0. \quad (1)$$

In order to find the optimal set of parameters, \mathbf{c} , we must minimize the total distance between \mathbf{X} and surface $f(\mathbf{c}, \mathbf{X}) = 0$. Hence, for this purpose we first need to find the corresponding set of points $\hat{\mathbf{X}} = \{\hat{p}_i\}_{i=1}^n$ over the surface in order to minimize the total distance:

$$\min_{\mathbf{c}} \left(\sum_{i=1}^n \min_{\hat{p}_i} d(p_i, \hat{p}_i) \right), \quad (2)$$

where $d(p_i, \hat{p}_i)$ is the Euclidean distance between p_i and \hat{p}_i .

Theoretically, both unknown surface parameters and correspondence between couples of points must be found simultaneously. In general, this problem is tackled by first assuming an initial set of parameters \mathbf{c} , and then minimizing the total distance between the original set points and the corresponding ones, on the *assumed* surface. This process is iterated till convergence is reached.

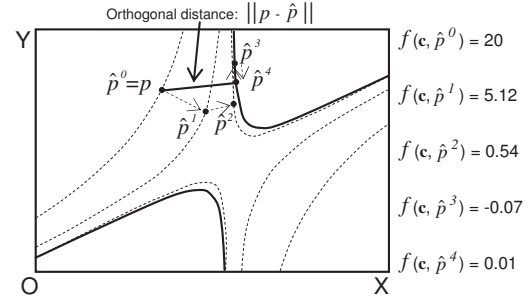


Figure 1: Orthogonal distance computed by means of the iterative approach proposed in [1]. Solid curve correspond to the $f(\mathbf{c}, \mathbf{X}) = 0$, while dashed ones show the level curves obtained after each iteration of eq. (4); \hat{p} converges to the curve ($f(\mathbf{c}, \mathbf{X}) \cong 0$) just after four iterations.

In order to find the optimal solution for (2) we split it up into two stages: 1) Point correspondence search 2) Surface parameter refinement. The first stage deals with the inner part of (2), while the second one with the outer part of (2). Both stages are detailed below.

3 Point correspondence search

As mentioned above the first stage consists in finding for every point p its corresponding one \hat{p} on the surface. The best couple is the one that minimize $d(p, \hat{p})$. Two different approaches have been proposed in the literature for solving this problem: a) find the shortest distance by solving a non-linear system through an iterative approach (e.g., [9], [1]); and b) compute an estimation of that shortest distance (e.g., [6], [2]). In the current work two recent approaches have been implemented and compared; they are summarized below.

3.1 Orthogonal Distance

In general, the orthogonal distance is computed by means of iterative algorithms. Recently, [1, 8] proposes the *direct method*, which is based on the general properties of the shortest distance point. The

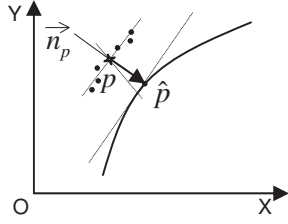


Figure 2: Shortest distance estimation between point $p(x,y)$ and a quadric curve; \hat{p} is obtained by computing the surface orientation through a PCA based approach [6].

necessary condition for the shortest distance is that on the one hand the line connecting \hat{p} with p should be parallel to the ∇f at \hat{p} . In other words, $\nabla f \times (\hat{p} - p) = 0$, where $\nabla = (\partial/\partial x, \partial/\partial y, \partial/\partial z)^T$. On the other hand, the contacting point \hat{p} is somewhere on the surface. Merging these two conditions, the following system of equations must be solved:

$$\mathbf{f}(\hat{\mathbf{p}}) = \begin{pmatrix} f \\ \nabla f \times (\hat{\mathbf{p}} - \mathbf{p}) \end{pmatrix} = \mathbf{0}. \quad (3)$$

The root of \mathbf{f} is found by means of an iterative approach based on a generalized Newton method. The iterative approach is initialized by assuming every \hat{p} to be equal to its corresponding p in \mathbf{X} . Hence at each iteration (k), every \hat{p}^k is updated as follow:

$$\begin{aligned} \hat{p}^{k+1} &= \hat{p}^k + \alpha \Delta p, \\ \frac{\partial \mathbf{f}}{\partial \hat{p}} \bigg|_{\hat{p}^k} \Delta p &= -\mathbf{f}(\hat{p}^k), \end{aligned} \quad (4)$$

where Δp is the refinement vector computed by solving the second system of equations; α is the refinement step. After convergence, the resulting \hat{p} is the one that minimize the inner part of (2), as we were looking for. Fig. 1 shows an illustration of this iterative scheme.

3.2 PCA-based Distance Estimation

Instead of computing the real shortest distance, [6] proposes to estimate it, avoiding thus iterative approaches. First a normal vector \vec{n}_p for each point

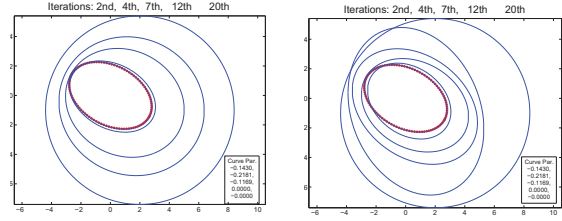


Figure 3: Ellipse fitted by using: (left) orthogonal distance and (right) PCA based distance estimation. In both cases the same initialization has been used (outer circle) and results obtained after 2nd, 4th, 7th, 12th, and 20th iteration are presented. Both approaches converge to the same set parameters.

p is computed by using principal components analysis (PCA) in a small $N \times N$ neighborhood centered at each point. A similar approach has been used in the past for surface reconstruction from unorganized points [7] (Fig. 2 shows an illustration in 2D space). In other words, \vec{n}_p is defined as the eigenvector of the local covariance matrix CV associated with the smallest eigenvalue:

$$CV = \frac{1}{s} \sum_{i=1}^s (p_i - \tilde{p}) \cdot (p_i - \tilde{p}), \quad (5)$$

where $\tilde{p} = \frac{1}{s} \sum_{i=1}^s p_i$ is the mean position of the neighboring points in the $N \times N$ region. Finally, \hat{p} is computed as the intersection of the surface $f(\mathbf{c}, \mathbf{X}) = 0$ with a line passing through p and parallel to $\vec{n}_p = (n_1, n_2, n_3)$:

$$\frac{x - x_p}{n_1} = \frac{y - y_p}{n_2} = \frac{z - z_p}{n_3}. \quad (6)$$

4 Surface parameter refinement

As a result from the previous stage the set of points $\{\hat{p}_i\}_{i=1}^n$, corresponding to every p_i in \mathbf{X} has been found. Since each \hat{p}_i lies on the surface, every distance $d(p_i, \hat{p}_i)$ can be easily expressed as a function of surface parameters:

$$d_i(\mathbf{c}) = \|p_i - \hat{p}_i\|, \quad (7)$$

more precisely, equation $f(\mathbf{c}, \hat{p}_i) = 0$ provides us a link between surface parameters and $\{\hat{p}_i\}_{i=1}^n$ set, and distances as a consequence. This matter will be later on used for computing the sensibility of distances.

The aim at this stage is to minimize the outer part of (2), $\sum_1^n d_i(\mathbf{c})$, with respect to surface parameters. In the current work this has been performed through the Levenberg-Marquardt algorithm [5]:

$$\begin{aligned} \mathbf{c}^{t+1} &= \mathbf{c}^t + \beta \Delta \mathbf{c}, \\ (J^T J + \lambda \text{diag}(J^T J)) \Delta \mathbf{c} &= J^T D, \end{aligned} \quad (8)$$

where β is the refinement step; $\Delta \mathbf{c}$ represents the refinement vector for the surface parameters; λ is the damping parameter in LM algorithm; vector $D = (d_1(\mathbf{c}^t), \dots, d_n(\mathbf{c}^t))^T$ corresponds to the distances; and $J = (J_{i,j})_{n \times 9}$ is the Jacobian matrix computed as indicated below:

$$\begin{aligned} J_{ij} &= \frac{\partial d_i}{\partial c_j} = - \frac{(p_i - \hat{p}_i)^T}{\|p_i - \hat{p}_i\|} \frac{\partial \hat{p}_i}{\partial c_j} \\ &= \frac{(p_i - \hat{p}_i)^T}{\|p_i - \hat{p}_i\|} \frac{\partial f / \partial c_j}{\partial f / \partial \hat{p}_i} \end{aligned} \quad (9)$$

these equalities are derived from the chain rule and shows the sensibility of distances with respect to surface parameters. Equations in (8) are iterated till convergence (or a maximum number of iterations) is reached, giving rise to the best set of parameters for fitting the given set of points $\mathbf{X} = \{p_i\}_{i=1}^n$.

5 Experimental Results

The two approaches presented above have been compared by fitting quadric surfaces using the LM algorithm. The same LM framework has been used in both cases; CPU time and accuracy results were used as comparison criteria. The accuracy of the results obtained by using both methods has been

Table 1: Fitting results after 100 LM iterations; ellipse defined by 100 noisy data (CPU time in sec)

	Orthogonal Distance Based	Estimated Distance Based
CPU time	2.62	0.31
Fit. Error	0.51	0.71

evaluated by measuring the fitting error. In both cases the provided fitting error is computed over the hole set of given points using the same metric (orthogonal distance).

Fig. 3 shows results obtained at different iterations when an ellipse, defined by 100 points, is fitted by using the LM minimization scheme. In both cases the same initialization has been used (outer circle) and 100 iterations of LM algorithm were performed. Results in Fig. 3(*left*) were obtained with an orthogonal distance computed with [1]; it took 1.80 sec. and the final fitting error was $6.2e^{-22}$. Fig. 3(*right*) corresponds to the case where the shortest distance is estimated by a PCA based approach [6], it took 0.31 sec. while the final fitting error was $6.27e^{-24}$. In this case, since a set of synthetic and uniformly distributed points has been used the fitting algorithm based on a PCA distance estimation has the best performance (i.e., similar fitting error has been obtained almost six time faster). On the contrary to the previous result, Table 1 shows CPU time and error when a set of 100 noisy data is fitted with the two approaches. It can be appreciated that in this case a better fitting error is obtained using the orthogonal distance based approach; however the distance estimation based approach is considerably faster.

Fig. 4(*left*) shows results obtained after fitting a set of 606 3D noisy points using a fitting approach based on the orthogonal distance; it took 20 sec. to reach a fitting error of 1.73. A similar result has been obtained using an estimated distance; in this case the fitting error was 2.5 and it took 12.11 sec. to compute the surface parameters. Finally, Fig. 4(*right*) shows a cylinder fit-

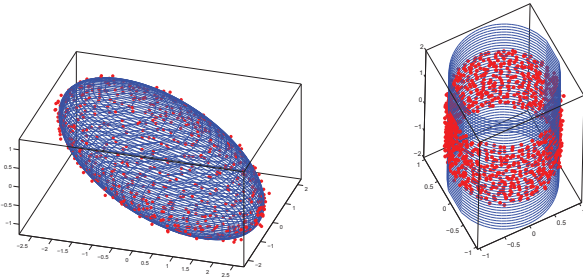


Figure 4: 3D surfaces obtained after 100 LM iterations using the orthogonal distance (both approaches converge to similar surface parameters).

ting a set of 1071 3D noisy points. This result has been obtained using an orthogonal distance based approach that took 41.8 sec. with a fitting error of 0.85. Again, a similar result was obtained using an estimation of the distance, but in this case six times faster, with a fitting error of 1.16. The set of 3D data points corresponding to the ellipsoid shown in Fig. 4(*left*) has been used for comparing the robustness of both approaches to noisy data. Fig. ?? illustrates the evolution of the fitting error as a function of the noise in the given set of 3D data points. The noise percentage, varying from 0 to 20%, shows the relative deviation from the ellipsoid centroid. As it was expected in all the cases the orthogonal based distance approach reach a smaller fitting error. However, more CPU time was required (on average more than six times).

The proposed approaches have been also compared using real range images obtained with the MSU's Technical Arts Arts 100× Range Scanner (images containing isolated quadric surfaces were selected). Due to space limitations only a single illustration is provided in Fig. 5. Figure 5(*left*) shows the intensity image corresponding to the patch fitted by both approach (1000 3D data points). Similarly to the synthetic cases, both fitting approaches converge to the same result, Fig. 5(*right*) (smaller fitting error was obtained with

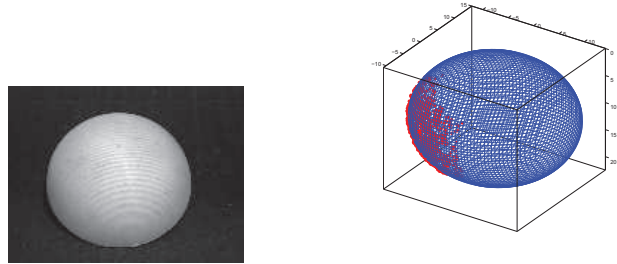


Figure 5: (*left*) Intensity image corresponding to the patch to be fitted. (*right*) Surface obtained with the distance estimation based approach (both approaches reach the same result).

the orthogonal distance based approach). The distance estimation based approach was almost five time faster.

6 Conclusions

This paper presents a comparison of two different fitting approaches. In both cases surface parameters are computed through the same minimization framework—LM algorithm. The shortest distance, referred as orthogonal distance, is used in the first case. It is computed by means of an iterative approach. Instead of relying on a costly iterative approach, in the second case an estimation of the shortest distance is considered. As a conclusion we can say that even though several algorithm have been proposed for quadric surface fitting in the context of computer vision there is a trade off between CPU time and accuracy of surface parameters for selecting the best one; this trade off get more evident when it is increased the number of points to be fitted or the percentage of noise. As a future work we will study the possibility of merging both approaches into a hybrid scheme, to exploit advantages of each one of them. Preliminary results are encouraging. The idea is to use the fastest one (estimated distance based) for computing a coarse approximation and then the most accurate one (orthogonal distance based) for reach-

ing the final result.

References

- [1] S. Ahn, W. Rauh, H. Cho, and H. Warnecke. Orthogonal distance fitting of implicit curves and surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(5):620–638, May 2002.
- [2] Y. Chen and C. Liu. Quadric surface extraction using genetic algorithms. *Computer-Aided Design*, 31(2):101–110, 1999.
- [3] P. Faber and R. B. Fisher. Euclidean fitting revisited. In *Proc. 4th Int. Workshop on Visual Form*, pages 165–175. Springer-Verlag LNCS, 2001.
- [4] P. Faber and R. B. Fisher. Pros and cons of euclidean fitting. In *Proceedings of the 23rd DAGM-Symposium on Pattern Recognition*, pages 414–420, London, UK, 2001. Springer-Verlag.
- [5] R. Fletcher. *Practical Methods of Optimization*. New York: Wiley, 1990.
- [6] P. Gotardo, O. Bellon, K. Boyer, and L. Silva. Range image segmentation into planar and quadric surfaces using an improved robust estimator and genetic algorithm. *IEEE Trans. on Systems, Man, and Cybernetics Part B: Cybernetics*, 34(6):2303–2316, 2004.
- [7] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle. Surface reconstruction from unorganized points. In *SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, pages 71–78, New York, NY, USA, 1992. ACM.
- [8] M. Rouhani and A. Sappa. Quadric surface fitting: Orthogonal versus estimated distances. In *LNCS Advanced Concepts for Intelligent Vision Systems*, pages 121–132, Bordeaux, France, 2009. Springer-Verlag.
- [9] G. Taubin. Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(11):1115–1138, 1991.

Author Index

al Haj, Murad	137	Mirza-Mohammadi, Mehdi.....	197
Aldavert, David.....	70	Moreno, Jaime	119
Álvarez, José Manuel.....	1,13	Mozerov, Mikhail	143
Amato, Ariel.....	143	Murray, Naila.....	76
Azpiroz, Fernando.....	131	Nourbakhsh, Farshad	204
Bagdanov, Andrew D....	137,149,167,210	Onkarappa, Naveen.....	173
Baldrich, Ramón	58	Otazu, Xavier	76,119
Barrera, Fernando	216	Pal, Umapada.....	112
Beigpour, Shida.....	34	Párraga, Alejandro	40
Benavente, Robert.....	46	Pedersoli, Marco	210
Bernal, Jorge	125	Ponsa, Daniel	7,13
Casale, Pierluigi	220	Pratim Roy, Partha.....	112
Chakraborty, Bhaskar	149	Pujol, Oriol	197,222
Cheda, Diego.....	7	Radeva, Petia	131,155,161,191,222
Ciampi, Francesco	197	Roca, Jordi	40
Clavelli, Antonio.....	100	Rojas Vigo, David A.	82
Diego, Ferran	13	Rojas, Mario A.	88
Drozdal, Michal.....	131	Rouhano, Mohammad	228
Elfiky, Noha.....	179	Rubio Ballester, José Carlos	23
Escalera, Sergio.....	155,197	SalahEldeen, Hany M.	46
Escudero, Alberto	155	Sánchez, Javier	125
Gatta, Carlo	161	Sappa, Angel.....	173,216,228
Gerónimo, David.....	29	Seguí, Santi	131
Gevers, Theo	191	Serrat, Joan	13,23
Gibert, Jaume	185	Shahbaz Khan, Fahad	52
Gómez, Juan Diego	161	Toledo, Juan Ignacio.....	94
Gong, Wenjuan	167	Toledo, Ricardo	70
Gonfaus, Josep M ^a	191	Valveny, Ernest.....	106,185,204
González, Jordi 137,143,149,167,179,191,210		van de Weijer, Joost	34,52,82,94
Gordo, Albert	106	Vanrell, Maria.....	40,46,52,64,76,119
Karatzas, Dimosthenis	100,204	Vázquez, David.....	29
Lladós, Josep.....	112	Vázquez, Eduard.....	58
López, Antonio	1,7,13,17,29	Vázquez Corral, Javier.....	64
Lumbreras, Felipe	216	Vilariño, Fernando.....	131
Malagelada, Carolina	131	Villanueva, Juan José.....	210
Marín, Javier	17	Vitrià, Jordi	88,131,155
Masip, David.....	88		