Proceedings of the Eighth CVC Workshop
CVCR&D2013

Centre de Visió per Computador
Bellaterra, Catalonia (Spain)
October 25, 2013

ISBN: 978-84-940902-2-6
Deposito Legal:
Printed by Ediciones Gráficas Rey, S.L.

Preface

This book contains the papers presented at the Eighth CVC Workshop on Computer Vision Trends and Challenges (CVCR&D'2013). The workshop was held at the Computer Vision Center (Universitat Autònoma de Barcelona), the October 25th, 2013. 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, and also, to discuss about challenges and future perspectives. In addition, the workshop is the welcome event for new people that recently have joined the institute.

The program of CVCR&D is organized in a single-track single-day workshop. It comprises several sessions dedicated to specific topics. For each session, a doctor working on the topic introduces the general research lines. The PhD students expose their specific research. A poster session will be held for open questions. Session topics cover the current research lines and development projects of the CVC: Medical Imaging, Medical Imaging, Color & Texture Analysis, Object Recognition, Image Sequence Evaluation, Advanced Driver Assistance Systems, Machine Vision, Document Analysis, Pattern Recognition and Applications. We want to thank all paper authors and Program Committee members. Their contribution shows that the CVC has a dynamic, active, and promising scientific community.

We hope you all enjoy this Eighth workshop and we are looking forward to meeting you and new people next year in the Ninth CVCR&D.

Bellaterra, October 2013

Jorge Bernal and David Vázquez
Workshops General Chairs

Proceedings of the Eighth CVC Workshop on
Computer Vision Trends and Challenges
CVCR&D 2013

Workshop Organization

GENERAL CHAIRS

Bernal, Jorge
Vázquez, David

ORGANIZATION COMMITTEE

Lladós, Josep
Vilariño, Fernando
Culleré, Montse
Martín, Mireia
Rionegro, Raquel

PROGRAM COMMITTEE

Amato, Ariel
Bagdanov, Andrew
Benavente, Robert
Diaz, Katerine
Escalera, Sergio
Fornés, Alicia
Gil, Debora
Gonfaus, Josep Mª
Gonzàlez, Jordi
Hernàndez, Aura
Igual, Laura
Karatzas, Dimosthenis
López, Antonio
Lumbreras, Felipe
Mas, Joan
Masip, David
Otazu, Xavier
Pàrraga, Alejandro
Ponsa, Daniel
Raducanu, Bogdan
Rusiñol, Marçal
Sabate, Anna
Sánchez, Gemma
Sánchez, Javier
Sappa, Angel
Valveny, Ernest
van de Weijer, Joost
Vanrell, Mari
Vazquez, Javier
Vitrià, Jordi

Table of Contents

3. Pattern Recognition and Computer Vision Methods

4. Medical Image Analysis

5. MSc's Thesis

5. Recent and Upcoming PhD's Thesis

Fast Structural Matching for Document Image Retrieval through Spatial Database

Hongxing Gao
Advisor: Marçal Rusiñol, Dimosthenis Karatzas,Josep Lladós
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail:hongxing@cvc.uab.es

Keywords: Document image retrieval, distance transform MSER,spatial database,large dataset

1 Summary of Previous and Current Work

We proposed distance transform based MSER (DTMSER) for document images that extracts multi-level semantical key-regions which basically corresponding to letters, words, paragraphs and columns. Meanwhile, DTMSER efficiently compute the hierarchy of the key-regions defining how letter-regions merge to be words regions and words to paragraphs. The comparison of DTMSER and other detectors like MSER, SIFT could be found in [1].

Another issue for structural document image retrieval is how to evaluate the similarity between the key-regions trees whose branches carry *contain* relations. On the other hand, spatial database holds advanced techniques for dealing with such spatial relations. Consequently, we propose spatial indexing framework as showed in Figure 1. SIFT is employed to compute feature and k-means is performed for quantization. Afterwards, spatial database stores the quantized key-region and builds spatial index. During query time, we decompose the key-region tree as a list of pairs which is then used to retrieve all the pairs that hold the same property(label and *contain* relation) . Voting process based on RANSAC is then performed. The paper on this idea is in preparation.

2 Future Work and Challenges

Future works may fall into querying more detailed spatial relation like 'overlap' 'neighboring' or 'left/right of' rather than 'contains' only. Another possible further research is improving the consistence of DTMSER detector and generalize the framework to part-based document image analysis.

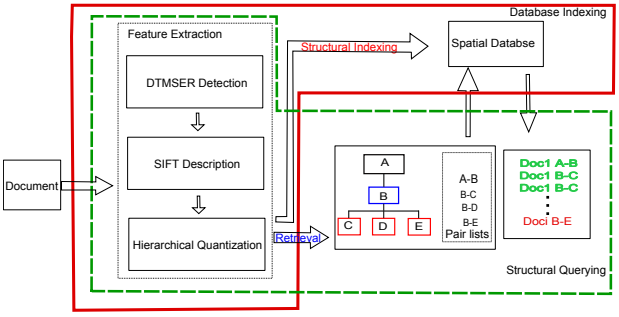


Figure 1: Example of graphical representation of the presented methodology.

Publications

[1] H.Gao, M.Rusiñol, D.Karatzas, J.Lladós, T.Sato, M.Iwamura and K.Kise, Key-region Detection for Document Images–Application to Administrative Document Retrieval. In *12th International Conference on Document Analysis and Recognition*, 230-234, 2013.

Hongxing Gao received his B.S. degree in Automation from Shandong University of Science and Technology, Shandong, China, in 2008. He received his M.S. degree in 2011 from East China University of Science and Technology, Shanghai, China. And he is currently a Ph.D. Candidate in the Department of Computer Science at the Universitat Autonoma de Barcelona. His research interests are in the areas of structure extraction for document image, feature detection and description, structural image retrieval, spatial indexing.

Language models for handwriting recognition

Núria Cirera Turigues
Advisor: Alicia Fornés, Josep Lladós
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: {ncirera,afornes,josep}@cvc.uab.cat

Keywords: Language Models, Handwriting Recognition, Document Analysis, Pattern Recognition

1 Summary of Previous and Current Work

Since I joined CVC as a PhD student I have been working on the development of language models for syntactically structured documents. The project I am involved with partially aims to transcribe the handwritten content of certain historical documents. Recognizing an out-of-vocabulary (OOV) word, this is a word that is not present in the training set, is not a triviality. This is illustrated in Figure 1.

We proposed a language model composed by a regular grammar (which adds syntactic structure to the recognition) that includes the option of recognizing a word letter by letter, based on character bi-gram probabilities estimated on the training set. This means that we restrain the recognition to fulfil a syntactic structure while in certain syntactic categories we can recognize OOV words by recognizing its characters. This work was published in [1].

In the same research direction, I am currently improving this proposed language model and investigating other ways to combine language models such as stochastic grammars.

2 Future Work and Challenges

During my research I have learned that handwriting recognition, independently of the recognition technique used, relies on the language model that we provide to the system as well as the quality of the character modelling/generation. Based on this, I plan to research on these two areas to obtain a better handwriting recognition system (better meaning faster and more accurate).

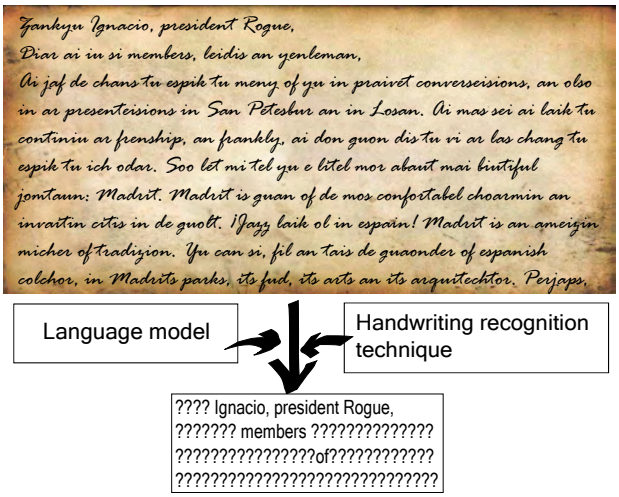
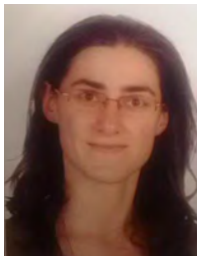


Figure 1: Representation of OOV words effect on recognition if the language model can not deal with this phenomena.

Publications

[1] Nuria Cirera, Alicia Fornes, Volkmar Frinken, Josep Llados, Hybrid grammar language model for handwritten historical documents recognition. *6th IbPRIA*, Volume 7887, pages 117-124, 2013.



Núria Cirera Turigues received her B.S. degrees in Mathematics and in Statistics from Autonomous University of Barcelona, Barcelona, Catalunya, in 2011. She received her M.S. degree in 2012 and is currently a Ph.D. Candidate in the Department of Computer Science at the Autonomous University of Barcelona. Her research interests are handwriting recognition, language models, document analysis, linguistics, hidden Markov models and pattern recognition.

Probabilistic Graphical Models for Layout Analysis

Francisco Cruz
Advisor: Oriol Ramos Terrades
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: fcruz@cvc.uab.es

Keywords: Layout Analysis, Segmentation

1 Previous and Current Work

This PhD thesis is focused in the tasks of layout analysis and segmentation applied to different types of documents (i.e., historical, administrative and contemporary documents). For this purpose, the ambit of the thesis will comprise the exploration of Probabilistic Graphical Models (PGMs), the analysis of contextual relationships between the different entities, as well as the treatment of hidden variables and the research of novel techniques to perform these tasks.

The first contribution of the thesis was focused in the research of different families of PGMs to perform the task of document segmentation. During this time, we studied the behavior of Conditional Random Fields (CRFs) using different inference algorithms as Belief Propagation, GraphCut or Junction Trees. We developed a method based in CRFs in combination with Relative Location Features to perform segmentation of historial and contemporary documents with promising results, see Fig 1. This work was published in the conference ICPR'12 [1].

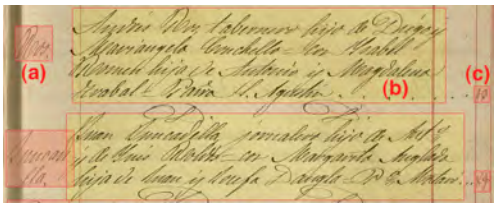


Figure 1: Example of historical document on which our segmentation method was applied. The different labels indicate the regions to detect (a)name, (b)body and (c)tax, for the detection of marriage licenses.

In views to explore another techniques further than the PGMs and be able to compare our approach, we also developed a layout analysis method based in 2D Stochastic Context-Free Grammars in collaboration with the ITI from The Universitat Politècnica de València. This work was published in the conference IbPRIA'13 [2].

As a participation in a project with an external company, the second contribution of this thesis was to develop a method for the detection of handwritten lines in administrative documents. We developed an EM based method to compute a set of regression lines through the textual elements as a way to extract the lines of handwritten documents, see Fig 2. The result of this work was published in ICDAR'13 and presented in the handwritting segmentation contest in the same conference [3].

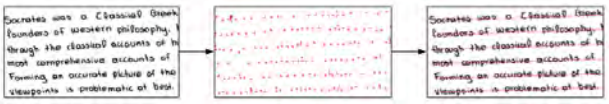


Figure 2: This image illustrate the main idea for the detection of the lines. (1) Input of handwritten text lines, (2) Extraction of the centroids of the connected components, (3) Resulting regression lines.

2 Future Work and Challenges

In the next months we plan to extend the EM scheme developed for the line segmentation task to the problem of layout analysis, so we will be able to compare the different approaches that we test for this task. Besides we are working in a full layout analysis method that integrates both the region and line segmentation modules for different types of documents. In addition, we plan to continue the research on PGMs and its training and inference algorithms applied to image documents a complex model structures.

Publications

[1] Francisco Cruz and Oriol Ramos Terrades. Document segmentation using relative location features. In *21st International Conference on Pattern Recognition (ICPR)*, pp. 1562-1565, 2012.

[2] Francisco Alvaro, Francisco Cruz, Joan-Andreu Sanchez, Oriol Ramos Terrades and Jose Miguel Bemedi. Page Segmentation of Structured Documents Using 2D Stochastic Context-Free Grammars. *In 6th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA*, Vol. 7887, pp. 133-40.

[3] Francisco Cruz and Oriol Ramos Terrades. Handwritten Line Detection via an EM Algorithm. *In 12th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 718-722, 2013.

Francisco Cruz received his B.S. degree in Computer Science from the Politechnical University of Valencia, Valencia, Spain, in 2010. He received his M.S. degree in 2012 and is currently a Ph.D. Candidate in the Department of Computer Science Engineering at the Autonomous University of Barcelona. His research interests are in the areas of layout analysis, segmentation, with emphasis on probabilistic graphical models and context analysis.

New advances on structural floor plan interpretation

Lluís-Pere de las Heras
Advisor: Gemma Sánchez
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: lpheras,gemma@cvc.uab.es

Keywords: Structural Interpretation, Graphics Recognition

1 The goal

The aim of my thesis is to create a general syntactic approach that, by taking into account the hierarchical and structural information between elements, will be capable to interpret and recognize floor plans. The main difficulty lies in the fact that there is no standard notation defined: elements such as walls, windows, furniture, indications, etc. are drawn distinctively depending on the architect and the country. Therefore, existing approaches are usually focused in one specific notation convention, and are not usable for the rest ones. These existing problems encourage us to construct a syntactic model capable to interpret every floor plan independently of its notation. The whole system is shown im Fig. 1.

2 Progress up to date

In the first part of the thesis we have focused on giving a solution to the problem of detecting the structural elements in floor plans independently to their notation; these are walls, windows, doors, and rooms. Later, a syntactic model will be created to account the spatial, semantic and hierarchical relations between these objects to achieve the proper final interpretation of the plans.

2.1 Wall detection

At the first stages of the thesis we focused on wall segmentation. Since walls convey inherent information of the building structure, they can be used to subsequently detect the rest of the elements. In this way, a first approach of detecting walls independently to their graphical notation was presented in [1], and an enhanced version in [2]. It was inspired by the appearance-based state-of-the-art strategies in Computer Vision for object detection in real scenes. As a result, a bag of visual patches able to learn the visual appearance of walls from a labeled collection of floor plans was presented as the first approach able to segment walls in multiple collections of real documents.

The second attempt was to tackle the problem under a structural point of view. The latent idea is to be able to drive the detection of potential elements belonging to walls by general structural properties of buildings and thus, without the need of any learning step for each notation. As a result, we presented in [3] an unsupervised approach driven by six structural statements of general properties of buildings, that combined fuzzily are able to segment walls in different collections of floor plans with successfully close results to the supervised approach. Lately, we proposed in [4] to learn the graphical appearance of [3]'s output to make the system more flexible and make it able to segment curved walls, beams, etc.

2.2 Floor plan interpretation

A complete method for floor plan interpretation has been built recently. This method, which is mainly inspired by the way engineers draw and interpret floor plans, applies two recognition steps in a bottom-up manner. First, basic building blocks, i. e., walls, doors, and windows are detected using the statistical patch-based segmentation approach in [2]. Second, a graph is generated and structural pattern recognition techniques are applied to further locate the main entities, i. e., rooms of the building. The proposed approach is able to analyze any type of floor plan regardless of the notation used. We have evaluated our method on two different publicly available datasets of real architectural floor plans with different notations. The overall detection and recognition accuracy is about 95%, which is significantly bet-

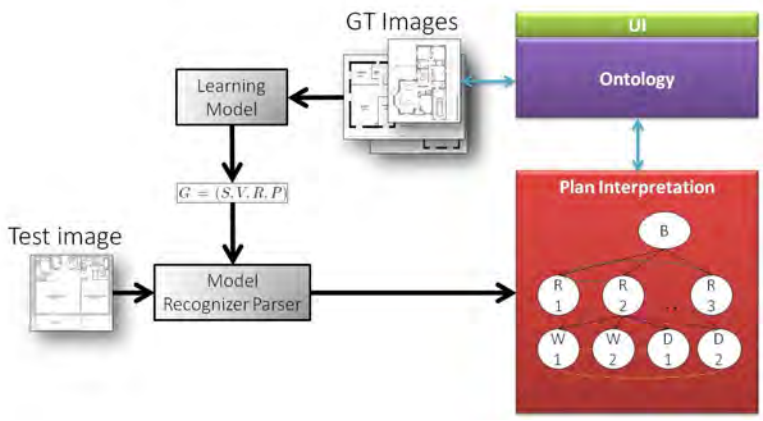


Figure 1: Automatic floor plan interpretation system.

ter than any other state-of-the-art method. Our approach is generic enough such that it could be easily adopted to the recognition and interpretation of any other printed machine-generated structured documents. This method has been recently submitted to IJDAR.

3 Future Work and Challenges

We are preparing a floor plan interpretation system that works semi-supervisedly and is capable to deal with a huge amount of different notations. This system is guided by a stochastic grammar built on an AND-OR Graph, that permits a recursive bottom-up/top-down structural element recognition to finally get the most probable floor plan interpretation. This model was validated as a good option to model floor plans under a structural and hierarchic point of view in [5].

Publications

[1] L.-P. de las Heras, J. Mas, E. Valveny, and G. Sánchez, Wall patchbased segmentation in architectural floorplans. In *Proceedings of the 11th International Conference on Document Analysis and Recognition*, 1270–1274, 2011.

[2] L.-P. de las Heras, E. Valveny, and G. Sánchez, Notation-invariant patch-based wall detector in architectural floor plans. In *Graphics Recognition. New Trends and Challenges, ser. Lecture Notes in Computer Science*, vol. 7423, 78–88, 2013.

[3] L.-P. de las Heras, D. Fernández, E. Valveny, J. Lladós, and G. Sánchez, Unsupervised wall detector in architectural floor plan. In *Proceedings of the 12th International Conference on Document Analysis and Recognition*, 1277–1281, 2013.

[4] L.-P. de las Heras, E. Valveny, and G. Sánchez, Combining structural and statistical strategies for unsupervised wall detection in oor plans. In *Proceedings of the 10th IAPR Conference on Graphics Recognition*, 123–128, 2013.

[5] L.-P. de las Heras, and G. Sánchez, And-or graph grammar for architectural floor plan representation, learning and recognition. A semantic, structural and hierarchical model. In *Pattern Recognition and Image Analysis, Lecture Notes in Computer Science*, 17–24, 2011.

Lluís-Pere de las Heras received both, his B.Sc. degree in Computer Science in 2009 and his M.Sc. in Artificial Intelligence and Computer Vision in 2010, from the Universitat Autònoma de Barcelona (UAB). He is currently a PhD student in the Computer Science Department of the UAB and the Computer Vision Center under the supervision of Gemma Sánchez. He is an active member of the Document Analysis Group and assistant professor at the Computer Science Department at the UAB. His research work is mainly focused on structural image interpretation, semantic understanding and graphics recognition.

Towards a Real Time Robust Scene Text Detection Method based in Perceptual Organization

Lluís Gómez i Bigordà
Advisor: Dimosthenis Karatzas
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: lgomez@cvc.uab.es

Keywords: scene text, perceptual organization, grouping

1 Summary of Previous and Current Work

The main goal of my research is to explore the bridges between the task of Text Extraction in Natural Scenes and the Perceptual Organization theoretical framework in order to build a successful text extraction method on the basis of their links. The research hypothesis arises from the observations that Gestalt laws of perception play an important role in the core of the writing systems, and thus has to be taken into account in order to solve the Scene Text Extraction problem in a perceptually inspired bottom-up approach.

After the first year of development of the thesis research most of the objectives for this initial phase have been already reached. An exhaustive review of the state of the art has been done, both in the context of scene text detection and of the Gestalt mathematical formulations, while different potentially useful existing techniques have been implemented. For example, the grouping principles of the Perceptual Organization framework Proposed by Desolneux et al., different feature descriptors used for text classification, or the Class Specific Extremal Regions algorithm proposed by Neumann and Matas. Moreover, I have identified many standard scene text datasets (covering different domains) and implemented their evaluation frameworks, while at the same time I've been collaborating in the creation of a new dataset for text localization in video sequences [3].

The work done during this year has already produced some initial results presented in our ICDAR2013 paper "Multi-script Text Extraction from Natural Scenes" [1], our CBDAR demo [2], and will be extended in a journal paper we are preparing.

Figure 1: Our method exploit the common perceptual organization laws always present in text, irrespective on scripts and languages.

Our text extraction method is based along two basic ideas: 1) A key characteristic of text is the fact that it emerges as a gestalt, a perceptually significant group of similar atomic objects. 2) In natural scenes many times we see different weak cues combining together synergetically providing evidence for a particular grouping. Hence, our method works by analysing different similarity feature spaces in parallel and combining them using the Evidence Accumulation Framework as a voting system. A perceptual organization analysis is performed on each feature-space based in the principle of non-accidentalness: A group of objects is considered as perceptually meaningful when it is not possible to appear merely by chance. This meaningfulness test is closely related with statistical hypothesis testing and can be formulated in terms of the expectation of a given group to be a realization of a random background process or not.

Even at this relatively early stage, our research has started showing the possible benefits of the perceptual organization approach. The main advantages

of the proposed method are its ability to deal with multi-script and multi-oriented text in a natural manner, and its near real time performance, being faster than many other methods in the literature.

2 Future Work and Challenges

One important limitation of our text extraction method is the high dependence of a good segmentation of text regions from the original image. For this reason we are currently moving from the MSER region detector to the Class Specific Extremal Regions algorithm proposed Neumann and Matas. This approach increases the recall for character segmentation because regions are extracted from the component tree of the image based in their conditional probability of being text characters, dropping the stability requirement of MSERs. However, the features used in the original paper for character vs. non-character classification are designed solely for latin script and horizontally aligned text, and thus we have to propose new features for our more general multi-script and multi-orientation scenario.

Furthermore, we have identified several possible extensions of our text extraction method which may lead to better results and/or alternative applications of our method. Most of this developments can be carried out as collaboration with other researchers as they require the use of techniques that fall out of the scope of our research.

For example, building a text detection method for video sequences requires the use of a tracking module in order to exploit temporal information. Thus, we plan to make this work in collaboration with researchers in the object tracking field.

Text line detection is an important post-processing step in our grouping method as many of the standard datasets are labeled at the text line level. Currently our method uses a simple method based in linear regression for the text line separation and validation. Using more complex and recently developed methods we expect to increase our results in those datasets.

Other interesting collaborations we are planning to do this year are: the combination of our localization algorithm with a state of the art word-spotting method, and the use of a dictionary based grammar for text recognition.

Apart from that, we are planning to delve into certain theoretical aspects of the perceptual organisation framework that we have identified, where original research outcomes are also expected.

Publications

[1] Lluis Gomez, and Dimosthenis Karatzas, "Multi-script Text Extraction from Natural Scenes". In *ICDAR*, 2013.

[2] Lluis Gomez, and Dimosthenis Karatzas, "Demonstration of a Human Perception Inspired System for Text Extraction from Natural Scenes". In *CBDAR* Workshop, 2013.

[3] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez, Sergi Robles, Joan Mas, David Fernandez Mota, John Almazàn Almazàn, Lluis Pere de las Heras, "ICDAR 2013 Robust Reading Competition". In *ICDAR*, 2013.

Lluís Gómez i Bigordà owns B.S. and M.S. degrees in Computer Science from Universitat Oberta de Catalunya. He received his M.S. degree in Computer Vision in 2010 from Universitat Autònoma de Barcelona where he is currently a PhD Candidate and a research assistant in the Computer Vision Center within the Document Analysis and Pattern Recognition Group. His research interests are on scene text detection and recognition.

Domain Adaptation for Pedestrian Detection

Jiaolong Xu
Advisor:Antonio M. López
Computer Vision Center & Computer Science Department at Universitat Autònoma de Barcelona
E-mail: jiaolong@cvc.uab.es

Keywords: pedestrian detection, domain adaptation

1 Summary of Previous and Current Work

In the first year of my PhD I demonstrated that the virtual synthetic data can be used to train deformable part-based model (DPM) with full supervision [1]. Based on this, I extended the star structure to a mixture of parts. Extensive experiments show its state-of-the-art performance on some benchmarks. This part of work is illustrated as S1 in Fig. 1. However, the accuracy may still get significant drop when testing on larger range of real-world datasets, which is known as domain shift problem. To address this, I focused on domain adaptation (DA) technique to adapt a pedestrian detector from virtual world to real world. The work was started with holistic model [2] and later mainly focused on DPM [3]. This part of work corresponds to S2 in Fig. 1.

2 Future Work and Challenges

The future work would explore various DA methods, *e.g.* feature transformation based methods (S2 in Fig. 1). Also multiple domains can be used to leverage the DA. More challenging scenarios in visual domain adaptation are to be addressed. For instance, the labels of training samples are not available (S3 in Fig. 1), training data are incrementally added, or multiple target (sub-target) domain potentially exist. We would seek solutions on unsupervised learning, multi-task learning, online learning in our future work.

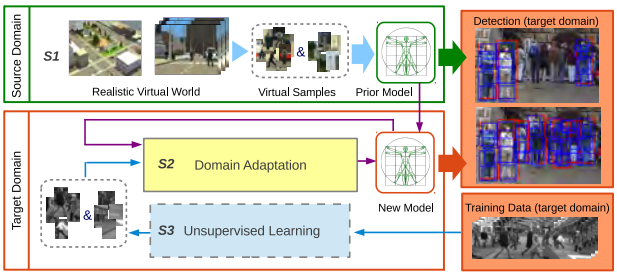


Figure 1: Domain adaptation framework.

Publications

[1] J. Xu, D. Vázquez, A.M. López, J. Marín and D. Ponsa, Learning a Multiview Part-based Model in Virtual World for Pedestrian Detection. In IEEE Intelligent Vehicles Symposium, Gold Coast, Australia, 2013.

[2] J. Xu, D. Vázquez, S. Ramos, A.M. López and D. Ponsa, Adapting a Pedestrian Detector by Boosting LDA Exemplar Classifiers. In CVPR Workshop on Ground Truth, Portland, OR, USA, 2013.

[3] J. Xu, S. Ramos, D. Vázquez, and A.M. López, Domain Adaptation of Deformable Part-Based Models. (PAMI under submission).

Jiaolong Xu received the B.Sc. degree and the M.Sc. degree from the National University of Defense Technology in 2008 and 2010 respectively. Currently, he is a PhD student at the Computer Vision Center (CVC) in Universitat Autònoma de Barcelona (UAB). His research interests include pedestrian detection, domain adaptation, and structured output learning.

Live and Semantic 3D Maps for Autonomous Driving Scenarios

German Ros
A. Sappa⋆, D. Ponsa⋆‡, J. Guerrero† and A.M Lopez⋆‡
⋆Computer Vision Center
‡DCC at Universitat Autonoma de Barcelona
†Dept. de Matematica Aplicada, Universidad de Murcia
E-mail: gros@cvc.uab.es

Figure 1: Dense 3D reconstruction of a neighbourhood in Karlsruhe, including buildings and urban objects.

Keywords: mapping, semantics, autonomous driving, VSLAM

1 Summary of Previous and Current Work

Autonomous navigation of vehicles is a passionate challenge that has led my research during the last years. The interest of this problem stands upon the future benefits of autonomous vehicles, which include safer traffic conditions, an overall better driving performance and a more intelligent control and distribution of urban fleets.

We have addressed the task of autonomous navigation from the point of view of the Visual Simultaneous Localization and Mapping (VSLAM) problem [3]. In our work, we defend that producing accurate localization of vehicles along with detailed maps of their environment are the required bases to build an appropriate solution to this problem. Visual SLAM is a challenging problem by itself, mainly due to the large amount of variables and constraints involved in the problem (i.e., thousands of locations and million of landmarks). An important part of our effort has been dedicated to develop novel VSLAM approaches that are more accurate and also computationally more efficient [1][2]. Such an effort has given rise to novel and general theoretical concepts such as the robust-compressed regression, a family of methods that are able to exploit the information contained in large amounts of data sets [1] while keeping a reduced computational budget. Other of our relevant efforts consisted in creating optimization appro-

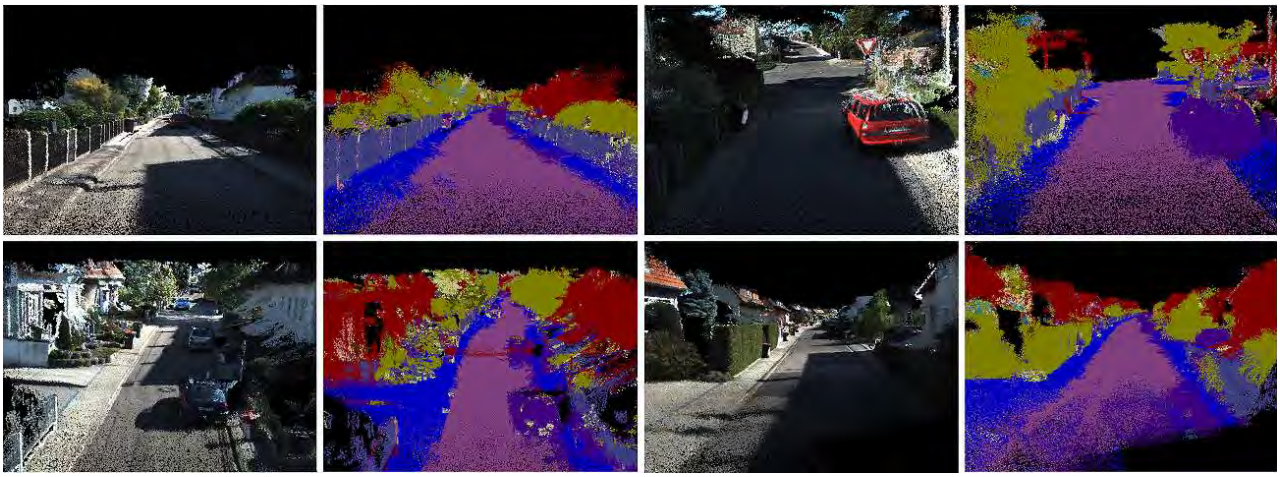


Figure 2: Example of a urban 3D map with semantic labels: road (purple), sidewalks (blue), etc.

aches based on novel manifold structures, as in the case of the K-fixed-rank manifold, a useful tool for cleaning up noisy signals.

Thanks to our advances in localization and mapping we are now able to create large and dense maps of cities, as depicted in Fig. 1, which allows us to pursuing bigger goals, like the concept of live and semantic 3D maps. This new idea consists in developing fully-semantic maps of cities. Here, we aim for extending the classical 3D geometry of maps with useful and essential information for the autonomous driving task, including: segmentation of the scene in semantic classes (e.g., road, side-walks, vegetation, buildings, etc.), position of traffic intersections, cross walks, traffic signs, terrain quality and many other features (see Fig. 2). The creation of this kind of maps require an enormous amount of computation, but we have proposed a new approach capable of building semantic maps offline and then use them online when needed. In this way, intelligent vehicles can access online to all the semantics of the current scene at very low computational cost.

2 Future Work and Challenges

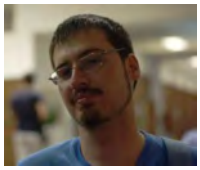
The concept of semantic mapping is very promising in the field of autonomous vehicles, but the current state-of-the-art still needs to be improved to produce reliable tools. These must include better 3D geometry in the process of mapping and more accurate results in the semantic segmentations process. Moreover, we have to face the problem of map updating to ensure maps are always coherent with reality.

3 Publications

[1] G. Ros, J. Guerrero, A.D. Sappa, D. Ponsa and M. Lopez, Fast and Robust L1-averaging-based Pose Estimation for Driving Scenarios. In *BMVC*, 2013.

[2] G. Ros, J. Guerrero, A.D. Sappa, D. Ponsa and M. Lopez, VSLAM pose initialization via Lie-groups and Lie-algebras optimization. In *ICRA*, 2013.

[3] G. Ros, A.D. Sappa, D. Ponsa and M. Lopez, Visual SLAM for Driverless Cars: A Brief Survey-In *IV Workshops*, 2012.



German Ros received his B.Sc. (Hons) degree in Computer Science in the modality of robotics and computer vision from Universidad de Murcia, in 2010. As an undergraduate student he carried out research activities in computer vision for the Applied Engineering group as a research intern. In 2011 he received his M.Sc degree by Kingston University of London while he was collaborating in the Robotics Vision Team (RoViT) and the Human Body Motion Group. Afterwards he obtained a M.Sc degree in 2012 from Universitat Autonoma de Barcelona, where he is currently pursuing a Ph.D. His research interests are related to computer vision, robotics and visual perception, with special interest and dedication to problems that have to do with visual geometry, optimization, compressive sensing, SLAM and 3D mapping.

Spatiotemporal Information for Pedestrian Detection

Alejandro González
Advisor:Jaume Amores, Antonio M. López
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: agalzate@cvc.uab.es

Keywords: Pedestrian Detection, Spatiotemporal, Volume Adaptation

1 Summary of Previous and Current Work

We've been working actively in the field of Pedestrian detection; more precisely we have been working on the usage of spatiotemporal (ST) information for improving the detection in video sequences. In order to achieve this goal, we have adapted the Stacked Sequential Learning (SSL) algorithm presented by *W. Cohen et al.*, in which the authors use the information extracted inside a sequential neighborhood (see Fig. 1a.) in order to reinforce the descriptor of a given sample. Using the same principle we define a ST-neighborhood around the sample (see Fig. 1b.). The samples included in this ST-neighborhood have a spatiotemporal relation with the given sample. In order to define the temporal relation we propose 2 different options, the projection across the temporal axis (fixed volume), or an adaptation in the position using optical flow values (flexible volume). In Fig. 2. we can see the volumes generated using the 2 options named previously. With the implementation of this ST-SSL algorithm we reach better performance over different static descriptors: HOG, HOGLBP, and HOGHOF (appearance and motion descriptor). In this work the main contribution is that based in a basic descriptor we can reach better performance by introducing appearance information extracted in a given ST-neighborhood.

2 Future Work

A first line of future work we are planning to do a study of the video frame rate impact in the ST-SSL algorithm using the different approaches of neighborhood definition. This study will allow us define

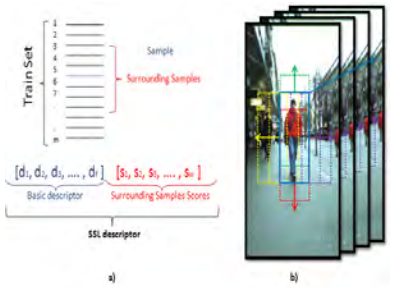


Figure 1: a. SSL algorithm description definition, b. ST neighborhood definition for ST-SSL.

the restriction introduced by the frame rate in term of which neighborhood definition approach we must use, for this goal we acquire a new data set with a high frame rate. This study is necessary because in a high frame rate (30 FPS) sequence the projection method may be enough but in a low frame rate (3 FPS) sequence the projection due to the egomotion is useless so in this case the optical flow method is a better choice. Also we propose a new neighborhood generation method based on a affine transformation. This new method takes advantage of the translation given by the optical flow and also calculates the scale change between frames. The scale change is due to the egomotion, while the car is going forward, the approaching objects in the scene seems to be in a large scale.

A second line we are planning the way to use the defined ST-neighborhood for the extraction of ST-features, this include the definition of new descriptors or the adaptation of the existing for being extracted in a non fixed volume. With this line we expect not only extract appearance features in the ST-neighborhood but also extract ST-features (motion, temporal coherence ...).

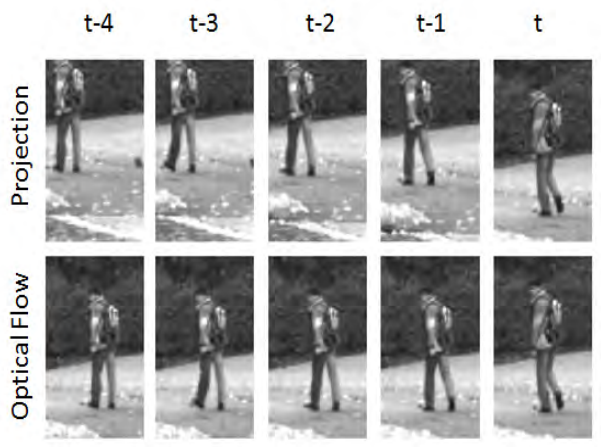


Figure 2: Different volumes generated applying projection and optical flow neighborhood generation approaches.

Alejandro González received his B.S. degree in Electronic Engineering from National University, Bogota, Colombia, in 2010. He received his M.S. degree in 2011 and is currently a Ph.D. Candidate in the Department of Computer Science at Universitat Autonoma de Barcelona. His research interests are in the areas of temporal information applied to detection, adaptation of spatiotemporal region for description calculation.

Detecting pedestrians in multi-spectral images

Yainuvis Socarrás Salas
Advisors: Antonio M. López Peña, Theo Gevers.
Computer Vision Center & Computer Science Department at Autonomous University of Barcelona
E-mail: ysocarras@cvc.uab.es

Keywords: pedestrian detection, image segmentation, thermal images

1 Summary of Previous and Current Work

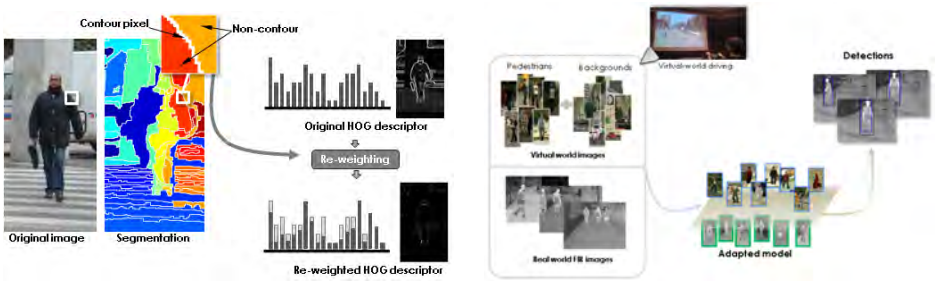
In recent years, I had focused my research in detecting pedestrians in images from real world scenarios by means of different computer vision techniques. For this purpose, I had studied the existing approaches of the state of the art in the object detection field, more specifically, pedestrian detectors.
I decided to improve the *histogram of oriented gradients* (HOG), a core descriptor for object detection, by the use of higher-level information coming from image segmentation. Here, during its computation, the HOG descriptor is re-weighted according to the information coming from image segmentation cues, but without increasing its size, see Figure 1 (a). This method was tested in the INRIA person dataset embedding it in a human detection system. The well-known segmentation method, mean-shift (from smaller to larger super-pixels), and different methods to re-weight the original descriptor (constant, region-luminance, color or texture-dependent) has been evaluated. We achieve performance improvements of 4.47% in detection rate through the use of differences of color among contour pixel neighborhoods as re-weighting function. This work was published in [1].
Additionally, the principal methods of the state of the art in pedestrian detection were tested in images beyond the visual spectrum, i.e., thermal images. Here, a different technique is presented. The aim is to adapt a pedestrian classifier trained with synthetic images (source domain) and the corresponding automatically generated to operate with far infrared (FIR) images (target domain). To this end, different methods have been tested to collect a few pedestrian samples from the target domain and to combine them with many samples from the source domain in order to train a domain adapted pedestrian classifier, see Figure 1 (b). The FIR dataset is composed by two sets of images collected in different moments of the day, *daytime* and *nighttime*. The information contained in this kind of images allows to develop a robust pedestrian detector invariant to extreme illumination changes.

2 Future Work and Challenges

Nowadays, detecting and segmenting humans in images are still hot topics in many computer vision research lines. Therefore, as future work I decided to improve the performance of the pedestrian detectors in FIR images, as well as the semantic segmentation of the detected object.
For this purpose, I will exploit low level features coming from an over-segmentation process combined with the information of the HOG descriptor in the detected window. Features from virtual objects will also be included since synthetic images can provide valuable information of the target object, e.g., shape prior. At this point, we will have a new framework that will be able to identify at pixel level a detected object in FIR images.

Publications

[1] Yainuvis Socarras, David Vazquez, Antonio Lopez, David Geronimo and Theo Gevers. Improving HOG with Image Segmentation: Application to Human Detection. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, Volume 7517, pages 178-189, year 2012.



(a) (b)

Figure 1: Overview of the presented methodologies. (a) show the HOG re-weighting process based on image segmentation. (b) present the adapted pedestrian detection process from synthetic to FIR images.

Yainuvis Socarrás Salas received his B.S. degree in Informatic Engineering from Instituto Superior Politécnico "José A. Echeverría" (ISPJAE), Havana, Cuba, in 2006. She received her M.S. degree in 2011 and is currently a Ph.D. student in the Department of Computer Science, Universitat Autónoma de Barcelona, Spain. Her research interests are in the areas human detection, image segmentation, object description and multi-modal imagery with emphasis on semantic segmentation in visual and thermal images.

Efficient Semantic Segmentation and Application for Scene Understanding

Sebastian Ramos
Advisor: Antonio M. López
Computer Vision Center
E-mail: sramosp@cvc.uab.es

Keywords: Visual Scene Understanding, Semantic Segmentation, Domain Adaptation.

1 Summary of Previous and Current Work

Most of my previous and current work address two principal lines: vision for robotics and autonomous driving. Before coming to CVC, I have spent 8 months researching within an European project named as Interactive Urban Robot, where I have conducted some research in visual scene understanding for mobile robots. Currently, I follow my research within a Spanish project denominated eCoDrivers, where I develop computer vision techniques for visual scene understanding of urban environments and visual domain adaptation.

One of my most recent accepted articles address the problem of how to make increase the efficiency of high accurate but costly semantic segmentation algorithms. State-of-the-art semantic segmentation pipelines often contain conditional random fields (CRF), where the inference process is done by maximum a posteriori probability (MAP) algorithms that optimize an energy function knowing all the potentials. We focus on CRFs where the computational cost of instantiating the potentials is orders of magnitude higher than MAP inference, as it is often the case in semantic image segmentation, where most potentials are instantiated by slow classifiers fed with costly features. We describe a novel technique for facing this problem through Active MAP inference (Fig. 1), achieving similar levels of accuracy but with major efficiency gains.

2 Future Work and Challenges

One of our principal goals is the creation of useful tools that allow to perform autonomous navigation of vehicles in urban scenarios. To this end we are currently developing a novel approach that consists in creating semantically rich 3D maps (Fig. 2), which encode all the information required by the navigation tasks. This map contains critical information such as: the segmentation of the voxels in classes (road, buildings, sidewalks, fences, etc.), traffic intersections, traffic signs, quality of the terrain, and much more. Since *semantics* are hard to compute, the creation of the maps is done offline and then, when a new vehicle needs this information it can access it online very efficiently.

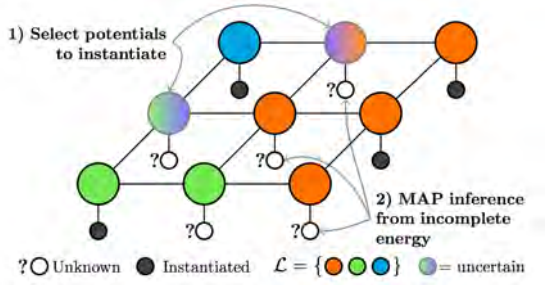


Figure 1: Active MAP inference. Example of CRF with unknown unary potentials.

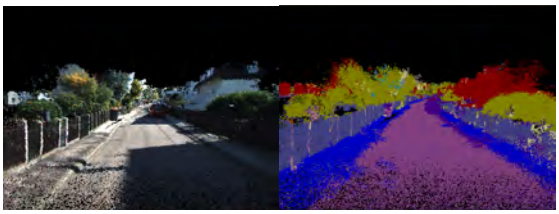


Figure 2: Semantic 3D Maps for Autonomous Driving Scenarios.

Publications

(*) indicates equal contribution.

[1] J. Xu, S. Ramos, D. Vazquez, A. M. Lopez. Domain Adaptation of Deformable Part-Based Models. T-PAMI (Under review), 2013.

[2] G. Roig(*), X. Boix(*), S. Ramos, R. de Nijs, L. Van Gool. Active MAP Inference in CRFs for Efficient Semantic Segmentation. In ICCV, 2013. (Oral Presentation)

[3] S. Ramos(*), R. de Nijs(*), G. Roig, X. Boix, L. Van Gool, K. Khnlenz. On-line Semantic Perception Using Uncertainty. In IROS, 2012.

[4] S. Ramos(*), Y. Socarras(*), D. Vazquez, A. M. Lopez, T. Gevers. Adapting Pedestrian Detection from Synthetic to Far Infrared Images. In ICCV Workshop, 2013.

[5] J. Xu, D. Vazquez, S. Ramos, A. M. Lopez, D. Ponsa. Adapting a Pedestrian Detector by Boosting LDA Exemplar Classifiers. In CVPR Workshop, 2013.

[6] D. Vazquez, J. Xu, S. Ramos, A. M. Lopez, D. Ponsa. Weakly Supervised Automatic Annotation of Pedestrian Bounding Boxes. In CVPR Workshop, 2013.

Sebastian Ramos received the B.Sc. degree in Electronic Engineering from the National University of Colombia in 2013. He is currently working towards the Ph.D. degree in Computer Science at the Computer Vision Center (CVC) in Universitat Autònoma de Barcelona (UAB). He was awarded a German scholarship to visit Technical University Munich from 2010 to 2012. His research interests include visual scene understanding, probabilistic graphical models and domain adaptation.

Accuracy and Permanency Limits for Calibration in Webcam Based Eye Trackers

Onur Ferhat
Advisor: Fernando Vilariño
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: oferhat@cvc.uab.es

Keywords: eye tracking, calibration, HCI

1 Summary of Previous and Current Work

Our work is focused on eye tracking and other modalities of human-computer interaction (HCI). Previously, we proposed several improvements over the open source eye tracker Opengazer which resulted in a more robust, easy-to-use application. We installed it on a Raspberry Pi to see the limits of cheap hardware in building an alternative standalone eye tracker. We presented this work in [1] where there was much interest in cheap eye tracker solutions.

In order to investigate the multi-modal HCI area, we worked with the Leap Motion controller to combine it with the gaze estimation as an input. The work resulted in a demo application where the gaze and the gestures allowed switching between the currently open windows in a Ubuntu system.

Currently, we are working on calibration strategies and analyzing how several factors affect the performance.

Our eye tracker is not invariant to head pose and the accuracy decreases in time as the head pose changes. We address this problem with our new correction mechanism as shown in Figure 1. When the estimations become inaccurate, the user is presented with a superimposition image indicating their current head pose and the head pose during calibration. They are expected to move their head so that the head poses are aligned. The eye tracker detects the face in the images and calculates the L2 norm in the face area. Thresholding this value into three categories, the system decides whether the alignment is good, average or bad and draws a frame with green, orange or red color on the screen, accordingly.

2 Future Work and Challenges

The biggest problem that is yet to tackle is head pose invariance of the eye tracker. Currently we are trying to reduce the effects of this issue without changing the base components of the system. However, as a future work we have the following lines of research:

- 3D model based head tracking
- Geometrically calculated gaze estimation
- Sub-pixel accurate iris segmentation methods

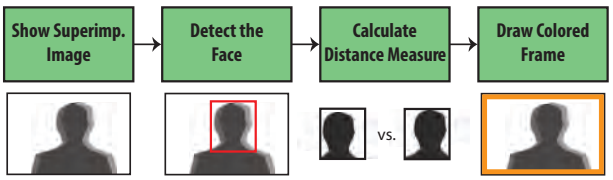


Figure 1: Head pose correction flow.

Publications

[1] Onur Ferhat, Fernando Vilariño, A Cheap Portable Eye-Tracker Solution for Common Setups. To appear in *Journal of Eye Movement Research*, 2013.

Onur Ferhat received his B.S. degree in Computer Engineering from Bogazici University, Istanbul, Turkey, in 2009. He received his M.S. degree in 2012 and is currently a Ph.D. Candidate in the Department of Computer Science at Universitat Autonoma de Barcelona. Currently his research is focused on eye tracking technology and multimodal HCI.

Depth-based Multi-part Body Segmentation

Meysam Madadi, Sergio Escalera, Jordi Gonzàlez, F. Xavier Roca, and Felipe Lumbreras
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
Dept. Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain.
E-mail: meysam.madadi@gmail.com

Keywords: 3D point cloud alignment, 3D shape context, depth maps, human body segmentation, soft biometry analysis

1 Summary of Current Work

Introduction Soft biometrics are traits of the human body which can be used to describe a person like height, weight, and skin. They have been used in video surveillance to track people [2], in combination to hard biometrics to increase reliability and accuracy [3], person re-identification [5], and supported diagnosis in clinical setups [6], just to mention a few.

In this paper, we propose a body segmentation approach in 3D space applying Kinect to compute accurate geometrical soft biometrics such as arm and leg lengths, and neck, chest, stomach, waist and hip sizes. Pixel labels are extracted in a model based multi-part approach. To compute and compare the results accurately, a user must stand face-to-Kinect such that the whole body can be seen.

Methodology In this section, we describe our system for human limb segmentation and soft biometrics computations. You can see the whole process in the Fig. 1.

To extract body pixel labels in the point cloud, we use an iterative approach which aligns the test point cloud to the nearest model in the training set. In fact, the process is divided into two parts: training and test. In the training step, we clusterize all poses using EM algorithm applying HOG descriptor [7] on depth images to put similar poses in the same Gaussian mixture and keep the centroids as representative models for those clusters. We optimize the number of mixtures using a combination of EM and k-means [4]. Finally, we keep EM parameters of clustering for future pose estimation as a classifier.

In the test step, we find the nearest model to the test image using HOG descriptor and the trained EM. In the next step, a random pixel selection is applied

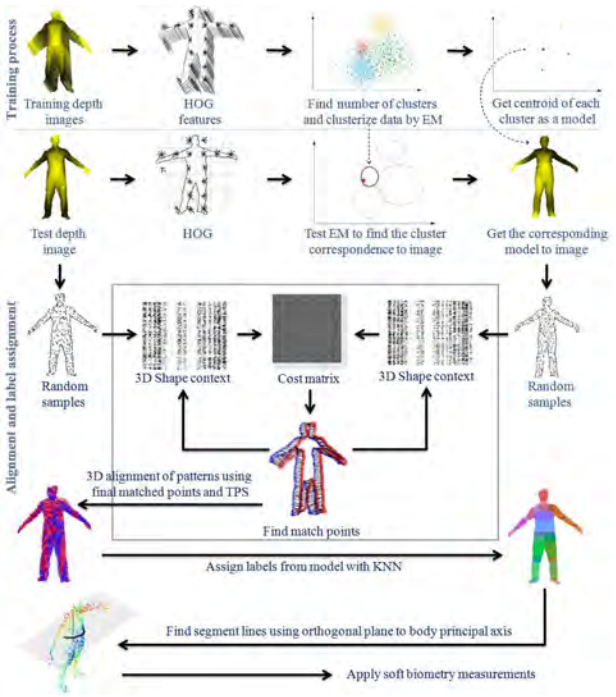


Figure 1: Process diagram of the system.

on test and model depth image and an iterative 3D alignment is performed applying the following process [1]:

1. 3D shape context descriptors are computed for selected points,
2. The cost matrix between all point descriptors is computed after adding some dummy points,
3. The best match points are extracted using linear assignment problem using Jonker-Volgenant algorithm,
4. The best match points are aligned using 3D thin plate spline algorithm.

This process is repeated to refine the matching points. Pixel labels are assigned from the nearest model pixels after final 3D alignment.

Having an accurate body segmentation, we are able to compute limb sizes using the hitting points of the orthogonal plane to the principal axes of the limb crossing from the mean points and body hull. The interpolation of such points makes the curve which can be used to measure the size. The length of the limbs like arm can be directly computed from the summation of distances between joint points.

Discussion We created a dataset containing 1061 frames of 31 individuals using Kinect to evaluate our method. We used a 10-fold cross validation over all frames to generate the results. We compare our results to the standard random forest pixel labeling approach [8]. Given the result in Fig. 2, we could get pure segmentation in edges showing an accurate segmentation also applicable in pose recovery. Our approach shows a low sensitivity to the number of training data vs. random forest.

Figure 2: Qualitative results. First row our method, and second row RF labeling approach. Black points correspond to segment lines. It can be seen that segment lines accuracy has a direct relation with the segmentation accuracy and purity.

2 Future Work and Challenges

The most critical part in this approach is point cloud registration. A more accurate registration will cause a better alignment and consequently better segmentation. In the future work, we apply a pose retrieval system to find nearest model besides using a global vs. local descriptor for registration. We apply our method on more complicated poses.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. PAMI, 24:509–522, 2002.

[2] S. Denman, A. Bialkowski, C. Fookes, and S. Sridharan. Identifying customer behaviour and dwell time using soft biometrics. Springer, 409:199–238, 2012.

[3] G. Guo, G. Mu, and K. Ricanek. Cross-age face recognition on a very large database: The performance versus age intervals and improvement using soft biometric traits. ICPR, pages 3392–3395, 2010.

[4] Y. Lee, K. Y. Lee, and J. Lee. The estimating optimal number of gaussian mixtures based on incremental k-means for speaker identification, 2006.

[5] A. Møgelmose, A. Clapés, C. Bahnsen, T. Moeslund, and S. Escalera. Tri-modal Person Re-identification with RGB, Depth and Thermal Features. CVPR, 2013.

[6] M. Reyes, A. Clapés, J. Ramírez, J. R. Revilla, and S. Escalera. Automatic digital biometry analysis based on depth maps. Computers in Industry, 2013.

[7] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. ICCV, 2:750–757, 2007.

[8] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Realtime human pose recognition in parts from single depth images. CVPR, 2011.

Meysam Madadi received his Bachelor degree in Software Engineering at BuAli Sina university of Hamedan and M.S. degree in Computer Vision and Artificial Intelligence at Universitat Autònoma de Barcelona (UAB) in 2007 and 2013, respectively. He has started his research activities by focusing on information retrieval and data mining since his bachelor project, continuing in master specifically on computer vision and image processing. He gave a special attention to pose recovery and human behavior analysis from his master thesis. He is interested in generating and developing new algorithms in these topics applying the knowledge in computer vision and retrieval systems besides machine learning, algorithms design in artificial intelligence, statistics, linear algebra, different geometries and many other relevant areas.

Fast face and eye centre detection in still images

Marc Oliu and Sergio Escalera
Dept. Applied Mathematics, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, Barcelona
Computer Vision Center, Campus UAB, Edifici O, 08193, Barcelona
E-mails: moliusimon@gmail.com, sergio@maia.ub.es

Keywords: face detection, eyes detection, Viola & Jones, fast radial symmetry

1 Problem Definition

Face and eye centre detection is usually implemented by means of the Viola & Jones [1] method for object detection and the fast radial symmetry transform [2] method. These allow the detection of face bounding boxes and posterior eye centre location. This type of approach is fast and tends to have a high recall or true positive rate, but has the problem of also having a low specificity, tending to return false positives. In order to increase the robustness of the system while maintaining a high detection speed, we propose a methodology that takes into account the contextual information of the problem.

The proposed methodology intends to speed up the eye centre detection process, which is the most costly part part of the method, by reducing the search space. It also intends to discard the false positives of both eyes and faces by means of a series of basic 'sanity' checks and a statistical estimation of the detected features probabilities.

2 Summary of the Developed Work

A methodology is proposed for reducing the false positive rate when performing faces and eye centres detection. First the face detection process is performed with the Viola & Jones face detection algorithm. Afterwards, the same technique is applied for eyes detection inside the face regions (Figures 1a-1b), thus reducing the search space. In order to discard the false positive eye detections inside each face region, first a set of sanity check rules are used, discarding eyes at the lower half of the face or too close to the face bounding box edges (Figure 1c). Afterwards, all possible face configurations are evaluated, and the configuration maximising the probability given a pair of eyes is selected, considering the rest of candidates as false positives (Figures 1d-1e).

In order to evaluate a face configuration, a set of normally distributed face features are selected and modelled into a multivariate normal distribution defining the faces space, and the inverse Cumulative Density Function (CDF) is used as the configuration probability. The face features selected are the eyes surface area, interpupillary distance, and eyes height with respect to the face. In the case no valid pair of eyes is found, that is, no eye is found at the left or right half of the face, the algorithm discards the face (Figure 1h) and the next face bounding box is processed.

Once the optimal pair of eyes is selected, the centres of the eyes are located (Figure 1f). This is done by first scaling each eye bounding box to a fixed size and applying a fast radial symmetry transform [2] on the image region inside the eyes bounding boxes for a set of test radius. This gives a set of maps of symmetry for the image, and the map with the highest peak symmetry value is selected and merged with the two contiguous maps. The merging is performed in order to account for imperfect radial symmetries caused by a low image resolution, image deformation or eye plane transform, making the iris edge appear not completely symmetric and making the radius from the centre to the edge variable. From the merged map, the location with the highest positive symmetry value is considered to be the eye centre. Finally, once determined the centre for the pair of eyes, the probability of these centres for the given pair is determined (Figure 1g).

The evaluation of the eyes pair centres is preformed using the same technique used for the face configuations. A bivariate normal distribution is modelled using the vertical and horizontal position variation of one eye centre with respect to the other,

and the inverse Cumulative Density Function (CDF) is used as an estimation of the probability of the pair of eye centres being correct.

3 Results and Future Work

The algorithm has been tested for a set of images containing 156 faces, and in all of the cases all false positives for the faces and eyes have been successfully discarded. Some examples of the discarded false positives can be seen in the examples in Figure 2. The true positives, on the other hand, have not been discarded, keeping a true positive rate of the 100. In the case of eyes, all the false positives are also correctly discarded, with no true positive discarded because of the implemented filtering technique.

In order to increase the accuracy of the method for both detecting eyes centres and distinguishing accurately between closed and open eyes, apart from training a closed eyes detector, a temporal analysis of video sequences could be used taking into account the measured state of the eye on each video frame to help determine the state on contiguous frames. This could be done by applying an on-line unsupervised clustering technique on a set of features measured on the eye bounding box, and allow for a speed-up of the algorithm by preventing it from calculating the symmetry maps on closed eyes, and help extrapolate the position of faces and eyes in frames where the detection fails.

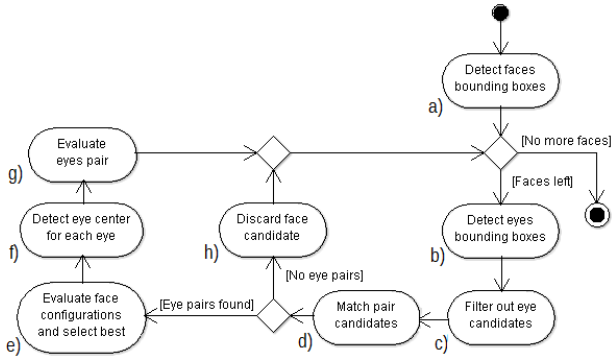


Figure 1: Activity diagram outlining the methodology of the system.

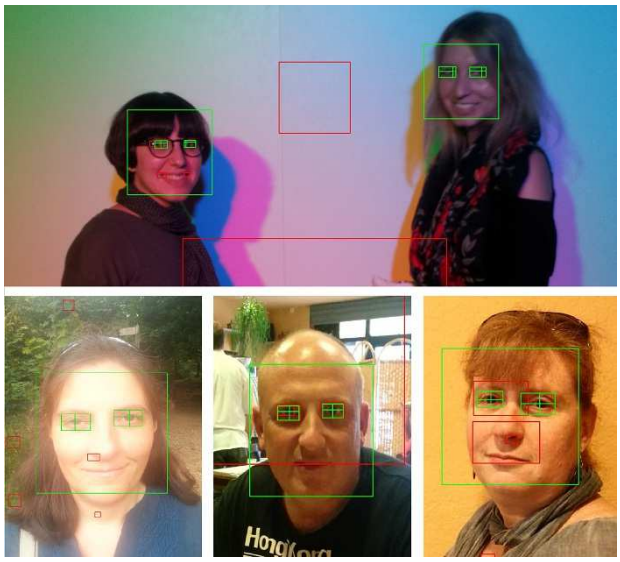
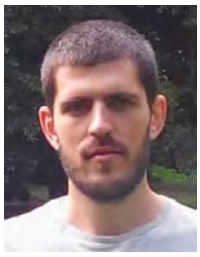


Figure 2: Results obtained by the developed methodology. Red rectangles represent the discarded elements, while green rectangles are the ones accepted as valid.

References

[1] P. Viola, M. Jones, ”Rapid object detection using a boosted cascade of simple features”, In *Conference on Computer Vision and Pattern Recognition*, 2001.

[2] G. Loy, A. Zelinsky, ”A fast radial symmetry transform for detecting points of interest”, *European Conference on Computer Vision*, Volume 1, 358-368, 2002.



Marc Oliu received the Technical Bachelor degree in Computer Science from Universitat de Girona, Girona, in 2010. He is currently finishing the final project for the Bachelor degree in Computer Science and studying his Master degree in Artificial Intelligence at Universitat Politcnica de Catalunya (UPC), Universitat de Barcelona (UB), and Universitat Rovira i Virgili (URV). He is interested in Computer Vision and Machine Learning in general.

Tri-modal Human Body Segmentation

Cristina Palmero[1], Albert Clapés[1,2], Chris Bahnsen[3], Andreas Møgelmose[3]
Advisors: Sergio Escalera[1,2], Tomas B. Moeslund[3]
E-mail: c.palmero.cantarino@gmail.com, aclapes@cvc.uab.cat, sergio@maia.ub.es, {am,cb,tbm}@create.aau.dk
[1] Dept. Matemàtica Aplicada i Anàlisi, UB, Gran Via de les Corts Catalanes 585, 08007, Barcelona
[2] Computer Vision Center, Campus UAB, Edifici O, 08193, Bellaterra, Barcelona
[3] Aalborg University, Sofiendalsvej 11, 9200 Aalborg SV, Denmark

Keywords: Body segmentation, Multi-modal features

1 Motivation

Segmentation of people in images is still nowadays a very challenging and difficult problem in computer vision. There exist lots of possible applications for people segmentation such as surveillance, patient caregiving, human-computer interaction, and so on and so forth. In the state-of-the-art, we mostly find the usage of color images recorded by cameras or, more recently, with the apparition of RGB-Depth devices – such as Microsoft® Kinect™ –, the usage of depth maps in combination with the information provided by the color cue. In this context, we propose adding a third modality that is the thermal imagery got from thermal infrared cameras and, thus, complementing other information sources and making easier the segmentation task [1]. Although thermal cameras are relatively expensive devices, their market price is lowering substantially every year (as it happens with other sensory devices).

The main contribution of this paper is a novel tri-modal database of people acting in three different scenes, consisting of more than 2,000 frames each one, in which three different subjects appear and interact with objects performing different actions such as reading, working on a laptop, speaking by phone, etc. In addition, a human segmentation baseline methodology is also proposed, consisting in segmenting first the people in each of the modalities separately and, finally, fusing the results in an optimization graph-cuts framework.

2 Method

Having the modalities already registered (from a previous work), background subtraction is initially performed in each of the modalities in order to extract candidate subject and object regions. Then, in the training phase, the subject regions are described at pixel-level using particular descriptors in each of the modalities. Once the subject pixels have been described in all the modalities, Gaussian Mixture Models (GMMs) are learnt. These GMMs are the ones used in the testing phase to compute the probabilities of being a subject pixel in the different modalities. Finally, the probability maps are combined in the Graph Cuts optimization step. A graphical representation of the proposed methodology is shown in Figure 1, which is explained in more detail below.

Background subtraction is performed separately in the different cues. In each modality, a model of the background is modeled given an initial set of F frames, learning a gaussian distribution for each pixel. Since alignment among dataset modalities is not accurate, foreground regions are fused and aligned at near pixel level by re-scaling of nearest detected regions. Thus, the segmented regions in the three different scenes can be merged to later describe them.

Then, the descriptors can be computed for all the pixels in all the modalities. Each modality involves its own specific descriptors: in the color cue, we have computed the histogram of oriented gradients (HOG) [2] in a $h \times w$ window centered at the pixel; the color cue also allows us to obtain motion information by computing dense optical flow and describing the distribution of the resultant vectors, known as histogram of oriented optical flow (HOOF); in the depth cue, histograms of oriented surface normals (HOSF) have been used; and, in the thermal cue, histograms of

thermal intensities concatenated with histograms of oriented gradients (HI-HOG). This implies having 4 different descriptions for each pixel.

At this point, the distributions of the previously computed descriptors are modeled by GMMs. Instead of learning only one GMM for the pixels in each kind of description, we divide the subject regions in a grid of cells and learn a GMM in each cell, thus having $4 \times \#cells$ GMMs.

Eventually, in the testing step, the new extracted regions are also divided in cells, their pixels described and the probabilities predicted from the corresponding GMM. Finally, these obtained probabilities are combined together with the extracted human body probabilities from Ramanan et. al. method [3], weighting each one, and used as the data-term in Graph Cuts optimization algorithm [4], so as to obtain a final segmentation of the human body.

3 Results

The results obtained so far are mainly qualitative but allow us to develop approaches to the problems that we face before starting to retrieve quantitative conclusions. Descriptive examples of the different stages are represented in Figure 2 and 3.

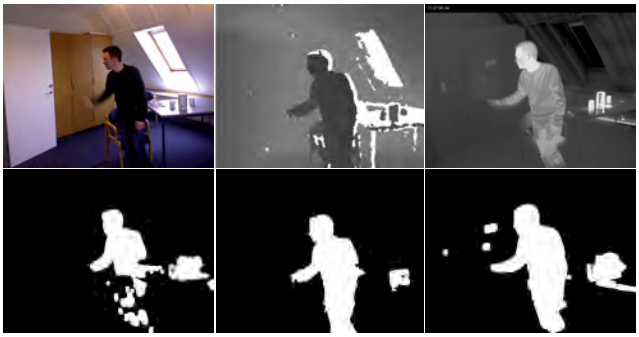


Figure 2: Background subtraction for different visual modalities of the same scene.

References

[1] Andreas Møgelmose, Albert Clapés, Chris Bahnsen, Thomas B. Moeslund and Sergio Escalera. Tri-modal Person Re-identification with RGB, Depth and Thermal Features. *9th IEEE Workshop on Perception Beyond the Visible Spectrum*, 2013

[2] Dalal Navneet and Bill Triggs. Histogram of oriented gradients for human detection. In *CVPR 2005. IEEE Computer Society Conference on.*, Vol. 1, p. 886-893, 2005.

[3] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. *CVPR 2011. IEEE Conference on.*, p. 1385-1392, 2011.

[4] Yuri Boykov and M-P. Jolly, Interactive graph cuts for optimal boundary region segmentation of objects in ND images, *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on.*, Vol. 1, p. 105-112, 2001.

Cristina Palmero received her Bachelor degree in Audiovisual Telecommunication Systems Engineering at Universitat Politècnica de Catalunya (UPC), Terrassa, Spain, in 2011. She is currently studying her Master degree in Artificial Intelligence at Universitat Politècnica de Catalunya (UPC) and Master degree in Computer Vision at Universitat Autònoma de Barcelona (UAB). She is mainly interested in signal and digital image processing and computer vision techniques applied to human behavior analysis, scene understanding and robotics.

Albert Clapés received his B.S. degree in Computer Science at Universitat de Barcelona in 2012. He is currently studying the interuniversity M.S. degree in Artificial Intelligence at Universitat Politècnica de Catalunya. He is a research fellow at Department of Applied Mathematics and Analysis in Universitat de Barcelona and an eventual member of the Computer Vision Center (Universitat Autònoma de Barcelona). His main interests in research are computer vision and machine learning applied to human pose recovery and behavior analysis, and also the human-machine natural interaction technologies.

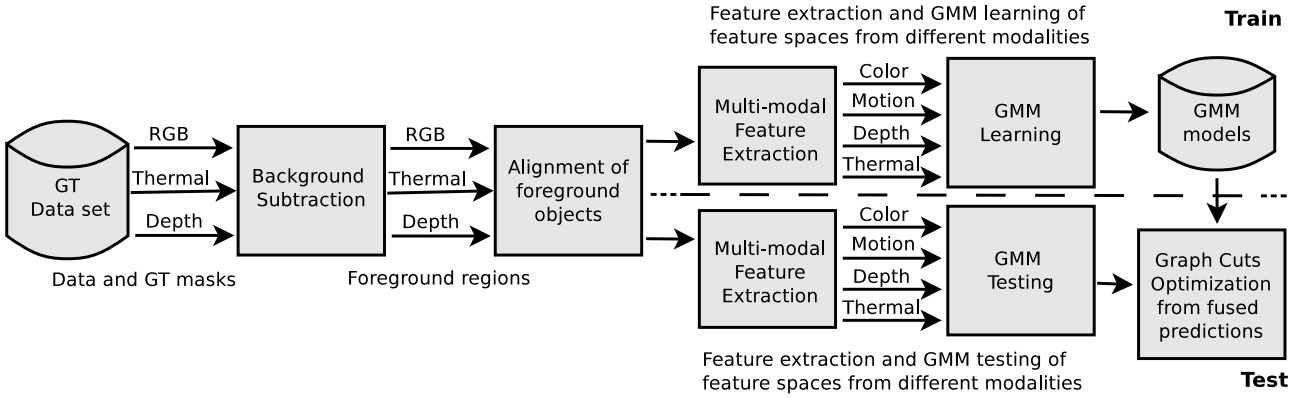
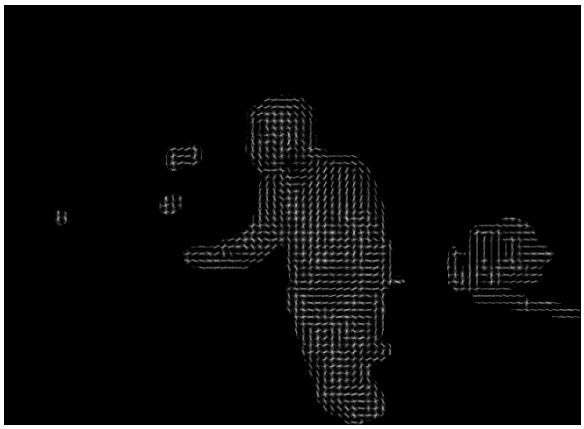
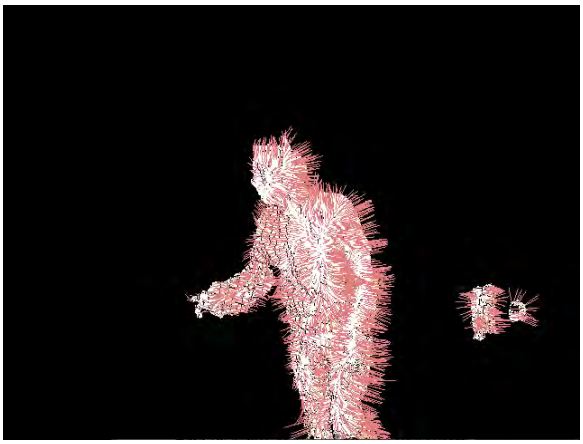


Figure 1: Pipeline of the presented methodology.

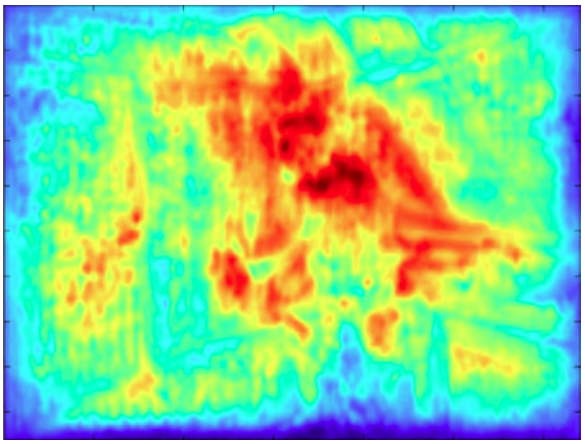


(a) HOG descriptors from thermal

(b) Optical flow from RGB



(c) Normals from depth



(d) Ramanan probability map

Figure 3: Example of descriptors from different modalities.

Intrinsic Image Characterization and Evaluation

Marc Serra
Advisors: Robert Benavente, Olivier Penacchio
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: mserra@cvc.uab.es

1 Summary of Previous and Current Work

I have been working in the field of intrinsic image characterization. We have developed a method for intrinsic image decomposition based on a graphical model framework which combines information from different color cues of a single image, namely color name categorization and gradual color surface variations, in order to estimate its intrinsic reflectance and shading images. Our model has been tested on the MIT dataset, achieving state-of-the-art performance. Figure 1 shows one of the objects in the dataset and its decomposition into its intrinsic reflectance and shading components. This work was published in [1].

We have also built a new dataset for intrinsic image decomposition. The novelty of this new dataset is the use of rendering algorithms and multispectral sensors to create synthetic images which properly emulate real world scenes, thus simplifying the process of building a dataset and allowing it to include more realistic scenes, with multiple objects, complex illumination and related effects such as interreflections. This work was published in [2].

We are currently working on the definition of a general framework for intrinsic image decomposition which extends and unifies all previous formulations and also includes other important factors in the image creation process such as information on the camera sensors or the light source.

2 Future Work and Challenges

We are also extending our dataset for intrinsic image characterization. There is a need in the intrinsic characterization field for an extensive dataset which includes not only more realistic scenes, but also ground truth on other intrinsic characteristics of the scene such as the shape of the objects, the color and geometry of the illuminants, and so on. We will extend our synthetic dataset and will show whether synthetic datasets are more useful than real datasets to train intrinsic image decomposition models.

One problem that we have to face is the extension of our intrinsic image decomposition method in order to combine input information from different color cues and infer other intrinsic properties of the image such as the color and geometry of the illuminant or the shape of the objects in the scene. This is a hard but interesting problem, since it unifies in a single problem many existing topics in the computer vision Literature which deal with intrinsic characterization of scenes, such as intrinsic image decomposition, color constancy, shape from shading or specularity removal among others.

Figure 1: Example of intrinsic image decomposition.

Publications

[1] Serra, M. and Penacchio, O. and Benavente, R. and Vanrell, M., **Names and Shades of Color for Intrinsic Image Estimation**. In *CVPR*, 278–285, 2012.

[2] Beigpour, S. and Serra, M. and van de Weijer, J. and Benavente, R. and Vanrell, M. and Samaras, D., **Intrinsic Image Evaluation On Synthetic Complex Scenes**. In *ICIP*, 2013.

Marc Serra received his B.S. degree in Mathematics from Universitat Autonoma de Barcelona, Catalonia, in 2009. He received his M.S. degree in Computer Vision and Artificial Intelligence in 2010 and is currently a Ph.D. Candidate in the Department of Computer Science at the same university. His research interests are in the areas of intrinsic image characterization and evaluation, study and combination of color cues, probabilistic models applied to image characterization, and related topics such as camera calibration, color constancy and so on.

RGB vs. Depth: Human Pose Recovery and Gesture Recognition

Antonio Hernández-Vela
Advisor: Sergio Escalera, Stan Sclaroff
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: ahernandez@cvc.uab.es

Keywords: gesture recognition, depth, pose recovery

1 Summary of Previous and Current Work

During the past year, I was working in the field of gesture recognition from RGB-D (RGB + depth) video sequences. More specifically, I participated in the ChaLearn gesture recognition challenge, with some other members of the Human Pose and Behaviour Analysis (HuPBA) group. This challenge consisted on developing a one-shot learning gesture recognizer which takes as input an RGB-D video sequence of a person performing a sequence of gestures in front of the camera. The system we developed consisted on a Bag of visual words model, combining features extracted from both RGB and Depth image modalities in a late fusion fashion (see Figure 1). In this framework, we compared different RGB and depth state-of-the-art descriptors, computed from Spatio-Temporal Interest Points (STIPs) extracted from the sequences. Furthermore, we proposed a new depth descriptor based on surface normals' angle distribution, which helped in our gesture recognition task. This work was published in [1, 2].

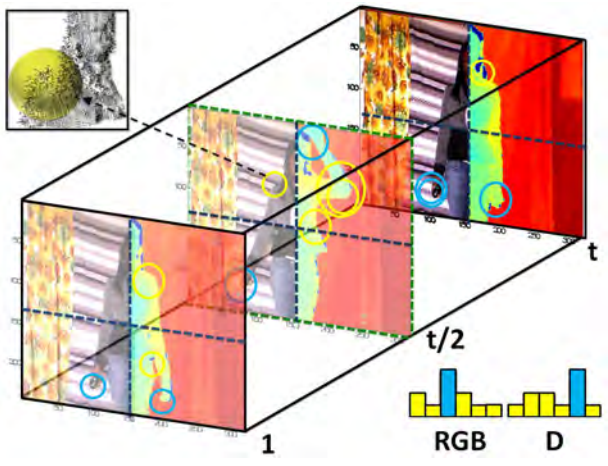


Figure 1: BoVDW framework depicting a RGB-D spatio-temporal volume from a gesture video sequence. Yellow and Blue circles depict STIPs (blue color indicates a particular vocabulary bin assignation).

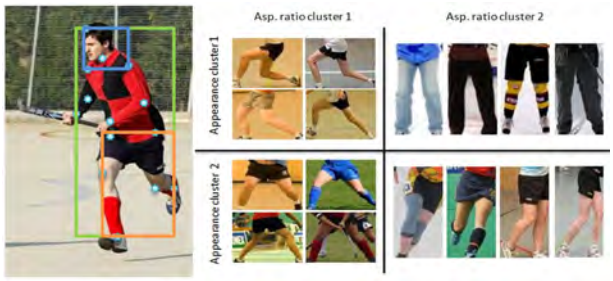


Figure 2: (Left) Different mid-level parts defined by bounding boxes containing different subsets of body joints. (Right) Different aspect ratio and appearance clusters for the lower body.

From February to June 2013 I visited the Image and Video Computing research group in Boston University, where I worked under the supervision of Prof. Stan Sclaroff. During this visit I worked in the problem of human pose estimation in RGB still images. Following recent works in the state-of-the-art, we defined a model of the human body formed by 14 different parts basic parts (one for each joint in the human body, e.g. ankle, knee, etc), but we also defined a set of mid-level parts (e.g. upper body, lower body, etc.), which would be used as context for improving the basic part detections (see Figure 2. For each one of these parts, different types were defined (depending on the appearance, and also on the bounding box aspect ratio in the case of mid-level parts), and a different detector was trained for each one of them, using HOG features. Once these detectors were trained, we could run them on a different set of images, and compute some relative features

between the basic and mid-level detections. Finally, these features were used to train a classifier which would re-score the detections of a given basic part detector.

2 Future Work and Challenges

First results of the basic part detections re-scoring are promising, and indicate that the mid-level part detections are useful for correcting confusion beween left and right limbs. As future work, we could define a richer mid-level parts model, following the Poselets philosophy.

Publications

[1] Antonio Hernndez-Vela, Miguel ngel Bautista, Xavier Perez-Sala, Vctor Ponce, Xavier Bar, Oriol Pujol, Cecilio Angulo, Sergio Escalera, BoVDW: Bag-of-Visual-and-Depth-Words for Gesture Recognition. In *ICPR*, pp. 449-452, 2012.

[2] Antonio Hernndez-Vela, Miguel ngel Bautista, Xavier Perez-Sala, Vctor Ponce-Lpez, Sergio Escalera, Xavier Bar, Oriol Pujol, Cecilio Angulo, Probability-based Dynamic Time Warping and Bag-of-Visual-and-Depth-Words for Human Gesture Recognition in RGB-D. In *PRL*, Available online 20 September 2013, ISSN 0167-8655, doi:10.1016/j.patrec.2013.09.009.

Antonio Hernández-Vela received the B.S. degree in computer science and the M.S. degree in computer vision and artificial intelligence, both from the Universitat Autònoma de Barcelona (UAB), Barcelona, Spain, in 2009 and 2010, respectively. He is currently working toward the P.h.D. degree in mathematics on human pose recovery and behavior analysis at the Universitat de Barcelona. He is a Research Member at the Computer Vision Center, UAB, and a member of the BCN Perceptual Computing Lab, which is a consolidated research group of Catalonia. His main research interests include the application of computer vision and artificial intelligence to projects that can help impaired people to improve their quality, specially in the field of human pose recovery and behavior analysis.

Human Body Pose Estimation Using Deformable Part-Based Models

Maedeh Aghaei Gavari
Advisor: Petia Ivanova Radeva
Department of Applied Mathematics and Analysis at Universitat de Barcelona
E-mail: maghaeigavari@ub.edu

Keywords: Human Interaction Analysis, Pose Estimation, Gaussian Mixture Models, Pictorial Structure

1 Summary of Previous and Current Work

During my master thesis I developed a strong technique for modeling and recognizing human behavior, focused on human pose detention. In particular the project is based on detecting interactions between people and classifying the type of interaction. This has been done by detecting the people in the scene and retrieving their corresponding pose and position sequentially in each frame of the video. We believe that body pose plays a crucial role in analysis of human interaction. To achieve this goal our work relies on robust object detection algorithm which is based on discriminatively trained part-based models. Gaussian Mixture Model (GMM) based method also has been used to extract the background and to accelerate pose estimation. we have also trained the part-based algorithm on a large and challenging computer vision dataset (PASCAL Visual Object Classes 2010) to obtain our specific pose detection filters. In this dataset, for every element in an object class (Person class in our case) exists specific piece of information, tagged as pose, which we explicitly used it for training of our classifiers. In the end, we have evaluated our method on a home-made database comprising depth data from Kinect sensors. After successfully collecting information corresponding to object's label as well as their pose and position over a video frames, movement understanding of them comes naturally which is an important step towards human interaction analysis. An intelligent analysis of these data together with other computer vision techniques, enables us to design an integrated solution for many other complicated computer vision problems. An example of these problems is to design an automatic algorithm for recognition of events. This project was presented end of 2012.

Our research currently is focused on developing a real-time multi-frame, multi-target tracking algorithm, based on the idea of Multiple Dimensional Assignment (MDA) problem. The two frame bipartite assignment problem, also known as linear assignment problem, can be solved exactly in polynomial time by methods such as Hungarian algorithm. However, these one-pass greedy algorithms do not work well when there is target interaction or occlusion in a scene. Moreover, multi-frame data association problem is combinatorial optimization problem of significant complexity. Developing multi-frame search algorithms that yield good quality approximate solutions in polynomial running time has therefore become a problem of considerable research interest in the field.

2 Future Work and Challenges

As our future work, we are focused on creation of an event recognition system which basically is an integrated system for recognition of physical movements (examples: standing, walking, etc.) and instrumental activities (examples: watching TV, writing letters, etc.). This enables us to extract and summarize daily life activities of a person, with the possibility of being used to the further diagnosis made by the physician, through observation and interpretation of patient records.

Our future work mainly is divided into two main modules: vision module and event recognition module. Vision module is a modular platform that allows us to study different algorithms for each step in the chain of computer vision. This module is where my

previous work links with the future goals. We need to improve, adjust and adapt some part of the previous work to the new goals. Developing a strong tracking technique is of high importance in this path.

Event recognition and analysis module assess what activities and scenes are of interest to the event recognition module. This module is responsible of determining the various scenes and actions which leads to explain different events and their subsequent classification. In this module we need to identify sort of sensible restrictions and levels of information in order to achieve intelligent understanding of events.

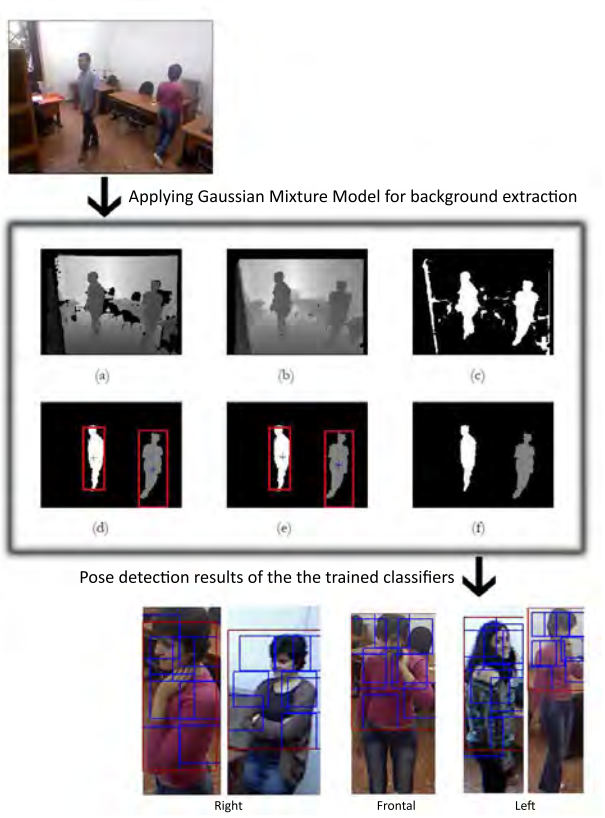


Figure 1: Sequential procedure of pose detection framework. Steps which are surrounded by a black border demonstrate image segmentation process: (a) raw depth image. (b) depth image after filling up the depth reading mistakes using SOR technique. (c) segmentation output, black pixels corresponding to the background vs. white pixels corresponding to the foreground. (d) labeling output from applying body detection algorithm on RGB images and extracting each human body GMM. (e) the segmented and labeled image from step (d) considering segmentation and labeling output from five previous frames. (f) the final segmented output.

Maedeh Aghaei received her B.S. degree in Computer Software Engineering from Science and Culture University, Tehran, Iran, in 2009. She received her M.S. degree in Artificial Intelligence in 2013 from Polytechnic University of Catalunya and is currently a Ph.D. Candidate in the Department of applied mathematics and analysis at the University of Barcelona. Her research interests are in the areas of gesture recognition, pose recognition, scene analysis and tracking, with emphasis on event modeling, recognition and analysis.

Cast Shadows and Self Shadows detection in Natural Images

Camp Davesa
Advisor: Ramon Baldrich
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: cdavesa@cvc.uab.es

Keywords: shadows, object recognition

1 Summary of Previous and Current Work

We investigate a new method for cast and self shadow detection. Detecting shadows would allow to create a new image representation where further methods could obtain better results, as multi illuminant estimation, object detection or scene understanding.

In figure 1 there is a scheme of the method. First of all, we segment the image in superpixels or regions using the method explained in [1]. In order to detect which region is a shadow or not, we assume that two regions from the same surface should have a similar chromatic value with different intensity, so we convert the images into opponent colours.

To calculate these values, we use the RAD method [2]. We calculate the ridges for each possible pair of opponent channels. Analysing the first and second channel, we obtain the chromatic information. If both regions belong to the same ridge and the mean value of each one is almost the same, the probability of one of them being a shadow region is high. Then we analyse the pairs of channels containing the intensity one. If in both cases, the regions belong to the same ridge but with some distance between the mean values, the probability of shadow is high. Otherwise, if the mean value is the same it means that they have the same intensity, so the probability of being a shadow is low.

Obviously, in the case that the regions do not belong to the same region in one of the tree pair of channels, the probability of one of them being the shadow of the other is automatically 0.

This process is repeated for all the regions of the image and in different scales. The shadow result is a combination of all the shadow regions obtained in all scales.

2 Future Work and Challenges

One problem that we have to face is the difficulty to find the correct ridges for each image. For each image, the parameters change, and there is no model we can follow to calculate them. In order to solve that, we are studding different ways to obtain the same information but in a more simple and efficient way.

We want to prove our method in different databases, until that moment, we only have proved it in some images, and the results are promising.

Publications

[1] Felzenszwalb, Pedro F and Huttenlocher, Daniel P, Efficient graph-based image segmentation. In *International Journal of Computer Vision*, 59, 167–181, 2004.

[2] Vazquez, Eduard and van de Weijer, Joost and Baldrich, Ramon, Image segmentation in the presence of shadows and highlights. In *Computer Vision–ECCV*, 1–14, 2008.

Camp Davesa received her B.S. degree in Computer Science from Universitat Autònoma de Barcelona, Barcelona, Spain, in 2010. She received her M.S. degree in 2011 and is currently a Ph.D. Candidate in the Department of Computer Science at the Universitat Autònoma de Barcelona. Her research interests are in the areas of computer vision, pattern recognition and color analysis.

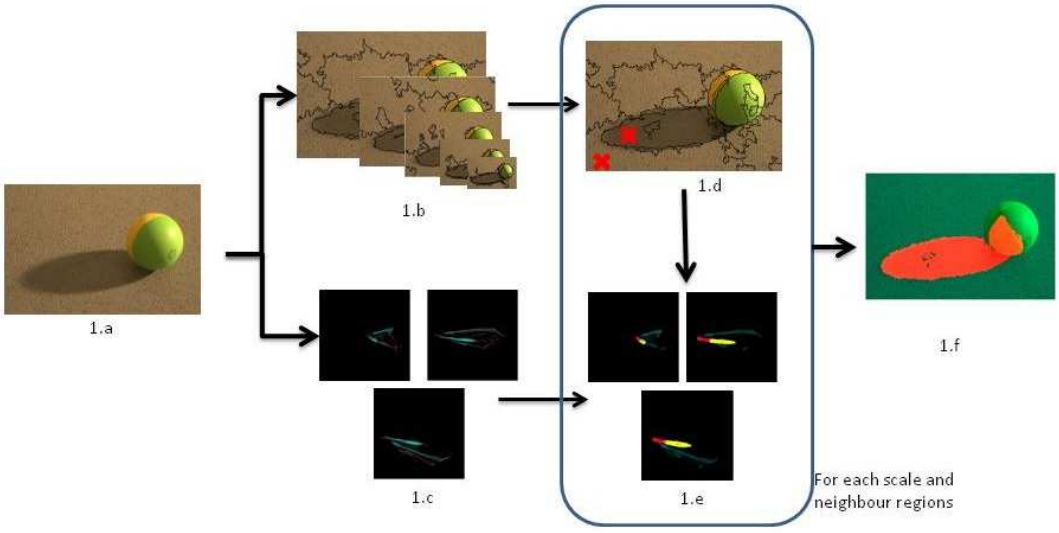


Figure 1: Example of graphical representation of the presented methodology. (a) Original image. (b) Segmented scaled images using the method in [1]. (c). Distribution of the opponent image and the ridges found using the method in [2]. Each distribution corresponds to the different combinations of the opponent channels. (d). Selection of two neighbour regions. This is done for every pair of neighbour regions. (e) Representations in the distribution of each region: Red represents the shadow region and yellow the non-shadow region with each mean value. (d) and (e) are repeated for each scale. (f) Result of the shadow detection. Red zones are the shadow ones.

Performance Analysis of Optical Flow in the Absence of Ground Truth

Patricia Márquez-Valle
Advisors: Debora Gil, Aura Hernàndez-Sabaté
Computer Vision Center & Computer Science Department at Universitat Autònoma de Barcelona
pmarquez@cvc.uab.es

Keywords: Optical Flow, Confidence Measures, Performance Evaluation.

1 Summary of Previous and Current Work

A correct optical flow computation is of prime importance for its further use in applications such as car driver assistance, 3D reconstruction, video compression, or medical support. However, current techniques may produce errors, induced wether by the method itself or by the computations. Consequently, we need to discard those wrong computations by means of a confidence measure. A good confidence measure should be able to provide an upper bound of the error. That is, for a given value of the confidence measure, the error should not be higher than a chosen threshold. Thus, we need tools to evaluate a confidence measure's performance and also tools to determine which is the value of the confidence measure that ensures that the error is bounded.

In the literature several confidence measures have been defined [6, 11, 7, 8, 9, 5], and some papers compare the different measures by means of the Sparsification plots [7]. However, as far as we know, there is no thorough evaluation of the confidence measures so that we know the errors they can detect. In addition, we have not found how to determine if a confidence measure is able to bound the error. Finally, we have not found any literature trying to learn a threshold of the confidence measure that ensures for a given risk, a bound of the error. So, until now we only have definition of confidence measures and comparison across them, but we still missing a deep analysis to truly evaluate their capabilities according the different optical flow methods and a further analysis to learn the threshold of the confidence measure.

As a first step, a confidence measure based on the numerical stability of Lucas-Kanade-based schemes [10] has been defined in papers [4, 2] (fig.1 Step 1). As a second step, in order to improve the Sparsification plots [7], a framework that assesses when a confidence measure is able to bound the error has been defined in [1] (fig.1 Step 2). However, such framework still lack of capabilities of assessing a threshold of the confidence measure that assures that the error is bounded for a given risk. Thus, in paper [3] the concept of risk is introduced, as well as a new framework that can assess the risk (fig.1 Step 2). In addition, there is a thorough analysis of the different sources of error of different optical flow algorithms.

Currently we are working on the definition of a new framework such that, for a given threshold of the error of the optical flow, we can predict for each value of the confidence measure the percentage of points that are above such threshold. Such framework combines the frameworks defined in [1, 3] and it is better than the previous ones because it integrates the concept of risk, and the capability of detect if the confidence measure can bound the error (fig.1 Step 3).

2 Future Work and Challenges

The current framework we are working on, will be tested for different confidence measures and optical flow techniques on databases with ground truth for optical flow. From the information obtained, we can learn the pair optical flow-confidence measure (OF-CM) that works better to discard erroneous computations according to a sequence. And thus, we can choose such OF-CM pair automatically from sequences without ground truth (fig.1 Step 3).

The final complete framework should assess, for a given sequence, risk, and maximum error expected, the best pair OF-CM and the value of the confidence

measure that ensures for the given risk, that the errors are not higher than the maximum error expected (fig.1 Step 4). Note that this framework can only be applied for sequences that have similar features since there is a process to chose the pair OF-CM and to learn the threshold of the confidence measure.

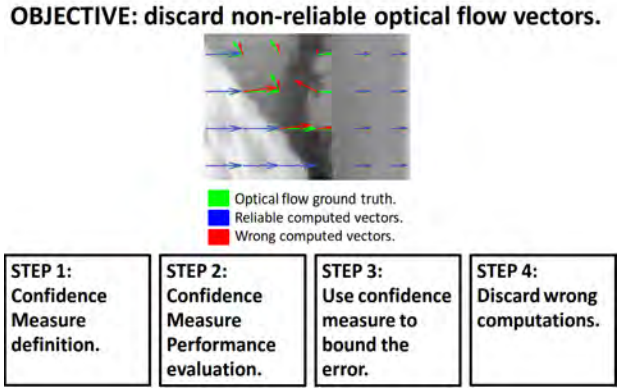


Figure 1: Graphical representation of the steps of the presented methodology.

Publications

[1] P. Márquez-Valle, D. Gil, and A. Hernàndez-Sabaté. A complete confidence framework for optical flow. In *European Conference on Computer Vision - Workshops*, pages 124–133, 2012.

[2] P. Márquez-Valle, D. Gil, and A. Hernàndez-Sabaté. Error analysis for lucas-kanade based schemes. In *International Conference on Image Analysis and Recognition*, pages 184–191, 2012.

[3] P. Márquez-Valle, D. Gil, A. Hernàndez-Sabaté, and D. Kondermann. When is a confidence measure good enough? In *International Conference on Computer Vision Systems*, pages 344–353, 2013.

[4] P. Márquez-Valle, D. Gil, and A. Hernàndez-Sabaté. A confidence measure for assessing optical flow accuracy in the absence of ground truth. In *International Conference on Computer Vision - Workshops*, pages 2042–2049, 2011.

References

[5] O. M. Aodha, A. Humayun, M. Pollefeys, and G. Brostow. Learning a confidence measure for optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1107–1120, 2013.

[6] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. *International Journal of Computer Vision*, 12:43–77, 1994.

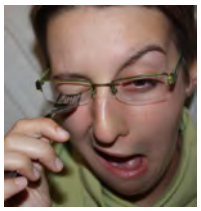
[7] A. Bruhn and J. Weickert. A confidence measure for variational optic flow methods. In *Geometric Properties for Incomplete Data*, pages 283–298, 2006.

[8] C. Kondermann, R. Mester, and C. Garbe. A statistical confidence measure for optical flows. In *European Conference on Computer Vision Workshops*, pages 290–301, 2008.

[9] J. Kybic and C. Nieuwenhuis. Bootstrap optical flow confidence and uncertainty measure. *Computer Vision and Image Understanding*, pages 1449–1462, 2011.

[10] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereovision. In *DARPA Image Understanding Workshop*, pages 121–130, 1981.

[11] J. Shi and C. Tomasi. Good features to track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.



Patricia Márquez-Valle has a background in mathematics and received her Msc in Computer Vision and Artificial Intelligence at the Universitat Autònoma de Barcelona in 2010. She is doing a PhD on Computer Science under the supervision of Debora Gil and Aura Hernàndez-Sabaté at the Computer Vision Center. She has a grant from the Spanish government within the Miocardia project. Her research interests are in the areas of optical flow techniques, confidence measures for optical flow, performance evaluation, 3D reconstruction, and medical imaging.

Mid-level descriptors foe object representation: stability and centrality

Ekaterina Zaytseva
Advisor: Jordi Vitria
Computer Vision Center & Department de Matematica Aplicada and Analisi of the University of Barcelona
E-mail: ezaytseva@cvc.uab.es

Keywords: mid-level descriptor, histogram of oriented gradients, object representation

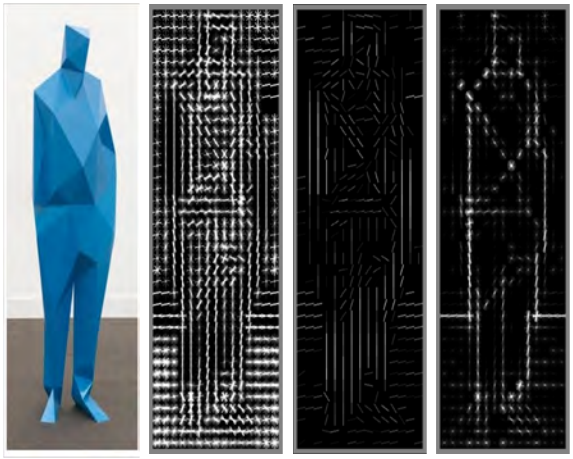
1 Summary of Previous and Current Work

Histograms of orientations and statistic derived from them have proven to be effective image representations for various recognition tasks. During last year we attempt to improve the accuracy of object detection systems by including the new features that explicitly capture mid-level information. More precisely we have developed two new feature descriptors, stability and centrality. The example of this descriptors can be seen in figure 2. Stability is a local statistical characteristic extracted from HOG and it assigns a value that represents the homogeneity of the gradient vector field that corresponds to a HoG cell. The concept behind the stability indicates that easier visual classes are characterized by higher stability values whilst harder classes have fewer admissible orientations. The continuity is a global characteristic and it represents the level of continuity of the object edges that are represented by that cell. The use of two additional mid-level characteristics should increase the performance of classifiers because more abstract information is readily accessible to it. This work was published in [1]

2 Future Work and Challenges

The objective of the future work will be:

- Including robust texture descriptors in the mid-level image representation, because one problem that we have to face is the difficulties when detecting objects with rich texture;



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

Figure 1: (a) Original image; (b) HoG features of (a); (c) Stable orientations of (b); (d) Cells of (b) with high continuity values.

- Use of the recently introduced paradigm of Structured Support Vector Machines for multi-class classification;
- Study of automatic methods for part definition in part-based representation of objects. This problem is now solved by manual definition of the number of parts, but it can be done automatically by using latent SSVM models.

Publications

[1] Ekaterina Zaytseva, Santi Segu, and Jordi Vitri, Sketchable Histograms of Oriented Gradients for Object Detection. In *17th Iberomerican Conference on Pattern Recognition. Buenos Aires, Argentina*, 2012. Lecture Notes Voluem. 7441

Ekaterina Zaytseva received her B.S. degree in Mathematics and Computer Science from Lomonosov Moscow State University, Moscow, Russia, in 2007. He received his M.S. degree in Telecommunications from Technical University of Catalonia, BarclonaTECH, in 2011 in and is currently a Ph.D. Candidate in the Department de Matematica Aplicada and Analisi at the University of Barcelona and Computer Vision Center in Bellaterra, Spain. Her research interests are in the areas of object detection and classification.

Error-Correcting Output Codes and Graph Cuts Optimization for Human Segmentation in Still Images

Daniel Sánchez[1], Tomás Pérez[1], Miguel Ángel Bautista[1,2], and Sergio Escalera[1,2]
[1]*University of Barcelona- Applied Math and Analysis Dept., 08007, Barcelona*
[2]*Computer Vision Center, Campus UAB, 08193, Barcelona*
E-mail: gammarl@gmail.com, topeya@gmail.com, mbautista@ub.edu, sergio@maia.ub.es

Keywords: Human limbs Segmentation, Error-Correcting Output Codes, Graph Cuts

1 Summary of Previous and Current Work

Recovering human pose in still images is a hard task because of the high variability in appearance produced by changes in the point of view, lighting conditions, and number of articulations of the human body. Even so, it has become one of main interest area of research because of its capabilities in final applications. Actually, state of the art approaches like [1] have achieved very good results for person detection and [2] in human pose estimation (detecting articulations).

We propose a two-level approach for the segmentation of the human body. In a first step, a set of human limbs were trained to be split in a tree-structure way and trained using a cascade of Adaboost classifiers with Haar-like features [3]. Then, it was included in a ternary Error-Correcting Output Codes (ECOC) [4] framework to improve classification performance by correcting errors of individual classifiers. This first classification step was applied in a windowing way on a new test image, defining a body-like probability map, which was used in a Graph Cuts optimization procedure. The proposed methodology is tested in a novel limb-labeled data set [5].

As present work, we follow the pipeline mentioned before by including a two-level refinement approach in order to segment human limbs. As a first step, we split in a tree-structure without background a set of 6 human limbs categories. Then, this tree-structure splitting is trained using SVM classifiers with HOG features and included in a ECOC framework. In order to perform multi-limb classification, we apply sliding windows on a new test image, defining in this case a set of limb-like probability maps which are used in a multi-limb human segmentation stage by means of alpha-beta swap Graph Cuts optimization [6]. The obtained results show that we are able to segment the human body limbs with high overlapping scores. The pipeline of the developed system is summarized in Figure 1.

2 Future Work and Challenges

We have to deal with the detection of tubular structure like arms, forearms, legs and thighs. Since they can be quite similar, in order to face the problem we plan to modify the descriptors and including contextual information. As future work, we also plan to use the achieved multi-segmentation results to define rich postural descriptors in time series and perform Human Action/Gesture recognition.

References

[1] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection" *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 1, 886–893 vol. 1, 2005.

[2] Y. Yang, D. Ramanan, "Articulated Pose Estimation using Flexible Mixtures of Parts" *IEEE Conference on Computer Vision and Pattern Recognition*, 1, 1385–1392 vol. 1, 2011.

[3] Viola, P., Jones, M., "Rapid object detection using a boosted cascade of simple features" *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE*

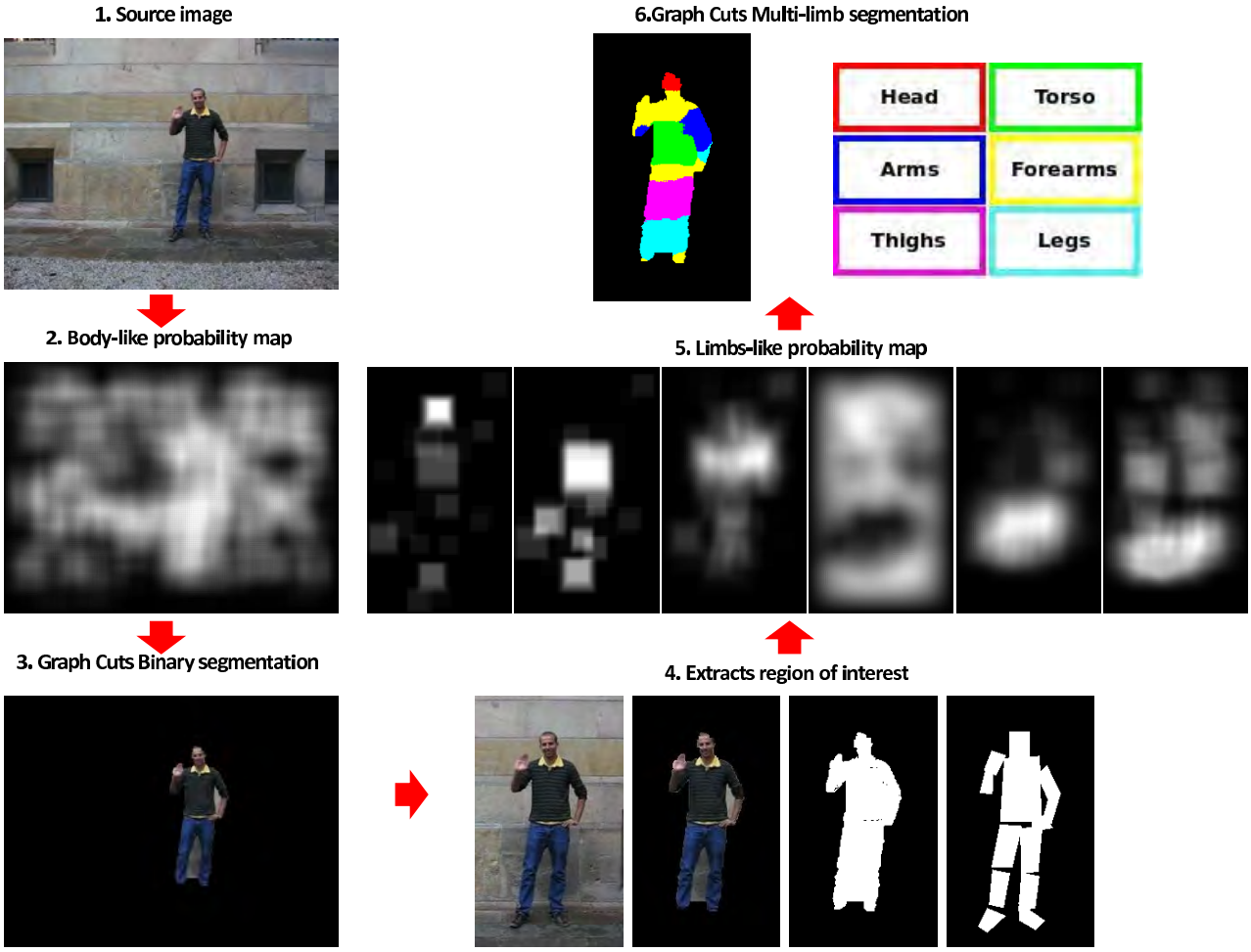


Figure 1: Summary of the proposed methodology for multi-limb human body segmentation.

Computer Society Conference on, 1, 511–518 vol. 1, 2001.

[4] Escalera, S., Pujol, O., Radeva, P., "On the decoding process in ternary error correcting output codes" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32, 120–134, 2010.

[5] Sánchez, D., Ortega, JC., Bautista, MÁ., Escalera, S., "Human Body Segmentation with Multi-limb Error-Correcting Output Codes Detection and Graph Cuts Optimization" *Iberian Conference on Pattern Recognition and Image Analysis*, 7887, 50-58, 2013.

[6] Hernández-Vela, A. and Zlateva, N. and Marinov, A. and Reyes, M. and Radeva, P. and Dimov, D. and Escalera, S., "Graph cuts optimization for multi-limb human segmentation in depth maps-Graph cuts optimization for multi-limb human segmentation in depth maps" *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 726–732, 2012.

Daniel Sánchez received his Bachelor degree in Computer Science at Universitat de Barcelona (UB) in 2012. He is currently studying his Master degree in Artificial Intelligence at UPC (Universitat Politècnica de Catalunya), UB and URV. He is mainly interested in computer vision applied to human pose recovery and behavior analysis.

Towards automatic colonoscopy quality assessment: Vascular content characterization.

Joan M. Núñez Do Rio
Advisor: Fernando Vilariño
Computer Vision Center & Computer Science Department at Universitat Autònoma de Barcelona
E-mail: jmnunez@cvc.uab.es

Keywords: Colonoscopy, Blood vessel, Anatomical structuring

1 Summary of Previous and Current Work

Colorectal cancer ranks in the third place in incidence and it is the fourth most common cause of cancer death worldwide. Based on demographic trends, the annual incidence is expected to increase by nearly 80% to 2.2 million cases over the next two decades, mostly in the less developed regions of the world. Fortunately, experience in Europe -where colorectal cancer is the second leading cause of cancer deaths with approximately 435.000 new cases diagnosed yearly- has shown that systematic early detection and treatment has the potential to improve control of the disease.

Colon cancer's survival rate depends on the stage it is detected on. Hence the importance of detecting it on its early stages by using screening techniques, such as colonoscopy. Colonoscopy screening technique has been proved to be the most effective method to reduce mortality rates caused by colorectal cancer. However, the effectiveness of the procedure in reducing colon cancer incidence depends on several reasons and some studies have shown a substantial variability of performance among different centers and endoscopists. Clinical studies have cleared some reasons and circumstances that can have consequences in colonoscopy effectiveness, such as colon preparation, cecal intubation or withdrawal time.

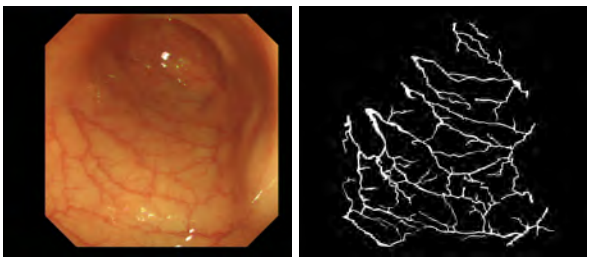
Developing standardized systems to assess colonoscopy efficiency involve many tasks. Several objects or regions appear in the endoluminal scene. The knowledge of these regions would clear up the path to an standard and effective screening protocol. After focusing on the study of polyp characterization and localization [1], we aim to improve our knowledge of the endoluminal scene by taking advantage of the information that blood vessels, lumen or colon wall folds can provide. One of the final goals we may accomplish would consist in being able to create time and position flags that would allow doctors to go back to a determined spot in the colon track. That is why our research aims to find and track anatomical markers like the vascular content in the scene [2].

2 Future Work and Challenges

We are creating an image database from several colonoscopy videos. A manual segmentation of the vascular content is provided for each selected frame as well as a manual selection of keypoints, such as vessel endpoints and junctions (see Figure 1). The accurate segmentation of the blood vessels in the scene has been showed to be a very difficult task. Actually, computing the thorough segmentation of the blood vessels is not strictly necessary for our goals. We are trying to develop a keypoint detector so that they could become landmarks to face future tasks. We intend to locate the most meaningful landmarks so that this knowledge may lead us to conform a characterization of the vascular tree. Besides, well-know and characterized landmarks may be tracked along the colonoscopy video so that this knowledge could become a good start point for the characterization of the other regions in the endoluminal scene.

Publications

[1] Bernal, J. and Sánchez, J. and Vilariño, F., Towards automatic polyp detection with a polyp appearance model. In *Pattern Recognition*, 45(9):3166-3182, 2012.



(a) (b)

Figure 1: (a) Input image. (b) Vascular content manual segmentation.

[2] Núñez, J.M. and Bernal, J. and Sánchez, J and Vilariño, F., Blood vessel characterization in colonoscopy images to improve polyp localization. In *International Conference on Computer Vision Theory and Applications*, 1:162-171, 2013.

Joan M. Núñez received his B.S. degree in Telecommunication Engineering from Universitat Politècnica de Catalunya, Barcelona, Spain, in 2007. He received his M.S. degree in 2010 and is currently a Ph.D. Candidate in the Computer Science Department at Universitat Autònoma de Barcelona. His research interests are in the areas of medical imaging, video understanding, object tracking and motion analysis.

Medical image sequence analysis by means of novel method for simultaneous registration and modeling

Simeon Petkov
Advisor: Carlo Gatta
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: spetkov@cvc.uab.es

Keywords: X-ray, registration, segmentation

1 Summary of Previous and Current Work

Segmentation, registration and modeling are three important aspects of general medical imaging processing. The advantage of tackling them simultaneously has been recognized, giving rise to several applied methods. However, the problem of segmenting blood vessels in cardiac X-ray sequences, tracking them and modeling of blood flow is not addressed. One of the main challenges, as illustrated in Fig. 1, is to distinguish between geometric variations (translation of catheter or blood vessel) and inherent signal changes due to physiological processes (injection of contrast liquid or change in its gray level). The current stage of development uses a-priori models of gray level variations for catheter, contrast liquid and background. The registration component of our method is based on the existing SPREF (Spatio-temporal regularity flow) method for non-rigid registration. In addition, we implemented a framework for generating synthetic gray level sequences together with registration ground truth between subsequent frames. The popular Free-Form Deformation method was applied to the synthetic data demonstrating its deficiency in discriminating between a contrast liquid flowing in a blood vessel and catheter translation.

2 Future Work and Challenges

Firstly, the implementation of the unified framework for simultaneous registration, segmentation and modeling needs to be finished and quantitatively evaluated. The next step is to obtain automatically the models for different objects and physiological processes using the input data.

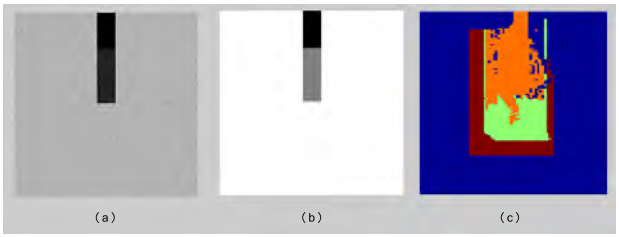


Figure 1: Exemplar frame of a synthetic sequence (a) and segmentation ground truth of catheter, contrast liquid and background (b). The third image (c) depicts latest segmentation results of our method.

Simeon Petkov graduated as a bachelor in Informatics in 2009 at Sofia University (Bulgaria). In 2010 he was enrolled in a Master course in Artificial Intelligence at Barcelona University. The next year he started working in Computer Vision Center on a project for automatic myocardial perfusion estimation. In 2012 he completed the master course and started a Ph.D. research on simultaneous modeling, registration and segmentation in medical imaging. The funding for the Ph.D. program is provided by FI national scholarship.

Towards airway characterization in respiratory endoscopy

Carles Sánchez
Advisor: F Javier Sánchez, Debora Gil
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: csanchez@cvc.uab.es

Keywords: Bronchoscopy, Tracheobronchial rings, Lumen detection

1 Summary of Previous and Current Work

Videobronchoscopy is a medical imaging technique that allows interactive navigation inside the respiratory pathways and minimal invasive interventions. Tracheal procedures are ordinary interventions that require measurement of the percentage of obstructed pathway for injury (stenosis) assessment. Visual assessment of stenosis in videobronchoscopic sequences requires high expertise of trachea anatomy and is prone to human error in 30% of the cases [1]. We propose two new techniques for a description of the airway:

In one hand accurate detection of tracheal rings [2] is the basis for automated estimation of the size of stenosed trachea. Processing of videobronchoscopic images acquired at the operating room is a challenging task due to the wide range of artifacts and acquisition conditions. We have presented a model of the geometric-appearance of tracheal rings for its detection in videobronchoscopic videos. Experiments on sequences acquired at the operating room, show a performance close to inter-observer variability.

In the other hand accurate lumen centre detection. The proposed method is based on the appearance and geometry of the lumen, which we defined as the darkest image region which centre is a hub of image gradients [3]. Experimental results validated on the first public annotated gastro-respiratory database prove the reliability of the method for a wide range of images (with precision over 95%).

Results of this two methods can be seen in Figure 1. In the first two images we can see how tracheal rings are accurately detected and not extra structures are included. The last two images show how we are detecting the centre of the lumen even if the image has multilumen. Another important point is the potential of our algorithm in detecting lumen presence.

2 Future Work and Challenges

Our future work will consist on getting objective measures of the airways using the information extracted from:

- Lumen segmentation using previously computed centre point of the lumen as seed (providing information about camera position).
- Fill tracheal ring discontinuities (providing information about shape of the bronchial tree).

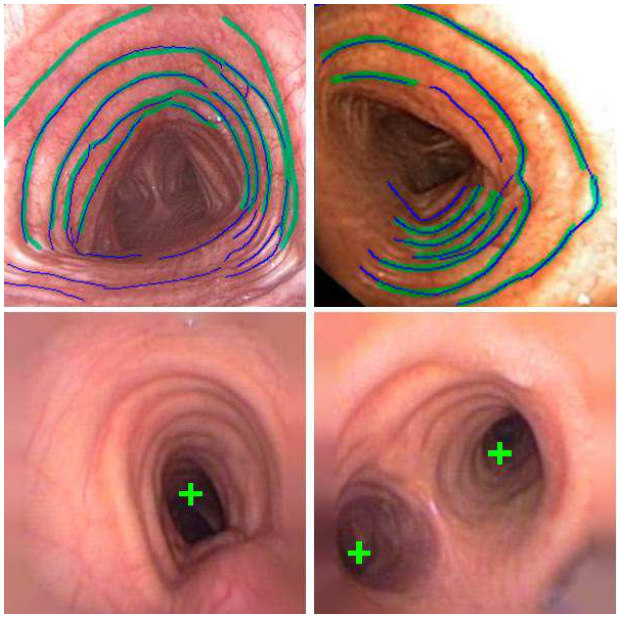


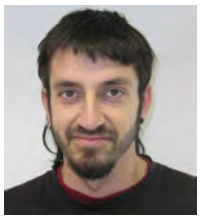
Figure 1: Results of the presented methodology in, both, tracheal ring segmentation (first row) and lumen centre detection (second row).

Publications

[1] Norwood S et al., Incidence of tracheal stenosis and other late complications after percutaneous tracheostomy. In *Anals of surgery*, vol. 2, pp. 232-233.

[2] C. Sanchez, D. Gil, A. Rosell, A. Andaluz, and F. J. Sanchez, Segmentation of tracheal rings in videobronchoscopy combining geometry and appearance. In *VISAPP 2013*, vol. 1, pp. 153-161.

[3] C. Sanchez, J. Bernal, D. Gil and F. J. Sanchez, On-line lumen centre detection in gastrointestinal and respiratory endoscopy. In *MICCAI-CLIP 2013*.



Carles Sanchez received his B.S. degree in Computer Science from Universistat Autonoma de Barcelona (UAB) in 2009 in Bellaterra, Barcelona. He received his M.S. degree in Computer Vision and Artificial Intelligence in 2011. He is currently a Ph.D. candidate and an assistant teacher in the Dep. of Computer Science in the UAB. His research interests are in the areas of Medical Imaging, Feature Detection and Extraction, Image Segmentation and Endoscopy images.

Artistic Heritage Motive Retrieval

Francesco Brughi
Advisor: Debora Gil, Oriol Ramos
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: fbrughi@cvc.uab.es

Keywords: CBIR, object detection, archaeology

1 Previous and current work

The work that has been carried out so far is a preliminary study on the possibility of applying content-based image retrieval (CBIR) techniques to the problem of dating archaeological art pieces, focusing in particular on Athenian painted pottery.

The goal of this work is to provide a system that may help archaeologists and art historians to classify unknown pottery pieces by searching for particular artistic motives within museum digital records. Such a new approach is motivated by the fact that nowadays methods are slow and highly inefficient [1]. With artistic motives, we refer to particularly recurring patterns that bring significant information about the time and place the artworks were crafted at.

Besides the typical challenges (variation in lighting condition, scale and perspective) offered by the common CBIR datasets [2], [3], we have to take into account the ambiguity introduced by the human hand. In fact, the strong human component that characterizes paintings produces a high variability of appearance, also within the same semantic class of artistic motives (Fig. 1). Moreover, the artistic motives the user might be interested in may be very different between each other (Fig. 2). This makes the motives hard to be treated all with the same method.

Figure 1: Examples of appearance variability within the same motive class.

Figure 2: Examples of shape difference between different motive classes.

First, we have addressed the problem with a bag-of-words (BOW) architecture [2], in order to check if current systems could tackle the critical aspects related with the dataset domain. The performance was poor, since, in most cases, the top-ranked results did not contain the searched query, whereas the last-ranked ones sometimes did. In order to determine the sources of errors, we have started by examining the image description stage, which is well-known to be a key issue. We have designed it as a modular system, decomposed in three main stages: interesting area detection, shape description and feature matching (Fig. 3). Within this framework, we have evaluated several existing methods [5],[4],[6] estimating their discriminative power when applied to the artistic motives of interest. To this end, we have considered the distribution of the matching distance values obtained both in positive cases (the target image contains the query) and negative cases (the target image does not). A method has been considered to be discriminative when it was possible to reject the hypothesis that the distributions of positive and negative samples have the same mean value.

With our experiments, we have showed that ad hoc image description approaches are necessary to treat different motives, given their peculiar characteristics. The results we have obtained are very different depending on which class of artistic motives was con-

sidered. For one of the three selected motives, we have found a suitable candidate as a feature extraction method. In the other two cases, all the evaluated techniques behaved poorly, providing us a basis to make hypothesis on such failure.

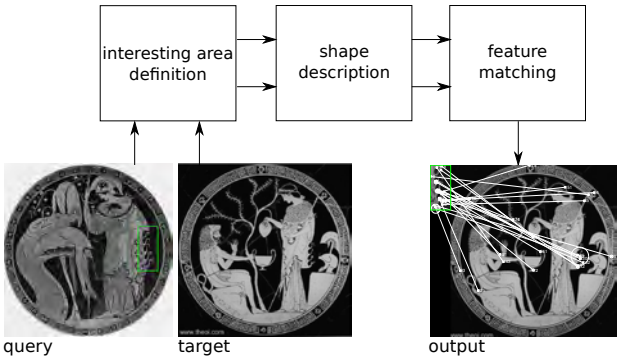


Figure 3: Example of graphical representation of the presented methodology.

2 Future work and challenges

A severe limitation to our research has been posed by the number and the quality of the available images. A larger dataset, possibly collected observing a specied protocol in order to uniform the image quality, is desirable for further studies.

The image description strategies we considered have to be improved in order to deal with the different semantic classes of artistic motives we treated. The future approach will consist in investigating their discriminative power by validating them against a ground truth of annotated images. The goal is to achieve a system which is capable to indicate the most suitable features to describe also unknown new motives.

A promising improvement to the image description stage would consist of introducing spatial consistency conditions, that may help when a signicant number of consistent matches are achieved, to discard the wrong ones.

Finally, in order to cope with the relevant variability within the artistic motive classes, in case a larger dataset was acquired, a good strategy might be to define a hierarchy where the main semantic classes, are divided in sub-classes by collecting the more similar patterns.

References

[1] R. M. Cook, *Greek Painted Pottery, 3rd ed.*, New York: Routledge, 1996.

[2] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," *ICCV*, pp. 1470–1477, 2003.

[3] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching", *CVPR*, 2007.

[4] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[5] J. Matas, O. Chum, U. Martin, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions", *BMVC*, vol. 1, pp. 384-393, 2002.

[6] R. Ortiz, "Freak: Fast retina keypoint," *CVPR*, pp. 510–517, 2012.

Francesco Brughi received his B.S. degree in Electronics Engineering from the University of Bologna, Italy, in 2008. He received his M.S. degree in Electronics Engineering in 2012 from the Politecnico of Torino, and his M.S. degree in Computer Vision and Artificial Intelligence in 2013 from the Univeritat Autonoma de Barcelona (UAB). He is currently a Ph.D. Candidate in the Computer Vision Center (CVC) at the UAB. His research interests are in the areas of CBIR.

Exploring low-level vision models. Case study: saliency prediction.

Ivet Rafegas
Advisor: Maria Vanrell
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: irafegas@cvc.uab.es

Keywords: bio-inspired, deep hierarchies, early vision models, visual cortex, saliency estimation

1 Summary of Previous and Current Work

We have been researching for several existing low-level vision models, that have proposed hierarchical schemes ([1],[2]) to simulate the first stages of the ventral stream. We have been analysing Malik-Perona (MP) [7], HMAX [8] and the Induction-Derived family (ID) models ([3], [4], [5], [6]), giving a unifying overview of them, but focusing my master dissertation in the ID, due to its generalisation properties shown in predicting both colour induction effects and saliency maps. Its main stages can be summarized as a linear filtering (L1) followed by a centre-surround mechanism (L2), a divisive normalisation (L3) and the application of a weighting function (ECSF) (L4) (see Figure 1). Although there exist a large number of hierarchical bio-inspired models, we are interested in understanding the functionality step by step. L1 consists of a frequency-orientation selectivity. In L2 there is applied the first non-linearity to refine the previous responses. L3 makes a second non-linearity to extend maximum values and threshold the minimum. As a particular layer, ID model apply L4 to weight previous responses in order to enhance or discard them depending on the visual problem. Finally, L5 is where a new representation is achieved, closely to the visual codes. We have been exploring each layer function by proposing alternative implementations to achieve more accurate responses. As case study, we have been working on saliency prediction since a large standard datasets are available allowing testing the effects of the studied alternatives. Additionally, we have started to explore how we can scale on these hierarchies, in view of a more complex task such as object recognition. In this line, we derive a new representation from the model output that can be the starting point for a trainable layer that could give a visual code for object recognition. Our work lead us to conclude that L1 is more accurate when DOOG (Differences of Ofset Gaussians) family of filters is used. The use of adapted center-surrounds in L2 provides also more accurate responses and opens the possibility to be adapted to detect more complex features. We have seen that the divisive normalisation step (L3) can be improved using sigmoid functions. Finally, the performed experiments appear to diminish the effects of L4, but the weighting function could be important to be learnt in other visual tasks.

2 Future Work and Challenges

The main idea is to extend the ID layers to complete V1 responses and also, to extend them beyond V1, facing higher-level visual tasks such as face recognition. Nevertheless, current stages can be also improved. In L1, we should make an study of the best settings to use for DOOG family of filters. In L2, the centre-surround regions follow a constraint size-relation between them and we propose to use a multi-scale. Regarding L3, the shape or the parameters of this function requires to be studied depending on the task is being faced. Finally, in L4, our hypothesis still remains that ECSF could correspond to a responses task-adaptation step related to the L3 parameters, and it could be defined using common machine learning techniques.

Bibliography

[1] J.J. DiCarlo, D. Zoccolan and N.C. Rust, How does the brain solve visual object recognition?.

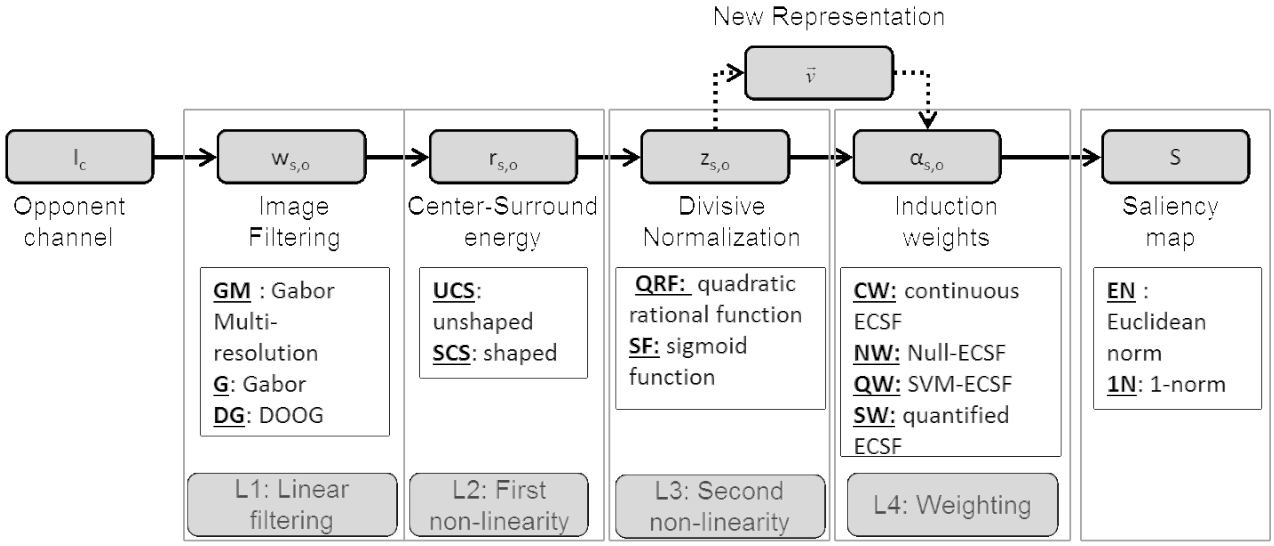


Figure 1: Example of graphical representation of the ID stages and the different alternatives that we have been exploring.

In *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol. 29 no. 3, 411-426, 2006.

[2] M. Riesenhuber and T. Poggio, Hierarchical models of object recognition in cortex. In *Nature Neuroscience*, vol. 2, 1019-1025, 1999.

[3] X. Otazu, M. Vanrell, and C. A. Parraga, Multiresolution wavelet framework models brigthness induction effets. In *Vision Research*, vol. 48, 733-751, 2008.

[4] X. Otaxu, C.A. Parraga, and M. Vanrell, Toward a unified chromatic induction model. In *Journal of Vision*, vol. 10(12), no, 6, 2010.

[5] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, Saliency estimation using a non-parametric low-level vision model. In *IEEE Conference onComputer Vision and Pattern Recognition (CVPR)*, 433-440, 2011.

[6] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, Low-level spatio-chromatic grouping for saliency estimation. In *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2003.

[7] J. Malik and P. Perona, Preattentive texture discrimination with early vision mechanisms. In *Journal of the Optical Society of America A*, vol. 7, 923-932, 1990.

[8] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, Robust object recognition with cortex-like mechanisms. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, 411-426, 2007.

Ivet Rafegas received her B.S. degrees in Computer Science and Mathematics from Universitat Autònoma de Barcelona, Spain, in 2012. She received her M.S. degree in Computer Science and Artificial Intelligence in 2013. Currently is a Ph.D. Candidate in the Computer Vision Center at the Universitat Autònoma de Barcelona. Her research interests are in the areas of bio-inspired low-level vision models based on the ventral stream process, enough accurate to be able to solve complex tasks such as object recognition.

Probabilistic Models for 3D Urban Scene Understanding

Prassanna Ganesh Ravishankar
Advisor: Gemma Sánchez, Antonio M. López
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: prassanna@cvc.uab.es

Keywords: visual scene understanding, probabilistic models, 3D reasoning, 3D semantic segmentation

1 Summary of Previous and Current Work

During the past few months, I have been conducting a literature review in the field of Visual scene understanding focused on the used of Probabilistic Models to face this open challenge. Visual scene understanding is an essential goal for developing autonomous vehicles. Emphasis is given to model the scene from input systems which are relatively cheaper and can be obtained *off-the-shelf*. The aim here is to use single or multiple cameras (visible or infra red) to capture the required information without having to use bulkier and expensive systems such as LIDAR.

Visual scene understanding is a problem that has been explored since the start of the computer vision research [1] and has been gaining large momentum over the last decade. Scene understanding (an example given in Fig. 1(a)), owing to its direct application in driverless automobiles has diverged in multiple branches, each using a unique (or different) method to achieve the end objective. Some mathematical tools that are quite useful to address this open question come from the probabilistic models as presented in [3].

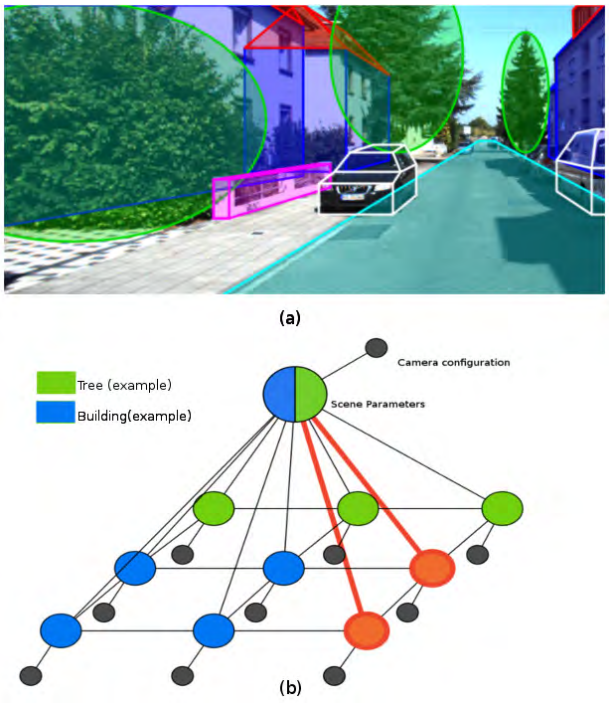


Figure 1: Example of a Urban Scene Understanding pipeline(a) *A labelled frame.* The labels are shown as polygons for simplicity (b) *Probablistic graphical model of elements that form the scene.* As a scene might be a combination of more than one class of objects or sub-scenes, this may be represented in the graphical model.

2 Future Work and Challenges

Since the final aim is to use this module for autonomous driving, emphasis is given to computational complexity (thereby speed - measured in frames per second) and performance. It is planned to make two versions of this system - offline (to build maps for later use) and online (to handle immediate on-road tasks). Studying the performance of systems which primarily rely on grammars and bottom-up/top-down approaches, it is found that they suffer in speed (computational complexity - due to the inference search space being large). In comparison, probabilistic graphical models (such as the one given in Fig. 1(b)) do well in such scenarios (outdoor scene understanding) [3]. The trade off between accuracy and computational complexity is a concern, however, the implementation of this system will be done with scalability in mind (following concepts similar to [7]). The major factors involved in the implemen-

tation of this project would include :- speed, training datasets (such as PASCAL [4]), feature selection (such as SIFT [5]), image segmentation (such as Superpixels [6]) and testing.

References

[1] Ohta, Yu-ichi, Takeo Kanade, and Toshiyuki Sakai, An analysis system for scenes containing objects with substructures. In *International Joint Conference on Pattern Recognitions*, 1978.

[2] Feng Han and Song-Chun Zhu, Bottom-up/top-down image parsing by attribute graph grammar. In *ICCV*, 2, 1778-1785, 2005.

[3] Geiger, Andreas, Christian Wojek, and Raquel Urtasun. Joint 3d estimation of objects and scene layout. In *Advances in Neural Information Processing Systems*, 1467-1475, 2011.

[4] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A., The PASCAL Visual Object Classes (VOC) Challenge. In *IJCV*, 88(2), 303-338, 2010.

[5] David G. Lowe, Distinctive image features from scale-invariant keypoints. In *IJCV*, 60(2), 91-110, 2004.

[6] Achanta, Radhakrishna, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Ssstrunk, Slic superpixels. In *cole Polytechnique Fdral de Lausssanne (EPFL)*, 60(2), Tech. Rep 149300, 2010.

[7] Yao, Jian, Sanja Fidler, and Raquel Urtasun, Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *In Computer Vision and Pattern Recognition (CVPR)*, 702-709, 2012.

Prassanna Ganesh Ravishankar received his B.E. degree in electrical engineering from Manipal Institute of Technology, India, in 2007. He received his M.S. degree in 2011 from the University of Sheffield, England and is currently a Ph.D. Candidate(since May 2013) in the Computer Vision Center at the Autonomous University of Barcelona. His research interests are in the areas of scene understanding, feature detection and machine learning.

Multiple Cues Integration for Query-by-String Word Spotting

David Aldavert
Advisor: Ricardo Toledo
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: aldavert@cvc.uab.cat

Keywords: Multiple Cue Integration, Word-Spotting, Bag of Visual Words

1 Summary of Previous and Current Work

Recently, I have been working in multi-modal integration applied in the field of document analysis. More precisely, we have developed a word spotting framework that follows the query-by-string paradigm where word images are represented both by textual and visual information. The textual representation is formulated in terms of character n-grams while the visual one is based on the bag-of-visual-words scheme. These two modalities are merged together and projected to a common topics space using a latent semantic model. This projection improves the word snippet representation by reducing ambiguities (e.g. uncertainties due to visual features used to represent different characters) and strengthening the relationship between the features used to represent the same concept (e.g. the different visual features used to represent the same character). Moreover, the mixture of textual and visual features into a common sub-vector space allows to, given a textual query, retrieve word instances that were only represented by the visual modality. This is very convenient property in scenarios where some modalities are more difficult to obtain than others.

2 Future Work and Challenges

The latent semantic model improves its performance as more samples are used to build the model. However, obtaining the transcription of handwritten documents can be a tedious task, so that, we are planning to use synthetic information to train the whole framework. Another problem is the vast possible parameter combinations of the bag-of-visual-words model. Consequently, we intend to use an adaptive framework which automatically generates the codebook and spatial representation based on the indexation errors obtained in the model training phase.

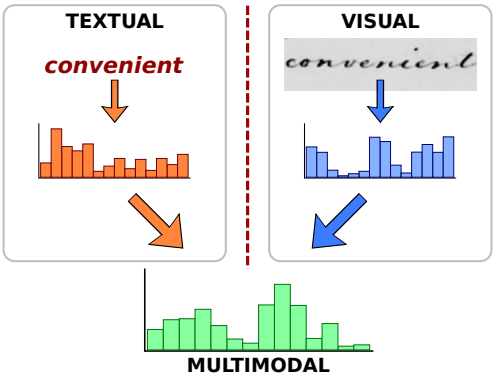


Figure 1: The method projects the textual and visual representations into a common space.

Publications

[1] David Aldavert, Marçal Rusiñol, Ricardo Toledo, Josep Lladós, "Integrating Visual and Textual Cues for Query-by-String Word Spotting". In *ICDAR*, 2013.

David Aldavert received his B.S. degree in Computer Science from Universitat Autònoma de Barcelona, Barcelona, Spain, in 2004. He received his M.S. degree in 2006 and is currently a Ph.D. Candidate in the Computer Vision Center at the Universitat Autònoma de Barcelona. His research interests are in the areas of mobile robotics, registration, semantic segmentation and historical document analysis with emphasis on real-time applications.

Looking at Faces: Detection, Tracking and Pose Estimation

Murad Al Haj
Advisors: Jordi Gonzàlez, F.Xavier Roca
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: murad@cvc.uab.es

Keywords: face detection and tracking, expression recognition, multiple-instance learning

1 Summary of Work

Humans can effortlessly perceive faces, follow them over space and time, and decode their rich content, such as pose, identity and expression. However, despite many decades of research on automatic facial perception, a complete solution remains elusive. Automatic facial perception encompasses many important and challenging areas of computer vision and its applications span a very wide range: video surveillance, human-computer interaction, content-based image retrieval, biometric identification, video coding and age/gender recognition.

In face detection, an initial simple model is presented that uses pixel-based heuristics to segment skin locations and hand-crafted rules to determine the locations of the faces present in an image. Different colorspaces are studied to judge whether a colorspace transformation can aid skin color detection. Experimental results show that the separability does not increase in other color spaces when compared to the RGB space. The output of this study is used in the design of a more complex face detector that is able to successfully generalize to different scenarios.

In face tracking, a framework that combines estimation and control in a joint scheme is presented to track a face with a single pan-tilt-zoom camera. An extended Kalman filter jointly estimates the object world-coordinates and the camera position. The output of the filter is used to drive a PID controller in order to reactively track a face, taking correct decisions of when to zoom-in on the face to maximize its size and when to zoom-out to reduce the risk of losing it. The applicability of this method is demonstrated on simulated as well as real-life scenarios.

Figure 1: In this Thesis, we address three problems in automatic face perception, namely face detection, face tracking and pose estimation.

The last and most important part of this thesis is dedicate to monocular head pose estimation. In most prior work on heads pose estimation, the positions of the faces on which the pose is to be estimated are specified manually. Therefore, the results are reported without studying the effect of misalignment. Regression, as well as classification, algorithms are generally sensitive to localization error. In this part, a method based on partial least squares (PLS) regression is proposed to estimate pose and solve the alignment problem simultaneously. The contributions of this work are two-fold: 1) demonstrating that the proposed method achieves better than state-of-the-art results on the estimation problem and 2) developing a technique to reduce misalignment based on the learned PLS factors that outperform multiple instance learning (MIL) without the need for any retraining or the inclusion of misaligned samples in the training process, as normally done in MIL.

Publications

[1] Murad Al Haj, Jordi Gonzàlez, Larry S. Davis, " On Partial Least Squares in Head Pose Estimation: How to simultaneously deal with misalignment ", in 25th IEEE Computer Vision and Pattern Recognition (CVPR2012), Providence, RI, June, 2012

[2] Murad Al Haj, Carles Fernández, Zhanwu Xiong, Ivan Huerta, Jordi Gonzàlez, F. Xavier Roca, " Beyond the Static Camera: Issues and Trends in Active Vision ", Visual Analysis of Humans: Looking at People, Chapter 2, Springer Netherlands, October, 2011.

[3] Murad Al Haj, Andrew D. Bagdanov, Jordi Gonzàlez and F. Xavier Roca, " Reactive object tracking with a single PTZ camera ", in 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August, 2010

[4] Murad Al Haj, Andrew Bagdanov, Jordi Gonzàlez, F. Xavier Roca, " Robust and Efficient Multipose Face Detection Using Skin Color Segmentation ", in 4th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA2009), Pvoa do Varzim, Portugal, June, 2009

[5] Murad Al Haj, Javier Orozco, Jordi Gonzàlez, Juan José Villanueva, "Automatic Face and Facial Features Initialization for Robust and Accurate Tracking", in 19th International Conference on Pattern Recognition (ICPR'2008), Tampa, Florida, USA, December, 2008

[6] Mikhail Mozerov, Ariel Amato, Murad Al Haj, Jordi Gonzàlez, " A simple Method of Multiple Camera Calibration for the Joint Top View Projection ", In 5th International Conference on Computer Recognition Systems (CORES'2007), Wroclaw, Poland, October, 2007

[7] Ariel Amato, Murad Al Haj, Mikhail Mozerov, Jordi Gonzàlez, " Trajectory fusion for Multiple Camera Tracking ", In 5th International Conference on Computer Recognition Systems (CORES'2007), Wroclaw, Poland, October, 2007

[8] Murad Al Haj, Ariel Amato, F. Xavier Roca, Jordi Gonzàlez, " Face Detection in Color Images using Primitive Shape Features ", In 5th International Conference on Computer Recognition Systems (CORES'2007), Wroclaw, Poland, October, 2007

[9] Murad Al Haj, Ariel Amato, Gemma Sánchez, Jordi Gonzàlez, " On-line One Stroke Character Recognition Using Directional Features ", in International Workshop on Advances in Pattern Recognition (IWAPR 2007), Plymouth, UK, July, 2007

[10] Ariel Amato, Murad Al Haj, Josep Lladós, Jordi Gonzàlez, " Computationally Efficient Graph Matching via Energy Vector Extraction ", in International Workshop on Advances in Pattern Recognition (IWAPR 2007), Plymouth, UK, July, 2007

Handwritten Word Spotting

Jon Almazán
Advisor: Ernest Valveny, Alicia Fornés
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: {almazan,afornes,ernest}@cvc.uab.es

Keywords: word spotting, segmentation-free, attributes

1 Single-writer Word Spotting with Exemplar-SVMs

We developed a system for efficient segmentation-free single-writer word spotting. The use of HOG templates provides a very natural model, and its discriminative power is improved through the use of Exemplar SVMs with SGD solver. The use of Product Quantization drastically improves the efficiency of the system at test time. Finally, the use of more informative features in combination with reranking and query expansion improves the final accuracy of the method at a reduced cost. A general scheme of the method can be seen in Figure 1. We refer the reader to [1] for further details.

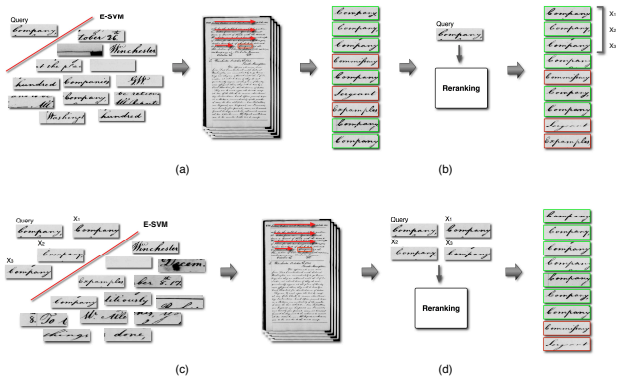


Figure 1: General scheme of the method proposed. (a) E-SVM training and sliding-window search. (b) First reranking of the best retrieved regions. (c) E-SVM retraining and sliding-window search applying query expansion with the first reranked regions. (d) Second reranking using the expanded training set.

2 Multi-writer Word Spotting with Embedded Attributes

We propose an approach to multi-writer word spotting, where the goal is to find a query word in a dataset comprised of document images. We propose an attributes-based approach (Figure 2) that leads to a low-dimensional, fixed-length representation of the word images that is fast to compute and, especially, fast to compare. This approach naturally leads to an unified representation of word images and strings, which seamlessly allows one to indistinctly perform query-by-example, where the query is an image, and query-by- string, where the query is a string. We also propose a calibration scheme to correct the attributes scores based on Canonical Correlation Analysis that greatly improves the results on a challenging dataset. We test our approach on two public datasets showing state-of-the-art results. We refer the reader to [2] for further details.

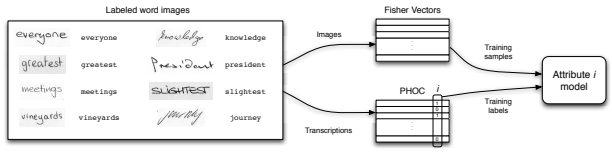


Figure 2: Training process for i-th attribute model. A classifier is trained using the FV representation of the images and the i-th value of the PHOC representation as label.

Publications

[1] Jon Almazán, Albert Gordo, Alicia Fornés, Ernest Valveny, Efficient Exemplar Word Spotting. In *BMVC*, 2012.

[2] Jon Almazán, Albert Gordo, Alicia Fornés, Ernest Valveny, Word Spotting with Corrected Attributes. In *ICCV*, 2013.

Jon Almazán received his B.Sc. degree in Computer Science from the Universitat Jaume I (UJI) in 2008, and his M.Sc. degree in Computer Vision and Artificial Intelligence from the Universitat Autònoma de Barcelona (UAB) in 2010. He is currently a PhD student in the Computer Science Department of the UAB and the Computer Vision Center under the supervision of Ernest Valveny and Alicia Fornés. He is an active member of the Document Analysis Group and assistant professor at the Computer Science Department in the UAB. His research work is mainly focused on image retrieval, shape recognition and attribute-based representations.

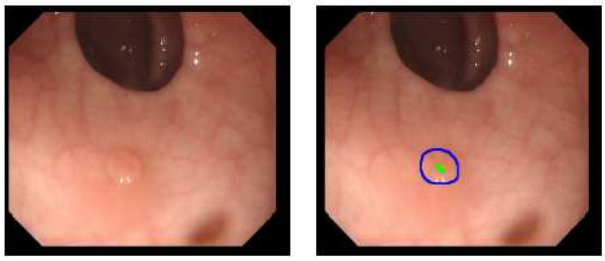
Polyp Localization and Segmentation in Colonoscopy Images by Means of a Model of Appearance for Polyps

Jorge Bernal
Advisors: F. Javier Sánchez, Fernando Vilariño
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: jbernal@cvc.uab.es

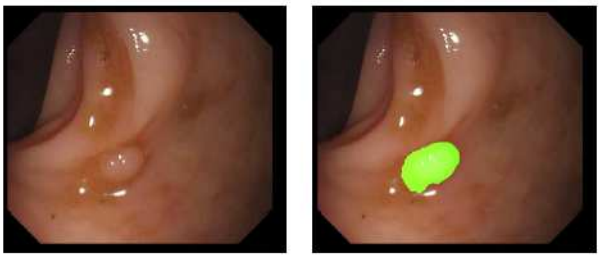
Keywords: Polyp Localization, Polyp Segmentation, Colonoscopy, Eye-tracking, Energy maps

1 Summary of Work

Colorectal cancer is the fourth most common cause of cancer death worldwide and its survival rate depends on the stage in which it is detected on hence the necessity for an early colon screening. There are several screening techniques but colonoscopy is still nowadays the gold standard, although it has some drawbacks such as the miss rate. Our contribution, in the field of intelligent systems for colonoscopy, aims at providing a polyp localization and a polyp segmentation system based on a model of appearance for polyps.



(a)



(b)

Figure 1: In this Thesis, we aim to: (a) locate and (b) segment the polyp in the image.

To develop both methods we define a model of appearance for polyps, which describes a polyp as enclosed by intensity valleys. The novelty of our contribution resides on the fact that we include in our model aspects of the image formation and we also consider the presence of other elements from the endoluminal scene such as specular highlights and blood vessels, which have an impact on the performance of our methods. In order to develop our polyp localization method we accumulate valley information in order to generate energy maps, which are also used to guide the polyp segmentation.

Our methods achieve promising results in polyp localization and segmentation. As we want to explore the usability of our methods we present a comparative analysis between physicians fixations obtained via an eye tracking device and our polyp localization method. The results show that our method is indistinguishable to novice physicians although it is far from expert physicians.

Publications

[1] Jorge Bernal, F. Javier Sánchez, Fernando Vilariño CURRENT CHALLENGES ON POLYP DETECTION IN COLONOSCOPY VIDEOS - From Region Segmentation to Region Classification. A Pattern Recognition-based Approach, Proceedings of the 2nd International Workshop on Medical Image Analysis and Description for Diagnosis Systems - MIAD 2011, pp. 62-71, Rome, Italy, January 2011

[2] Jorge Bernal, F. Javier Sánchez, Fernando Vilariño ”A Region Segmentation Method for Colonoscopy Images Using a Model of Polyp Appearance”, Proceedings of IbPRIA 2011, Las

Palmas de Gran Canaria, LNCS Vol. 6669, pp. 134-143. June 2011

[3] Jorge Bernal, Fernando Vilariño, F. Javier Sánchez ”Towards Intelligent Systems for Colonscopy”, In-Tech Colonoscopy Book, Chapter 16, pp. 245-270, July 2011.

[4] Jorge Bernal, F. Javier Sánchez, Fernando Vilariño. Integration of Valley Orientation Distribution for Polyp Region Identification in Colonoscopy. In In MICCAI 2011 Workshop on Computational and Clinical Applications in Abdominal Imaging (Vol. 6668, pp. 7683). Lecture Notes in Computer Science. Springer Link

[5] Jorge Bernal, F. Javier Sánchez, Fernando Vilariño ”Towards Automatic Polyp Detection With a Polyp Appearance Model”, Pattern Recognition (Vol. 45 Num. 9, pp. 3166-3182), 2012

[6] Joan Manel Nuñez, F. Javier Sánchez, Fernando Vilariño ”Blood vessel characterization in colonoscopy images to improve polyp localization”, In Proceeding of the 8th International Conference on Computer Vision Theory and Applications (Vol. 1, pp. 162171). SciTePress.

[7] Jorge Bernal, F. Javier Sánchez, Fernando Vilariño ”Impact of Image Preprocessing Methods on Polyp Localization in Colonoscopy Frames”, Proceedings of EMBC 2013, Osaka, Japan, July 2013

Jorge Bernal received the B.Sc. degree in Telecommunications Engineering from Universidad de Valladolid in 2008. He received his M.Sc. in Computer Vision and Artificial Intelligence in 2009 and his Ph.D. degree in 2012 at the Computer Vision Center (CVC/UAB). He is currently a postdoctoral researcher at CVC. His research interests include low level image processing, medical image analysis, image segmentation and biological structures characterization.

Model free approach to human action recognition

Bhaskar Chakraborty
Advisors: Jordi Gonzàlez, F.Xavier Roca
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: bhaskar@cvc.uab.es

Keywords: human action recognition, spatio-temporal interest points, large-scale surveillance analysis

1 Summary of Work

Automatic understanding of human activity and action is a very important and challenging research area of Computer Vision with wide scale applications in video surveillance, motion analysis, virtual reality interfaces, robot navigation and recognition, video indexing, content based video retrieval, HCI, health care, choreography and sports video analysis etc.

Our first approach towards this goal is based on a probabilistic optimization model of body parts using Hidden MarkovModel (HMM). This strong model based approach is able to distinguish between similar actions by only considering the body parts having major contributions to the actions, for example legs for walking and jogging; arms for boxing and clapping. Next approach is based on the observation that the action recognition can be done by only using the visual cue, i.e. human pose variation during the action, even with the information of few frames instead of examining the whole sequence. In this method, actions are represented by a Bag-of-key-poses model to capture the human pose changes during an action.

To tackle the problem of recognizing the action in complex scenes, we propose a model free approach which is based on the Spatio-temporal interest points (STIPs) and local feature. To this end, a novel selective STIP detector is proposed which uses a mechanism similar to that of the non-classical receptive field inhibition that is exhibited by most orientation selective neurons in the primary visual cortex. An extension of the selective STIP based action recognition is applied to the human action recognition in multi-camera systems. In this case, selective STIPs from each camera view point are combined using the 3D reconstructed data, to form 4D STIPs (3D space + time) for multi-view action recognition. The concluding part of the thesis dedicates to the continuous visual event recognition (CVER) on large scale video dataset. This is an extremely challenging problem due to high scalability, diverse real environment state and wide scene variability. To address these issues, a motion region extraction technique is applied as a preprocessing step. A max-margin generalized Hough Transform framework is used to learn the feature vote distribution around the activity center to obtain an activity hypothesis which is verified by a Bag-of-words + SVM action recognition system.

Figure 1: In this Thesis, we present a series of techniques to solve the problem of human action recognition in video.

We validate our proposed approaches on several benchmark action recognition datasets as well as small scale and large scale activity recognition datasets. We obtain state-of-the results which shows a progressive improvement of our proposed techniques to solve human action and activity recognition in video.

Publications

[1] Bhaskar Chakraborty, Jordi Gonzàlez, F. Xavier Roca, " Large scale continuous visual event recognition using max-margin Hough transformation framework ", Computer Vision and Image Understanding, in press., December, 2012

[2] Michael Holte, Bhaskar Chakraborty, Jordi Gonzàlez, Thomas B. Moeslund, " A Local 3-D Motion Descriptor for Multi-View HumanAction Recognition from 4-D Spatio-Temporal Interest Points ", IEEE Journal of Selected Topics in Signal Processing, in press, April, 2012

[3] Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund, Jordi Gonzàlez, " Selective Spatio-Temporal Interest Points ", Computer Vision and Image Understanding, in press, March, 2012

[4] Bhaskar Chakraborty, Andrew D. Bagdanov, Jordi Gonzàlez, F.Xavier Roca, " Human action recognition using an ensemble of body-part detectors ", Expert Systems, in press, doi: 10.1111/j.1468-0394.2011.00610.x, October, 2011

[5] Bhaskar Chakraborty, Michael B. Holte, Thomas B. Moeslund, Jordi Gonzàlez, and F. Xavier Roca, " A Selective Spatio-Temporal Interest Point Detector for Human Action Recognition in Complex Scenes ", in 13th International Conference in Computer Vision (ICCV2011), Barcelona, Spain, November, 2011

[6] Bhaskar Chakraborty, Andrew Bagdanov, Jordi Gonzàlez, " Towards Real-Time Human Action Recognition ", in 4th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA2009), Pvoa do Varzim, Portugal, June, 2009

[7] Bhaskar Chakraborty, Ognjen Rudovic, Jordi Gonzàlez, "View-Invariant Human-Body Detection with Extension to Human Action Recognition using Component-Wise HMM of Body Parts", in 8th IEEE International Conference on Automatic Face and Gesture Recognition (FG'2008), Amsterdam, The Netherlands, September, 2008

[8] Bhaskar Chakraborty, Marco Pedersoli, Jordi Gonzàlez, "View-invariant human action detection using component-wise HMM of body parts", 5th International Workshop on Articulated Motion and Deformable Objects (AMDO'2008), Andratx, Mallorca, Spain, July, 2008

[9] Marco Pedersoli, Jordi Gonzàlez, Bhaskar Chakraborty, Juan Jose Villanueva, " Enhancing Real-time Human Detection based on Histograms of Oriented Gradients ", In 5th International Conference on Computer Recognition Systems (CORES'2007), Wroclaw, Poland, October, 2007

Contributions to the Intestinal Motility Analysis by means of Wireless Capsule Endoscopy

Michał Drożdżal
Advisor: Petia Radeva
Computer Vision Center & Departament de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Barcelona, Spain
E-mail: michal@cvc.uab.es

Keywords: Medical Imaging, Capsule Endoscopy, Intestinal Motility

1 Summary of Previous and Current Work

This paper is a summary of the work that will appear in the PhD thesis. The thesis is in the field of automatic analysis of Wireless Capsule Endoscopy (WCE) data. More precisely we have been developing computer vision tools for automatical analysis and characterization of intestinal motility events acquired by WCE camera.

The graphical representation of the work carried out in the last years is presented in the Fig. 1. The work is divided into four main parts (chapters).

In the first chapter, we have tested various methods for motility visualization from WCE data and proposed a novel one called Motility Bar based on cost minimization by Dynamic Programming [2]. Motility Bar has permitted to see intestinal motility holistically as a single image.

In the second chapter, we have developed several models for automatic detection of different intestinal events from both 1) individual video frames, e. g. intestinal content [3] or wrinkles [1] and 2) Motility Bar, e. g. motility rhythms. Moreover, we have used concentration inequalities to define intestinal segments of similar motility. In particular, we have been working with the concentration inequalities for multivariate data streams.

In the third chapter, we have developed two types of efficient labeling schemes based on concepts from Active Learning that allow for quick labeling of huge amounts of WCE data [5, 6]. Finally, the automatic methods have been incorporated into a single model to detect abnormal intestinal motility [4].

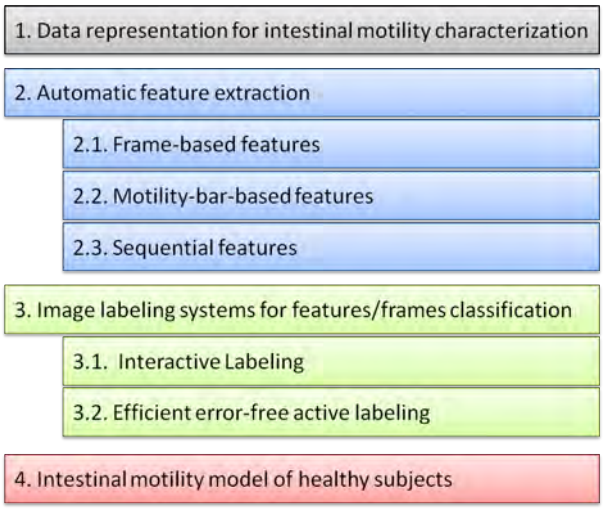


Figure 1: Graphical representation of work done during PhD studies.

Publications

[1] S. Seguí, M. Drozdzal, E. Zaytseva, C. Malagelada, F. Azpiroz, P. Radeva, and J. Vitrià, Detection of wrinkle frames in endoluminal videos using betweenness centrality measures for images. In *IEEE Journal of Biomedical and Health Informatics*, Under submission.

[2] M. Drozdzal, S. Seguí, J. Vitrià, C. Malagelada, F. Azpiroz, P. Radeva, Adaptable image cuts for motility inspection using WCE. In *Computerized Medical Imaging and Graphics*, Volume 37, Issue 1, January 2013, Pages 72-80.

[3] S. Seguí, M. Drozdzal, F. Vilarino, C. Malagelada, F. Azpiroz, P. Radeva, J. Vitrià, Categorization and Segmentation of Intestinal Content Frames for Wireless Capsule Endoscopy. In *Information Technology in Biomedicine, IEEE*

Transactions on, vol.16, no.6, pp.1341,1352, Nov. 2012.

[4] C. Malagelada, S. Seguí, S. Mendez, M. Drozdzal, J. Vitrià, P. Radeva, J. Santos, A. Accarino, J.R. Malagelada, F. Azpiroz, Functional gut disorders or disordered gut function? Small bowel dysmotility evidenced by an original technique. In *Neurogastroenterology & Motility*, Volume: 24, Issiue:3.

[5] M. Drozdzal, S. Seguí, C. Malagelada, F. Azpiroz, J. Vitrià, P. Radeva, Interactive labeling of WCE images. In *IbPRIA*, 2011: 143-150.

[6] M. Drozdzal, S. Seguí, P. Radeva, C. Malagelada, F. Azpiroz, J. Vitrià, An Application for Efcient Error-Free Labeling of Medical Images. In *Multimodal Interaction in Image and Video Applications*, pp 1-16.

Michal Drozdzal received his M.S. degree in Teleinformatics from Wroclaw University of Technology, Wroclaw, Poland, in 2008. Currently he is a Ph.D. Candidate in the Departament de Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Barcelona, Spain. His research interests are in the areas of wireless capsule endoscopy and medical imaging.

Symbol spotting in graphical documents with serialized subgraph hashing

Anjan Dutta
Advisors: Josep Lladós and Umapada Pal
E-mail: adutta@cvc.uab.es

Keywords: Graph Matching, Symbol Spotting, Graphic Recognition

1 Serialized subgraph hashing

Graphs are an efficient data structure for representing line drawings and (sub)graph matching naturally fits with the problem of symbol spotting. On the other hand, (sub)graph matching is an NP-hard problem. But the number of graphical documents is increasing day by day. So applying symbol spotting method to any real life application demands it to be time efficient. We propose a symbol spotting technique in graphical documents with hashing the subgraphs. We propose a graph serialization to reduce the usual computational complexity of graph matching. Serialization of graphs is performed by computing acyclic graph paths between each pair of connected nodes. Graph paths are one dimensional structures of graphs which are less expensive in terms of computation. At the same time they enable robust localization even in the presence of noise and distortion. Indexing in large graph databases involves a computational burden as well. We propose a graph factorization approach to tackle this problem. Factorization is intended to create a unified indexed structure over the database of graphical documents. Once graph paths are extracted, the entire database of graphical documents is indexed in hash tables by locality sensitive hashing (LSH) of shape descriptors of the paths. The hashing data structure aims to execute an approximate k-nearest neighbour search in a sub-linear time. We have performed detailed experiments with various datasets of line drawings and compared our method with the state-of-the-art works. The results demonstrate the effectiveness and efficiency of our technique. The detailed of the work can be found in [1].

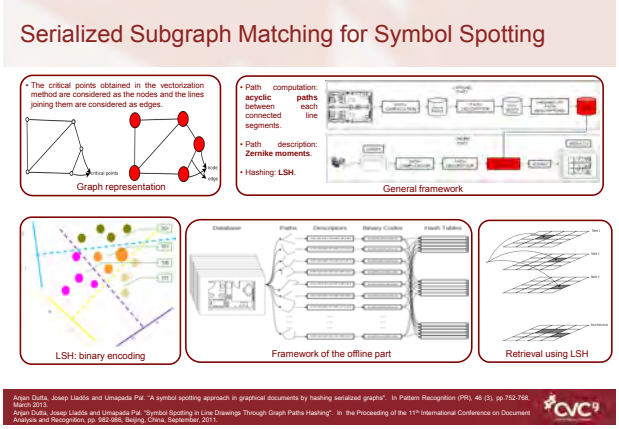


Figure 1: Example of graphical representation of the presented methodology.

Publications

[1] Anjan Dutta, Josep Lladós, and Umapada Pal. A symbol spotting approach in graphical documents by hashing serialized graphs. *Pattern Recognition*, 46(3):752–768, 2013.

Anjan Dutta received his Masters degree in Computer Vision and Artificial Intelligence from the Universitat Autònoma de Barcelona, Barcelona, Spain in the year of 2010. Currently he is a final year PhD student in the Centre de Visió per Computador, Barcelona, Spain under the supervision of Dr. Josep Llads and Dr. Umapada Pal. In his PhD he is working on subgraph matching applied for symbol spotting in graphical documents. His main research interests include efficient subgraph matching, graph indexing, graphics recognition, structural pattern recognition.

Word Spotting in Historical Handwritten Documents

David Fernández
Advisor: Josep Lladós, Alicia Fornés
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: dfernandez@cvc.uab.es

Keywords: Handwritten documents, Document image processing, Historical document analysis, Word-Spotting

1 Previous Work

We developed a an approach for word spotting in handwritten document images. We stated the problem from a focused retrieval perspective, i.e. locating instances of a query word in a large scale dataset of digitized manuscripts. We combined two approaches, namely one based on word segmentation and another one segmentation-free. The first approach uses a hashing strategy to coarsely prune word images that are unlikely to be instances of the query word. This work was published in the 13th International Conference on Frontiers in Handwritten Recognition (ICFHR 2012).

We also developed an algorithm to segment lines in any kind of handwritten documents. The algorithm is robust to different document structure, skew and warping disturbs. It is able to solve the problem of touching lines finding the optimal way to separate the touching components. The main idea of our algorithm is to formulate line segmentation as finding the optimal paths with minimum cost between two consecutive lines. This work is under review in IJDAR journal

We proposed a word segmentation algorithm which involves replacing the line segmentation of the work developed by Manmatha et. al with a much better line segmentation process inspired in the work explained above. The new approach substantially improves the final accuracy. The new line segmentation process for off-line handwriting searches a minimum energy path along the medial axis which divides consecutive text lines. Our method allows to segment the lines coping with the difficulties of multi-skew and touching components. This work was published in the 13th International Conference on Frontiers in Handwritten Recognition (ICFHR 2012).

2 Current Work: Word Spotting using structural information

The books of the Barcelona Cathedral dataset present always the same structure: in the first pages appears the index of the book and in the rest of the document appears the marriage records. The index presents the first surname of the husband and the wife and the page where is the record. The record has, in the left part, the surname of the husband. Both appear in the same order, so, we can use methods of alignment to do the matching between the indexes and the records.

We are developing two different Word Spotting approaches where we use the structural/semantical information.

The first one is centered in the Markov Logic Networks. We propose the use of Markov Logic Networks (MLN) to improve the results of word spotting according to the stated hypothesis (Fig. 1). MLN is a very powerful statistical relational learning model that provides a very rich representation. The use of MLN to model a grammatical structure offers more flexibility in the definition of the rules, incremental and simple learning,with respect to traditional language models used in handwriting recognition. As experimental setup, a database of handwritten marriage licenses of the Barcelona Cathedral Archive has been used. The documents are semi-structured in records (paragraphs). Each record contains the information of a marriage using a regular structure, but with some variations from one period to another, or from one social status to another.

The second word spotting approach we are developing is based in the structural information. The dataset (*Llibre d'Esposalles*) used in this work is composed by 254 volumes, which each one has to

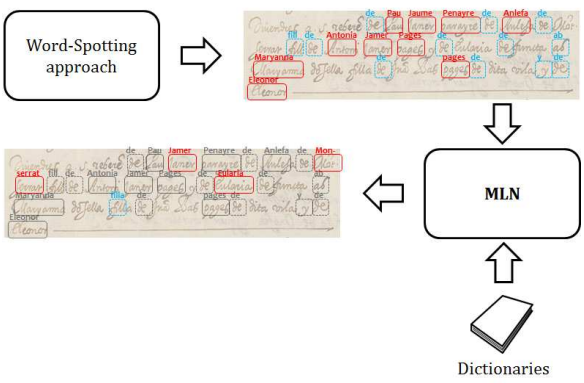


Figure 1: Architecture of the MLN approach.

parts: the index of the book and the records. The index contains a relation of surnames of the husband of the marriage and the page where appears the marriage record. The records is compose by three parts: on the left part the surname of ht husband, in the central part the information of the record and on the right part the tax to pay. The objective of this work is to do the alignment between the words of the indexes (surname of the husbands) and the surnames that appear in the left column of the registers (Fig. 2). A characteristic of these documents is the order in which the words appear, they have the same order, but in the pages of the records appears words interspersed. So, the idea is to do a word-spotting using as input queries the words of the indexes, and using the structural information that the words appears in the same order, and discarding the words that appears in the records between the true positives words.

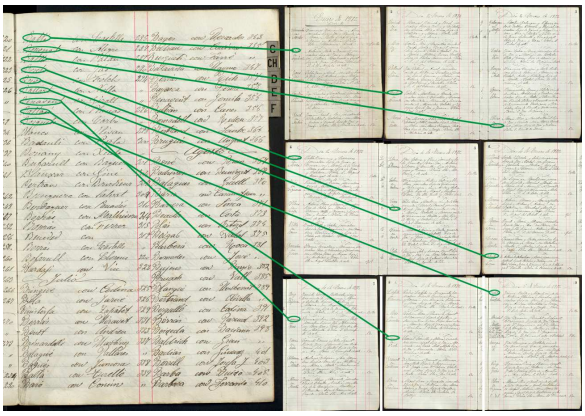


Figure 2: Alignment of the indexes.

Publications

[1] Fernández D., Lladós J., Fornés A., A Graph based Approach for Segmenting Touching Lines in Historical Handwritten Documents. In *IJDAR*, Under review.

[2] Fernández D., Marinai S., Lladós J., Fornés A., Contextual Word Spotting in Historical Manuscripts using Markov Logic Networks. In *HIP*, 36–43, 2013.

[3] Lladós J., Rusiñol M., Fornés A., Fernández D., Dutta A., On the influence of word representations for handwritten word spotting in historical documents. In *IJPRAI*, 26, 2012.

[4] Fernández D., Manmatha R., Lladõs J., Fornés A., On influence of line segmentation in efficient word segmentation in old manuscripts. In *ICFHR*, 759-764, 2012.

[5] Almazán J., Fernández D., Fornés A., Lladós J., Valveny E., A Coarse-to-Fine Approach for Handwritten Word Spotting in Large Scale Historical Documents Collection. In *ICFHR*, 453-458, 2012.

[6] Fernández D., Lladós J., Forns A., Handwritten Word Spotting in Old Manuscript Images using a Pseudo-Structural Descriptor Organized in a Hash Structure.le. In *IbPRIA*, 628-635, 2011.

David Fernández graduated in Computer Science from the Universitat Jaume I of Castellón. He received his M.S. degree in Computer Vision and Artificial Intelligence from the Universitat Autònoma de Barcelona, Barcelona, Spain in 2010. Currently he is pursuing his Ph.D. in the Centre de Visió per Computador, Barcelona, Spain under the supervision of Dr. Josep Lladós and Dr. Alicia Fornés. In his Ph.D. he is working on historical handwritten documents. His main research interests include enhancement and segmentation of documents and word spotting.

Towards Deep Image Understanding: From pixels to semantics

Josep M. Gonfaus
Advisors: Jordi Gonzàlez, Theo Gevers, F.Xavier Roca
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: gonfaus@cvc.uab.es

Keywords: image segmentation, object detection, multi-class recognition

1 Summary of Work

Understanding the content of the images is one of the greatest challenges of computer vision. Recognition of objects appearing in images, identifying and interpreting their actions are the main purposes of Image Understanding.

The image reflected on the retina of the eye (human or robotic), or that, by extension, of a video or a camera taking a picture, enables the user to conceptualize its surroundings and, therefore, to interact with it. For example, for an intelligent robot or a smart car to function effectively, it is essential that they recognize their environment to navigate safely. Similarly, in the near future, web browsers will also need to recognize the image contents in order for indexation to take place. This thesis seeks to identify what is present in a picture. Our objective is to categorize and locate all objects within an image.

Firstly, to deepen the knowledge on the creation of images, we suggest a method to recognize the physical properties used to produce the image. By combining photometric with geometric information, we learn to distinguish between material edges and scene alterations caused by shadows or light reflections.

Semantic segmentation focuses on resolving the ambiguity within categories at the pixel-level. This task is essentially done by adding contextual information. We propose three scale levels in order to resolve such ambiguity. At low level, we learn whether the appearance of a pixel resembles the object or not. At middle level, we add information about the object as a whole entity. At top level, we enforce consistency with the rest of the scene, introducing the concept of semantic co-occurrence.

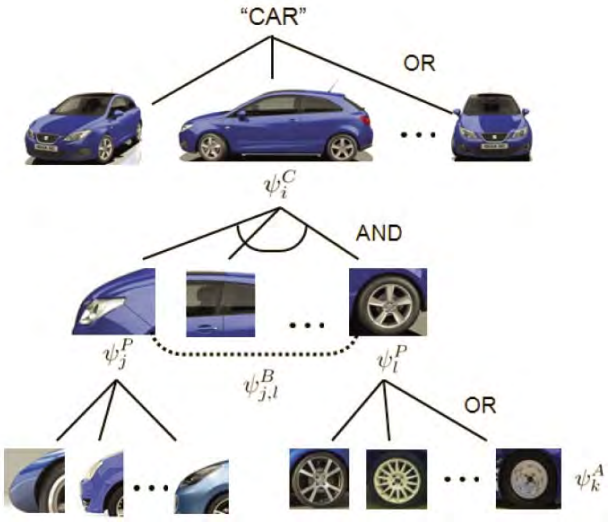


Figure 1: In this Thesis, we seeks to identify what is present in a picture, i.e. to categorize and locate all objects within an image.

Finally, regarding object detection, we present two new methods. The first one is focused on improving the object representation at local level with the concept of factorized appearances. An object is represented by several parts. Each of those can then be represented by more than one local appearance. The second method addresses the computational problem of identifying and locating thousands of categories of objects in an image. The main advantage of this method is to create representations of objects which can be reused for other objects, which reduces the computational cost for the other categories.

The results given have been validated on several commonly used datasets, reaching international recognition and state-of-the-art within the field.

Publications

[1] Marc Castelló, Jordi Gonzàlez, Ariel Amato, Pau Baiget, Carles Fernández, Josep M. Gonfaus, Ramón A. Mollineda, Marco Pedersoli, Nicolás Pérez de la Blanca, F. Xavier Roca, " Exploiting Multimodal Interaction Techniques for Video-Surveillance ", Multimodal Interaction in Image and Video Applications, Intelligent Systems Reference Library Vol. 48, Chapter 8, Springer Netherlands, February, 2013

[2] Josep M. Gonfaus, Theo Gevers, Arjan Gijsenij, F. Xavier Roca, Jordi Gonzàlez, " Edge Classification Using Photo-Geometric Features ", in 21st International Conference on Pattern Recognition (ICPR'2012), Tsubuka, Japan, November, 2012

[3] Xavier Boix, Josep M. Gonfaus, Joost van de Weijer, Andrew D. Bagdanov, Joan Serrat, Jordi Gonzàlez, " Harmony Potentials Fusing Global and Local Scale for Semantic Image Segmentation ", International Journal of Computer Vision, doi:10.1007/s11263-011-0449-8, in press , April, 2011

[4] Jordi Gonzàlez, Josep M. Gonfaus, Carles Fernández, F. Xavier Roca, " Exploiting Natural-Language Interaction in Video Surveillance Systems ", in V&L Net Workshop on Vision and Language, Brighton, UK, September, 2011

[5] Josep Gonfaus, Xavier Boix, Joost Van de Weijer, Andrew Bagdanov, Joan Serrat, and Jordi Gonzàlez, " Harmony Potentials for Joint Classification and Segmentation ", in 23rd IEEE Conference on Computer Vision and Pattern Recognition (CVPR2010), San Francisco, CA, June, 2010

[6] Josep M. Gonfaus, Xavier Boix, Fahad S. Khan, Joost van de Weijer, Andrew D. Bagdanov, Marco Pedersoli, Joan Serrat, Xavier Roca, Jordi Gonzàlez, " Harmony Potentials: Fusing Global and Local Scale for Semantic Image Segmentation ", the PASCAL Visual Object Classes Challenge Workshop, in conjunction with ECCV2010, Crete, Greece, September, 2010

3D Motion Data aided Human Action Recognition and Pose Estimation

Wenjuan Gong
Advisors: Jordi Gonzàlez, F. Xavier Roca
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: wenjuan@cvc.uab.es

Keywords: action recognition, pose estimation, bag-of-poses

1 Summary of Work

In this Thesis, we explored human action recognition and pose estimation problems. Different from traditional works of learning from 2D images or video sequences and their annotated output, we seek to solve the problems with additional 3D motion capture information, which helps to fill the gap between 2D image features and human interpretations.

We first compare two different schools of approaches commonly used for 3D pose estimation from 2D pose configuration: modeling and learning methods. By looking into experiments results and considering our problems, we fixed a learning method as the following approaches to do pose estimation. We then establish a framework by adding a module of detecting 2D pose configuration from images with varied background, which widely extend the application of the approach. We also seek to directly estimate 3D poses from image features, instead of estimating 2D poses as a intermediate module. We explore a robust input feature, which combined with the proposed distance measure, provides a solution for noisy or corrupted inputs. We further utilize the above method to estimate weak poses,which is a concise representation of the original poses by using dimension deduction technologies, from image features. Weak pose space is where we calculate vocabulary and label action types using a bog of words pipeline. Temporal information of an action is taken into consideration by considering several consecutive frames as a single unit for computing vocabulary and histogram assignments.

To validate the proposed methods, we use HumanEva data set, IXMAS data set and TUM kitchen data set. The experiments we conducted includes: compare the performances of modeling and learning methods for estimating 3D poses from 2D poses with the training set of HumanEva data set and TUM kitchen data set under different conditions, like different performers, viewpoints, and action types; using state-of-art body part detectors, we detect 2D pose configurations from HumanEva data set and take 2D pose configurations as inputs for the pose estimation framework; for action recognition, we use cross validation to fix the dimension of weak poses and the size of temporal steps; also in action recognition experiments, we compare action recognition accuracies from only 2D image features and incorporating 3D motion information.

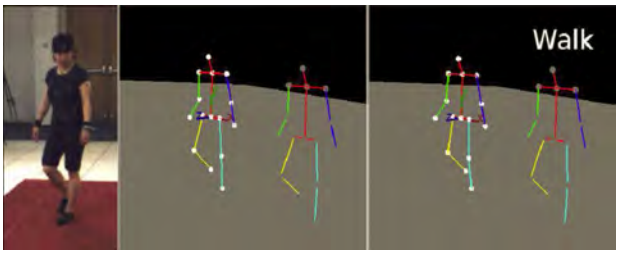


Figure 1: In this Thesis, we explore the effect of 3D motion data in 3D pose estimation and action recognition problems.

From the work, we conclude that 3D motion data, which solve the ambiguity of 2D representation itself, could be utilized directly for accurate pose estimation and aids to enhance action recognition accuracies from 2D image sequences compared with using solely 2D image features. In our future work, we would like to explore how to improve the mapping mechanism from feature space to pose or action space that would hopefully fill the semantic gap.

Publications

[1] Wenjuan Gong, Jordi Gonzàlez and F. Xavier Roca, " Human Action Recognition based on Estimated Weak Poses ", EURASIP Journal on Advances in Signal Processing, in press, June, 2012

[2] Adela Barbulescu, Wenjuan Gong, Jordi Gonzàlez, Thomas Moeslund, F. Xavier Roca, " 3D Human Pose Estimation Using 2D Body Part Detectors ", in 21st International Conference on Pattern Recognition (ICPR'2012), Tsubuka, Japan, November, 2012

[3] Wenjuan Gong, Jordi Gonzàlez, João Manuel R.S. Tavares and F. Xavier Roca, " A New Image Dataset on Human Interactions ", in 7th Conference on Articulated Motion and Deformable Objects (AMDO'2012), Andratx, Mallorca, Spain, July, 2012

[4] Nataliya Shapovalova, Wenjuan Gong, Marco Pedersoli, F. Xavier Roca and Jordi Gonzàlez, " On Importance of Interactions and Context in Human Action Recognition ", in 5th Iberian Conference on Pattern Recognition and Image Analysis (ibPRIA2011), Las Palmas de Gran Canaria, Canary Islands, Spain, June, 2011

[5] Wenjuan Gong, Jürgen Brauer, Michael Arens, Jordi Gonzàlez, " Modeling vs. Learning Approaches for Monocular 3D Human Pose Estimation ", in 1st IEEE International Workshop on Performance Evaluation on Recognition of Human Actions and Pose Estimation Methods (PERHAPS2011), in conjunction with ICCV2011, Barcelona, Spain, November, 2011

[6] Jürgen Brauer, Wenjuan Gong, Jordi Gonzàlez, Michael Arens, " On the Effect of Temporal Information on Monocular 3D Human Pose Estimation ", in 2nd IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS2011), in conjunction with ICCV2011, Barcelona, Spain, November, 2011

[7] Wenjuan Gong, Andrew D. Bagdanov, F. Xavier Roca, and Jordi Gonzàlez, " Automatic Key Pose Selection for 3D Human Action Recognition ", 6th International Conference on Articulated Motion and Deformable Objects (AMDO2010), Andratx, Mallorca, Spain, July, 2010

Static and dynamic tumor quantification in PET scans

Frederic Sampedro[1] and Sergio Escalera[2,3]

[1] *Autonomous University of Barcelona, Faculty of Medicine, 08193 Barcelona, Spain*
[2] *Computer Vision Center, Campus UAB, Edifici O, 08193, Barcelona, Spain*
[3] *Dept. Applied Mathematics, University of Barcelona, 08007, Barcelona, Spain*
E-mail: fredsampedro@gmail.com, sergio@maia.ub.es

Abstract In this work we present an automatic tumor volume segmentation system of whole body PET scans, which would provide a relevant quantitative and objective framework in clinical nuclear medicine settings, specially in cancer response assessment scenarios. We focus on supervised learning schemes and contextual learning strategies.

Keywords: Tumor segmentation, PET scans

1 Introduction

Positron Emission Tomography (PET) is a 3D nuclear medicine metabolic imaging technology widely used in cancer management. Whole-body FDG-PET/CT scans allow nuclear medicine physicians to possibly identify any malignant activity at any anatomical location throughout the patient's body, obtaining a global picture of the patient's cancer state and spread [1]. Quantitative information regarding the tumor volume, aggressiveness (related to its metabolic activity) and spread is an important complement in the clinical setting, since it offers an objective way to summarize the overall patient's cancer state (Fig. 1). However, expert-guided tumor segmentation of whole body PET/CT scans is highly time-inefficient and suffers from inter-and intra-observer variabilities.

2 Method

To perform automatic segmentation of tumor regions is an extremely complex task due to the highly variable anatomical and physiological properties of human bodies. Fig. 2 summarizes our supervised learning framework design to deal with this problem.

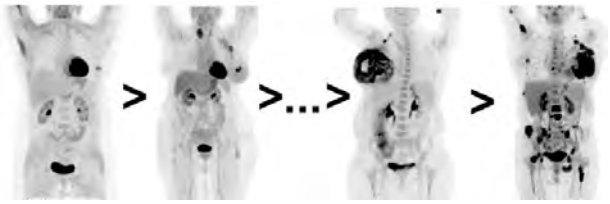


Figure 1: Several whole body PET scans with different degree of tumor presence and spread.

One key clinical application of oncological whole body PET technology is cancer evolution assessment, where physicians conclude about the patient's cancer progression or response condition (Fig. 3). In the clinical setting, there is a lack of objective quantitative information regarding the severity of the cancer progression, which would prove very useful in cancer treatment response analysis scenarios [2]. Fig. 4 shows our proposed framework for addressing this problem in a completely automatic manner.

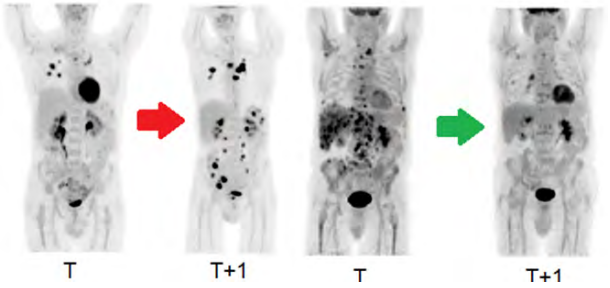


Figure 3: Patient's cancer progression (left) and response (right) clinical scenarios given two time consecutive (T and T+1) whole body PET scans.

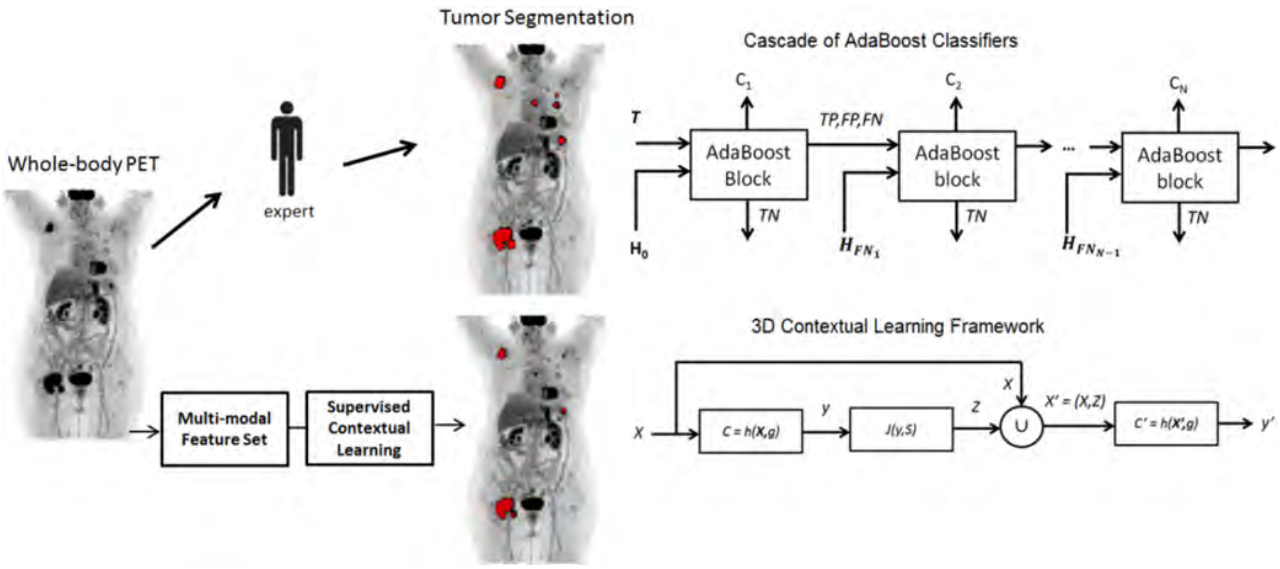


Figure 2: Automatic whole body PET tumor segmentation supervised learning framework.

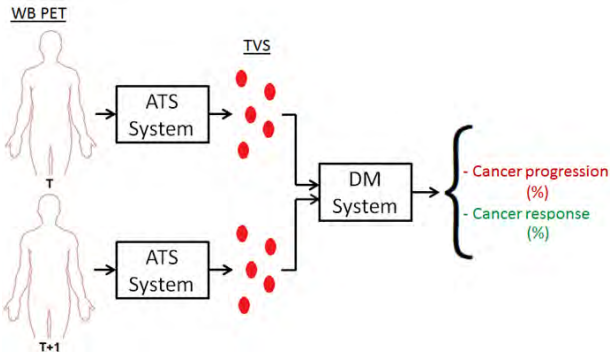


Figure 4: Automatic cancer evolution assessment framework. WB PET: Whole Body PET scan, ATS: Automatic Tumor Segmentation, TVS: Tumor Volume Segmentation, DM: Decision Making.

References

[1] M. Okada, N. Sato , K. Ishii, K. Matsumura, M. Hosono, T. Murakami, FDG PET/CT versus CT, MR Imaging and 67Ga Scintigraphy in the Posttherapy Evaluation of Malignant Lymphoma., RadioGraphics 30, 939-957 (2010).

[2] H. Zhang, K. Wroblewski, S. Liao, R. Kampalath, B. Penney, Y. Zhang, Y. Pu, Prognostic value of metabolic tumor burden from 18F-FDG PET in Surgical Patients with Non-small-cell Lung Cancer., Academic Radiology 20, 32-40 (2013).

Frederic Sampedro received his Bachelor degree in Computer Science (2010), Electrical Engineering MSc (2012), Biomedical Engineering MSc (2012). Current he is a PhD candidate at Hospital de Sant Pau (Autonomous University of Barcelona): whole body PET/CT automatic tumor quantification. Research interests: applied biomedical signal and image processing.

Sergio Escalera obtained the P.h.D. degree on Multi-class visual categorization systems at Computer Vision Center, UAB. He leads the Human Pose Recovery and Behavior Analysis Group. He is a lecturer at the Universitat de Barcelona. He is a partial time professor at Universitat Oberta de Catalunya. He is member of the Computer Vision Center at Campus UAB. He is Editor-in-Chief of American Journal of Intelligent Systems and advisor of ChaLearn Challenges in Machine Learning.

Domain Adaptation of Virtual and Real Worlds for Pedestrian Detection

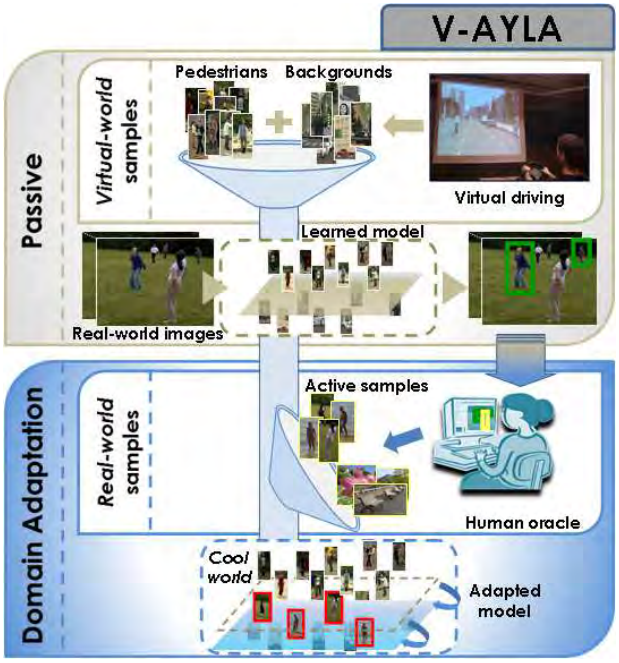
David vázquez
Advisors: Antonio M. López, Daniel Ponsa
Computer Vision Center & Computer Science Department at Universitat Autonoma de Barcelona
E-mail: dvazquez@cvc.uab.es

Keywords: Domain Adaptation, Pedestrian Detection, ADAS, Machine Learning

1 Summary of Work

Pedestrian detection is of paramount interest for many applications, e.g. Advanced Driver Assistance Systems, Intelligent Video Surveillance and Multimedia systems. Most promising pedestrian detectors rely on appearance-based classifiers trained with annotated data. However, the required annotation step represents an intensive and subjective task for humans, what makes worth to minimize their intervention in this process by using computational tools like realistic virtual worlds. The reason to use these kind of tools relies in the fact that they allow the automatic generation of precise and rich annotations of visual information. Nevertheless, the use of this kind of data comes with the following question: *can a pedestrian appearance model learnt with virtual-world data work successfully for pedestrian detection in real-world scenarios?*. To answer this question, we conduct different experiments that suggest a positive answer. However, the pedestrian classifiers trained with virtual-world data can suffer the so called dataset shift problem as real-world based classifiers does.

Accordingly, we have designed different domain adaptation techniques to face this problem, all of them integrated in a same framework (V-AYLA). We have explored different methods to train a domain adapted pedestrian classifiers by collecting a few pedestrian samples from the target domain (real world) and combining them with many samples of the source domain (virtual world). The extensive experiments we present show that pedestrian detectors developed within the V-AYLA framework do achieve domain adaptation. Ideally, we would like to adapt our system without any human intervention.



(a)

Figure 1: Domain adaptation framework (V-AYLA:*virtual*-world *a*nnotations *y*et *l*earning *a*daptively) overview. Passive + domain adaptation training.

Therefore, as a first proof of concept we also propose an unsupervised domain adaptation technique that avoids human intervention during the adaptation process. To the best of our knowledge, this Thesis work is the first demonstrating adaptation of virtual and real worlds for developing an object detector. Last but not least, we also assessed a different strategy to avoid the dataset shift that consists in collecting real-world samples and retrain with them in such a way that no bounding boxes of real-world pedestrians have to be provided. We show that the generated classifier is competitive with respect to the counter-

part trained with samples collected by manually annotating pedestrian bounding boxes. The results presented on this Thesis not only end with a proposal for adapting a virtual-world pedestrian detector to the real world, but also it goes further by pointing out a new methodology that would allow the system to adapt to different situations, which we hope will provide the foundations for future research in this unexplored area.

Publications

[1] David Vazquez. "Domain Adaptation of Virtual and Real Worlds for Pedestrian Detection"(Vol. 1). Ph.D. thesis, Ediciones Grficas Rey, Barcelona. 2013

[2] David Vazquez, Antonio Lopez, Daniel Ponsa, & David Geronimo. "Interactive Training of Human Detectors". In Multimodal Interaction in Image and Video Applications Intelligent Systems Reference Library (Vol. 48, pp. 169182), 2013.

[3] David Vazquez, Javier Marin, Antonio Lopez, Daniel Ponsa, & David Geronimo. "Virtual and Real World Adaptation for Pedestrian Detection". In IEEE T-PAMI, 2013.

[4] David Vazquez, Jiaolong Xu, Sebastian Ramos, Antonio Lopez, & Daniel Ponsa. "Weakly Supervised Automatic Annotation of Pedestrian Bounding Boxes" In CVPR Workshop on Ground Truth What is a good dataset?, 2013.

[5] Javier Marin, David Vazquez, Antonio Lopez, Jaume Amores, & Bastian Leibe. "Random Forests of Local Experts for Pedestrian Detection". In ICCV, 2013

[6] Javier Marin, David Vazquez, Antonio Lopez, Jaume Amores, & Ludmila I. Kuncheva. "Occlusion handling via random subspace classifiers for human detection". In IEEE Transactions on Systems, Man, and Cybernetics (Part B), 2013.

[7] Jiaolong Xu, David Vazquez, Antonio Lopez, Javier Marin, & Daniel Ponsa. "Learning a Multiview Part-based Model in Virtual World for Pedestrian Detection". In IEEE Intelligent Vehicles Symposium, 2013.

[8] Jiaolong Xu, David Vazquez, Sebastian Ramos, Antonio Lopez, & Daniel Ponsa. "Adapting a Pedestrian Detector by Boosting LDA Exemplar Classifiers". In CVPR Workshop on Ground Truth What is a good dataset?, 2013.

[9] Yainuvis Socarras, Sebastian Ramos, David Vazquez, Antonio Lopez, & Theo Gevers. "Adapting Pedestrian Detection from Synthetic to Far Infrared Images" In ICCV Workshop on Visual Domain Adaptation and Dataset Bias, 2013

David Vázquez received the B.Sc. degree in Computer Science from the Universitat Autònoma de Barcelona (UAB) in 2008. He received his M.Sc. in Computer Vision and Artificial Intelligence in 2009 and his Ph.D. degree in 2013 at the Computer Vision Center (CVC/UAB). He is currently a research scientist at CVC. His research interests include pedestrian detection, virtual worlds, domain adaptation and active learning.

Author Index