



INTELLIGENT SYSTEMS REFERENCE LIBRARY
Volume 48

Angel D. Sappa
Jordi Vitrià

Multimodal Interaction in Image and Video Applications

 Springer

Editors-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

Dr. Lakhmi C. Jain
Adjunct Professor
University of Canberra
ACT 2601
Australia

And

University of South Australia
Adelaide
South Australia SA 5095
Australia
E-mail: Lakhmi.jain@unisa.edu.au

Angel D. Sappa and Jordi Vitrià

Multimodal Interaction in Image and Video Applications

 Springer

Authors

Dr. Angel D. Sappa
Computer Vision Center
Edifici O Campus UAB
Barcelona
Spain

Dr. Jordi Vitrià
Departament de Matemàtica
Aplicada i Anàlisi
Facultat de Matemàtiques
Universitat de Barcelona
Barcelona
Spain

ISSN 1868-4394

ISBN 978-3-642-35931-6

DOI 10.1007/978-3-642-35932-3

Springer Heidelberg New York Dordrecht London

e-ISSN 1868-4408

e-ISBN 978-3-642-35932-3

Library of Congress Control Number: 2012955291

© Springer-Verlag Berlin Heidelberg 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

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

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Preface

Traditional Pattern Recognition (PR) and Computer Vision (CV) technologies have mainly focused on full automation, even though full automation often proves elusive or unnatural in many applications, where the technology is expected to assist rather than replace the human agents. However, not all the problems can be automatically solved being the human interaction the only way to tackle those applications.

Recently, multimodal human interaction has become an important field of increasing interest in the research community. Advanced man-machine interfaces with high cognitive capabilities are a hot research topic that aims at solving challenging problems in image and video applications. Actually, the idea of computer interactive systems was already proposed on the early stages of computer science. Nowadays, the ubiquity of image sensors together with the ever-increasing computing performance has open new and challenging opportunities for research in multimodal human interaction.

This book aims to show how existing PR and CV technologies can naturally evolve using this new paradigm. The chapters of this book show different successful case studies of multimodal interactive technologies for both image and video applications. These case studies were developed in the framework of the Spanish research program “Multimodal Interaction in Pattern Recognition and Computer Vision (Consolider Ingenio 2010)”. This program started in 2007 and last for 5 years involving more than 100 researchers from ten research institutions.

Each chapter discusses its theoretical foundations as well as practical questions related to the implementation issues. The book covers a wide spectrum of applications, ranging from interactive handwriting transcriptions to human-robot interactions in real environments.

To summarize, this book provides a coherent and well-founded description of practical systems that make use of pattern recognition and computer vision technologies for developing advanced multimodal interactive systems.

Barcelona,
November 2012

Angel D. Sappa
Jordi Vitrià

Acknowledgements

This book would not have been possible without the collaboration and support of many people. First and foremost, we would like to thank Prof. Enrique Vidal who devised the “*multimodality, interactivity and adaptability*” concepts all together under the same theoretical framework. These concepts were the pillars of the research program Multimodal Interaction in Pattern Recognition and Computer Vision (Consolider Ingenio 2010). Prof. Vidal motivated all the team members with the idea of doing research on these topics even a couple of years before the project started; after all these years, that idea becomes a reality that has been validated in different prototypes related to image and video applications. Great thanks also to all the researchers and engineers from the different institutions that contributed to the developments of prototypes and on-line demos. Finally, the editors want to thank the Spanish Government for the financial support through the Research Program Consolider Ingenio 2010: MIPRCV (CSD2007-00018) as well as national projects TIN2009-14404-C02 and TIN2011-25606



Contents

An Application for Efficient Error-Free Labeling of Medical Images	1
<i>Michal Drozdal, Santi Seguí, Petia Radeva, Carolina Malagelada, Fernando Azpiroz, Jordi Vitrià</i>	
1 Introduction	2
1.1 Motivation and Novelty	2
1.2 Real-Life Problem	2
1.3 Problem Set-Up	3
1.4 Intuition of Our Approach	4
1.5 Contributions	7
2 Related Work	7
3 Methodology	8
3.1 Algorithm for Data Structure Explortaion/Explotation . .	9
3.2 Displaying Strategy	11
3.3 Interface for Interactive Labeling of Endoscopic Frames	12
4 Experiments	12
5 Conclusions	14
References	15
Interactive Document Retrieval and Classification	17
<i>Ernest Valveny, Oriol Ramos, Joan Mas, Marçal Rossinyol</i>	
1 Introduction	17
2 Basic Document Classification and Retrieval	19
2.1 Document Classification	19
2.2 Logo Detection	21
3 Interactive Document Retrieval	23
3.1 Interactive Document Classification	23
3.2 Interactive Logo Detection	24
3.3 Interactive Class-Based Logo Detection	25
4 Prototype	26
4.1 Experiments	27
5 Conclusions	30
References	30

Interactive Visual and Semantic Image Retrieval 31
Joost van de Weijer, Fahad Khan, Marc Masana

- 1 Introduction 31
- 2 Interactive Visual and Semantic Image Retrieval 33
- 3 Image Representations 35
 - 3.1 Semantic Image Representation 35
 - 3.2 Visual Image Representation 38
- 4 Image Retrieval Application 39
 - 4.1 Technical Implementation 39
 - 4.2 User Interface 40
- 5 Demonstration and Experiment Results 41
 - 5.1 Semantic Image Description 42
 - 5.2 Interactive Visual and Semantic Retrieval 42
- 6 Conclusions 43
- References 44

Coloresia: An Interactive Colour Perception Device for the Visually Impaired 47
Abel Gonzalez, Robert Benavente, Olivier Penacchio, Javier Vazquez-Corral, Maria Vanrell, C. Alejandro Parraga

- 1 Introduction 47
 - 1.1 Colour Vision and Colour Visual Deficiencies 48
 - 1.2 Perceptual Interaction between Colour and Sound 50
- 2 State of the Art 51
- 3 From Colour Signal to Sound 52
 - 3.1 Properties of Colour 54
 - 3.2 Colour Constancy 56
 - 3.3 Properties of Sound 57
 - 3.4 Colour Sonification: Our Proposal 58
- 4 A Multimodal Device for the Visually Impaired 60
 - 4.1 Test and Results 63
- 5 Conclusions 64
- References 64

Interactive Pansharpener and Active Classification in Remote Sensing 67
Pablo Ruiz, Javier Mateos, Gustavo Camps-Valls, Rafael Molina, Aggelos K. Katsaggelos

- 1 Introduction 67
- 2 Interactive Pansharpener Based Classification 69
 - 2.1 Pansharpener 70
 - 2.2 Classification 71
 - 2.3 Interactive Pansharpener Based Classification 71
- 3 Active Learning 72
 - 3.1 Bayesian Modeling and Inference 72
 - 3.2 Classification 73

- 3.3 Active Learning Approaches 74
- 4 Prototypes Description 75
 - 4.1 Interactive Pansharpening Based Classification 75
 - 4.2 Bayesian Active Learning Remote Sensing 77
- 5 Conclusions 80
- References 80

Interactive Image Retrieval Based on Relevance Feedback 83

Mauricio Villegas, Luis A. Leiva, Roberto Paredes

- 1 Introduction 83
- 2 Methodology 84
 - 2.1 Relevance Feedback 84
 - 2.2 Dynamic Visual-Textual Fusion 86
 - 2.3 Query Refinement 89
 - 2.4 Tag Cloud 90
- 3 Prototype Implementation and Description 92
 - 3.1 System Architecture 93
 - 3.2 User Interface 93
 - 3.3 Web Server 95
 - 3.4 Image Database 95
 - 3.5 Indexed Database 98
 - 3.6 Hardware Configuration 99
- 4 Experiments and Results 99
 - 4.1 Relevance Feedback Evaluation 100
 - 4.2 Dynamic Linear Fusion Evaluation 101
 - 4.3 Query Refinement Evaluation 102
 - 4.4 Tag Cloud Evaluation 104
- 5 Conclusion 106
- References 107

An User-Driven Tool for Interactive Retrieval of Non Annotated Videos 111

M. Angeles Mendoza, Tomás Arnau, Isabel Gracia, Filiberto Pla, Nicolás Pérez de la Blanca

- 1 Introduction 111
- 2 Related Works 114
- 3 Prototype Functional Definition 115
 - 3.1 The Representation and Updating Model 116
 - 3.2 The Feedback 117
 - 3.3 The Algorithm 118
 - 3.4 User Interaction 119
- 4 Prototype Design 120
 - 4.1 Architecture 120
 - 4.2 Prototype GUI 121
- 5 Database and Experiments 123

- 5.1 The CCV Database 123
- 5.2 Video Preprocessing 124
- 5.3 Experiments 125
- 6 Discussion and Conclusion 131
- References 132

Exploiting Multimodal Interaction Techniques for Video-Surveillance .. 135

Marc Castelló, Jordi González, Ariel Amato, Pau Baiget, Carles Fernández, Josep M. Gonfaus, Ramón A. Mollineda, Marco Pedersoli, Nicolás Pérez de la Blanca, F. Xavier Roca

- 1 Introduction 135
- 2 Methodology and Theoretical Background 137
 - 2.1 Adaptation: Anomaly Detection Based on Trajectory Analysis 137
 - 2.2 Feedback: Interaction Based on Natural-Language Generation and Understanding 139
 - 2.3 Multimodality: Animation of Virtual Avatars for Communication with End-Users 142
- 3 The VID-HUM Demonstrator 144
- 4 Discussion 148
- 5 Conclusions 149
- References 149

Interactive Video Surveillance for Perimeter Control 153

Javier Ortells, Henry Anaya-Sánchez, Raúl Martín-Félez, Ramón A. Mollineda

- 1 Introduction 153
- 2 Related Work 154
- 3 Interactive Learning Strategy 156
- 4 Prototype Description 158
 - 4.1 Functional Scope 158
 - 4.2 System Architecture 158
 - 4.3 Hardware and Software Resources 164
- 5 Conclusions and Future Work 165
- References 166

Interactive Training of Human Detectors 169

David Vázquez, Antonio M. López, Daniel Ponsa, David Gerónimo

- 1 Introduction 169
- 2 HaarEOH-Based Pedestrian Detection 172
 - 2.1 Detection Architecture 172
 - 2.2 Datasets: Real- and Virtual-World Samples 175
 - 2.3 Training with Virtual- vs Real-World Samples 175
 - 2.4 Testing with Real-World Images: Dataset Shift 176
- 3 Semi-supervised Domain Adaptation 177
- 4 Experimental Results 179

5	Conclusion	182
	References	182
	Robot Interactive Learning through Human Assistance	185
	<i>Gonzalo Ferrer, Anaís Garrell, Michael Villamizar, Iván Huerta, Alberto Sanfeliu</i>	
1	Introduction	185
2	Interactive Motion Learning for Robot Companion	187
2.1	People Detection and Tracking	188
2.2	Human Motion Prediction and the Social-Force Model Applied to Robot Companion	189
2.3	Interactive Robot Motion Learning	191
2.4	Experimental Results	193
3	Autonomous Mobile Robot Seeking Interaction for Human-Assisted Learning	193
3.1	Robot’s Proactively Seeking Interaction	194
3.2	On-Line Face Learning Approach	196
3.3	Experiments	199
4	Conclusions	200
	References	201

An Application for Efficient Error-Free Labeling of Medical Images

Michal Drozdal, Santi Seguí, Petia Radeva, Carolina Malagelada,
Fernando Azpiroz, and Jordi Vitrià

Abstract. In this chapter we describe an application for efficient error-free labeling of medical images. In this scenario, the compilation of a complete training set for building a realistic model of a given class of samples is not an easy task, making the process tedious and time consuming. For this reason, there is a need for interactive labeling applications that minimize the effort of the user while providing error-free labeling. We propose a new algorithm that is based on data similarity in feature space. This method actively *explores* data in order to find the best label-aligned clustering and *exploits* it to reduce the labeler effort, that is measured by the number of “clicks. Moreover, error-free labeling is guaranteed by the fact that all data and their labels proposals are visually revised by an expert.

Michal Drozdal · Petia Radeva · Jordi Vitrià
Dept. Matemàtica Aplicada i Anàlisi,
Universitat de Barcelona,
Barcelona, Spain
Computer Vision Center (CVC),
Universitat Autònoma de Barcelona,
Barcelona, Spain
e-mail: {michal.drozdal,petia.ivanova,jordi.vitria}@ub.edu

Santi Seguí
Computer Vision Center (CVC),
Universitat Autònoma de Barcelona,
Barcelona, Spain
Dept. Matemàtica Aplicada i Anàlisi,
Universitat de Barcelona,
Barcelona, Spain
e-mail: ssegui@cvc.uab.es

Carolina Malagelada · Fernando Azpiroz
Hospital de Vall d’Hebron,
Barcelona, Spain

1 Introduction

1.1 Motivation and Novelty

For many machine learning applications, the compilation of a complete training set for building a realistic model of a given class of samples is not an easy task. The motto "there is no data like more data" suggests that the best strategy for building this training set is to collect as much data as possible. But in some cases, when dealing with problems of high complexity and variability, the size of this data set can grow very rapidly, making the learning process tedious and time consuming. Time is expended in two different processes: 1) the labeling process, which generally needs human intervention, and 2) the training process, which in some cases exponentially increments computational resources as more data is obtained.

Labeling is a human activity domain¹, in many learning-based tasks an expert knowledge is needed to label the data samples and expert time and effort are expensive. Moreover, when spending long time on data labeling an oracle/expert gets tired and errors can be introduced thus as a result the labeling can be inconsistent. In the literature the problem of effort minimization in data labeling was addressed by using active learning techniques [19]. It is important to note that in active learning the main focus is put on classification of the data samples. In this sense, the learner collects the labels that are best from point of view of the future data generalization problem.

However, there are learning-based applications where all training data must be checked by the oracle to ensure the correctness of the labeling. In this case the main motivation is to minimize the oracle's effort. This is not a case of direct application of active learning techniques, as they are focused on maximizing classifier performance in test set and not on minimizing the oracle effort when labeling the whole training set. So, active learning techniques are not designed for efficient error-free labeling² and thus a need for interactive labeling applications appear that will minimize the effort of the user while providing error-free labeling. An example of application where the correctness of the labeling is of high importance are medical data, where no error can be committed in the training set construction in order to ensure an optimally correct labeling in training process of designed computer-aided classification systems.

1.2 Real-Life Problem

Wireless Capsule Endoscopy (WCE) image analysis (see Fig. 1(a)) is a clear scenario of medical imaging device where it is highly desirable to construct a labeling

¹ Here it is assumed that humans are highly accurate while labeling the data.

² The term error-free labeling refers to the fact that all data and its labels proposals are revised by an expert, it is a difference with Active Learning where only some samples are being displayed to the oracle. Clearly it is possible that an expert will miss-label some data according to the criteria of different expert.

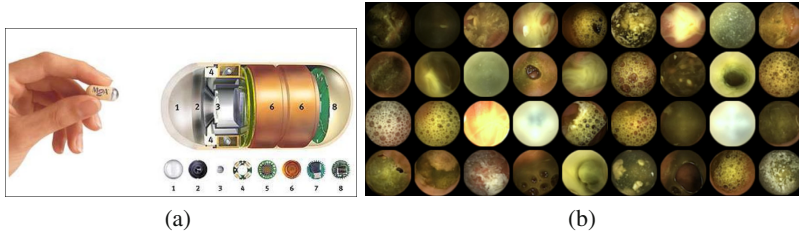


Fig. 1 (a) The wireless video capsule; (b) Non informative frames

error-free training set to learn to classify frames of the endoscopic video. The WCE contains a camera and a full electronic set which allows the radio frequency emission of a video movie in real time. This video, showing the whole trip of the pill along the intestinal tract, is stored into an external device which is carried by the patient. These videos can have duration from 1h to 8h, what means that the capsule captures a total of 7.200 to 60.000 images. WCE videos have been used in the framework of computer-aided systems to differentiate diverse parts of the intestinal tract [11], to measure several intestinal disfunctions [22] and to detect different organic lesions (such as polyps [13], bleeding [12] or general pathologies [4]). In most of these applications, machine learning plays a central role to define robust methods for frame classification and pattern detection.

A common stage in all these research lines is the discrimination of informative frames from non-informative frames. Non-informative frames are defined as frames where the field of view is occluded. Mainly, the occlusion is caused by the presence of intestinal content, such as food in digestion, intestinal juices or bubbles (see Fig. 1(b)). The ability of finding non-informative frames is important since: 1) generally, it helps to reduce time of video analysis, and 2) the majority of non-informative frames are frames with intestinal content that is an indicator for intestinal disfunctions [16, 15].

The main strategy in the search of non-informative frames is the application of machine learning techniques in order to build a two-class classifier. Generally, non-informative frames are characterized by their color information [1]. Robust classifiers can be built when the training set is representative of the data population. The problem that occurs when dealing with WCE videos is the high color variability of non-informative frames in different videos. Therefore, to overcome this problem, we need to construct a wide set of labeled training set of informative and non-informative frames. Here, a naive approach for the labeling process of these video frames could mean the manual annotation of thousands of video frames.

1.3 Problem Set-Up

The real life example motivates the need for efficient real error-free labeling. In order to ensure that all data are correctly labeled the oracle has to revise visually all

the elements in the data set. A new element will be labeled only if it has been seen by the oracle. The effort of the oracle can be highly minimized by convenient organization of the data before displaying them. First, the system has to come up with a proposal of the label. This proposal is based on the knowledge gained by the system during the labeling process. Then, the human operator faces two possible decisions: to accept the model proposal or to change the label of the sample. In practice, these two options have a non symmetric cost for the human operator: accepting the model proposal can be efficiently implemented with a low cognitive load for the operator assuming it by default, while changing a label has a larger cost consisting in manual intervention while labelling the frame. This cost can be evaluated by the number of its interventions with the system during the labeling process [7]. As a result the oracle has seen all the data samples and has confirmed or has changed the system proposed label so it can be assumed that all labels are correctly assigned. Moreover, in the algorithm discussed in this chapter it is assumed that 1) all data come in a single batch, 2) data distribution is known before starting the labeling process and 3) data can be of more than two classes (possible labels).

1.4 Intuition of Our Approach

The system that minimizes the oracle's effort should address two issues: 1) how the data should be organized while being displayed and 2) which rule the system should follow when giving the proposition of the label.

A natural way to display the data is driven by data similarity and not by randomness (see Fig. 2(a)). It is convenient to display similar samples together. If data was obtained by sequential, in time, process it can be assumed that data acquired in instance i and $i + 1$ are similar (see Fig. 2(b)). In case when the data have not been sampled in sequential process or when the samples come from highly dynamic events the similarity can be defined in some feature space. Other possibility is to group similar images using distances between data in some feature space into cluster-structure (see Fig. 2(c) for an example of data grouping in color histogram feature space). The assumption done here is that, in some well defined feature space, it is more probable that the similar samples share same labels than the samples far away in the feature space.

The question is how to find the optimal labeled-aligned cluster structure (see Fig. 3). If we knew all the labels, our problem would become trivial (see Fig. 3(a)): we could present the data to the labeler in an optimal cluster-based organization to minimize the effort. For example: if a cluster is pure (only one class), we can label it (all elements inside the cluster) using only one intervention (e. g. one click). But at the beginning of the labeling process, the labels are unknown (see Fig. 3(b))! This problem, of finding optimal label-aligned cluster structure, can be seen as joint cluster *exploration* and *exploitation* task. *Exploration* is responsible for a discovery of data structure while *exploitation* is in charge of finding an optimal discrimination between classes in the data structure [10].

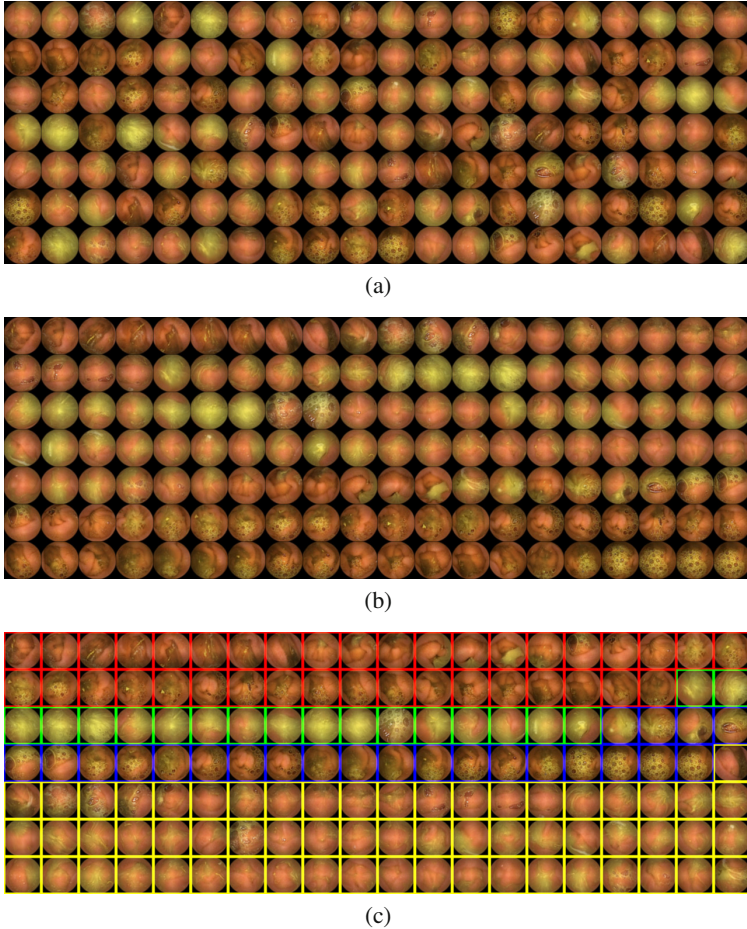


Fig. 2 An example of image ordering in case of Wireless Capsule Endoscopy data. Two class classification problem: informative vs. non-informative frames. a) random order, b) sequential in time, c) order according to the frames similarity with respect to the color features (color mark indicates different clusters obtained with k-means algorithm - similar images).

Recently, a novel approach to active label-aligned³ cluster discovery that does not require any classifier training step has been proposed by Dasgupta and Hsu in [5]. This algorithm, called hierarchical sampling, is a specific case of a more general paradigm called *partition-based sampling* (in statistics also referred as *cluster-based sampling*), characterized by the use of a data structure that represents all possible data clusters or partitions. These methods are aimed to first explore and then to

³ We will also use the notation of pure cluster(s) referring to well aligned cluster(s) to its label(s).

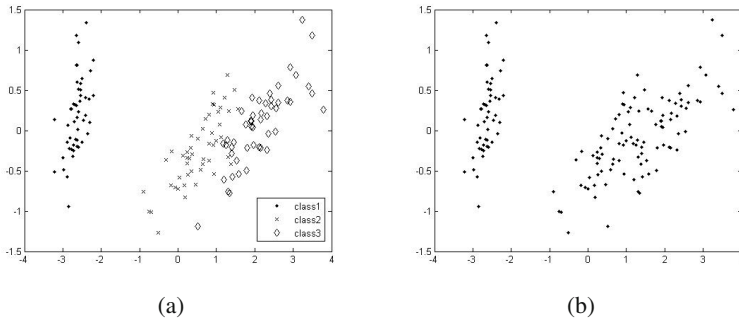


Fig. 3 Some example data distribution in 2D (data from IRIS dataset after applying PCA). a) data with all labels uncovered, b) data with unknown labels.

exploit the cluster structure in data. This class of methods, takes advantage of data-structure and aims in active search for an optimal label-aligned partition of the data. Under this assumption, the most critical part is how to find pure clusters as fast as possible by efficiently exploring the data set.

In [5] the chosen data structure is a hierarchical clustering tree based on the Ward's method [23]. This tree, which is computed offline, is a static data structure that is used to guide the active sampling strategy by navigating among all possible prunings of the tree. It has been shown that this active sampling strategy for pure cluster discovery is clearly better than the random sampling strategy in the presence of label-aligned data clusters, and it is not worse (except maybe for some pathological cases) than random sampling when these structures are not present in data.

One of the main advantages of the hierarchical sampling [5] method for active label-aligned data structure discovery is that it is not based on retraining multiple times a classifier (in contrast to many active sampling strategies), but in a sampling process that makes only one assumption on data distribution, that the data are distance-clusterizable in some distance space. The classifier is trained once the exploration of the data structure has produced fairly pure clusters that can be exploited in labeling process resulting in a satisfactory solution.

In the literature there is a vast variety of methods for label proposals where the criteria for choosing the proposal is classifier dependent [19]. Moreover, in general, these strategies are not designed to take advantage of data cluster structure. A simple way to come up with the label proposal in label-aligned cluster structure is the majority label. For each cluster, the system can estimate the majority label using the labels seen so far and treat this label as the most probable label for all the elements inside the cluster. It has been shown in [5] that the cluster-adaptive sampling strategy has theoretical bounds on the empirical estimation of the majority label. These bounds give, at any time, the interval in which the true labeling error lies and the bounds holds with high probability. This property of the sampling algorithm allows for labeling complexity analysis, indicating the number of queries to be seen

in order to obtain a given labeling accuracy. The ability for complexity analysis is important since the information about labeling error is a direct estimation of the expected effort needed for labeling of the whole data set.

1.5 Contributions

In this chapter, an application for efficient error-free labeling is presented. The method minimizes the oracle's effort, by using a strategy that minimizes the number of oracle's intervention in the labeling process. This minimization holds when the data can be represented by some label-aligned structure. The algorithm comes with label proposition and the oracle is asked to confirm/change the label. What differs our approach from more classical one, based on active learning, is the necessity of error-free labeling, needed for medical applications, what means that the oracle has to visually revise all the data and their labels. This strategy is build to minimize the number of oracle's interventions.

The chapter is organized as follows, in the next section we reveal some related works. In section 2 we discuss the sampling algorithm and the strategy of showing the images to the oracle. Section 4 presents some experimental results and finally, section 5 concludes.

2 Related Work

The majority of the work that is similar in spirit to our set-up are based around active sampling techniques for active learning. As said before, the main difference is the final objective of each strategy, for active learning it is to get the optimal performance of some classifier with low number of labeled samples. In our case the goal is to label all samples with minimum oracle's effort. Thus, in our case, all samples have to be revised and the goal of the system is two-fold: 1) the data should be displayed in the way that improves oracle performance and 2) the data should come-up with some label proposition. Anyway it is worth revising the literature in the field of active sampling for active learning while it is conceptually similar to our approach. Both are trying to find the data samples that are the most "favorable" to be displayed and manually labeled.

The active learning strategies vary in the selection strategy in sampling process. For example, *density sampling* methods sample from maximal-density unlabeled regions [17, 18]. On the other hand, *uncertainty sampling* methods sample the regions where the trained classifier is least certain [21]. The combination of density and uncertainty criteria has also been explored and it is called *representative sampling*. This approach explores the clustering structure of uncertain samples for selecting the most representative ones [24]. Using a different heuristic, *instability sampling* approaches are based on sampling from those regions that maximally change the classifier decision boundary [8].

In order to increase the efficiency of these sampling strategies, several research lines have been proposed. One of the most successful lines is to take benefit from the use of ensemble learning methods [14]. For example, *Query-by-committee* [9] selects samples that cause maximum disagreement amongst an ensemble of hypotheses. Another interesting line has been the development of hybrid methods, such as the method presented in [6], where the parameters selection strategy can be adaptively updated after each actively sampled point.

There are works that are not based implicitly on active learning but on concept of interactive labeling with the objective on efficient data labeling. Usage of Self-Organizing Maps (SOMs) to group similar elements in data set and to display them jointly to the user is investigated in [2]. In [25] the authors analyse the use of Bayesian networks to produce refined labeling. There is also a group of works that are based on refining the initial clustering structure with the help of user interaction [20, 3].

3 Methodology

In order to build an application for efficient label-free labeling, the following elements should be considered:

1. An engine responsible for *exploration* and *exploitation* of the data structure (e. g. partition space).
2. A strategy for choosing the elements to be displayed, for both, individual data elements due to structure *exploration* step and a group of elements representing some part of the data structure to enable to the user the possibility to exploit the data structure.
3. An interface to show the data according to displaying strategy and to accept user interactions with the system.

Fig. 4 shows the application flowchart. The core of our approach is based on the Dasgupta’s algorithm [5] that is presented in Section 3.1. Section 3.2 presents how to adapt the algorithm output to display to the expert/oracle. In Section 3.3 the interface is presented.

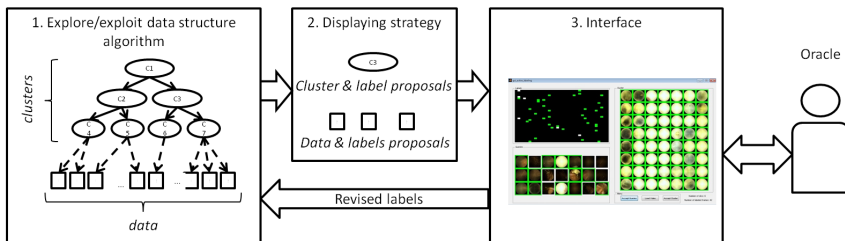


Fig. 4 Application flowchart

3.1 Algorithm for Data Structure Explortation/Explotation

We propose to base our approach on a modified formulation of *partition based active learning* [5]. The paradigm can be described in its most general terms in the following way: Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be the set of unlabeled data, \mathbf{C} be a possible partition set of \mathbf{X} and C be an element of this partition such that:

$$\bigcup_{C_k \in \mathbf{C}} C_k = \mathbf{X}$$

The objective under this paradigm is to *efficiently search through the space of partitions* for a partition \mathbf{C}^{opt} of \mathbf{X} such that all data points belonging to an element $C_k \in \mathbf{C}^{opt}$ can be automatically labeled with high confidence by using small number of labeled data. Its basic steps are represented by Algorithm 1.

The algorithm has four associated procedures that need some explanation: `Select`, `Display samples`, `Bound` and `Search`.

The procedure `Bound` estimates the mislabeling error within a given element C of the partition given the samples seen so far. An elegant way of implementing this procedure is to use the bounds derived from the tails of the binomial or multinomial distribution [5].

Suppose there are η possible labels and that their proportions in a given element C of the partition are $p_{C,l}$ for $l = 1, \dots, \eta$. Then, the error induced by assigning all points in C to the majority label is $\epsilon_C = 1 - \max_l(p_{C,l})$. At any given iteration k of the algorithm, we can associate with each element C of the partition \mathbf{C}^k a confidence interval within which we expect the true probability $p_{C,l}$ to lie: $[p_{C,l}^{LB}, p_{C,l}^{UB}] = [\max(p_{C,l} - \frac{1}{m} - \sqrt{\frac{p_{C,l}(1-p_{C,l})}{m}}, 0), \min(p_{C,l} + \frac{1}{m} + \sqrt{\frac{p_{C,l}(1-p_{C,l})}{m}}, 1)]$, where m is the number of points sampled from C up to that moment. In this way, we have defined a statistical measure about the error introduced if we label automatically the samples of an element C after seeing m labels. The conservative estimate of this error is defined as `Bound`(C) = $1 - p_{C,l}^{LB}$ [5].

The procedure `Select` determines which elements of the partition are sampled. There are several alternatives to implement this procedure, but the most evident choice, taking into account the objective of minimizing the number of queries, is to focus the selection process on those regions of the space that are still under-sampled. A simple implementation of this idea is to choose an element C with probability proportional to $\omega_C \text{Bound}(C)$, where ω_C is the fraction of the unrevised dataset covered by element C (proportional to the number of data points inside an element C). This sampling rule from one side, will reduce querying in elements C of the current partition \mathbf{C} that has a low `Bound`(C), thus reducing labeling efforts in those regions that can be automatically labeled with high confidence. From the other side, the factor ω_C permits to explore the large and fairly pure elements of the data structure where small “missing-clusters” can be hidden.

The most critical part of the algorithm is the definition of a searching strategy for finding \mathbf{C}^{opt} . After seeing sufficient number of samples, the algorithm must be able to generate a new, probably better partition \mathbf{C}^{k+1} from \mathbf{C}^k . In this framework,

Algorithm 1. *Exploration/exploitation algorithm*

Require: A data structure \mathbf{C} to represent any partition of \mathbf{X} .

Require: A number s representing the number of samples to be queried at each algorithm step.

```

1:  $\mathbf{C}^0 \leftarrow \mathbf{X}$ 
2:  $\text{Bound}(\mathbf{C}^0)$ 
3:  $\hat{\mathbf{L}}(\mathbf{X}) \leftarrow 1$  {Arbitrary label for label proposal list.}
4:  $\mathbf{L}(\mathbf{X}) \leftarrow \text{empty}$  {Revised labels list.}
5:  $i \leftarrow 0$ 
6:  $k \leftarrow 0$ 
7: while unseen labels do
8:    $j \leftarrow 0$ 
9:    $\mathbf{x} \leftarrow \text{empty}$  {List of data points to query.}
10:  while  $j < s$  do
11:    Select an element  $C$  from  $\mathbf{C}^k$ .
12:    Sample a random point  $p$  from  $C$ .
13:     $\mathbf{x} \leftarrow \mathbf{x} \cup p$ 
14:     $j \leftarrow j + 1$ 
15:  end while
16:  Find the purest cluster  $C^p$  in  $\mathbf{C}^k$ 
17:  Display samples with label proposals  $\hat{\mathbf{L}}(\{\mathbf{x}, C^p\})$  and get true labels in case of
  exploration  $\mathbf{L}(\mathbf{x})$  xor exploitation  $\mathbf{L}(C^p)$ .
18:   $i \leftarrow i + s$ 
19:  if explore then
20:    for each  $C \in \mathbf{C}^k$  do
21:      Compute  $\text{Bound}(C)$ , a conservative estimate of the mislabeling error within  $C$ .
22:      Update label proposals  $\hat{\mathbf{L}}(C)$ 
23:    end for
24:    Search for a new better partition  $\mathbf{C}^{k+1}$ 
25:  else if exploit then
26:     $\mathbf{C}^{k+1} \leftarrow \mathbf{C}^k / C^p$ 
27:  end if
28:   $k \leftarrow k + 1$ 
29: end while
30: return the full labeled set  $\{\mathbf{X}, \mathbf{L}\}$  to train a classifier.

```

we can define the following order relation between partitions: A partition \mathbf{C}^{k+1} is better than a partition \mathbf{C}^k if $\sum_{C \in \mathbf{C}^{k+1}} \omega_C \text{Bound}(C) < \sum_{C \in \mathbf{C}^k} \omega_C \text{Bound}(C)$, that is, if the fraction of the dataset that can be automatically labeled with confidence in \mathbf{C}^{k+1} is larger than the one in \mathbf{C}^k .

A simple but efficient alternative to limit the size of the search space over possible data partitions when implementing *exploration/exploitation* algorithm, used in [5], is to consider a reduced partition space: instead of considering all possible partitions⁴ of \mathbf{X} , the authors consider only the subspace formed by the partitions defined by a given pruning of the tree representing a hierarchical clustering of \mathbf{X} .

⁴ The number of ways a set of n elements can be partitioned into nonempty subsets is called a Bell number and is denoted by B_n . For example, $B_{10} = 115975$.

In this case, the navigation strategy can be also simplified and the `Search` procedure in Algorithm 2 consists of selecting a good pruning of the pre-calculated hierarchical clustering tree at each active learning step (see [5] for more details).

The procedure `Display samples` sends the data samples with labels proposals to the display and waits for the user interaction. After executing this procedure, the algorithm receives a group of true labels that can be used either for structure exploitation either for exploration step. The details on displaying strategies are discussed in section 3.2

3.2 Displaying Strategy

At each step of the *exploration/exploitation* of data structure step, the algorithm produces three outcomes: 1) the samples from impure clusters with the label proposals $\{\mathbf{x}, \hat{\mathbf{L}}\}$, 2) the current clustering structure grouping the similar data samples and their labels proposals $\{\mathbf{C}, \hat{\mathbf{L}}\}$ and 3) the purity measure of each cluster $\{\mathbf{C}, \text{Bound}(\mathbf{C})\}$. The question that arises is how to present all this information to the user (see Fig. 5)?

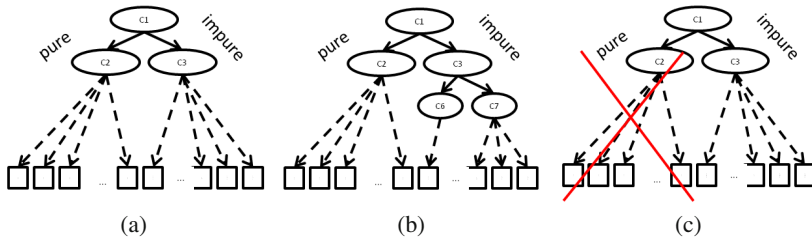


Fig. 5 An illustration of possible actions on data structure. a) hypothetical data structure with one pure and one impure cluster, b) *exploration* - user decides to label samples from impure cluster, as an effect the algorithm descends in hierarchical structure and uncover new clusters, and c) *exploitation* - user decides to revise the labels in pure cluster, and as an effect all data from this cluster are labeled and no further samples will be displayed in sampling process.

It is straight-forward to present to the oracle the samples from impure clusters $\{\mathbf{x}, \hat{\mathbf{L}}\}$ while it is a necessity in data structure exploration step (see Fig. 5(b)). If the user decides to label those samples, the algorithm will learn about true labels $\{\mathbf{x}, \mathbf{L}\}$. This results in dividing the current cluster structure into more refined one (data structure exploration).

When it comes to the clusters it is infeasible to display all the partitions at once. In this case two different criteria can be applied: one that maximizes the information gain or other that minimizes the oracle's effort. The first one means to display the most impure cluster from the current data structure and to ask the user to correct the labels. But this displaying strategy is contrary to the problem set-up while it is not minimizing the oracles effort. Moreover, the algorithm already provides to the oracle some samples from impure clusters due to data structure exploration step.

The second approach is to present to the user the purest cluster and its labels proposals $\hat{L}(C^p)$ and asking him/her to correct, hopefully, a few samples and accept the whole cluster. Once the oracle accepts the cluster, all the samples have been assigned a correct label and this part of the space will no longer be sampled by the algorithm. If the data can be represented by the label-aligned data structure with large clusters in this step with a few (or non) oracle's interventions, a whole cluster can be labeled (see Fig. 5(c)). Moreover, if the data can be divided in a few such clusters, the labeling process will be very efficient and cheap in oracle's effort. This is a step of the algorithm that exploits the current data structure.

To be able to jointly *explore/exploit* data structure and minimize the oracle's effort, both data, 1) samples from impure clusters and 2) the purest cluster, should be presented in parallel to the user. In this set-up, the user decides whenever it is convenient to exploit the current data structure or to continue exploring to get finer cluster - label alignment.

3.3 Interface for Interactive Labeling of Endoscopic Frames

The interface is shown on Fig. 6. The interface is composed of 3 fields, one to display the data's labels that have been revised by the oracle. One for displaying the samples from impure clusters, and one to display the purest cluster. The user can choose in which field he/she is willing to interact. If the purest cluster is homogeneous, it is favorable to change a few (or none) labels and to accept a large number of labeled samples. Otherwise the oracle should interact in the field of samples from impure clusters and wait for a pure cluster to appear. Once the oracle has revised the labels in the field of samples from impure clusters (or in the purest cluster field) he/she should press the button accept queries (or cluster) and the algorithm will come-up with new data and their proposals.

With respect to the interface two questions remain open:

- Number of images to be displayed in each field.
- Optimal resolution of the image.

These questions are not treated in this chapter while the answers should be adjusted individually to the data set that is being labeled and to the screen's resolution. The general remark is that the parts of the image that are being subject to labeling should be well visible to the user. The user should not spend too much time on visual inspection of a single image and should be able to quickly spot the discriminative (in context of labeling) parts of the image like: color, structure, shape etc. .

4 Experiments

The purpose of the experiments section is to evaluate three possible scenarios:

1. *Random order* - a label has to be provided for each frame individually, the number of oracle's interventions is proportional to the number of samples.

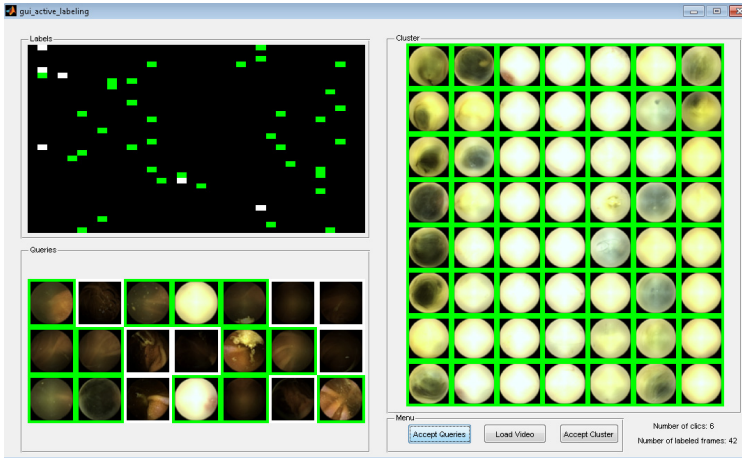


Fig. 6 An example of interface, used to display images during labeling process of Wireless Capsule Endoscopy video. Left-top assigned labels (green intestinal content frames, white clear frames). Left-down file displaying samples from all clusters with the label proposal. Right file showing the purest cluster with label proposal.

2. *Sequential (in time) order* - the label is activated and last until it is changed, the number of oracle's interventions depends on the dynamics of the process that is being observed, if the process is slow in time the number of interventions is proportional to the number of classes in the data, if the process is highly dynamic the number of interventions is proportional to the number of samples.
3. *Proximity in feature space (hierarchical clustering)* - the data are organized into clusters in the feature space, the number of oracle's intervention depends on the cluster structure, if some large fairly-pure clusters are present, the number of interventions is proportional to the number of clusters. If the data can not be organized into fairly-pure label-aligned clusters, the number of interventions is proportional to the number of samples.

To evaluate the scenarios the data from informative vs. non-informative frame problem of Wireless Capsule Endoscopy is used. The scenarios 2 and 3 are further analyzed, in case of scenario 1 it is assumed that the number of oracle's intervention is equal to the number of samples.

In the evaluation one WCE video of 55156 frames was sub-sampled every 50th frame producing a string of images of length 1104 frames. First, the video was presented to the expert according to the Scenario 2 asking her/him to label the informative and non-informative frames. Each sample was displayed in a sequential

order with a label proposal, if the proposal was incorrect the user could change the label. In this scenario of reviewing and labeling of all the frames, the oracle needed 57 clicks.

Second, the data was presented to the same user using the application presented in Section 2 asking her/him to label the informative and non-informative frames. In the field of samples from impure clusters, 27 frames were displayed. In this scenario the user needed 45 interventions to revise and label 1104 frames of WCE.

In order to fully appreciate the utility of the application, we tested the proposed labeling scheme in the task of face database creation, where we labeled examples of facial and not-facial images. First, the face hypothesis was generated using the Viola-Jones classifier containing true face detections and some hard false positives cases. The goal was to separate the true from the false detections. In order to do so, each face detection image was represented by using the Histogram of Gradients (HoG) feature. The structure representing the image similarity was calculated in the HoG feature space. In the experiments we used 1064 images (of faces and no-faces). At the beginning, all images were assigned to face class. Using our approach the user was able to review all labels and get perfect labeling with 87 clicks.

5 Conclusions

In this chapter, a new application for error-free labeling with the user in the loop has been presented. The application is based on data similarity in feature space. This method actively *explores* data in order to find the best label-aligned clustering and *exploits* it to reduce the oracle's effort. At each step of the method, the oracle can decide whether it is more convenient to go for data exploitation (the displayed cluster is fairly pure) or for further data structure exploration. The algorithm for each data sample presents a label proposal, based on majority label estimation in current cluster. The error-free labeling is guaranteed by the fact that all data and their labels proposals are visually revised by an expert. Thanks to the clustering structure this revision can be done in an efficient way reducing significantly the time of constructing a wide set of training samples.

This strategy has been compared to the sequential (in time) ordering of the data. This strategy should be used when the data come from steady (or even static) process. On the other hand, the strategy based on proximity in feature space is favorable for the data where some large fairly-pure clusters are expected to be found.

Acknowledgments. This work was supported in part by a research grant from the MICINN Grants TIN2009-14404-C02 and CONSOLIDER-INGENIO 2010 (CSD2007-00018).

References

1. Bashar, M.K., et al.: Automatic detection of informative frames from wireless capsule endoscopy images. *Medical Image Analysis* 14(3), 449–470 (2010)
2. Bekel, H., Heidemann, G., Ritter, H.: Interactive image data labeling using self-organizing maps in an augmented reality scenario. *Neural Networks* 18(5-6), 566–574 (2005)
3. Biswas, A., Jacobs, D.: Active image clustering: Seeking constraints from humans to complement algorithms. In: *CVPR* (2012)
4. Coimbra, M.T., Cunha, J.P.S.: MPEG-7 visual descriptors: Contributions for automated feature extraction in capsule endoscopy. *IEEE TCSVT* 16(5), 628–637 (2006)
5. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 208–215. ACM (2008)
6. Donmez, P., Carbonell, J.G., Bennett, P.N.: Dual Strategy Active Learning. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) *ECML 2007. LNCS (LNAI)*, vol. 4701, pp. 116–127. Springer, Heidelberg (2007)
7. Drozdal, M., Seguí, S., Malagelada, C., Azpiroz, F., Vitrià, J., Radeva, P.: Interactive Labeling of WCE Images. In: Vitrià, J., Sanches, J.M., Hernández, M. (eds.) *IbPRIA 2011. LNCS*, vol. 6669, pp. 143–150. Springer, Heidelberg (2011)
8. Dwyer, K., Holte, R.: Decision Tree Instability and Active Learning. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) *ECML 2007. LNCS (LNAI)*, vol. 4701, pp. 128–139. Springer, Heidelberg (2007)
9. Freund, Y., Sebastian Seung, H., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. *Mach. Learn.* 28, 133–168 (1997)
10. Hospedales, T.M., Gong, S., Xiang, T.: Finding rare classes: Active learning with generative and discriminative models. *IEEE Transactions on Knowledge and Data Engineering* 99(PrePrints) (2011)
11. Igual, L., et al.: Automatic discrimination of duodenum in wireless capsule video endoscopy. *IFMBE Proceedings* 22, 1536–1539 (2008)
12. Jung, Y.S., et al.: Active blood detection in a high resolution capsule endoscopy using color spectrum transformation. In: *ICBEI*, pp. 859–862 (2008)
13. Kang, J., Doraiswami, R.: Real-time image processing system for endoscopic applications. In: *IEEE CCECE 2003*, vol. 3, pp. 1469–1472 (2003)
14. Kuncheva, L.I.: *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience (2004)
15. Malagelada, C., De Iorio, F., Seguí, S., Mendez, S., Drozdal, M., Vitria, J., Radeva, P., Santos, J., Accarino, A., Malagelada, J.R., Azpiroz, F.: Functional gut disorders or disordered gut function? small bowel dysmotility evidenced by an original technique. *Neurogastroenterology and Motility* 24(3), 223-e105 (2012)
16. Malagelada, C., et al.: New insight into intestinal motor function via noninvasive endoluminal image analysis. *Gastroenterology* 135(4), 1155–1162 (2008)
17. McCallum, A., Nigam, K.: Employing em and pool-based active learning for text classification. In: *Proceedings of the Fifteenth International Conference on Machine Learning, ICML 1998*, pp. 350–358. Morgan Kaufmann Publishers Inc., San Francisco (1998)
18. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: *Proceedings of the Twenty-First International Conference on Machine Learning, ICML 2004*, p. 79. ACM, New York (2004)
19. Settles, B.: *Active learning literature survey*. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)

20. Tian, Y.: A face annotation framework with partial clustering and interactive labeling. In: International Conf. on Computer Vision and Pattern Recognition, p. 7 (2007)
21. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* 2, 45–66 (2002)
22. Vilarino, F., et al.: Intestinal motility assessment with video capsule endoscopy: Automatic annotation of phasic intestinal contractions. *IEEE TMI* 29(2), 246–259 (2010)
23. Ward Jr., J.H.: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58(301), 236–244 (1963)
24. Xu, Z., Yu, G., Tresp, V., Xu, X., Wang, J.: Representative Sampling for Text Classification Using Support Vector Machines. In: Sebastiani, F. (ed.) *ECIR 2003*. LNCS, vol. 2633, pp. 393–407. Springer, Heidelberg (2003)
25. Zhang, L., Tong, Y., Ji, Q.: Active Image Labeling and Its Application to Facial Action Labeling. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part II*. LNCS, vol. 5303, pp. 706–719. Springer, Heidelberg (2008)

Interactive Document Retrieval and Classification

Ernest Valveny, Oriol Ramos, Joan Mas, and Marçal Rossinyol

Abstract. In this chapter we describe a system for document retrieval and classification following the interactive-predictive framework. In particular, the system addresses two different scenarios of document analysis: document classification based on visual appearance and logo detection. These two classical problems of document analysis are formulated following the interactive-predictive model, taking the user interaction into account to make easier the process of annotating and labelling the documents. A system implementing this model in a real scenario is presented and analyzed. This system also takes advantage of active learning techniques to speed up the task of labelling the documents.

1 Introduction

Huge amounts of documents are being stored currently as digital images at private and public organizations. However, for these raw digital images to be really useful, they need to be annotated with informative content. Document Image Analysis and Pattern Recognition techniques are at the heart of current solutions to this problem. However, when dealing with difficult unconstrained documents (see figure 1), standard solutions (for instance, commercial OCR products) are simply not usable since, in the vast majority of these documents, elements can by no means be isolated automatically. Given the high error rates involved in post-editing solutions, only semi-automatic or computer-assisted alternatives can be currently foreseen.

In this context, interactive tools emerge as a very appealing alternative to reduce the cost of labelling and annotating documents and, at the same time as a way of obtaining user feedback to improve the model for classification and retrieval.

Ernest Valveny · Oriol Ramos · Joan Mas · Marçal Rossinyol
Computer Vision Center / Computer Science Dept.,
Universitat Autònoma de Barcelona,
Building O, Campus UAB, 08193 Bellaterra (Spain)
e-mail: {ernest, oriolrt, jmas, marcal}@cvc.uab.es



Fig. 1 Examples of unconstrained documents

Hence, in this chapter we describe an interactive tool to annotate documents with semantic information, such as the category of the document or the location of relevant elements of the document which are difficult to automatically isolate. This tool follows an adaptation of the multimodal interactive-predictive approach (see figure 2) using adaptive learning techniques to reduce the human effort required to annotate these document images. The user can validate the initial labelling of the documents and, if necessary, edit this initial labelling by choosing among a set of alternatives. Using active learning methods, the system automatically proposes the optimal set of samples to validate and/or label. The information obtained by validation or edition of the labelling is used to update the model defined to annotate the documents. Once the labelling has been validated, this semantic information can be used for interactive retrieval of the documents stored in the database. This tool is used to annotate and retrieve difficult documents, specifically unstructured documents or documents with heterogeneous contents (containing printed and handwritten text, graphics, symbols, etc) such as administrative or ancient documents (see figure 1).

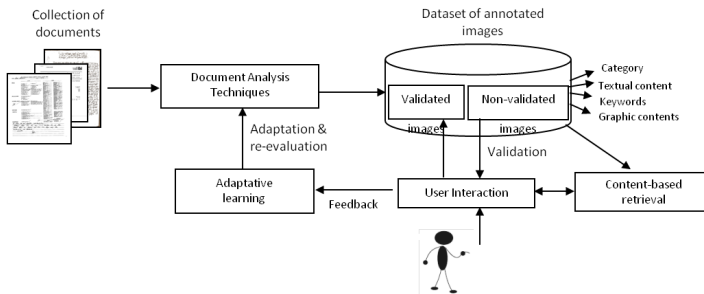


Fig. 2 Adaptation of the interactive-predictive framework in our prototype

We focus on two particular problems of document classification and retrieval where the use of an interactive tool can be especially appropriate: appearance-based document classification and logo-based retrieval. In both cases, we use state-of-the-art techniques to obtain the initial labelling of the documents. These techniques are described in section 2. Then, in section 3 we adapt them to be used within

interactive-multimodal framework. In section 4 we explain the implementation details of the prototype and show its main functionalities and some results obtained with its application to a set of documents. Finally, in section 5 we draw the main conclusion of the work presented.

2 Basic Document Classification and Retrieval

2.1 Document Classification

Document classification is performed by means of global visual appearance descriptors computed on the whole document image. We have used three different state-of-the-art descriptors that permit to capture document appearance using different techniques. In the following we, first give a brief description of the three descriptors and then, we explain the classification framework.

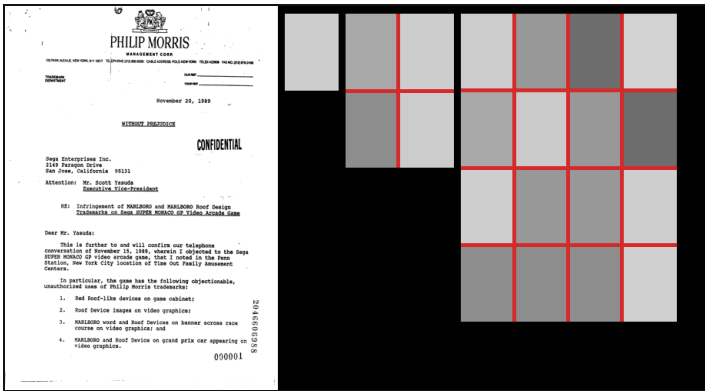


Fig. 3 Example of document image description by means of a pyramid of pixel densities

2.1.1 Visual Descriptors

Blurred Shape Model (BSM)

The BSM descriptor was proposed by Escalera et al. in [1] for the recognition of segmented graphical symbols. The BSM descriptor is a zoning-based method [6]. A regular grid of cells with a fixed number of rows and columns is overlaid on the document and then the grid is scanned from left to right and top to bottom where pixel density features are extracted from each cell in order to form the feature vector. In the original formulation of the BSM descriptor, each cell encodes pixel density from the foreground pixels that fall within it but also from the pixels assigned to the neighboring cells. Thus, each pixel contributes to a density measure of its cell, but also to its neighbors in order to achieve robustness to local variations

the document. This contribution is weighted in terms of the distance between the pixel and the centroid of the cell. The output descriptor is a histogram where each component corresponds to the amount of pixels in the context of the cell. Finally, the resulting histogram is $L1$ -normalized.

Pyramid of Pixel Densities (PYR)

The PYR encodes pixel densities at different scales. It is a hierarchical zoning descriptor introduced in [3]. In order to remove small details and noise from the incoming images, a Gaussian smoothing operator is first used to blur the images before computing the visual descriptor. Each document image is recursively split into rectangular regions forming a pyramid. At each new partition level l , 4^{l-1} dimensions are added to the feature vector. In each of the regions the pixel density computed as the average intensity value is stored in the corresponding position of the feature vector. We can see an example of the first levels of the pyramid in Figure 3. In our experimental setup, we use four levels, yielding to an 85-dimensional visual descriptor. The visual feature vectors are finally $L2$ -normalized.

Runlength Descriptor (RLD)

The Runlength descriptor [2] is based on a histogram of the length of the runs of consecutive pixels in the four main directions (horizontal, vertical and diagonals). The length of the runs is quantized in different intervals and thus, every bin of the histograms corresponds to one of these quantization intervals. The histograms in the four directions are concatenated to obtain the final descriptor. In addition, the image can be split into several regions at different levels of resolution (for instance, 2×2 and 4×4). In this case, a histogram for each region is computed and all of them are concatenated to obtain a multi-scale document representation that can capture information about the structure or the layout of the document.

2.1.2 Classifier

The process of classifying a document can be defined in the following way. Let I be a random variable describing the set of visual features (BSM, PYR or RLD) extracted from the document image. Let C be a random variable describing the set of classes. Then, the probability of classifying the document image as belonging to a certain class can be defined as $P(C|I)$. Applying Bayes' rule we get the following:

$$P(C|I) = \frac{P(I|C) \cdot P(C)}{P(I)} \quad (1)$$

where $P(I)$ is assumed to be equal for all images and can be ignored. $P(I|C)$ stands for the probability of the observed visual features given that the document belongs to a certain class. $P(C)$ can be seen as the a priori probability of every class. Initially, this probability is assumed to be equal for all classes. However, in the interactive scenario that will be described in the next section, it will change according to the user interaction.

We have explored two different ways of obtaining $P(I|C)$. The first one, in the framework of a Bayesian classifier, assumes that this probability follows a normal distribution with mean and variance computed from the training samples assigned to every class. The second one uses a k -NN classifier to classify the documents and obtain this probability. There are basically two ways of returning the probability of a class using a k -NN classifier, taking into account the classes of the k -th nearest samples to the unknown document: (a) using the relative frequencies of classes and (b) weighting each element by the inverse of its distance to the unknown element and normalizing to ensure a total mass of 1. Both methods asymptotically converges to the true probability when k increase but (b) is more robust and this is the option we have used.

In any of both classification schemes, an accurate estimation of the probability depends on having good labeled datasets. The construction of them are done taking the user feedback after a first unsupervised document classification. The samples validated by the user with the interactive framework described in the next section will be included in the training set and will be used to modify the parameters of the probability distribution used to compute $P(I|C)$.

2.2 Logo Detection

Logo detection consists in finding possible locations of a given query logo in the set of documents. In order to spot the position of logos appearing within document images we use a sliding window framework together with the blurred shape model (BSM) descriptor introduced in the previous section, but modified to take into account that we are working with non-segmented images. In the original formulation of the BSM descriptor, pixel density was computed over a regular $n \times n$ grid, assuming that the shapes to compare have been previously segmented. In our case we would like to locate a logo within a cluttered document image. Thus, we reformulate the BSM descriptor by forcing the spatial bins to have a fixed size (100x100 pixels in our experimental setup). Images having different sizes will result in feature vectors of different lengths. By using this reformulation of the BSM descriptor, the chosen size of the buckets will define the level of blurring and subsequently the information reduction for both the logos and the documents.

In order to locate a logo within a document image we use a sliding-window approach computed as a normalized two-dimensional cross-correlation between the BSM description of the model logo and the BSM description of the complete document. We use the two-dimensional cross-correlation proposed in [4] and computed as follows. Let t be the sought template represented by the BSM descriptor of the model logo and $f(x,y)$ the BSM description of the whole document image. The mean values of the template and a particular zone of the document descriptor are formulated as \bar{t} and $\bar{f}_{u,v}$ respectively. The correlation coefficient is then computed as

$$\gamma(u,v) = \frac{\sum_{x,y} [f(x,y) - \bar{f}_{u,v}] [t(x-u, y-v) - \bar{t}]}{\left\{ \sum_{x,y} [f(x,y) - \bar{f}_{u,v}]^2 \sum_{x,y} [t(x-u, y-v) - \bar{t}]^2 \right\}^{0.5}} \quad (2)$$

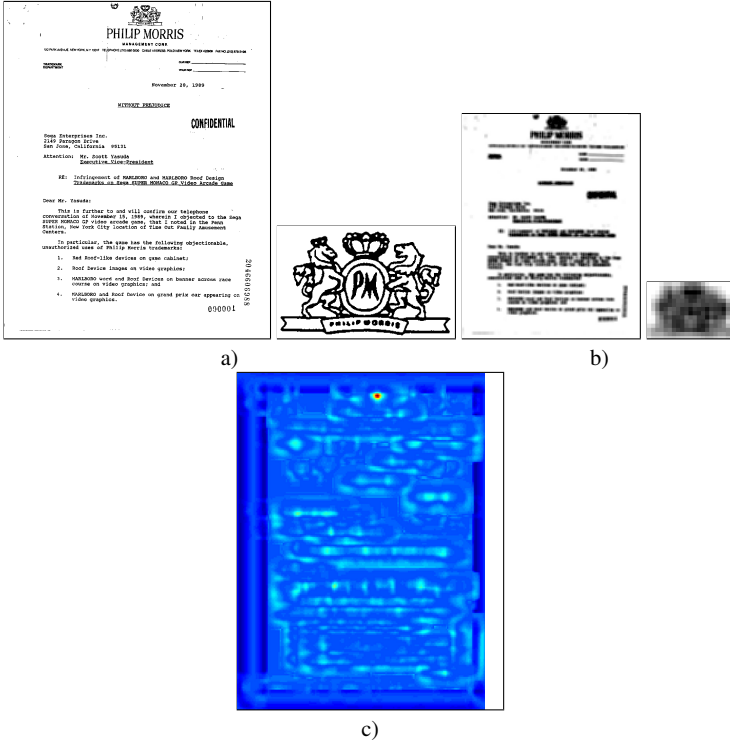


Fig. 4 Logo detection example using a sliding window over BSM descriptors. a) Original document and sought logo, b) BSM descriptors of the document and the logo, c) obtained correlation coefficients by the normalized two-dimensional cross-correlation.

As the result of the cross correlation between the BSM models and the BSM descriptor from the document, a peak should be formed in the correlation coefficient image in the location (u, v) where there is a high probability to find a something similar to the given logo. We can see in Figure 4 an example of the obtained output given a document and a logo to search for.

Using this basic set of techniques, the process of detecting a logo in an image can be defined in the following way. Let I be a random variable describing the set of features (BSM descriptor) extracted at every position of the sliding window over the image. Let X be a random variable standing for a given position of the sliding window (x and y coordinates of the bounding box). Let L be a random variable describing the possible set of logos. Then, the probability of detecting a logo at a certain location of the image, given the set of features extracted from the image can be defined as $P(X, L|I)$. Applying Bayes' rule we get the following:

$$P(X, L|I) = \frac{P(I|L, X) \cdot P(L, X)}{P(I)} \quad (3)$$

where $P(I)$ is assumed to be equal for all images and can be ignored. $P(I|L,X)$ stands for the probability of the observed image features given that we are trying to find a given logo at a certain location of the image. This probability can be modelled by the score of the cross correlation between the model of the logo and the BSM descriptor at every location of the image, as described before. $P(L,X)$ can be seen as the a priori probability of every logo and location. This expression can be further developed in the following way:

$$P(X,L) = P(X|L) \cdot P(L) \quad (4)$$

In a retrieval scenario where we are searching for a specific given logo, $P(L)$, the a priori probability of every logo, can be ignored and therefore, $P(X,L)$ only depends on the probability $P(X|L)$ that stands for the a priori probability of finding a given logo at a certain location of the image. This probability can be estimated from a set of learning documents, just by counting the frequency of appearance of a given logo at every location of the image. User interaction, as introduced in the next section, will permit to modify this a priori probability. The new logo detections validated by the user will be used to update $P(X|L)$ accordingly.

3 Interactive Document Retrieval

In this section we will explain how the document classification and logo detection tasks described in the previous section can be adapted to take into account the interactive-predictive framework introduced in [5]. In particular, we will consider the history of previous user interaction steps in order to help the system when taking a decision in the current step.

3.1 Interactive Document Classification

For document classification, user interaction will permit to validate or reject the last hypothesis made by the system concerning the class of the document. Thus, equation 1 must be re-formulated to take this interaction into account. In particular, the class of a document will not only depend on the visual features extracted from the image, I , but also on two new random variables, c' standing for the last hypothesis made by the system concerning the class of the document, and d , corresponding to the decoding of the user interaction, that is, accepting or rejecting the last hypothesis c' . Putting all together, the class of a document will be determined according to the following expression:

$$c = \arg \max_c P(C|I, c', d) \quad (5)$$

This equation can be expanded in a similar way as we did in equation 1 and then, we get that the class c assigned to a document is obtained in through this expression:

$$c = \arg \max_c P(C|I, c', d) = \arg \max_c P(I|c) \cdot P(C|c', d) \quad (6)$$

Thus, the class c assigned to a given document depends on two terms. The first one, $P(I|C)$ the same as in equation 1, only considers the visual appearance of the document and thus, it is constant. The second one, $P(C|c', d)$, accounts for the user interaction and therefore, it will change at each interaction step. It permits to update the a priori probability of each class according to the user feedback d and the previous hypothesis made by the system C' . The user feedback will consist in validating or rejecting the hypothesis C' . We will assume that d takes the value 0 if the previous hypothesis is rejected and 1 if it is validated. Under these assumptions the update of the a priori probability can be expressed in the following way:

$$P(C|c', d) = \left\{ \begin{array}{l} d = 1 \Rightarrow \left\{ \begin{array}{l} 1 \quad c = c' \\ 0 \quad c \neq c' \end{array} \right\} \\ d = 0 \Rightarrow \left\{ \begin{array}{l} 0 \quad c = c' \\ \frac{P(c)}{\sum_{c_i \neq c'} P(c_i)} \quad c \neq c' \end{array} \right\} \end{array} \right\} \quad (7)$$

being $P(c)$ the a priori probability of every class at the previous iteration step. If the previous hypothesis is validated ($d = 1$), then the class corresponding to the last hypothesis c' is assigned probability 1 while all other classes are assigned probability 0. However, if the previous hypothesis is rejected ($d = 0$), the class corresponding to c' is assigned probability 0, while its previous probability value is equally re-distributed among all the remaining classes.

3.2 Interactive Logo Detection

In a similar way as in document classification, user interaction can be included for logo detection in the form of validation or rejection of the hypothesis that a given logo appears in an image. Then, the probability of finding a logo at a certain location of an image (defined in equation 3) is modified to include a new random variable d accounting for the user feedback. In this way, this probability can be expressed as $P(X, L|I, d)$. User feedback we will consist in validating or rejecting the presence of a given logo in the image. Thus, d will consist of a sequence of pairs (l, y) where l will be any of the possible logos and y will take the value 0 if the logo does not appear in the image and 1 if it appears.

The probability $P(X, L|I, d)$ of finding a logo in the image can be expanded in a similar way as in equations 3 and 4 leading to the following (figure 5):

$$P(X, L, |I, d) = P(I|X, L) \cdot P(X|L) \cdot P(L|d) \quad (8)$$

$P(L|d)$ is the probability of finding a logo in the document giving the history of user feedback d . Initially all logos are assigned the same probability. As user provides feedback this probability is updated in a way that for all logos l for which a pair $(l, 1)$ appears in the sequence d their probability $P(L|d)$ is set to 1, while for all logos for which a pair $(l, 0)$ appears in d , the probability is set to 0.

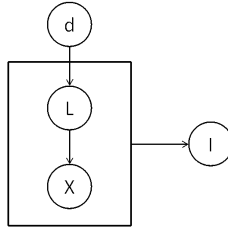


Fig. 5 Graphical model corresponding to $P(X, L|I, d)$

3.3 Interactive Class-Based Logo Detection

We can combine document classification and logo retrieval to propose a new scenario where we can use class information to help in the detection of logos. We can assume that logos appearing in an image will be different depending on the class of the document. Thus, we can relate logos with classes of documents through a new probability distribution, $P(L|C)$, that assigns to every logo a certain probability of appearing in documents of a given class.

Then, the probability of defining a logo in an image (equation 3) can be modified to include this new dependency in the following way:

$$P(X, L|I, C) = P(X|L, C) \cdot P(L|C) \cdot P(C) \quad (9)$$

Up to this point the system is fully automatic: given an image of the document, first, the class of the document is determined and depending on this, the probability of finding each logo in the image can be computed using the previous expression and, therefore, possible location of logos can be retrieved.

Going one step further, the interactive-predictive framework can also be used to take into account the user interaction for document classification. In this way, as described in section 3.1, given the initial classification of a document, the user can validate or reject this hypothesis. Then, the system uses this feedback to re-evaluate the classification of the document and, as a result of this evaluation, the probability of each logo is also modified and thus, a new set of possible logo detections is retrieved. All this process can be expressed in terms of probabilities combining equations 6 and 9 in the following way (see figure 6):

$$P(X, L|I_L, I_C, C, C', d) = P(I_L|X, L) \cdot P(X|L, C) \cdot P(L|C) \cdot P(I_C|C) \cdot P(C|C', d) \quad (10)$$

Note that in this expression we distinguish between image features used to find logos, I_L , and image features used for document classification, I_C . The whole process is illustrated in figure 7.

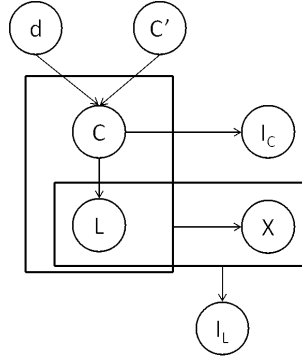


Fig. 6 Graphical model corresponding to the probability of finding a logo using class information and user interaction

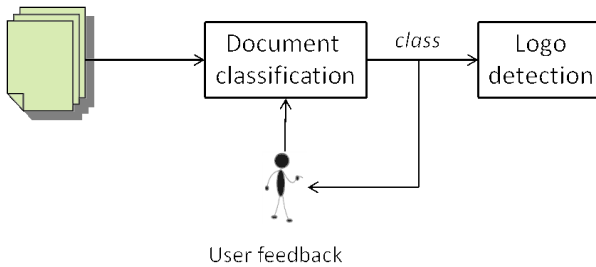


Fig. 7 Interactive process for detecting logos taking into account user interaction for document classification

4 Prototype

The prototype in its current release includes two document analysis applications: document classification and logo detection, which permits to annotate documents with different kinds of semantic information¹. We have uploaded several collections of documents and we have also prepared a set of configurations for demonstration purposes.

We have included the prototype functionalities in a web-based application structured in three different layers: (a) a graphical front-end taking care of displaying information and interaction with the user, (b) a back-end where the user sets the application configuration and (c) a set of tools for document analysis and learning.

The front-end of the prototype is in charge of the user interaction and displaying results. In the following we will illustrate how it works for the case of document

¹ Available at <http://dag.uab.es/documents> (write miprcv as username and password to login in this demo).

classification. The process for logo detection is very similar, just changing classes of documents by logo detections.

The general view always shows a list of all the classes of documents plus an additional class *Not Assigned* for those documents not classified in any of the previous classes. As a first step, the system performs an unsupervised classification of all the documents using the K -means algorithm in order to get a first assignment of documents to classes. The user only has to give a semantic name to the created classes. Then, user interaction is done in a simple way using the mouse and the keyboard. The user selects one of the predefined set of document categories and validates or rejects the documents assigned to it by just clicking on the green ticks and cross red buttons that appear on the right side of a thumbnail image of each document (see figure 8 (a) and (b)). Alternatively, the user can also globally validate or reject all the documents assigned to the class by clicking on the green tick button or the red cross underneath the class label. Additionally, the user can also select the class of documents not assigned to any class because their probability were very low. In this case, the user can directly assign each of them to the correct document class, instead of validating or rejecting them (see figure 8 (c)).

At every interaction step, samples pending of validation are shown to the user sorted using an active learning strategy based on uncertainty sampling. The samples with maximum entropy are selected as the first ones to be validated. After each interaction step, the interactive-predictive framework described in section 3 is used to update all the probability distributions concerning the model. Accordingly, the system modifies the classification hypothesis for all the documents pending of validation or still not assigned to any class.

The back-end view of the prototype controls the basic user functionalities for managing document collections. There, the user is able to create, manage and organize collections of documents. He is able to select the collection or collections to be used at any moment for a given application. In figure 9 we can see the check-list of all available datasets and the first labeled classes. We have selected the public and standard NIST dataset of forms for document classification and the Tobacco logo dataset for logo detection.

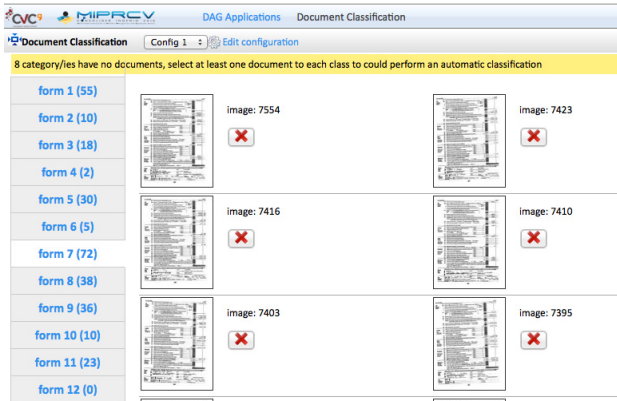
Once the user has selected the datasets and defined the semantic classes partitioning the dataset, the user can also select the configuration of the basic techniques (visual descriptor and base classifier) that will be used by the system, as it can be seen in figure 10. If the user clicks over the *Add configuration* button a new window, with a list of available descriptors and classifiers, appears on the top of the application. By default, each descriptor and classifier works with default parameters that can be changed by means of *Descriptor params* text box and *Classifier params* text box, respectively.

4.1 Experiments

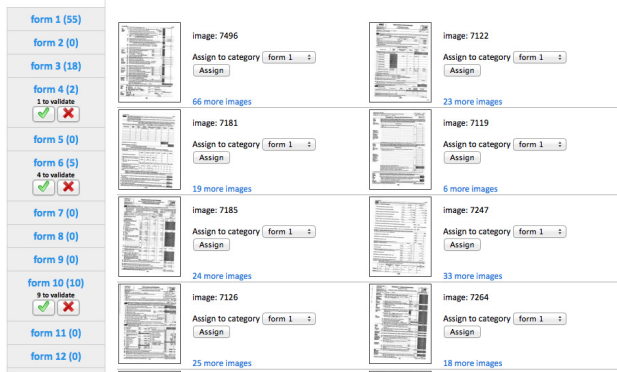
In order to evaluate the performance of the proposed approach we have simulated a series of user interaction steps in the task of document classification. We have



(a)



(b)



(c)

Fig. 8 Front-end of the prototype. On the left side, the list of classes of documents with the number of examples validated (in brackets). In the main frame, (a) a list of documents of one class to be validated. In (b) a list of validated documents for a given class. In (c) the set of documents not assigned to any class.

Fig. 9 Documents configuration menu. The user can check the list of datasets available (SFD in this demo) and define the document semantic classes (labeled as *form 1* to *form 20*).

ID	Configuration name	Descriptor	config.	Classifier	config.
27	Config 1	BSM - Pixel Distribution in 4 Rectangles		Bayes	
28	Config 2	PYR - Pixel Distribution in 4 Rectangles		Bayes	
29	Config 3	RLD - Run-Length Histogram		Bayes	

Fig. 10 User configurations menu. At each configuration the user can choose the type of descriptor (BSM, PYR, RLD) and the classifier.

used a subset of 3200 images from the NIST dataset of forms which is composed of 20 different classes. We have divided the set of instances in two blocks: training and test. The training set is composed of 200 images randomly selected, not equally distributed among the 20 classes. The test set is composed of the remaining 3000 images. We have used a configuration composed of the BSM descriptor and the k -NN classifier. We have followed this protocol: first, one instance of each class is selected to train the classifier. Then, at each new step, all the images in the training set that are not classified yet are separated in clusters using the K -means algorithm. For each cluster we select the most uncertain sample in terms of entropy and we re-train the model adding this new sample from every cluster according to the model explained in section 3.1. At each step the classification accuracy is determined. Results are shown in table 1. The first column shows the number of labeled samples used at every step to train the classifier. The second column shows the classification accuracy after each step. It can be seen that the accuracy improves at every step. It is worth noting that in this dataset the state-of-the-art in classification is very close to 100%. Thus, we start with already high accuracy rates and very fast (with a few samples and interaction steps) we converge to rates very close to the state-of-the-art.

Table 1 Classification accuracy after several steps of user interaction

N° of instances	Accuracy
27	99.1
36	99.2
47	99.0
56	99.4
66	99.5
75	99.6
84	99.5
94	99.6
106	99.6

5 Conclusions

In this chapter we have shown the adaptation of classical document analysis problems, such as document classification and logo retrieval to the interactive predictive model that permits to take advantage of the user interaction to improve retrieval results and re-train the models. We have seen how, in this scenario, both problems, document classification and logo retrieval can be easily related so that logo retrieval can take advantage of the class information obtained. We have also shown a practical implementation of this framework and some results that confirm that the user interaction can speed up the training process. However, a more exhaustive evaluation should be conducted in the future to establish the real power of combining both tasks under the same framework.

References

1. Escalera, S., Fornés, A., Pujol, O., Radeva, P., Sánchez, G., Lladós, J.: Blurred shape model for binary and grey-level symbol recognition. *Pattern Recognition Letters* 30(15), 1424–1433 (2009)
2. Gordo, A.: Document Image Representation, Classification and Retrieval in Large-Scale Domains. PhD thesis, Universitat Autònoma de Barcelona (2012)
3. Héroux, P., Diana, S., Ribert, A., Trupin, E.: Classification method study for automatic form class identification. In: *Proceedings of the Fourteenth International Conference on Pattern Recognition*, pp. 926–928 (1998)
4. Lewis, J.P.: Fast normalized cross-correlation. *Vision Interface* 10, 120–123 (1995)
5. Toselli, A., Vidal, E., Casacuberta, F.: *Multimodal Interactive Pattern Recognition and Applications*. Springer (2011)
6. Zhang, D., Lu, G.: Review of shape representation and description techniques. *Pattern Recognition* 37(1), 1–19 (2004)

Interactive Visual and Semantic Image Retrieval

Joost van de Weijer, Fahad Khan, and Marc Masana

Abstract. One direct consequence of recent advances in digital visual data generation and the direct availability of this information through the World-Wide Web, is a urgent demand for efficient image retrieval systems. The objective of image retrieval is to allow users to efficiently browse through this abundance of images. Due to the non-expert nature of the majority of the internet users, such systems should be user friendly, and therefore avoid complex user interfaces. In this chapter we investigate how high-level information provided by recently developed object recognition techniques can improve interactive image retrieval. We apply a bag-of-word based image representation method to automatically classify images in a number of categories. These additional labels are then applied to improve the image retrieval system. Next to these high-level semantic labels, we also apply a low-level image description to describe the composition and color scheme of the scene. Both descriptions are incorporated in a user feedback image retrieval setting. The main objective is to show that automatic labeling of images with semantic labels can improve image retrieval results.

1 Introduction

One direct consequence of recent advances in digital visual data generation and the direct availability of this information through the World-Wide Web, is a urgent demand for efficient image retrieval systems. The disclosure of the content of these millions of photos available on the internet is of great importance. The objective of image retrieval is to allow users to efficiently browse through this abundance

Joost van de Weijer · Marc Masana
Computer Vision Center, Barcelona, Spain
e-mail: {joost, marc.masana}@cvc.uab.es

Fahad Khan
Computer Vision Laboratory, Linköping University, Sweden
e-mail: fahad@cvc.uab.es

of images. Due to the non-expert nature of the majority of the internet users, such systems should be user friendly, and therefore avoid complex user interfaces.

Traditionally, two sources of information are exploited in the description of images on the web. The first approach, called text-based image retrieval, describes images by a set of labels or keywords [5]. These labels can be automatically extracted from for example the image name (e.g. 'car.jpg' would provide information about the presence of a car in the image), or alternatively from the webpage text surrounding the image. Another, more expensive way would be to manually label images with a set of keywords. Shortcomings of the text-based approach to image retrieval are obvious: many objects in the scene will not be labeled, words suffer from the confusions in case of synonyms or homonyms, and words often fall short in describing the esthetics, composition and color scheme of a scene. However, until recently many image retrieval systems, such as e.g. Google-image search, were exclusively text based.

A second approach to image description is called content-based image retrieval (CBIR). Here users are provided with feedback from an image-query purely based on the visual content of images. These methods are better able to describe the scene composition and color scheme of images. However, they suffer from the semantic gap, which is the gap between low-level image features and high level semantics of the image [21]. Features which are popular in such systems range from global color description [9], to texture descriptions [7], to precise shape descriptions [18]. Due to their different nature, CBIR and text-based image retrieval were found to be complementary [5].

Given the complexity of the image retrieval problem, researchers have acknowledged that user feedback should be an integral part of any image retrieval system [21, 26]. Therefore, relevant feedback mechanisms have been a popular research subject in image retrieval. Users are asked to provide the system with some form of feedback, for example by selecting images which match or do not match the target image. The system then reorders the images given the user feedback. Interactive image retrieval provides a way to approach the inherent ambiguities which exist in image retrieval. Furthermore, it allows adapting the results to be user-dependent. Important is that these systems should operate in real-time which excludes the use of complex learning algorithms. From user studies we know that adding interactive feedback significantly improves the efficiency of retrieval systems [26].

In recent years object recognition and scene categorization have made significant advances [6, 16], especially due to the usage of machine learning techniques in combination with a local feature description of images. The combination of highly discriminative features [15], and the bag-of-words framework have resulted in significant progress [6, 20]. Also the introduction of standard benchmark data sets, such as the VOC PASCAL challenge [3], have further contributed to fast developments in the field of object recognition. One could say that these advances have significantly reduced the semantic gap, and state-of-the-art is currently able to automatically label images with semantically labels.

In the image retrieval system, described in this chapter, we investigate the usage of recent developments in object recognition to bridge the semantic gap. We are especially interested to investigate how such high-level information can improve interactive image retrieval. We will apply a bag-of-word based image representation method to automatically classify images in a number of categories. These additional labels are then applied to improve the image retrieval system. Next to these high-level semantic labels, we also apply a low-level image description to describe the composition and color scheme of the scene. Both descriptions are incorporated in a user feedback image retrieval setting. In conclusion, the novelty of our prototype for image retrieval can be summarized as follows:

- Apply bag-of-word based image classification to bridge the semantic gap by automatically labeling images with a set of semantic labels.
- Improve user feedback by allowing the user to select images to resemble the target image according to semantic or esthetic (color composition) content.

The main objective is to show that automatic labeling of images with semantic labels can improve image retrieval results.

This chapter is organized as follows. In Section 2 an overview of our approach is given. In Section 3 the details of the both the semantic and the visual image representation are discussed. In section 4 the technical details and the user interface are discussed. In Section 5 a demonstration of the image retrieval system is given, and Section 6 finishes with concluding remarks.

2 Interactive Visual and Semantic Image Retrieval

A typical user of an image retrieval system is looking for images to use in a presentation, a report, or his webpage. Examples of images could be "A city-scene during the night", or "A living room in retro style". Communicating the desired image to other humans already can be a difficult task. To facilitate communicating these desires into queries for a computer, we differentiate between two sources of communication: semantic queries in the form of text, and visual queries in the form of images. Text queries, typically allow the user to communicate objects or buildings which should be present in the scene such as "car", or "town" (e.g. Google image search is known to be mainly text-based). Visual queries allow the user to steer the composition, color arrangement and general atmosphere (e.g. cold or warm) of the query.

Due to the inherent ambiguity in the initial query (e.g. different users could envisage different images but use the same query) user feedback will be crucial for successfully navigation of the system. The propose system is given in Figure 1. The user will initialize the system with a text query. Based on an image classification system a number of relevant images to the query will be presented to the user in the form of a ranked list. At this time, the user can precise his query by choosing visually and semantically relevant images. Furthermore, the user can leverage the importance of the text and visual query. Based on the combined query the system will re-rank the images and present them to the users. This loop can be repeated

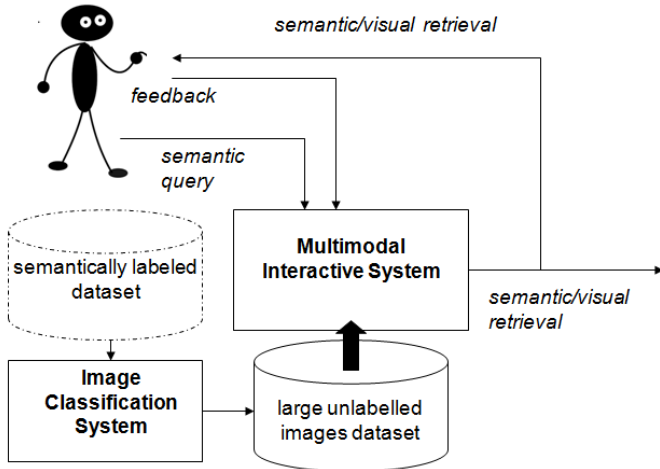


Fig. 1 Overview of image retrieval prototype combining both, semantic and visual queries. Adaptation of general model for multimodal iterative systems ([22]).



Fig. 2 Example of combination of visual and semantic query. The semantic query indicates that the user wants a "horse" to be present in the image, whereas the visual query suggest that the horse should be situated in a green outdoor setting.

until the user is satisfied with the returned results. An example of a combination of visual and semantic query is provided in Figure 2.

In the following we give a more precise overview of our approach. Each image is defined by a visual \mathbf{d}_v and a semantic descriptor \mathbf{d}_s according to

$$\mathbf{d} = [\mathbf{d}_v, \mathbf{d}_s]. \quad (1)$$

The semantic query is coded by the vector \mathbf{q}_s^0 which is 1 for the classes which are indicated in the semantic query and zero otherwise. Initially the system returns a ranked list according to the following distance equation for all images (indexed by i)

$$\varepsilon_0^i = \mathbf{F}(\mathbf{q}_s^0, \mathbf{d}_s^i) \quad (2)$$

where \mathbf{F} is a distance function. Throughout this chapter we will use the following distance measure

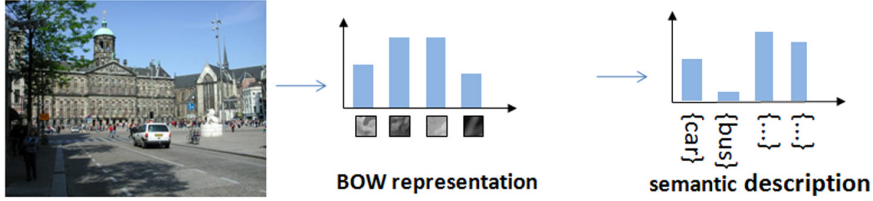


Fig. 3 We propose to use image classification methods based on bag-of-words to automatically label the image with semantic information, which are then applied for semantic image retrieval

$$F(\mathbf{a}, \mathbf{b}) = \frac{\sum_i (a_i - \min(a_i, b_i))}{\sum_i a_i} \quad (3)$$

where a_i denotes the element i of vector \mathbf{a} . Note that this distance measure is equal to histogram intersection in case of normalized vectors a and b (as we will see this is the case for the visual descriptors \mathbf{d}_v). However, it can also be used for unnormalized vectors, which the semantic descriptors \mathbf{d}_s will turn out to be.

Next the user can improve his query by selecting relevant images. The selected images are contained in the set D_r . Given the selected relevant images the user is provided with new results based on the following distance measure

$$\varepsilon^i = \lambda \left(\mathbf{F}(\mathbf{q}_s^0, \mathbf{d}_s^i) + \frac{1}{|D_r|} \sum_{j \in D_r} \mathbf{F}(\mathbf{d}_v^j, \mathbf{d}_v^i) \right) + (1 - \lambda) \left(\frac{1}{|D_r|} \sum_{j \in D_r} \mathbf{F}(\mathbf{d}_s^j, \mathbf{d}_s^i) \right) \quad (4)$$

where the parameter λ allows to leverage the relative influence of both cues and will be set by the user. Compared with Eq. 2 this equation has two additional parts, corresponding to the semantic and the visual distance to the relevant image set D_r . In the following section we explain how the visual description \mathbf{d}_v and the semantic description \mathbf{d}_s are computed. The images with lowest distance ε^i to the query are presented to the user for further evaluation.

3 Image Representations

In this section we shortly describe the two image representation methods which are used to describe the semantic and the visual content of the image.

3.1 Semantic Image Representation

In recent years object recognition has advanced significantly. As a direct consequence the semantic gap which exists between low-level image features and high-level semantic content of the images has been narrowed. The main idea is to use

Table 1 Overview of properties for several methods to combine multiple cues into the bag-of-word framework. Only Portmanteau vocabularies combine all three desirable properties. See text for discussion of table.

Method	Cue Binding	Cue Weighting	Scalability
Early Fusion	Yes	No	Yes
Late Fusion	No	Yes	Yes
Color Attention	Yes	Yes	No
Portmanteau	Yes	Yes	Yes

image classification methods to automatically label the image with semantically relevant labels (see Figure 3). It is important to note that our approach differs from existing bag-of-word based image retrieval methods (e.g. [17]), in that these methods do not transform the histogram into semantic classes. In this section, we shortly describe our approach to semantic image representation. In particular, we will discuss in detail how we combined several cues, in particular shape and color, into a single image representation. More details on our bag-of-words implementation can be found in [19, 2].

The bag-of-words approach which represents an image as a histogram of local features is currently the most successful approach for object and scene recognition [20, 15, 6, 16]. The approach works by constructing a visual vocabulary of local features after which a histogram is built by counting the occurrences of each visual word in an image. The histogram is then used to train a classifier. Consequently, given a test image the classifier is used to predict the category label of the image.

Introducing multi-modality, i.e. multiple cues, in bag-of-words image representations is an active field of research. Existing approaches used to combine color and shape information often provide below-expected results on a wide range of object categories. The inferior results obtained might be attributed to the way color is incorporated. Traditionally, there exist two approaches to combining color and shape features. The first approach, termed early fusion, combines color and shape features locally before the vocabulary construction stage. Therefore this representation has the *cue binding* property, meaning that the cue information is combined at the same location in the image. The second approach, called late fusion, combines the two visual cues after the vocabulary construction stage. In late fusion, separate visual vocabularies are constructed for color and shape and the two representations are then concatenated to construct the image representation. This representation lacks cue-binding, but possesses *cue weighting*, meaning that the relative weight of the cues can be balanced.

Recently, a method for combining multiple features, called color attention [10], has been introduced. This method combines both cue binding and cue weighting into a single representation. However, a disadvantage of this representation is that it does not possess *scalability* with the number of categories. Therefore, this method is not suitable for large class problems. An overview of the properties of the several methods to combine various cues in bag-of-words is given in Table 1. In the

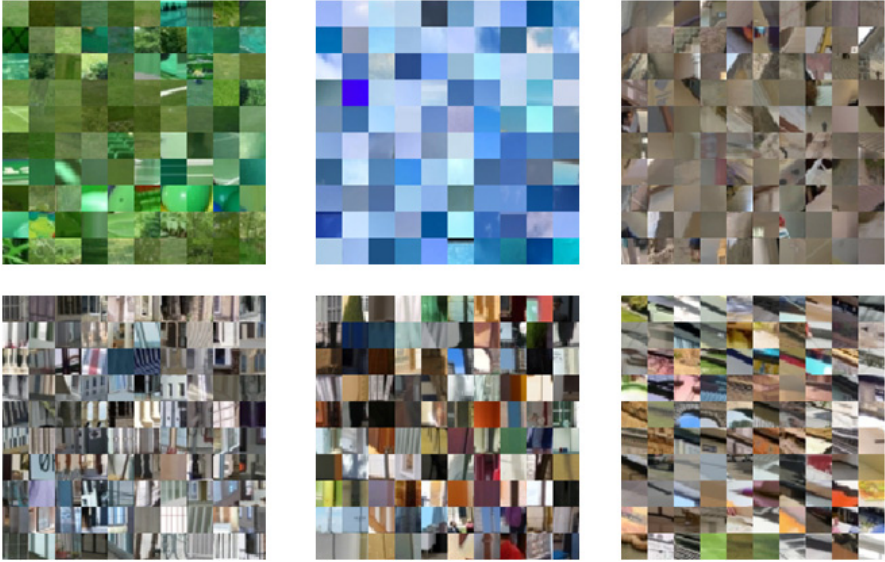


Fig. 4 Example of Portmanteau vocabulary: six different clusters are shown for the SUN data set, where every cluster is represented by 100 randomly sampled patches which are assigned to the cluster. Some clusters show constancy over color, whereas others are more constant over shape.

following, we will shortly describe the Portmanteau representation which we apply in our prototype [12]. This representation combines the desired properties cue binding, cue weighting and scalability.

A straightforward method to obtain the binding property is by considering a product vocabulary that contains a new word for every combination of shape and color terms. Assume that $S = \{s_1, s_2, \dots, s_M\}$ and $C = \{c_1, c_2, \dots, c_N\}$ represent the visual shape and color vocabularies, respectively. Then the product vocabulary is given by

$$\begin{aligned} W &= \{w_1, w_2, \dots, w_T\} \\ &= \{\{s_i, c_j\} \mid 1 \leq i \leq M, 1 \leq j \leq N\}, \end{aligned} \quad (5)$$

where $T = M \times N$. A drawback of the product vocabularies is that they result in a very large vocabulary size. As result this yields inefficient image representation, and in addition it is often difficult to obtain sufficient training data to prevent overtraining. Because of these drawbacks, compound product vocabularies have, not been pursued in literature. However, in recent years, several algorithms have been proposed which compress large vocabularies into small ones [8, 11].

Portmanteau vocabularies [12] are constructed by applying these algorithms to reduce the size of the product vocabularies. As a result we obtain a compact, multi-cue image representation. The algorithm joins words which have similar discriminative power over the set of classes in the image categorization problem. An example of the Portmanteau vocabulary for the SUN data set [25] is shown in Fig 4.

In conclusion, we use the Portmanteau image representation in our prototype because it combines multiple cues, namely color and shape, it is compact, and it was shown to obtain state-of-the-art results. Having the Portmanteau representation of the images we learn a SVM classifier with intersection kernel to label the images with the probabilities over a set of class labels. These probabilities constitute the semantic description \mathbf{d}_s of the image. When we compare different semantic descriptors with Eq. 3 we see that this has the desired property that the presence of objects which are not in the query does not increase the distance. However, the absence of objects which are in the query does increase the distance.

3.2 Visual Image Representation

Here we describe the visual image representation which is applied in the prototype. The aim of the visual representation is to capture both the color sensation and the composition of the image.

For the color description of the image we use color names. Color names are linguistic terms which humans use to communicate colors, such as ‘red’, ‘green’ and ‘blue’. We use the eleven basic color names of the English language, which are black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow. The mapping from RGB values to a probability over color names was learned from Google images (for a detailed description see [23] [24]).

Color names have the advantages that they are intuitively understandable to humans and they provide a very compact color description of an image. Furthermore color names possess a certain degree of photometric invariance, since many different shades of green are all captured by the single color name ‘green’. In addition, color names also describe the achromatic content of image, by using the color names ‘black’, ‘grey’ and ‘white’. This information is normally lost when working with photometric invariants such as *hue*, and normalized *RGB*. Because of these properties, color names were found to be excellent color descriptors [24] [11].

In addition, we use a weak composition descriptor image, by computing separate histograms over the color names for the bottom, the middle and the top of the image. This is similar to the spatial pyramids of Lazebnik [13] but was found to obtain better results. An overview of our visual image representation is given in Fig. 5. The final representation \mathbf{d}_v is only 33 bins, i.e. a concatenation of the colors in the bottom, middle, and top image represented in the eleven color names.

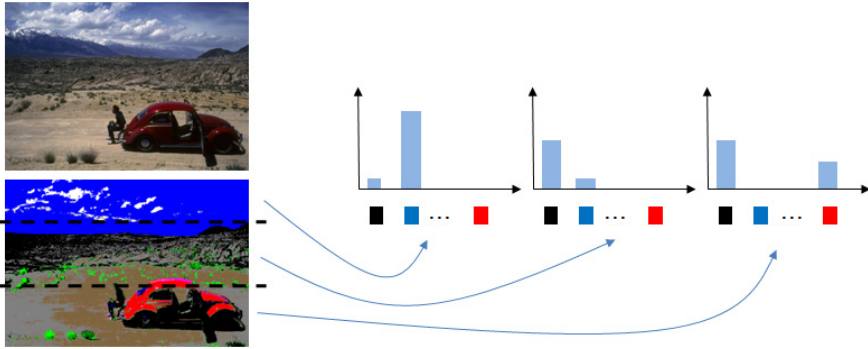


Fig. 5 The bottom image shows the color name assignment for the input image (superimposed lines indicate the three parts which are used to construct the final representation). For the visual representation of the image we concatenate the color histogram, over the eleven basic color terms of the English language, for the bottom, middle and top of the image.

4 Image Retrieval Application

In this section we provide the technical details of our system, and explain the user interface of our system.

4.1 Technical Implementation

The main challenges when implementing large scale image retrieval are user interaction and reaction speed. In our case, we need to provide visual feedback in the form of preferred images. Also, the user should be able to balance the strength of both semantic and visual queries which will result in a refined new query which can be repeated for further improvement.

To provide our system with these features, it is accessed with a web browser which is implemented in HTML. In addition to HTML we use PHP for the interactive functionality of the webpage. Furthermore, a combination of javascripts, AJAX and JSON are used to interact and perform computations on the client side. Finally, CSS is used to modify the style to make it more intuitive and pleasant for the user. PHP and HTML are interchangeable within the webpage, allowing the user to interact with the queries.

Each database is implemented into two PHP files. The first one contains the main part of the code to generate the webpage and relevant image ordering. The second one is a module file that contains functions to generate the remaining part of the website, mostly user interaction, but is not involved with the relevant image calculation.

Calculating distances for all images each time a query is performed, consumes a lot of time and slows the system down. To avoid this, pre-calculated distance values are stored which relieves the server side from laborious calculations. Each image has

different distance values to all the other images in the database stored. Then, when dealing with a group of relevant images the user has provided, these stored values are loaded resulting in a fast response time. The system has been tested to function with large datasets up to 40,000 images. The system can be tested at the following website: www.cat.uab.cat/Software/Image_Retrieval/index.php.

4.2 User Interface

An example of the user interface of our system is given in Figure 6. The user can select a semantic category, in the example 'bicycle' has been selected. The user can further decide the number of images which should be returned (set to 15 in the example). In addition the user can select images which are considered relevant for the query. In the example the user has already selected three bike images on a grass background. Finally, the user can select the relevance of the semantic content versus the visual content with a slider in the right top of the interface. Based on these inputs the system will return the most relevant images, four of which are given at the bottom of the example. In the experimental section we will investigate if the semantic image description leads to an improved image retrieval system.

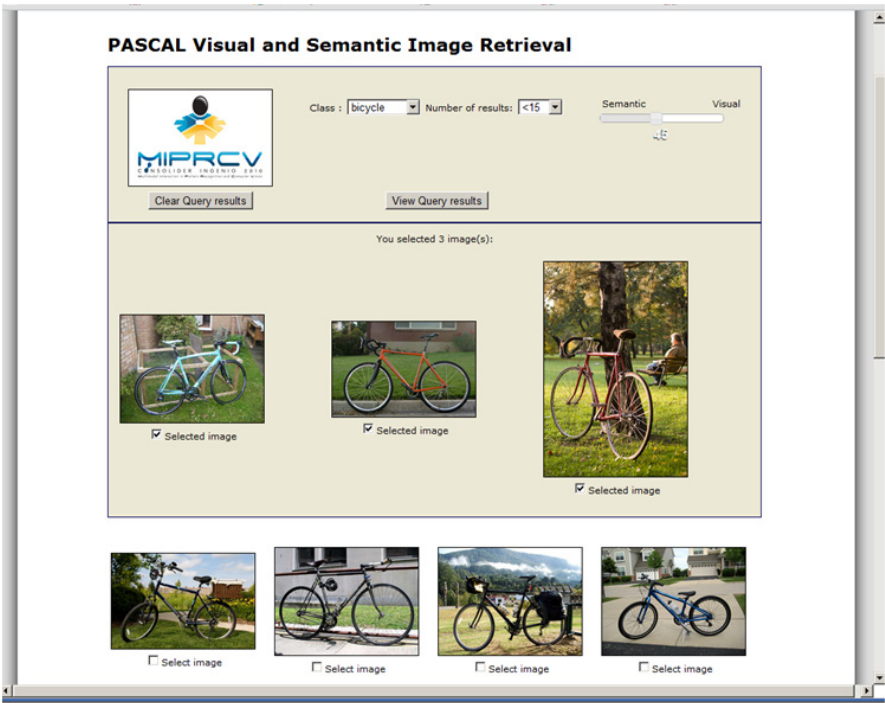


Fig. 6 Interface of the image retrieval system. See text for further explanation.

5 Demonstration and Experiment Results

In the introduction we pointed out that the main objective of our image retrieval application is twofold: 1. Apply bag-of-words based image classification to bridge the semantic gap by automatically labeling images with a set of semantic labels, 2. Improve user feedback by allowing the user to select images to resemble the target image according to semantic or esthetic (color composition) content. In this section, we provide two experiments to evaluate these objectives.

To test our image retrieval system we use two large datasets. The PASCAL VOC 2009 dataset consists of 13704 images. The images are divided into 20 different object categories. In our experiment we train on 3473 images and test the retrieval system on 3581 images of the validation set. The SUN dataset consists of 39700 images of 397 different scene categories. The dataset is divided into 19850 training and 19850 test images. Both datasets are difficult owing to large amount of variations both within an object category and across different object categories.

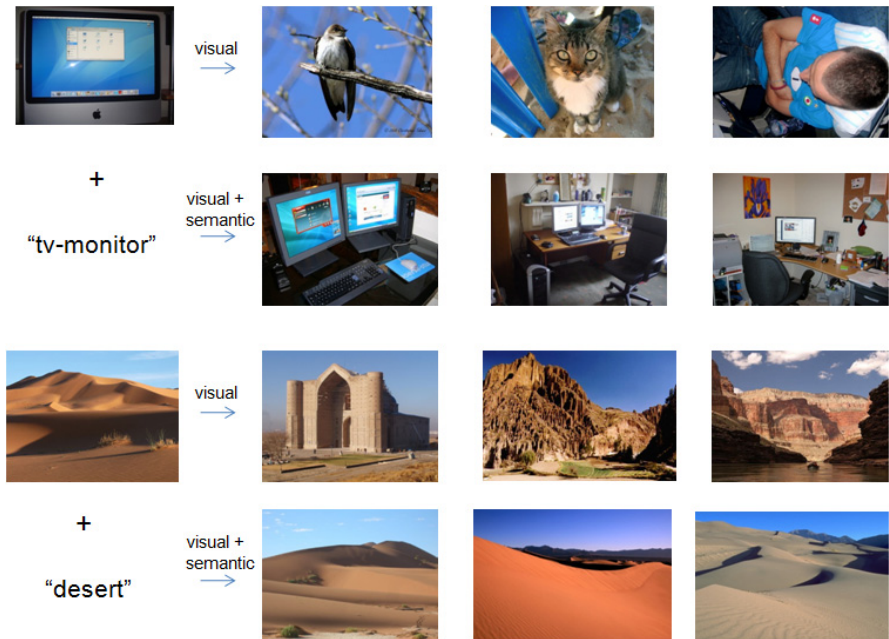


Fig. 7 Retrieval results for two different images. Top image from VOC PASCAL and bottom image from SUN data set. For both images the query is performed twice once only based on visual descriptor, and once on the combined visual and semantic descriptor. Note how the semantic description helps to improve the query results.

5.1 Semantic Image Description

In the first experiment we aim to evaluate if the semantic image description improves the overall image retrieval results. Two example retrievals which illustrate the importance of the semantic description are provided in Figure 7. To quantify the improvement we performed a small user study. Users were given a target image together with six retrieved images. The six images contained two images which were similar only in a visual sense, two in a semantic sense and two which are similar in both visual and semantic description. The images are randomly presented and the user is unaware which algorithm is related to which image. Next, the user is asked to select the image from the six which is most similar to the target image.

In Table 2 the results of the experiment are summarized. The results are based on a total of ten test person which provided ten preferences each. The visual and semantic description is significantly more often selected than the results returned by visual only. In 56% of the queries on PASCAL and in 42% of the queries on SUN the combined description was preferred. This clearly shows the importance of the semantic description for image retrieval.

Table 2 User preference for 'visual', 'semantic' or 'visual+semantic' description of images when asked which image is most similar to a target image. Results are provided in percentage of times the user selected the description.

Data Set	Visual	Semantic	Visual+Semantic
VOC PASCAL 2009	17%	27%	56%
SUN	24%	34%	42%

5.2 Interactive Visual and Semantic Retrieval

In the second experiment we desire to establish if the semantic description within the interface provided by Figure 1 is beneficial. To evaluate the user interface we designed a user experiment to measure the speed with which a user finds the desired image. Users are asked to find a given target image with the image retrieval system. The test is performed on the PASCAL data set. We compare the retrieval system with only visual image description (V-system) to the system with both visual and semantic information (VS-system). As an evaluation measure we compare the number of target images which were found within X rounds of interaction. Since in the VS-system the user can also balance the relative weight of semantic and visual information, we allow more rounds of interaction to the V-system. We choose X to be ten for the V-system and five for the VS-system.

The results of the experiment are presented in Table 3. Eight subjects have performed the experiment (five searches for the V-system and five for the VS-system). The same random set of images was evaluated by both systems. The results show that about double the amount of images were found by the VS-system, indicating

Table 3 Number of target images which were found by the system within X rounds of interaction for V-system and VS-system. The parameter X was set to ten for the V-system and to five for the VS-system.

Data Set	Visual	Visual+Semantic
VOC PASCAL 2009	5	11

that the additional semantic information does significantly improve the retrieval system. The fact that only eleven out of 40 images queries were found within five interactive rounds reveals that the user interaction can still be improved significantly. Users identified that they would have appreciate an additional feature which allows users to indicate whether the selected image is relevant for its semantic content or for its color and composition.

6 Conclusions

In this chapter we have investigated the usage of image classification methods to bridge the semantic gap. We apply image classification to automatically label images with semantic terms. These terms are then used to facilitate image retrieval. Users can start their query with a semantic term, and subsequently improve the query by selecting relevant images. Queries can be considered relevant with respect to their visual or semantic content. The user interface allows users to leverage between these two cues. Initial results are promising and show that semantic queries improve retrieval quality.

As future work we see incorporating automatic semantic labeling in more developed retrieval systems such as RISE [14]. In this case terms which are contributed by the automatic labeling can be handled similarly as terms extracted from the surrounding webpage of images or the filename. Another extension in which we are interested is further improving the semantic labeling by using object detectors [4]. This would allow users to further specify which part of the image they consider relevant for the query. In conclusion, we expect that in the near future object recognition techniques will be an integral part of most image retrieval systems. The gained semantic descriptions of images will improve the quality of the retrieval system. Furthermore, knowledge of the location of the semantic content in the image will open up new ways of improved user feedback.

Acknowledgements. We acknowledge the help and advice of Marc Paz in the implementation of this project. We also thank Carles Sánchez for his initial implementation of the system. We acknowledge the Spanish Research Program Consolider-Ingenio 2010: MIPRCV (CSD200700018); and Spanish project TIN2009-14173. Joost van de Weijer acknowledges support of Ramon y Cajal fellowship.

References

1. Dhillon, I., Mallela, S., Kumar, R.: A divisive information-theoretic feature clustering algorithm for text classification. *Journal of Machine Learning Research (JMLR)* 3, 1265–1287 (2003)
2. Elfiky, N., Khan, F.S., van de Weijer, J., Gonzalez, J.: Discriminative compact pyramids for object and scene recognition. *Pattern Recognition (PR)* 45(4), 1627–1636 (2012)
3. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2007 results (2007)
4. Felzenszwalb, P.F., McAllester, D.A., Ramanan, D.: A discriminatively trained, multi-scale, deformable part model. *IEEE Computer Vision and Pattern Recognition* (2008)
5. Ferencatu, M., Boujemaa, N., Crucianu, M.: Semantic interactive image retrieval combining visual and conceptual content description. *ACM Multimedia Systems* 13(5-6), 309–322 (2008)
6. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 264–271 (June 2003)
7. Fernandez, S.A., Salvatella, A., Vanrell, M., Otazu, X.: Low dimensional and comprehensive color texture description. *Computer Vision and Image Understanding* 116(1), 54–67 (2012)
8. Fulkerson, B., Vedaldi, A., Soatto, S.: Localizing Objects with Smart Dictionaries. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I. LNCS*, vol. 5302, pp. 179–192. Springer, Heidelberg (2008)
9. Gevers, T., Smeulders, A.: Color based object recognition. *Pattern Recognition* 32, 453–464 (1999)
10. Khan, F.S., van de Weijer, J., Vanrell, M.: Modulating shape features by color attention for object recognition. *International Journal of Computer Vision (IJCV)* 98(1), 49–64 (2012)
11. Khan, F.S., Anwer, R.M., van de Weijer, J., Bagdanov, A.D., Vanrell, M., Lopez, A.M.: Color attributes for object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2012)
12. Khan, F.S., Van de Weijer, J., Bagdanov, A.D., Vanrell, M.: Portmanteau vocabularies for multi-cue image representation. In: *Twenty-Fifth Annual Conference on Neural Information Processing Systems, NIPS 2011* (2011)
13. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178 (2006)
14. Leiva, L.A., Villegas, M., Paredes, R.: Query refinement suggestion in multimodal interactive image retrieval. In: *Proceedings of the 13th International Conference on Multimodal Interaction (ICMI)*, pp. 311–314 (2011)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)* 60(2), 91–110 (2004)
16. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10), 1615–1630 (2005)
17. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006*, vol. 2, pp. 2161–2168. IEEE Computer Society (2006)
18. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)

19. Rojas-Vigo, D., Khan, F.S., van de Weijer, J., Gevers, T.: The impact of color on bag-of-words based object recognition. In: *Int. Conference on Pattern Recognition, ICPR* (2010)
20. Schmid, C., Mohr, R.: Local grayvalue invariants for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 19(5), 530–534 (1997)
21. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval: the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(12), 1349–1380 (2000)
22. Toselli, A.H., Vidal, E., Casacuberta, F.: *Multimodal Interactive Pattern Recognition and Applications*. Springer (2011)
23. van de Weijer, J., Schmid, C.: Applying color names to image description. In: *IEEE International Conference on Image Processing (ICIP)*, San Antonio, USA (2007)
24. van de Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. *IEEE Transactions on Image Processing* 18(7), 1512–1524 (2009)
25. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2010)
26. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst.* 8(6), 536–544 (2003)

Coloresia: An Interactive Colour Perception Device for the Visually Impaired

Abel Gonzalez, Robert Benavente, Olivier Penacchio,
Javier Vazquez-Corral, Maria Vanrell, and C. Alejandro Parraga

Abstract. A significant percentage of the human population suffer from impairments in their capacity to distinguish or even see colours. For them, everyday tasks like navigating through a train or metro network map becomes demanding. We present a novel technique for extracting colour information from everyday natural stimuli and presenting it to visually impaired users as pleasant, non-invasive sound. This technique was implemented inside a Personal Digital Assistant (PDA) portable device. In this implementation, colour information is extracted from the input image and categorised according to how human observers segment the colour space. This information is subsequently converted into sound and sent to the user via speakers or headphones. In the original implementation, it is possible for the user to send its feedback to reconfigure the system, however several features such as these were not implemented because the current technology is limited. We are confident that the full implementation will be possible in the near future as PDA technology improves.

1 Introduction

Colour is an important feature of everyday life. Although highly saturated objects are not abundant in nature, we build and paint objects with highly saturated colours in an attempt to grab each other's attention, please each other, and transmit information. In the natural environment, colour helps organising scenes (blue is predominant in the sky, green in chlorophyll, brown in earth, grey in rocks, etc.) and crucially, it aids important survival tasks such as finding ripe fruit and leaves, detecting poisonous animals, and breaking luminance camouflage. In cities, colour highlights

Abel Gonzalez · Robert Benavente · Olivier Penacchio ·
Javier Vazquez-Corral · Maria Vanrell · C. Alejandro Parraga
Computer Vision Center / Computer Science Dept.,
Universitat Autònoma de Barcelona, Building O, Campus UAB, 08193 Bellaterra, Spain
e-mail: agonzgar@gmail.com,

[{robert, penacchio, jvazquez, maria, aparraga}@cvc.uab.cat](mailto:{robert,penacchio,jvazquez,maria,aparraga}@cvc.uab.cat)

or simplifies important information (red for danger or stop, green for way-out or go, fast identification of known products, understanding of train/metro maps, etc.) and this fact has been exploited to such degree that we are surrounded by advertising, fashion, traffic signalling, etc. that relies on colour to transmit distinctive visual information. However, colour processing is not an easy feat: years of research and technology development have shown that to extract reliable colour and texture information in lexical form from natural images is far from trivial. The main problems to be addressed are not related to the technology available (medium to high-quality colour portable digital cameras are ubiquitous nowadays) but instead are related to the way humans sample and perceive the wavelength distributions of visible light. The human visual system has several mechanisms to extract meaningful information from the light that reaches the eye, filtering out the less important, more redundant patterns. These include a bias towards representing the reflecting characteristics of objects rather than the chromatic content of the illumination (colour constancy) [32], a tendency to enhance or suppress the perceived richness (saturation) of a colour according to the variability of its extended surrounds [2], and several other mechanisms which alter the perceived hue of an object according to its immediate surroundings (chromatic induction) [3]. On top of this, there are various complex cultural issues that affect the way we transmit to others the information about what we perceive (language). For example, not everybody agrees on which semantic labels to assign to the same wavelength signal, and everybody is familiar with the experience of arguing about the colour of a piece of clothing or a newly painted wall. However, anthropologists have found a set of 11 basic colour terms that are common to most evolved cultures (white, black, red, green, blue, yellow, grey, brown, orange, pink, purple) [4] which are a good starting point to model the universal attributes of colour naming.

1.1 Colour Vision and Colour Visual Deficiencies

Colour is everywhere, and its very ubiquitousness and vividness makes us forget that it does not exist in the world "per se" but it is constructed by our brains from a few highly specialised neurons in our retinas. The delicate equilibrium of this neural construction becomes apparent when something goes wrong and our perception of the world becomes impaired. There are many forms of visual chromatic handicap, but some of the most common are impairments linked to deficiencies (or loss) of a given retinal photoreceptor. According to statistics compiled by the American Academy of Ophthalmology "*red-green colour vision defects are the most common form of colour vision deficiency. Approximately 8% of men and 0.5% of women among populations with Northern European ancestry have red-green colour defects. The incidence of this condition is lower in almost all other populations studied*" [5]. The rate of incidence of blue-yellow colour vision defects is the same for males and females (fewer than 1 in 10,000 people worldwide). Complete achromatopsia (a rare type of impairment where subjects do not see colours and only perceive shades of grey) affects an estimated 1 in 30,000 people. People with achromatopsia almost

always have additional problems with vision including reduced visual acuity, increased sensitivity to light (photophobia), etc. When visual acuity impairments are higher than 20/200 (10% of normal vision in Spain) or the visual field is less than 20 degrees in diameter, sufferers are considered legally blind. In the U.S., there are more than one million legally blind people aged 40 or older (0.3% of the population) and only 10% of those are totally blind [5] (see Figure 1).

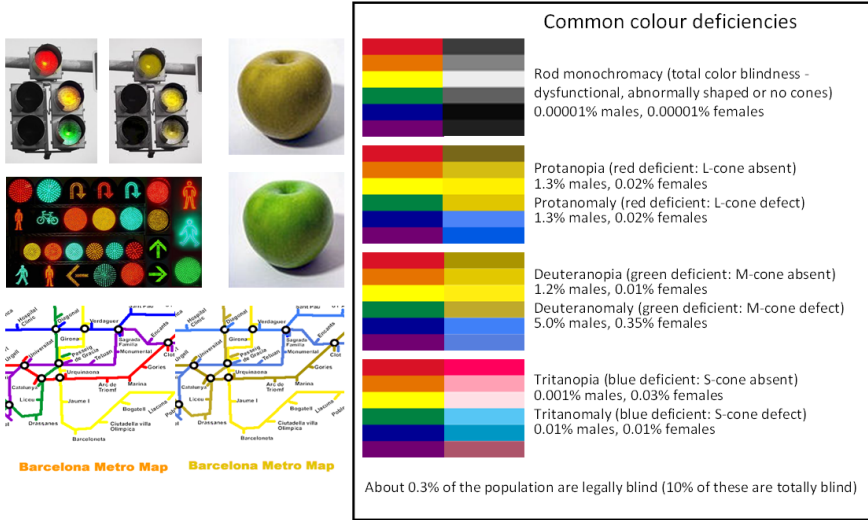


Fig. 1 Dichromats (People with impaired colour vision) find it difficult to perform basic tasks that involve detection and semantic labelling of different colours. These range from detecting danger signals at pedestrian crossings, discrimination of ripeness in fruit, discrimination of colour-coded train and tube lines in maps, etc. The right panel introduces some statistics about common deficiencies and their prevalence in the U.S. population [5]. (See color version of the figure at: <http://www.cic.uab.cat/Publications/>)

Visually impaired people face a number of everyday problems, ranging from the mild to the severe. In particular, they may experience problems recognising different bi-colour or tri-colour Light Emitting Diodes (LED) traffic lights and in some physical arrangements, the position may not be a cue to their colours, as in the case of horizontal traffic lights. There is also the inconvenience of not being able to navigate the coloured maps of motorways, trains and tube lines, either printed on paper or on electronic media. Dichromats also complain that other people "think that their choice of colours is strange" and that they cannot tell whether a piece of meat is raw or well done, or if a fruit is mature among other everyday problems.

1.2 *Perceptual Interaction between Colour and Sound*

Hearing is arguably the second most important way by which humans sense information about the world, and consequently sound is another important feature of everyday life. As with colour, we use sound to capture each other's attention, transmit information, and please each other.

Although they are processed by mainly separate neural mechanisms (and therefore studied by different disciplines), there is evidence that the mammalian visual and auditory systems may have many areas of overlapping. For instance, both systems share the ability to determine the speed and direction of a moving object, and to produce a unified percept of movement. Therefore, both types of sensory information have to merge or coordinate at some point. In addition, both systems have to coordinate and interact to direct attention to one modality or the other to control subsequent action [20]. More evidence of this neural mechanism overlap is provided by the involuntary cross-activation of the senses that occurs for a handful of individuals, in sound-colour synaesthesia, where auditory sensations spontaneously elicit visual experience. For example, when a key is struck on a piano a sound-colour synaesthete experiments a vivid colour sensation (see [42]) and this sensation may be different if another key is struck. However, if the same note is played the sensation elicited is internally very consistent over time. Many musicians experience this phenomenon [41].

Although individuals with sound-colour synaesthesia differ in their cross-modal associations, the sound-to-colour mapping they experience is not necessarily arbitrary. For example, the vast majority of them associate high pitch with light colour [42]. In addition, both non-synaesthete and synaesthete people share the same heuristics for matching colour and sound. The difference is that the cross-modal sensation is elicited involuntary for synaesthetes, whereas it involves a conscious initiative/effort for non-synaesthetes. All in all, it seems that sound-colour synaesthesia uses some common mechanisms of cross-modal perceptual interaction [42]. Accordingly, sound-colour cross-modal perception by synaesthetes is of interest for defining a colour-to-sound correspondence because it seems not to recruit privileged pathways between auditory and visual modalities.

Indeed extreme cases of synaesthesia are rare, however researchers studying how the brain combines information from different sensory modalities (i.e. cross-modal perception and multisensory integration) hypothesise whether it might be the case that all humans are synaesthetes to some degree and whether these naturally biased correspondences may influence the development of language [19].

Synaesthetic individuals seldom complain about their condition, and in many cases they claim that their lives have been enhanced by this ability to relate colour to sound or haptic information. This apparent "enhancement" has motivated us to apply current multimodal interactive techniques to deliver the information that is missing in one sense (vision) as a pleasant stream to other sense (hearing). In other words, we created a portable device (Android platform) that extracts semantic colour information from images in a manner compatible with the human visual system and conveys this information as a pleasant stream of music which does not overwhelm or

both the user (see figure 2). We also wanted to make the device “interactive”, i.e. capable of receiving input from the user and “adaptive”, i.e. capable of learning from the user input to improve its inherent properties. Unfortunately, some of the work towards this aim was not implemented in the prototype due to current limitations of the portable device technology. However, we are confident that at the current rate of technological improvement suitable devices will be available in the near future.

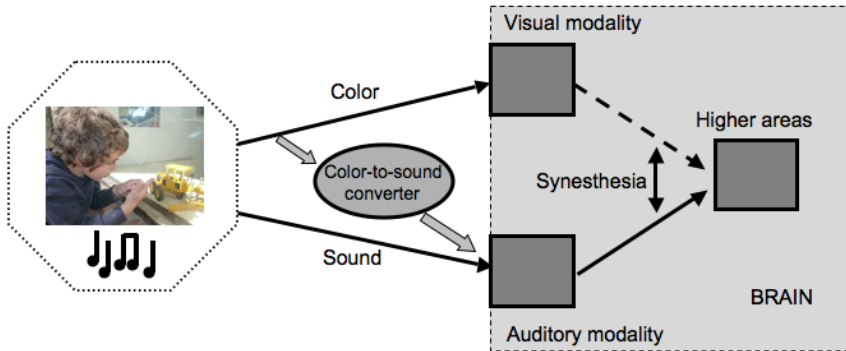


Fig. 2 Similarly to what happens in synaesthesia, the developed device converts colour information to sound

2 State of the Art

Until recently, conversion from colour to sound and vice versa had received more attention from visual arts than from science. Several techniques aiming to convert sounds or music to a visual presentation are included in what is known as visual music [24]. Although these first approaches did not exactly transform colour into sound, they were a first step towards the goal of expressing colour as music (see [21] for an historical account of colour-to-sound correspondences). In the last years, the idea of implementing aid devices for helping blind and visually impaired people to perceive colour through the representation of colour as music has received an increasing attention.

Cronly-Dillon *et al.* [25] showed the viability of representing some features from an image using music to describe its content to blind people. Their method selected different features of an image and represented each of them with a sound. The sounds for each part of the image were combined as a polyphonic melody that encoded the basic content of the image. Their experiments showed that blind people were able to interpret some images by hearing their associated melodies.

Following a similar line, Bologna *et al.* [26] proposed a method to transform coloured pixels into musical notes in order to describe image content for blind users. To this end, hue was divided in several sectors and was represented by timbre (see below), saturation was divided in four levels and was represented with different

notes, and luminosity was represented by bass for dark colours and a singing voice for bright colours. Using this transform, the input image was segmented and the sounds corresponding to the colours of the main parts of the image were reproduced. Bologna *et al.* also proposed to use saliency detection techniques to focus the description on the most salient parts of the image.

A similar idea was proposed by Rossi *et al.* [27], who developed a prototype of a device that transformed colours into melodies. The system was developed as a game for children and was implemented in a portable bracelet with a small camera installed on a pointer that allowed users to select any point of the scene. The system was able to identify six colours (red, green, blue, yellow, purple, and orange) by dividing the hue circle of the HSV colour space in six sectors. Each of these colours was assigned to a musical instrument that played a melody that could be chosen from a set of five melodies. Additionally, for each colour, three to five divisions were set on the value dimension, and each of these subdivisions was identified by a different tone. Black and white were also considered as additional cases on this system. As in the approach of Bologna *et al.* the initial identification of colour names was not perceptual and this fact might be a drawback of both systems.

The approach which is closer to our purpose is the one by the visual artist and composer Neil Harbisson [22]. Harbisson suffers from achromatopsia, a visual condition that allows him to only see the world in shades of grey. To overcome his lack of colour perception, he designed a device called Eyeborg, which consists of a sensor that he wears on his head and points towards the direction he is looking at. Using a chip fixed to the back of his neck, the frequencies of light are converted into audible frequencies, which he interprets as a colour scale. Harbisson has developed two different conversion algorithms. The first one directly transforms seven light frequency ranges into seven sound frequencies. His second approach, divides the light frequency scale in 12 ranges corresponding to different colours and converts them in 12 musical notes. Both methods result in unpleasant and even heady sounds.

As we stated in the introduction, our goal is to develop a personal assistant implemented on a mobile device running under the Android platform. Several applications that acquire images with the device camera and are related to colour detection and identification can be found at the online shop for the Android platform, *Google Play* [23]. Some examples are ‘*This Color What Color?*’, ‘*Color Detector*’, ‘*Color Picker*’, and ‘*Color Blend*’. Although some of them give the name of the colours using synthesised voice, to the best of our knowledge, there is no application implementing a colour-to-sound transform algorithm to specifically aid visually impaired people.

3 From Colour Signal to Sound

We have built a prototype for colour name extraction that is able to, given a digital image, provide a list of the main colours of the objects present in it, in a manner consistent with the behaviour of human observers (see prototype schematics in Figure 3).

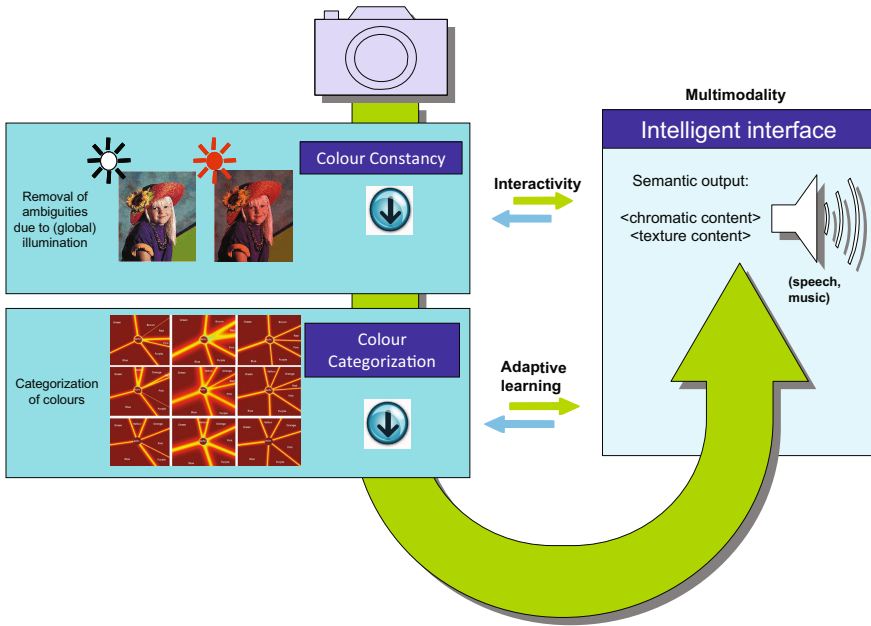


Fig. 3 Schematics of the prototype (See color version of the figure at: <http://www.cic.uab.cat/Publications/>)

The prototype is able to communicate this information to a visually impaired user in two modalities: words and music. The definition of visually impaired here ranges from dichromats to low vision or even blind users. In other words, we have built a portable system that acts on the output of a digital camera and reproduces the basic mechanisms that a human observer employs to identify the names of the colours of the objects present in the scene. The colour names are communicated to the user by means of synthesised music or alternatively, an automated voice system. We achieved this aim by:

- developing a human-based colour perception model to account for changes in perceived chromatic characteristics of the illuminant.
- developing a set of image descriptors to identify and label the main colours in images, in a manner similar to human observers.
- developing an interface based on natural language that is able to handle colour names.
- developing an interface based on sound that is capable to convert colour names into music

Our prototype was conceived as a portable device, based on a state-of-the-art personal digital assistant (PDA) with an embedded digital camera. Such devices are relatively inexpensive and provide the necessary capabilities to develop a

software-based model that uses the digital camera (input device) as a first stage and delivers its results through the sound system (speakers/headphones). They have also an adequate user interface hardware (touch screen) for entering the necessary user corrections to improve the colour-naming algorithm. Figure 4 provides the schematics of the prototype design. The input data comes via the PDA camera’s uncalibrated camera and the system applies an illumination removal algorithm to produce an image free of the colouring imposed by the illuminant. We use this representation to classify the content of the scene according to its colour names. The output of this algorithm comes in two alternative forms: as a voice through a voice synthesiser/speaker combination or as music.

In the following sections we explain in more detail the physical and perceptual properties of colour and sound that we are about to simulate and manipulate to achieve the “sonification” of the image, i.e. the transfer of colour information to the auditory system.

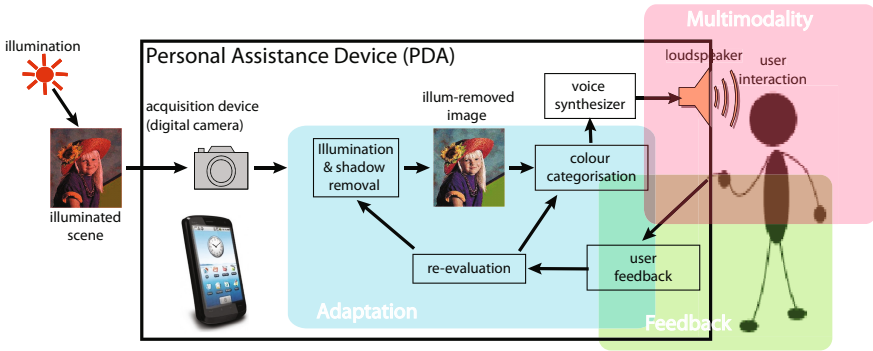


Fig. 4 Feedback, multimodality and adaptation and their role in the prototype (See color version of the figure at: <http://www.cic.uab.cat/Publications/>)

3.1 Properties of Colour

The wavelength content of the electromagnetic radiation that reaches our eyes is sampled in the retina by specialised neurons (cones), converted into neural information and transferred throughout different stages in the visual pathway. In the latest stages, the information is categorised. Categorisation is the process by which objects are differentiated and grouped, softening differences and favouring similarities among them, reducing an extremely complex world into cognitively tractable proportions. This reduction is extremely evident in the colour domain: from the nearly 2 million colours that can be distinguished perceptually we recover only about 30 colour categories which can be named by average subjects [7]. Although many colours can be distinguished and named, there is a group of 11 colour categories that are common to all advanced languages. They were defined by Berlin and Kay in their seminal work [4] and are thought to be inherent of the human neural machinery of colour categorisation [16, 17, 18]. These are black, white, red, green, yellow,

blue, brown, purple, pink, orange, and grey, and they appear in a language in this particular order as the language becomes more complex. More complex languages tend to have more categories, but these are the most primitive.

To model this categorisation process as accurately as possible is a goal of many disciplines, from colour image reproduction to computer vision. Recent computational models of colour space segmentation are based on either natural scene statistics [8] or psychophysical data [9, 10, 11, 12, 13, 14, 15]. We implemented a colour space segmentation model on the model of Benavente *et al.* [11] because it has several advantages over others: it is implemented in CIELab colour space (a perceptually uniform space that has its lightness dimension built from relative luminance) and is parametric, i.e. can be easily adjusted depending on the user feedback. The model is built from fuzzy sets segmenting CIELab space in 11 regions and in its current implementation, it assigns to each pixel $\mathbf{p} = (L, a, b)^T$ a membership value between 0 and 1 to each colour category. Hence, for each pixel \mathbf{p} , a 11-dimensional colour descriptor $CD(\mathbf{p})$ is defined as

$$CD(\mathbf{p}) = [\mu_{C_1}(\mathbf{p}), \dots, \mu_{C_{11}}(\mathbf{p})] \quad (1)$$

where each component of this 11-dimensional vector describes the membership of \mathbf{p} to a specific color category and the component with highest membership value determines to which category the pixel belongs.

The value of each of the components of the colour descriptor is obtained from a triple-sigmoid with elliptical center (TSE) function given by

$$TSE(\mathbf{p}, \theta) = DS(\mathbf{p}, \mathbf{T}, \theta_{DS})ES(\mathbf{p}, \mathbf{T}, \theta_{ES}), \quad (2)$$

where ES is an elliptical-sigmoid function which models the central achromatic region and is defined as

$$ES(\mathbf{p}, \mathbf{T}, \theta_{ES}) = \frac{1}{1 + \exp\left\{-\beta_e \left[\left(\frac{\mathbf{u}_1 \mathbf{R}_\phi \mathbf{T} \mathbf{p}}{e_x} \right)^2 + \left(\frac{\mathbf{u}_2 \mathbf{R}_\phi \mathbf{T} \mathbf{p}}{e_y} \right)^2 - 1 \right] \right\}}, \quad (3)$$

and DS is a double-sigmoid function defined as the product of two oriented 2D-sigmoids given by

$$DS(\mathbf{p}, \mathbf{T}, \theta_{DS}) = S_1(\mathbf{p}, \mathbf{T}, \alpha_y, \beta_y) S_2(\mathbf{p}, \mathbf{T}, \alpha_x, \beta_x) \quad (4)$$

$$S_i(\mathbf{p}, \mathbf{T}, \alpha, \beta) = \frac{1}{1 + \exp(-\beta \mathbf{u}_i \mathbf{R}_\alpha \mathbf{T} \mathbf{p})}, \quad i = 1, 2 \quad (5)$$

In equations 2 to 5, $\theta = (\mathbf{t}, \theta_{DS}, \theta_{ES})$, θ_{DS} , and θ_{ES} are the set of parameters of the TSE, the DS, and the ES functions, respectively, \mathbf{T} is a translation matrix, \mathbf{R}_ϕ is a rotation matrix of angle ϕ , $\mathbf{u}_1 = (1, 0, 0)^T$, $\mathbf{u}_2 = (0, 1, 0)^T$, e_x and e_y are the semiminor and semimajor axis of the central ellipse, β_e is the slope of the sigmoid curve that forms the central ellipse boundary, α_i is an angle with respect to axis i , β_i

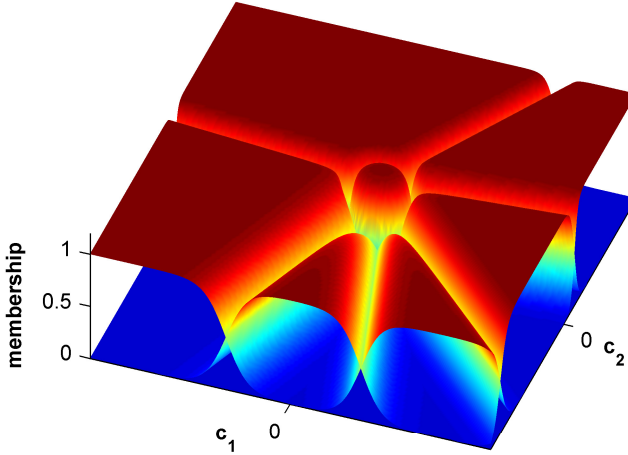


Fig. 5 TSE function fitted to the chromatic categories defined on a given lightness level. In this case, only six categories have memberships different than zero. (See color version of the figure at: <http://www.cic.uab.cat/Publications/>)

is the slope of a sigmoid function defined over axis i , and R_α is a rotation matrix of angle α .

Figure 5 shows an example of how the model divides a specific chromatic plane of the CIELab space.

3.2 Colour Constancy

Colour constancy is usually defined as the tendency of objects to appear the same colour even under changing illumination [28]. This is important due to the big variability of illumination in our real life (indoor/outdoor situations, midday/sunset daytime, etc.) For example, we will perceive as white a white piece of paper both in an indoor scenario or in an outdoor scenario at midday. However the information reaching the eye will be yellowish in the first case (tungsten illumination) and bluish in the second one. Several studies widely agree that human colour constancy is not based on a single mechanism [29].

In computational colour we simplify the human colour constancy property to convert it into a tractable problem. In particular, computational colour constancy tries to convert the captured scene under an unknown illumination into the same scene viewed under a white illumination (that is, we suppose that under white light, the perceived colours mimic the physical values). From a mathematical point of view, the problem is regarded as the search of a 3×3 matrix. However, for simplicity, researchers have widely used the Von Kries model [30] to simplify the problem. Von Kries model states that illumination change is a process which operates in each sensor response channel independently. Then, the 3×3 original matrix is converted to

a diagonal one, greatly simplifying colour constancy computation. Mathematically, let us suppose we have an object with reflectance $S(\lambda)$ viewed under two illuminants $E_1(\lambda)$, $E_2(\lambda)$, and captured by a camera with sensitivities $R_i(\lambda)$, $i \in \{1, 2, 3\}$. Then, the colours captured by the camera are denoted as $\underline{\rho}^1$ and $\underline{\rho}^2$, where their components are given by

$$\begin{aligned}\rho_i^1 &= \int S(\lambda)E_1(\lambda)R_i(\lambda)d\lambda \\ \rho_i^2 &= \int S(\lambda)E_2(\lambda)R_i(\lambda)d\lambda\end{aligned}\quad (6)$$

Then, in computational colour constancy we search for α, β , and γ fulfilling

$$\rho^1 = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \gamma \end{pmatrix} \cdot \rho^2 \quad (7)$$

There are several methods trying to solve for this equation. The simpler ones (that actually give quite good results in real databases) are Grey-World [31] and MaxRGB [32]. Basically, GreyWorld assumes that the average of the scene is grey, while MaxRGB assumes the highest intensity values of the scene as a white point. These two methods were generalised by Shades-of-Gray [33] where the Minkowski norm was added and Grey-Edge [34] where image derivatives were also added. Some other methods deal with physical properties, such as mutual reflections [37], highlights and shading [36], and specular highlights [35]. Finally, another set of colour constancy methods are probabilistic such as Color-by-Correlation [38] and Illumination-by-Voting [39].

Recently, a new voting method [40] has been defined. This method follows the category hypothesis: Feasible illuminants can be weighted according to their ability to anchor the colours of an image to basic colour categories. In particular, it chooses the focals of colour names to behave as anchor categories. In this way, it returns as a solution the scene maximising the number of nameable colours. For example, if we have an outdoor scene in a field, it will return the image that converts both the sky and the green colours into the prototypical blue and green that have evolved with humans. Due to the naming nature of this approach, it would be the most suitable for our system, however, for limitations of the current mobile devices, a simpler method, the MaxRGB algorithm, has been used as a preprocessing step.

3.3 *Properties of Sound*

Physically, sound corresponds to mechanical vibrations transmitted through an elastic medium (gas, liquid, or solid) and is composed of longitudinal waves characterised by their frequency (or wavelength) and amplitude. Humans with normal hearing are capable of perceiving frequencies between 20 and 20,000Hz and

intensities within a range of 12 orders of magnitude. When talking about sound, we refer to wavelength frequency as *pitch* and amplitude as *loudness* and interpret sound as a perceptual experience, in a way similar to how we interpret colour. When a key on a piano is struck, for example, we can identify both the pitch and loudness of the sound produced. The pitch is well defined and corresponds to physical properties of the wire struck (tension, linear mass density, and length), therefore we construct instruments manipulating these properties to produce different pitches. We can produce a louder sound by giving the key a bigger pull. In that case, the amplitude of the vibrations of the corresponding wire is bigger. Other attributes of sound events are *duration*, *spatial position* and *timbre*. Duration simply refers to the time span of a single sound event. On the other hand, the auditory system is capable of discerning the spatial localisation of a sound source. Localisation of sound events is by far less precise than localisation of objects by the visual system but not limited by the lighting conditions and in addition, hearing is omnidirectional.

By asking human subjects to tell the difference or express similarity judgement when listening to different sound excerpts corresponding to different musical instruments, one can derive timbre spaces. These spaces are perceptual and represent similarities between sounds. They are the counterpart in psychoacoustics of the perceptual colour spaces in vision, which are derived using psychophysics. However, giving a constructive definition of timbre is not easy and instead, timbre is often referred to a combination of qualities of sound that allow the distinction between sounds of the same pitch and loudness. To put it plainly, timbre is what allows us to tell the difference between a piano and a cello when both are playing the same note (pitch) with the same loudness (for the same duration and at the same position). Unlike pitch and loudness, which are characterised by frequency and amplitude, there is no single physical characteristic that directly relates to timbre. However, the main attributes of timbre are harmonic content and dynamic characteristics such as vibrato and the intensity *envelope* (attack, sustain, release, and decay).

3.4 Colour Sonification: Our Proposal

The central question is to find a systematic way to encode colour into sound. Such a mapping should have the following features:

- i easy to use
- ii not heady
- iii coherent with synaesthesia (main features of)
- iv perceptual isometry

Let us explain property (iv) in greater detail. Let \mathcal{C} be a perceptual colour space and \mathcal{S} a sound space. Suppose now that both spaces are endowed with a perceptual metric (denoted by $\|\cdot\|_{\mathcal{C}}$ and $\|\cdot\|_{\mathcal{S}}$, respectively). A mapping $\Phi : \mathcal{C} \rightarrow \mathcal{S}$ is said to be a *perceptual isometry* if the following property holds: for any two colours C_1, C_2 in \mathcal{C} , if $\|C_1 - C_2\|_{\mathcal{C}} = T_{\mathcal{C}}(C_1, C_2)$, where $T_{\mathcal{C}}(C_1, C_2)$ is the discrimination threshold

in the region of C_1, C_2 in \mathcal{C} , then $\|\Phi(C_1) - \Phi(C_2)\|_{\mathcal{S}} = T_{\mathcal{S}}(\Phi(C_1), \Phi(C_2))$, where $T_{\mathcal{S}}(\Phi(C_1), \Phi(C_2))$ is the discrimination threshold at $\Phi(C_1)$ in \mathcal{S} . Such a property would ensure no loss of discriminative power in the translation of colour into sound.

The first step in the construction of a timbre space is the extraction of physical characteristics. Sound events are expressed in terms of several time-frequency representations (harmonic sinusoidal components, short-term Fourier transform, energy envelope). Next, a large number of descriptors are derived which capture spectral, temporal, spectrotemporal, and energetic properties of sound events [43]. The information provided by these descriptors is highly redundant. Often, multidimensional scaling is applied to the space of descriptors to get a 3D space. The acoustic correlates of the three dimensions vary from a proposal to another. The spectral centroid receives a wide support in the literature and is often considered as the first and principal dimension (see [44] for a review on this issue). Another important dimension is provided by the attack time. The temporal variation of the spectrum is often adopted as the third dimension, but is less consensual. Note that describing sound using a three dimensional space \mathcal{S} is a requisite if we are to define a perceptual isometry from a three dimensional colour space \mathcal{C} to \mathcal{S} . Both spaces should have the same dimension.

For computational reasons, we have implemented a simplified approach of the colour sonification which is mainly based on pitch for characterising sound. The input to the sonification algorithm is the output of the colour naming model described in section 3.1, that is, an 11-dimensional vector containing the membership values to the eleven colour categories considered. Hence, a colour is described by the 11 membership values of the colour naming descriptor.

In our approach, each chromatic colour category¹ is characterised by a different pitch (note) of a violin sound. The loudness of the sound is varied according to the membership value of the pixel to each colour category. To avoid noise, only membership values higher than 0.1 are considered. Therefore, given a colour, the generated sound will be a mixture of the sounds corresponding to the categories with membership values higher than 0.1, with different loudness each.

To differentiate between chromatic and achromatic² categories, timbre is used. Thus, achromatic colours are converted to a violoncello sound instead of the violin sound used to represent the chromatic categories. The differentiation among the three achromatic categories is done by assigning a specific pitch (note) to each of them: black is mapped to note C (do), grey is mapped to F (fa), and white is mapped to B (si). Table 1 summarizes the colour sonification scheme used.

Finally, the lightness of the colour, which depends on the value of CIELab coordinate L , is represented by different octaves. Hence, the lightness axis L is divided in two parts (low/high lightness) and colours in each part are represented by sounds on a specific octave.

¹ Red, green, yellow, blue, brown, purple, pink, and orange.

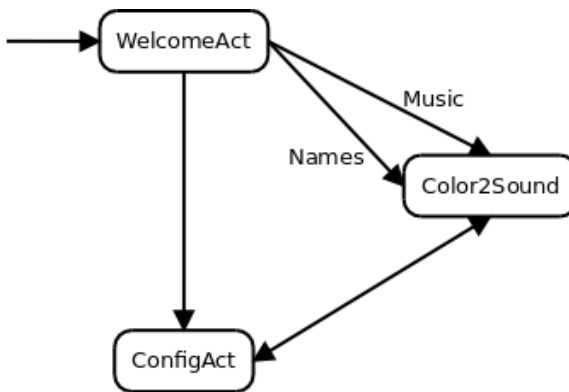
² Black, white, and grey.

Table 1 Summary of the conversion provided by the colour sonification algorithm

Colour	Pitch (note)	Timbre (instrument)
pink	E	violin
purple	D	violin
blue	C#	violin
green	A	violin
yellow	G#	violin
brown	G	violin
orange	F#	violin
red	F	violin
white	B	violoncello
grey	F	violoncello
black	C	violoncello

4 A Multimodal Device for the Visually Impaired

The mobile application developed is called **Coloresia** (i.e. a mixture between the words color and synaesthesia) and has three main modules, which are implemented as an Android activity³. WelcomeAct shows the initial interface of the application, Color2Sound is the main activity of the application and performs most of the tasks, such as acquiring images from the camera, displaying information on screen, or playing sounds, and ConfigAct allows the user to control the configuration of the application. Figure 6 shows a module diagram of the three activities of the application.

**Fig. 6** Schematics of the main modules of the mobile application Coloresia

³ In the Android platform, activities denote the basic components of applications. An activity corresponds to an interface of the application where the user can do some actions.

When the application is started, the user accesses to WelcomeAct, the initial activity of the application, which presents three buttons to the user. Two of these buttons take the user to the colour identification application in the two available modes, namely, music and voice. The third button calls the configuration module where the user can set different parameters of the application.

Figure 7(a) shows the interface of this initial activity. As it can be seen, the interface has been designed to facilitate the accessibility to visually impaired people: a large size font and colours with high differences in lightness have been used to highlight the text and make it easy to read.

From the WelcomeAct activity, the user can access to Color2Sound, the main activity of the application. When Color2Sound is started, the application acquires a sequence of images with the device camera and displays them on the screen. On one out of two frames of the sequence a region of interest (ROI) on the center of the image is selected. The dimensions of the ROI can be set by the user in the configuration activity.

The pixels' values in the ROI are averaged to obtain the mean RGB of the region. This mean RGB is the input to the colour naming method explained in section 3.1 to obtain the 11-dimensional vector with the membership values to the 11 colour categories considered. Then, this 11-dimensional vector is the input to the colour sonification algorithm presented in section 3.4.

Finally, the result of the conversion algorithm, i.e. a sound defined as a mixture of notes played by one or two instruments, is played on the device to allow the visually disabled users to know the colour of the objects at the center of the images they are acquiring with their device.

Besides the final sound played by the application, it also provides some information displayed on the screen of the device. This information is:

- The rectangle containing the region of interest.
- The colour name with the highest membership value corresponding to the mean RGB in the ROI.
- The mean RGB and CIELab values in the ROI.

Figure 7(c) shows the interface of the Color2Sound activity with all the information displayed on screen while the activity is working.

The Color2Sound activity also captures the events generated by the user on the touch screen. While this activity is working, the user can move the ROI through the image to identify the colour of a different image area. The user can also modify the size of the ROI, which can be set between a minimum size of 4×4 pixels and a maximum of 16×16 . The size of the ROI can also be modified at the configuration activity as detailed below.

The user can also access the application menu from the menu key of the device. The options in this menu allow the user to save images on the device memory card, to access the configuration tool, to change the operation mode, and to exit the application. Figure 7(d) shows a screen shot of the menu layout.

The last module of the application is the configuration activity ConfigAct. In this activity, the user can set the three main parameters of the application. The first one is

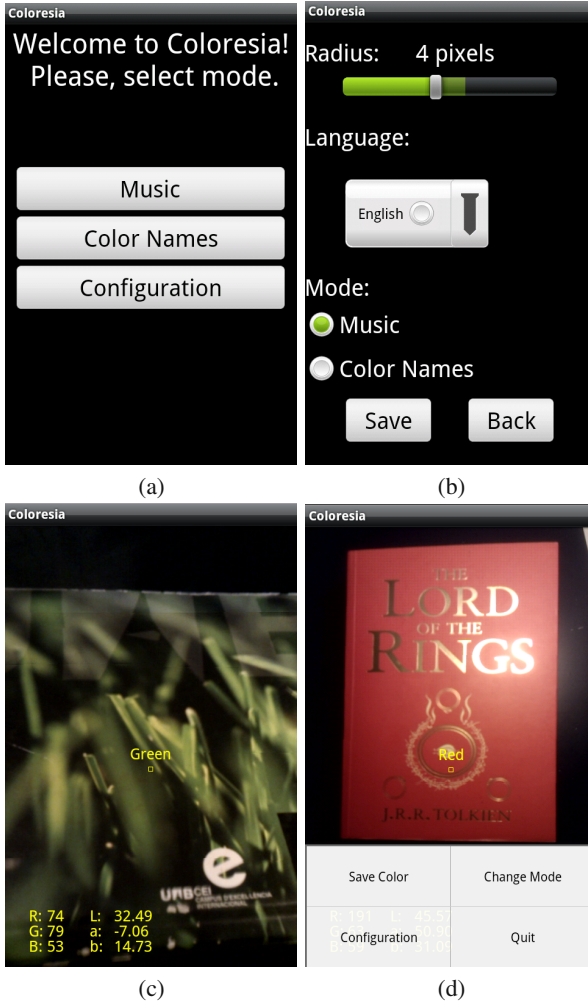


Fig. 7 Coloresia interfaces. (a) WelcomeAct activity. (b) ConfigAct activity. (c) Main interface of the Color2Sound activity. (d) Auxiliary menu of the Color2Sound activity. (See color version of the figure at: <http://www.cic.uab.cat/Publications/>)

the radius of the region of interest, with a minimum of 2 pixels (i.e. a 4×4 window) and a maximum of 8 pixels (i.e. a 16×16 window). The value of this parameter can be adjusted by means of a sliding bar.

The second parameter is the language of the application. The selected language will be used in all the messages at the interface and by the voice synthesiser. The selection can be done by a *spinner* among the three supported languages: English, Spanish, and German. By default, English is initially selected. If the device does not have the language selected by the user installed on the device, the application

proceeds to its installation. If, for any reason, the installation is not possible, the application warns the user by a message on the screen.

The third parameter that can be modified is the operation mode, where the user can choose between the default music output to represent the colours or a voice indicating the colour name of the stimulus detected by the application.

Finally, ConfigAct has two buttons to save the settings or going back discarding the changes. Figure 7(b) shows the layout of the activity that follows the same aesthetics as the previous activities.

4.1 Test and Results

The application has been tested on a HTC Desire mobile, with operative system Android v.2.2, a 1GHz processor, and 576Mb of RAM memory. The test of the application has been focussed on the processing time and the robustness against illumination conditions.

To test the speed of the colour identification part, the processing time of the 30 first detections on each test were averaged. The mean processing time was 123.18ms, with a standard deviation on 74.89ms. The test was only performed on the first executions to test the worst case, because after the initial colour detections the processing times reduce considerably to a mean processing time of 90ms.

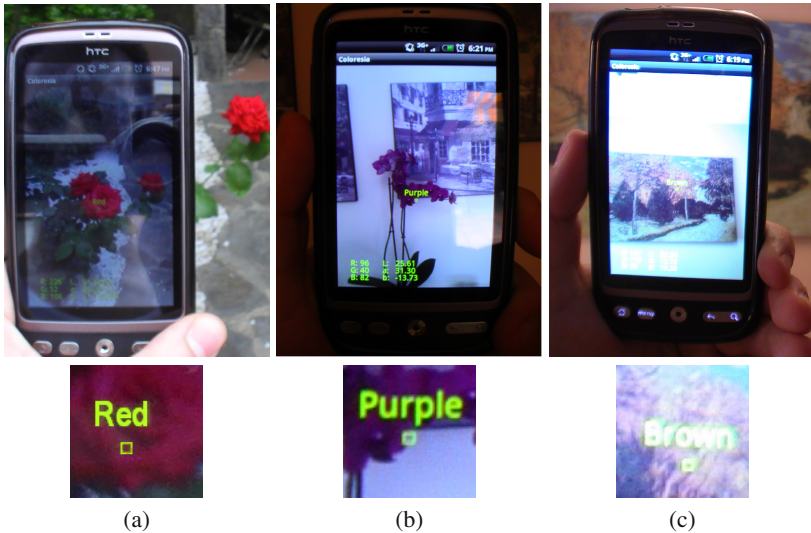


Fig. 8 Examples of detections performed by the application. In the lower row, the central part of each image is zoomed. (a) Under natural daylight. (b) Under a reddish tungsten bulb light. (c) Under a mixture of daylight and tungsten bulb light. (See color version of the figure at: <http://www.cic.uab.cat/Publications/>)

Regarding the robustness against illumination changes, the application has been tested on three different illumination conditions: daylight, reddish tungsten bulb light, and a mixture of both. Although the application has more problems with low-illuminated environments, the application is able to correctly describe colours in most cases on the tested illumination conditions. Figure 8 shows three examples of the application working under the three illumination conditions.

5 Conclusions

In this chapter we have presented a prototype to help visually impaired people who is not able to see colour properly. The application is implemented on a mobile device and acquires images with the device camera. From this image, a region of interest is selected and the mean colour of the region is converted to a sound that is played by the device. Therefore, the users of this application are able to interpret these sounds and can identify the colours in the scene.

The method to represent colour as a musical sound is based on two steps. The first one transforms the input colour stimulus to a 11-dimensional vector representing the membership value of the colour to the eleven basic colour categories. The second step converts each membership value to a sound, and these sounds are combined to produce the final output of the system. From this output, the user can interpret the colour of the stimulus he or she has in front of.

With this application colour-blind people have an easy-to-use and low-price assistant for everyday tasks such as choosing the clothes to wear, understanding an underground map, or even interpreting a piece of art.

Acknowledgements. Authors are grateful for support from TIN 2010-21771-C02-1 and ConsoliderIngenio 2010 CSD2007- 00018 of Spanish MEC (Ministry of Science). They also acknowledge support from GRC 2009-669 of Generalitat de Catalunya. OP acknowledges Perfecto Herrera-Boyer and Emilia Gómez for their input during the preparation of the manuscript.

References

1. Foster, D.H.: Color constancy. *Vision Research* 51, 674–700 (2011)
2. Brown, R.O., MacLeod, D.I.A.: Color appearance depends on the variance of surround colors. *Current Biology* 7, 844–849 (1997)
3. Otazu, X., Parraga, C.A., Vanrell, M.: Towards a unified model for chromatic induction. *Journal of Vision* 10(12), article 5 (2010)
4. Berlin, B., Kay, P.: *Basic color terms: their universality and evolution*, Berkeley, Oxford (1969)
5. Genetics Home Reference: Color vision deficiency, National Library of Medicine, <http://ghr.nlm.nih.gov/condition/color-vision-deficiency>
6. Vision Problems in the U.S.: Prevalence of Adult Vision Impairment and Age-Related Eye Disease in America. Prevent Blindness America and the National Eye Institute (2008), http://www.preventblindness.org/vpus/2008_update/

7. Derefeldt, G., Swartling, T.: Color Concept Retrieval by Free Color Naming - Identification of up to 30 Colors without Training. *Displays* 16, 69–77 (1995)
8. Yendrikhovskij, S.N.: A Computational Model of Colour Categorization. *Color Research and Application* 26, S235–S238 (2001)
9. Seaborn, M., Hepplewhite, L., Stonham, J.: Fuzzy colour category map for the measurement of colour similarity and dissimilarity. *Pattern Recognition* 38, 165–177 (2005)
10. Mojsilovic, A.: A computational model for color naming and describing color composition of images. *IEEE - Transactions on Image Processing* 14, 690–699 (2005)
11. Benavente, R., Vanrell, M., Baldrich, R.: Parametric fuzzy sets for automatic color naming. *Journal of the Optical Society of America A* 25, 2582–2593 (2008)
12. Menegaz, G., Troter, A.L., Sequeira, J., Boi, J.M.: A discrete model for color naming. *EURASIP J. Appl. Signal Process.* 2007(1), 113 (2007)
13. Wang, Z., Luo, M.R., Kang, B., Choh, H., Kim, C.: An Algorithm for Categorising Colours into Universal Colour Names. In: 3rd European Conference on Colour in Graphics, Imaging, and Vision (2006)
14. Hansen, T., Walter, S., Gegenfurtner, K.R.: Effects of spatial and temporal context on color categories and color constancy. *Journal of Vision* 7 (2007)
15. Moroney, N.: Unconstrained web-based color naming experiment. In: *SPIE Color Imaging VIII: Processing, Hardcopy, and Applications* (2003)
16. Boynton, R.M., Olson, C.X.: Saliency of Chromatic Basic Color Terms Confirmed by 3 Measures. *Vision Research* 30, 1311–1317 (1990)
17. Hardin, C.L., Maffi, L.: Color categories in thought and language. Cambridge University Press, Cambridge (1997)
18. Webster, M.A., Kay, P.: Individual and population differences in focal colors. In: MacLaury, R.E., Paramei, G.V., Dedrick, D. (eds.) *Anthropology of Color: Interdisciplinary Multilevel Modeling*, pp. 29–54. J. Benjamins Pub. Co., Amsterdam (2007)
19. Maurer, D., Pathman, T., Mondloch, C.J.: The shape of boubas: sound-shape correspondences in toddlers and adults. *Developmental Science* 9, 316–322 (2006)
20. Lewis, J.W., Beauchamp, M.S., DeYoe, E.A.: A comparison of visual and auditory motion processing in human cerebral cortex. *Cerebral Cortex* 10, 873–888 (2000)
21. Visual Music by Maura McDonnell (2002), <http://homepage.tinet.ie/~musima/visualmusic/visualmusic.htm>
22. Neil Harbisson. Sonochromatic cyborg, <http://www.harbisson.com> (cited July 01, 2012)
23. Google Play, <http://play.google.com/store> (cited July 01, 2012)
24. Evans, B.: Foundations of a visual music. *Computer Music Journal* 29, 11–24 (2005)
25. Cronly-Dillon, J., Persaud, K., Gregory, R.P.F.: The perception of visual images encoded in musical form: a study in cross-modality information transfer. *Proc. Roy. Soc. B* 266, 2427–2433 (1999)
26. Bologna, G., Deville, B., Pun, T., Vickenbosch, M.: Transforming 3D coloured pixels into musical instrument notes for vision substitution applications. *EURASIP J. Im. Video Process.* 2007, 76204 (2007)
27. Rossi, J., Perales, F.J., Varona, J., Roca, M.: COL.diesis: transforming colour into melody and implementing the result in a colour sensor device. In: *International Conference on Information Visualisation* (2009)
28. Hurlbert, A.: Colour constancy. *Current Biology* 21(17), 906–907 (2007)
29. Hurlbert, A., Wolf, K.: Color contrast: a contributory mechanism to color constancy. *Progress on Brain Research* 144 (2004)

30. Worthey, J.A., Brill, M.H.: Heuristic analysis of von kries color constancy. *Journal of the Optical Society of America A* 3, 1708–1712 (1986)
31. Buchsbaum, G.: A spatial processor model for object colour perception. *Journal Franklin Institute* 310, 1–26 (1980)
32. Land, E.H.: The retinex. *American Scientist* 52, 247–264 (1964)
33. Finlayson, G.D., Trezzi, E.: Shades of gray and colour constancy. In: *Color Imaging Conference* (2004)
34. van de Weijer, J., Gevers, T., Gijssenij, A.: Edge-based color constancy. *IEEE Transactions on Image Processing* 16, 2207–2214 (2007)
35. Lee, H.: Method for computing the scene-illuminant chromaticity from specular highlights. *Journal of the Optical Society of America A* 3, 1694–1699 (1986)
36. Klinker, G., Shafer, S., Kanade, T.: A physical approach to color image understanding. *International Journal of Computer Vision* 4, 7–38 (1990)
37. Funt, B.V., Drew, M.S., Ho, J.: Color constancy from mutual reflection. *International Journal of Computer Vision* 6, 5–24 (1991)
38. Finlayson, G.D., Hordley, S.D., Hubel, P.M.: Color by correlation: A simple, unifying framework for color constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 1209–1221 (2001)
39. Sapiro, G.: Color and illuminant voting. *IEEE Transactions on Image Processing* 21, 1210–1215 (1999)
40. Vazquez-Corral, J., Vanrell, M., Baldrich, R., Tous, F.: Color Constancy by Category Correlation. *IEEE Transactions on Image Processing* 21, 1997–2007 (2012)
41. Changeux, J.P.: *Du vrai, du beau, du bien: Une nouvelle approche neuronale*, Odile Jacob (2010)
42. Ward, J., Huckstep, B., Tsakanikos, E.: Sound-colour synaesthesia: to what extent does it use cross-modal mechanisms common to us all? *Cortex* 42, 264–280 (2006)
43. Peeters, G., Giordano, B.L., Susini, P., Misdariis, N., McAdams, S.: The Timbre Toolbox: extracting audio descriptors from musical signals. *Journal of the Acoustic Society of America* 130, 2902–2916 (2011)
44. Herrera-Boyer, P., Klapuri, A., Davy, M.: Automatic Classification of Pitched Musical Instrument Sounds. *Signal Processing Methods for Music Transcription, Part II*, 163–200 (2006)

Interactive Pansharpening and Active Classification in Remote Sensing

Pablo Ruiz, Javier Mateos, Gustavo Camps-Valls,
Rafael Molina, and Aggelos K. Katsaggelos

Abstract. This chapter presents two multimodal prototypes for remote sensing image classification where user interaction is an important part of the system. The first one applies pansharpening techniques to fuse a panchromatic image and a multispectral image of the same scene to obtain a high resolution (HR) multispectral image. Once the HR image has been classified the user can interact with the system to select a class of interest. The pansharpening parameters are then modified to increase the system accuracy for the selected class without deteriorating the performance of the classifier on the other classes. The second prototype utilizes Bayesian modeling and inference to implement active learning and parameter estimation in a binary kernel-based multispectral classification schemes. In the prototype we developed three different strategies for selecting the more informative pixel to be included in the training set. In the experimental section, the prototypes are described and applied to two real multispectral image classification problems.

1 Introduction

Remote sensing images are of great interest in numerous applications. Map drawing, delimitation of parcels, studies on hydrology, forest or agriculture are just a few examples where these images are used [10, 11, 7]. Many of these applications

Pablo Ruiz · Javier Mateos · Rafael Molina
Dept. Computer Science and Artificial Intelligence, University of Granada
e-mail: {mataran, jmd, rms}@decsai.ugr.es

Gustavo Camps-Valls
Image Processing Laboratory (IPL), Universitat de València
e-mail: gustavo.camps@uv.es

Aggelos K. Katsaggelos
Dept. of Electrical Engineering and Computer Science,
Northwestern University, Evanston, IL
e-mail: aggk@eecs.northwestern.edu

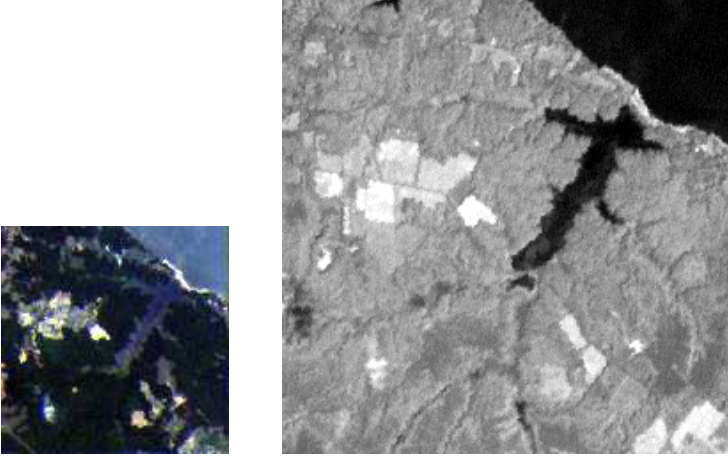


Fig. 1 Region of interest of the observed multispectral image (left) and the corresponding region of interest of the observed panchromatic image (right)

involve the classification of pixels in an image into a number of classes. In supervised classification, the user provides the label of a set of samples to train the classifiers. Usually, the bigger the training set, the better the classification results but more expensive (in time or money) the construction of such a set is.

In this paper we present two prototypes that use multimodal data and user interactivity to obtain more accurate classification results or obtain them at a lower cost. Both prototypes deal with the same multimodal remote sensing image classification problem although they consider user interaction from two different but complementary points of view: user feedback and system adaptation.

On one hand, the first prototype deals with the problem of obtaining more accurate classification results on remote sensing images for specific classes. Due to physical and technological constraints, satellites usually have multimodal sensors that capture two types of images. One sensor captures a multispectral (MS) image composed of several spectral bands with low spatial resolution (LR). For instance, Landsat ETM+ captures 6 spectral bands in the visible and infrared spectrum with each pixel covering an area of 30×30 meters. The other sensor captures a high spatial resolution (HR) image with a low spectral resolution, named panchromatic (PAN) image, that in the case of Landsat has a resolution of 15×15 meters per pixel. Figure 1 shows an example of an Landsat LR MS and the HR PAN images. The LR MS and PAN images are fused using pansharpening techniques to obtain an image with the spectral resolution of the MS image and the spatial resolution of the PAN image. A complete review of techniques to carry out the pansharpening procedure can be found in [2].

Bruzzone *et al.* [5] showed that the use of pansharpening methods that do not introduce significant spectral distortion helps the classifier to obtain higher accuracy, especially for pixels at the borders of objects. While pansharpening has traditionally

only been used as a preprocessing step, that is, before using supervised classification, in the first prototype we address the problem of interactively modifying the pansharpening parameters to improve the figures of merit of the supervised classification of a class of interest, selected by the user. Hence, in this case, the user interaction consists of selecting a class of interest and then the system adapts the pansharpening to obtain better classification results for the selected class.

On the other hand, the second prototype implements *active learning* concepts exploiting the Bayesian modeling and inference paradigm to tackle the problem of binary kernel-based multispectral image classification.

Active learning aims at building efficient training sets by *iteratively* improving model performance through sampling. When applied to remote sensing image classification active learning methods start with a few pixels from each class whose labels are known and, then, iteratively select, using a given criterion, pixels from the rest of the image. The classifier interactively queries the oracle, i.e., the user, for the class of the selected pixels. The feedback obtained at each step of the interaction process is converted into new, fresh training information, useful for improving the classifier. In the presented prototype we developed three different strategies for selecting the most informative pixel to be included in the training set. Additionally, if the user inputs a set of test samples, that is, a set of pixels of known class that can be used to evaluate the quality of the classification, the system will provide some numerical performance information. The objective is to attain the best classification accuracy with the minimum number of queries to the oracle. A survey of active learning algorithms for supervised remote sensing image classification can be found in [20].

The remainder of the paper is outlined as follows. Section 2 describes the theory behind the interactive pansharpening based classification prototype. The Bayesian active learning techniques used in the second prototype are described in section 3. Section 4 presents the prototypes description and experimental results and, finally, section 5 concludes the paper.

2 Interactive Pansharpening Based Classification

Multispectral images allow for an accurate recognition of several land cover classes but, due to their low resolution, information on the objects shape and texture may be lost. In contrast, panchromatic images allow for a better recognition of the objects in the image and their textures but provide no information about their spectral properties. Pansharpening is a technique that jointly processes multispectral and panchromatic images in order to obtain a new multispectral image that, ideally, exhibits the spectral characteristics of the observed multispectral image and the resolution of the panchromatic image.

In this prototype, we propose the use of pansharpening techniques to increase the performance of a classification system. The pansharpening method provides a high resolution multispectral image that is used, along with the observed panchromatic image, as input of the interactive classification algorithm. The pansharpening

method depends on a set of parameters that are automatically estimated from the data and which determine the quality of the obtained pansharpened image.

Even when using a sophisticated pansharpening technique, the classification results may not fulfill the expectations of the user. Hence, we propose to adapt the pansharpening method to improve the classification of a specific class of interest, chosen by the user, by adjusting its parameters to perform this specific class. This represents an application-specific pansharpening approach, where the application of interest is binary classification.

2.1 Pansharpening

The pansharpening methods included in the prototype are formulated within the Bayesian paradigm because it provides good reconstructions and allows for the adaptation of the pansharpening to the needs of the user. More precisely we have implemented the following two pansharpening methods:

- **SAR (Simultaneously Autoregressive) Based Pansharpening:** The method proposed in [13] assumes that each band of a MS image is a degraded (blurred and decimated) version of the original HR MS image, and the PAN image is a linear combination of HR MS image bands. It uses a classical Simultaneous Autoregressive (SAR) model [14] to impose smoothness on the HR MS image. The method depends on a series of parameters that are modeled as Gamma distributions since they allow for the incorporation of information from the data as well as prior information about the value of the parameters. Both the HR MS image and the associated parameters are automatically estimated by the method without user intervention.
- **CONTOURLET Based Pansharpening:** It is a novel method, proposed in [1], that assumes that the observed LR MS image has the same spectral properties than the HR MS image but with a lower level of details. The PAN image, on the other hand, contains the details of the HR MS image but lacks the spectral information. The method uses the non-subsampled contourlet transform (NSCT) to decompose the images into residual and detail bands and models the relations between the details of the HR MS image and those of the observed LR MS and PAN images in the contourlet domain. Since the NSCT detail bands are composed of relatively smooth regions separated by strong edges, Total Variation (TV) [15, 21] is used as prior model. The needed parameters are automatically estimated at each level of decomposition and direction for each band, providing a sound way to control the noise and preventing color bleeding.

It is worth noting that in both cases the estimation of the parameters and the HR MS image is performed in an iterative manner where, at each iteration, a new set of parameters is estimated from all the pixels of the current estimation of the HR MS image, and those parameters are used to obtain a new estimation of the pansharpened image.

2.2 Classification

Once the pansharpened image has been obtained, the classification procedure is carried out using the pansharpened and PAN images as input data. The user has to provide the class of a number of pixels to form the training set. Additionally, a validation/test set is also provided by the user to evaluate the performance of the classification process. In the presented prototype we have implemented two different classification techniques: linear discriminant analysis (LDA) and support vector machines (SVM), which have been largely exploited in remote sensing applications [6, 3].

LDA is an effective subspace technique that optimizes Fisher's score [9]. Subspace methods are algorithms focused on finding projections of an original hyper-dimensional space to a lower dimensional space where the classes have maximum separation. LDA is related to Fisher's linear discriminant and, roughly speaking, both aim at finding a linear combination of features that characterize or separate two or more classes.

SVM is one of the most successful examples of kernel methods, being a linear classifier that implements maximum margin separation between classes in a high dimensional Hilbert space \mathcal{H} . Kernel methods embed the data observed in the input space \mathcal{X} into a higher dimensional space, the feature space \mathcal{H} , where the data are more likely to be linearly separable. Therefore, it is possible to build an efficient linear classifier in \mathcal{H} , that translates into a nonlinear classifier in the input space. Kernel methods compute the similarity between training samples $\{\mathbf{x}_i\}_{i=1}^n$ using inner products between mapped samples instead of computing the dot product in the higher dimensional space explicitly. The so-called kernel matrix $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ contains all necessary information to perform many classical linear algorithms in the feature space, which are non-linear in the input space [19]. The radial basis function (RBF), $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_K^2)$, $\sigma_K \in \mathbf{R}^+$ is selected in this work. To implement SVM for multiclass problems we used the one-versus-all strategy given the particular characteristics of the proposed scheme.

Utilizing one of the described classification methods, the user obtains a classification map. We used the precision and recall scores to assess the model's accuracy:

$$recall = \frac{TP}{TP + FN}; \quad precision = \frac{TP}{TP + FP} \quad (1)$$

where TP is the number of pixels in the class correctly classified, FN is the number of pixels in the class incorrectly classified and FP is the number of pixels not belonging to the class incorrectly classified.

2.3 Interactive Pansharpening Based Classification

Using one of the classifiers, chosen by the user, the classification performed on the HR MS image and panchromatic images usually obtains higher accuracy than

the one performed using the observed multispectral and panchromatic images, especially for pixels at the borders of objects, as showed by Bruzzone *et al.* in [5]. However, it is possible that the outcome does not fulfill the user's expectations. This may be due to a suboptimal performance in a class of interest. Then, by examining the classification results, both visually and numerically, the user can select a class of interest to be improved.

Following the method in [17], the parameters needed by the pansharpening method are estimated again. Using the already estimated pansharpened image, the parameters for the new reconstruction are estimated utilizing only the pixels belonging to the class of interest in this image. Note that no iteration is required for this case. The new estimated parameters result in a new estimation of the HR MS image whose spectral and spacial characteristics are more accurate for the pixels in the class of interest and, hence, with a boosted classification performance for the elements of the class. Note however that this may imply that for other classes the user is not interested in, the classifier may perform slightly worse over the new pansharpened image [17].

3 Active Learning

An alternative way of handling user interaction is related to the emerging field of *active learning*. Let us assume that we have access to a set of samples for which the corresponding class, although not already known, can be provided by an oracle, i.e., the user. The key is to decide which elements to acquire from the set of possible samples in order to build an optimal compact classifier. Active learning aims at building efficient training sets by *iteratively* improving the model performance through sampling.

The prototype implements the Bayesian active learning procedure described in [16] where the Bayesian modeling and inference paradigm are applied to a binary kernel-based classifier tackling both active learning and parameter estimation for infinite dimensional feature spaces.

3.1 Bayesian Modeling and Inference

The Bayesian active learning procedure implemented in the prototype aims at solving the general two-classes supervised classification problem [4] that uses the classification function

$$y(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\mathbf{w} + b + \varepsilon, \quad (2)$$

where the mapping $\boldsymbol{\phi} : \mathcal{X} \rightarrow \mathcal{H}$ embeds the observed $\mathbf{x} \in \mathcal{X}$ into a higher dimensional (possibly infinite) feature space \mathcal{H} , \mathbf{w} is the vectors of parameters to be estimated, b represents the bias in the classification function, and ε are independent realizations of Gaussian distributions $\mathcal{N}(0, \sigma^2)$.

Hence, we can model the classification output $y(\mathbf{x}_i)$ associated with the feature samples $\phi(\mathbf{x}_i), i = 1, \dots, M$, with M the number of samples, as

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^M \mathcal{N}(y(\mathbf{x}_i)|\phi^\top(\mathbf{x}_i)\mathbf{w} + b, \sigma^2), \quad (3)$$

where $\mathbf{y} = (y(\mathbf{x}_1), y(\mathbf{x}_2), \dots, y(\mathbf{x}_M))^\top$. Since $\mathbf{x}_i, i = 1, \dots, M$, will always appear as conditioning variable, for the sake of simplicity, we have removed the dependency on $\mathbf{x}_1, \dots, \mathbf{x}_M$ in the left-hand side of the equation. We note that, for infinite dimensional feature vectors $\phi(\mathbf{x}_i)$, \mathbf{w} is infinite dimensional. Following [4], we assume as prior distribution that each component of \mathbf{w} independently follows a Gaussian distribution $\mathcal{N}(0, \gamma^2)$.

To perform the inference tasks, that is, parameter estimation, prediction and active learning, we will mainly use the marginal distribution of the observations. The marginal distribution of \mathbf{y} can be obtained by integrating out the vector of adaptive parameters \mathbf{w} . It can easily be shown, see for instance [4], that

$$p(\mathbf{y}|\gamma^2, \sigma^2) = \mathcal{N}(\mathbf{y}|b\mathbf{1}, \mathbf{C}), \quad (4)$$

with

$$\mathbf{C} = \gamma^2 \Phi \Phi^\top + \sigma^2 \mathbf{I}, \quad (5)$$

where Φ is the design matrix whose i -th row is $\phi^\top(\mathbf{x}_i)$. The estimation of the parameters b , γ^2 and σ^2 is carried out by using the evidence Bayesian approach [12] which amounts to maximizing the marginal distribution in Eq. (4).

It is important to note that we do not need to know the form of Φ explicitly to calculate this distribution. We only need to know the Gram matrix $\mathbf{K} = \Phi \Phi^\top$, which is an $M \times M$ symmetric matrix with elements $\mathbf{K}_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \phi^\top(\mathbf{x}_n)\phi(\mathbf{x}_m)$, which has to be a positive semidefinite matrix (see [18]).

3.2 Classification

Once the system has been trained, we want to assign a class to a new value of \mathbf{x} , denoted by \mathbf{x}_* . The conditional distribution $p(y(\mathbf{x}_*)|\mathbf{y})$ is a Gaussian distribution with mean $m(\mathbf{x}_*)$ and variance $v(\mathbf{x}_*)$ given by:

$$m(\mathbf{x}_*) = b + \gamma^2 \phi^\top(\mathbf{x}_*) \Phi^\top \mathbf{C}^{-1} (\mathbf{y} - \mathbf{1}b) \quad (6)$$

$$v(\mathbf{x}_*) = \sigma^2 + \gamma^2 \phi^\top(\mathbf{x}_*) \phi(\mathbf{x}_*) - \gamma^4 \phi^\top(\mathbf{x}_*) \Phi^\top \mathbf{C}^{-1} \Phi \phi(\mathbf{x}_*). \quad (7)$$

So, we classify \mathbf{x}_* using $m(\mathbf{x}_*)$ defined in Eq. (6) and write

$$\mathbf{x}_* \text{ is assigned to } \begin{cases} \mathcal{C}_1 & \text{if } m(\mathbf{x}_*) > 0.5 \\ \mathcal{C}_0 & \text{if } m(\mathbf{x}_*) < 0.5 \end{cases}. \quad (8)$$

3.3 Active Learning Approaches

As already explained, active learning starts with a small set of pixels whose class is already known. From these observations, the marginal distribution of \mathbf{y} , and all the parameters are estimated using the procedure described in the previous sections.

In order to improve the performance of the classifier we want to select, from the pixels of the image, the most informative sample, \mathbf{x}_+ , and the user will be asked about its label, $y(\mathbf{x}_+)$. Then the classifier is updated using the new information provided by the user. The process continues until a given number of samples has been included in the training set. To select which of the available samples is added to the training set, the prototype implements three methods described in [16], which are reviewed in the following sections.

3.3.1 Maximum Differential of Entropies

For a sample \mathbf{x} not already present in the training set, the distribution $p(y(\mathbf{x}_*)|\mathbf{y})$ can be calculated using Eqs. (6) and (7), and consequently we can select the new training sample as the one maximizing the variance of the prediction, that is,

$$\mathbf{x}_+ = \arg \max_{\mathbf{x}} v(\mathbf{x}). \quad (9)$$

This criterion amounts to select the sample the classifier is less certain about the class it belongs to.

3.3.2 Minimum Distance to Decision Boundary

In our classification problem the decision boundary corresponds to the set

$$\Pi = \{\mathbf{x} \in \mathcal{X} : m(\mathbf{x}) - 0.5 = 0\}. \quad (10)$$

We can then select the next sample to be included in the training set as the one with minimum distance to the decision boundary by using

$$\mathbf{x}_+ = \arg \min_{\mathbf{x}} d^2(\mathbf{x}, \Pi) = \arg \min_{\mathbf{x}} (m(\mathbf{x}) - 0.5)^2. \quad (11)$$

Note that this method provides a Bayesian formulation of the SVM margin sampling heuristic (see [20]).

3.3.3 Minimum Normalized Distance

The two active learning methods described above take into consideration only partial aspects of the conditional distribution $p(y(\mathbf{x}_*)|\mathbf{y})$. While maximum differential of entropies utilizes the variance of this distribution, it does not use the distance to the decision boundary. On the other hand, the minimum distance to the decision boundary criterion is based on the mean of this conditional distribution and does not take into account the uncertainty of the distribution. It is obviously very easy

to imagine scenarios where these two criteria will not select the best sample, either because it is too far from the decision boundary and, hence, having large variance does not represent a problem, or because, although the sample is the closest to the decision boundary, its uncertainty is very small and consequently it may not be the best sample to be included in the training set.

We can then use the following active learning procedure which combines precision and proximity to the decision boundary

$$\mathbf{x}_+ = \arg \min_{\mathbf{x}} \mathbb{E} \left[\frac{(y(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})} \right], \quad (12)$$

where the expected value is calculated utilizing the conditional distribution $p(y(\mathbf{x})|\mathbf{y})$. Notice that since

$$\mathbb{E} \left[\frac{(y(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})} \right] = 1 + \frac{(m(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})}, \quad (13)$$

we can rewrite this criterion as

$$\mathbf{x}_+ = \arg \min_{\mathbf{x}} \frac{(m(\mathbf{x}) - 0.5)^2}{v(\mathbf{x})}. \quad (14)$$

4 Prototypes Description

We have developed two prototypes implementing the proposed methodologies, that is, interactive pansharpening based classification and Bayesian active learning. The purpose of these prototypes is to show how adaptation and human interaction can be used to improve the performance of classification techniques. Let us now describe in detail each one of the mentioned prototypes.

4.1 Interactive Pansharpening Based Classification

In this prototype, developed in Matlab[®], whose graphical user interface is shown in Fig. 2, we address the problem of adaptively modifying the pansharpening parameters in order to improve the precision and recall figures of merit of the classification of a given class without significantly deteriorating the performance of the classifier over the other classes. The workflow of the prototype is as follows: First, the input LR MS and HR PAN images are loaded. Currently, the prototype only classifies Quickbird images, which have four spectral bands covering the blue, green, red, and infrared bands at 2.44×2.44 meters per pixel and a panchromatic band at 0.61×0.61 meters per pixel, although it can be easily adapted to use other remote sensing images. Then, a pansharpening method is selected from the pull-down list (marked with the number 2 in Fig. 2) and the pansharpening is performed by pressing the button ‘‘Pansharpening’’. The user can select between the two pansharpening methods, SAR and CONTOURLETS, described in section 2.1. We used the

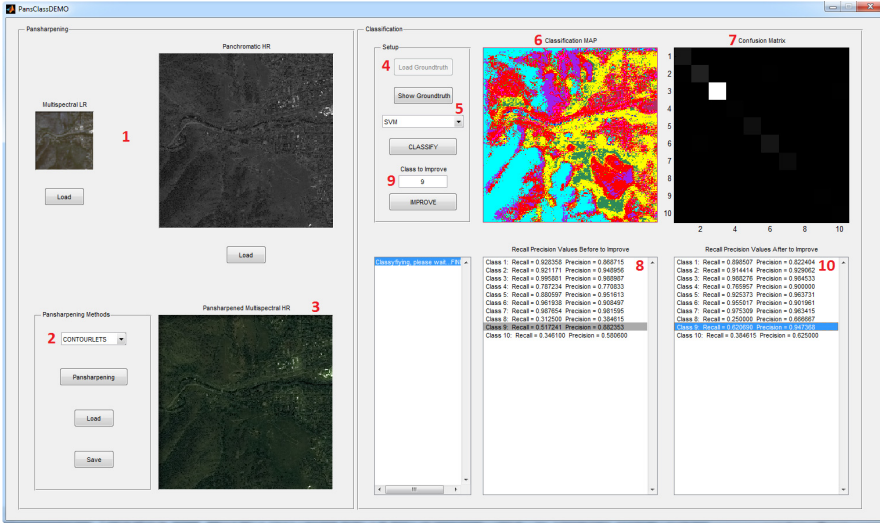


Fig. 2 Interactive Pansharpening prototype interface

Matlab nonsubsampling contourlets toolbox¹ for the contourlets based method. The computed pansharpened image is depicted in the area marked as 3 in Fig. 2. This pansharpened image can be saved to a file, together with some metadata as the used pansharpened method and the value of the parameters used to obtain the image, for its later use. Alternatively, the user can load a previously computed image by pressing the “Load” button in the “PanSharpening Method” area.

The next step is to select the training and test sets. In our prototype this task is achieved by loading, using the button marked as #4 in Fig. 2, a groundtruth image, that is, an image of the same size of the PAN and pansharpened image containing, at each pixel position, the class of this pixel. In the example shown in Fig. 2, ten different classes (cars, water, forest, ...) were used. Note that only a few pixels may be classified in this groundtruth image. Pixels with unknown class have label equal to zero. From this groundtruth image, the PAN and the pansharpened images, a training and a test set are obtained by randomly selecting, among the pixels with known class, 20% for training and the rest for testing. The features of each pixel consist of its panchromatic value and its four band values (that is, a vector with five components) together with the same five components of its four nearest neighbours. So, at each pixel location the corresponding feature vector has 25 components. In the pull-down list marked as #5 in Fig. 2 the classification method can be selected between two options: LDA and SVM, implemented by the Statistics toolbox in Matlab and libSVM², respectively. By pressing the button “CLASSIFY”, the classifier produces a Classification Map by classifying all the pixels in the image (marked as

¹ <http://www.mathworks.com/matlabcentral/fileexchange/10049>
² <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

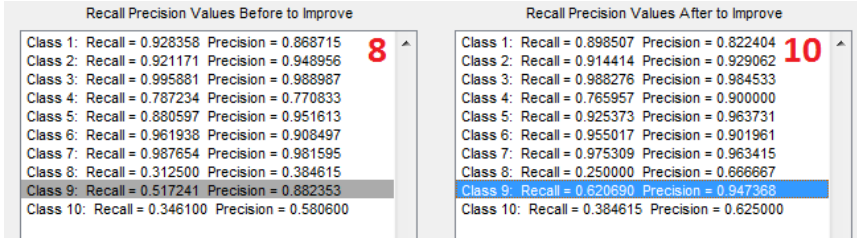


Fig. 3 Detail of the recall and precision values before and after improvement

#6), and a Confusion Matrix (marked as #7) and the precision-recall values for each class (marked as #8) from the pixels in the test set.

As described in section 2.3, the user can examine the classification results, both visually and numerically, and can select a class of interest to be improved by typing its number in the text box marked as #9 in Fig. 2. By pressing the button “IM-PROVE”, the pansharpening procedure is repeated estimating the parameters from the pixels of the selected class in the training set and the new classification map and confusion matrix are obtained and displayed. The new precision-recall values are shown in the window marked as #10 in Fig. 2 side by side to the previously obtained values to help the user compare them. For a better visual evaluation of the results, Figure 3 depicts both precision and recall windows. Notice that the precision-recall values for the class of interest (class 9) have increased from (0.517, 0.882) to (0.621, 0.947). Notice also that some other classes also increased their precision or recall (see, for instance, classes 5 and 10) without significantly decreasing the figures of merit for the other classes.

4.2 Bayesian Active Learning Remote Sensing

The second prototype implemented exploits the Bayesian modeling and inference paradigm to tackle the problem of kernel-based remote sensing image classification. The particular problem of active learning is addressed by proposing an incremental/active learning approach based on three different methods: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance as described in section 3.3.

The Matlab® prototype, whose interface is shown in Fig. 4 is designed to guide the user on the challenging problem of remote sensing land cover classification from multispectral data, and in particular for urban monitoring applications. It can be easily adapted to handle other binary classification problems. Initially, the user loads a Landsat TM MS image whose bands RGB are show in window “RGB Image” (marked as #1 in Fig. 4). Then, the user labels as urban or non-urban an initial set of pixels that are used as initial training set by pressing the button “Label Training Set” (marked as #2 in Fig. 4). For this task, the labeler-tool depicted in Fig. 5 has been developed. This tool allows to enlarge the image, move through it, select a

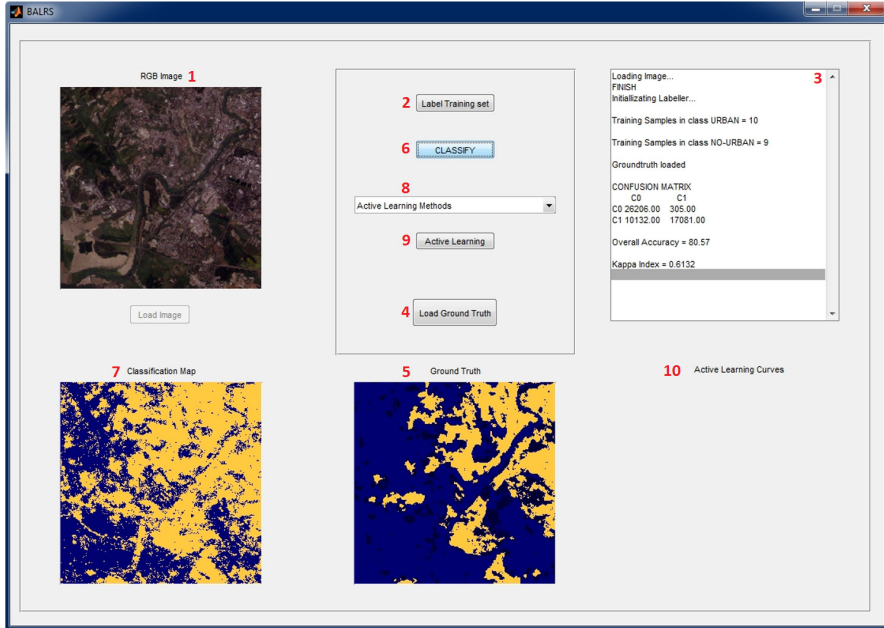


Fig. 4 Bayesian Active Learning prototype interface

pixel, its class and label it by pressing the “Label” button. In addition, it allows to load a training set from a file utilizing the button “Load Training Set”. Once a small number of pixels of each class are labeled, the user can close the labeler tool, returning to the interface main window that will show the number of labeled samples for each class in log window (marked as #3 in Fig. 4).

Optionally, a groundtruth image can be loaded by clicking on the button “Load Ground Truth” (marked as #4 in Fig. 4) allowing to obtain, in addition to the classification map, numerical results about the classification performance. If loaded, the Ground Truth image is depicted in the area marked as #5 in Fig. 4. In the image, the urban class is shown in bright color, the non-urban class in dark color and the background in black.

Utilizing the button “CLASSIFY”, marked as #6 in Fig. 4, the classifier is trained with the initial set and the whole image is classified. The “Classification Map” area (#7) shows the obtained classification map and, if the ground truth was loaded, the log window shows the confusion matrix, and overall accuracy and the estimated Cohen’s kappa statistic [8] as measures of accuracy and class agreement, respectively.

From this initial classifier, the implemented methods help the user to improve the classifier performance by using one of the active learning method described in section 3.3. They can be selected from the pull-down list marked as #8 in Fig. 4. By clicking the “Active Learning” button (#9) the most informative pixel according to the chosen active learning method is selected and the labeler automatically shows

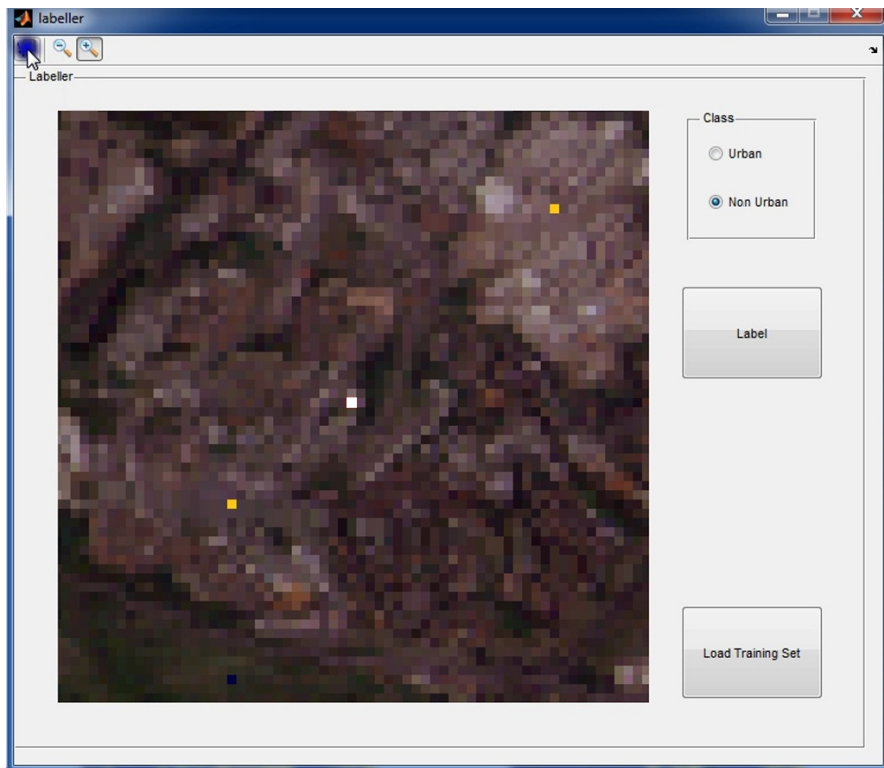


Fig. 5 Manual labeling tool interface

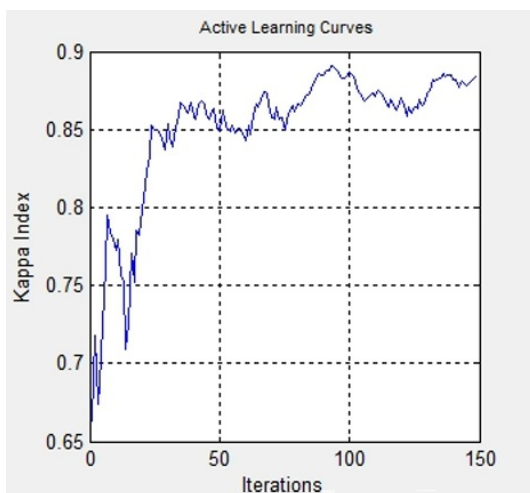


Fig. 6 Obtained Learning Curve. The kappa statistic is measured after each user query.

up to allow the user to label the pixel, shown in white as can be seen in Fig. 5. Once the user labels the pixel and closes the labeler, the sample is incorporated into the training set and the classifier is updated. The process continues by alternatively clicking on the “Active Learning” button and labeling the sample. For each iteration, the learning curve, plotted in the area “Active Learning Curves”, marked as #10 in Fig. 4 is updated. An example of the learning curve after 150 iterations is shown in Fig. 6. As can be seen in the figure, the learning curve grows significantly after a few samples have been included into the training set.

5 Conclusions

In this chapter we presented two prototypes to multimodal interaction in remote sensing image classification problems. The first one proves that pansharpening techniques can be used to increase the performance of classification methods when applied to MS images. We have addressed the problem of adaptively modifying a pansharpening method in order to improve the precision and recall figures of merit of the classification on a given class without deteriorating the performance of the classifier over the other classes. The validity of the proposed technique has been demonstrated using a real Quickbird image. The second prototype implements a non-parametric Bayesian active learning approach based on kernels for remote sensing image classification. We presented three different approaches for active learning: the maximum differential of entropies, the minimum distance to decision boundary, and the minimum normalized distance. The proposed prototype, dealing with the urban monitoring problem from multispectral data, show the validity of the proposed approach.

Acknowledgements. This work has been supported by the Spanish Ministry for Education and Science under projects, TIN2010-15137, EODIX/AYA2008-05965-C04-03 and the Spanish research program Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

References

1. Amro, I., Mateos, J., Vega, M.: Parameter estimation in the general contourlet pansharpening method using Bayesian inference. In: 2011 European Signal Processing Conference (EUSIPCO 2011), pp. 1130–1134 (2011)
2. Amro, I., Mateos, J., Vega, M., Molina, R., Katsaggelos, A.K.: A survey of classical methods and new trends in pansharpening of multispectral images. *EURASIP Journal on Advances in Signal Processing* 2011(79) (2011)
3. Bandos, T.V., Bruzzone, L., Camps-Valls, G.: Classification of hyperspectral images with regularized linear discriminant analysis. *IEEE Transactions on Geoscience and Remote Sensing* 47(3), 862–873 (2009)
4. Bishop, C.M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer (2007)
5. Bruzzone, L., Carlin, L., Alparone, L., Baronti, S., Garzelli, A., Nencini, F.: Can multiresolution fusion techniques improve classification accuracy? In: *Image and Signal Processing for Remote Sensing XII*, vol. 6365, p. 636509 (2006)

6. Camps-Valls, G., Bruzzone, L.: Kernel-based methods for hyperspectral image classification. *IEEE Trans. on Geoscience and Remote Sensing* 43(6), 1351–1362 (2005)
7. Camps-Valls, G., Tuia, D., Gómez-Chova, L., Jiménez, S., Malo, J. (eds.): *Remote Sensing Image Processing*. Morgan & Claypool Publishers, LaPorte (2011); Bovik, A. (ed.): *Collection ‘Synthesis Lectures on Image, Video, and Multimedia Processing’*
8. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37–46 (1960)
9. Duda, R., Hart, P.: *Pattern classification and scene analysis*. Wiley, New York (1973)
10. Liang, S.: *Quantitative Remote Sensing of Land Surfaces*. John Wiley & Sons, New York (2004)
11. Lillesand, T.M., Kiefer, R.W., Chipman, J.: *Remote Sensing and Image Interpretation*. John Wiley & Sons, New York (2008)
12. Molina, R., Katsaggelos, A.K., Mateos, J.: Bayesian and regularization methods for hyperparameter estimation in image restoration. *IEEE Transactions on Image Processing* 8, 231–246 (1999)
13. Molina, R., Vega, M., Mateos, J., Katsaggelos, A.K.: Variational posterior distribution approximation in Bayesian super resolution reconstruction of multispectral images. *Applied and Computational Harmonic Analysis, Special Issue on “Mathematical Imaging”, Part II* 24(2), 251–267 (2008)
14. Ripley, B.D.: *Spatial Statistics*. Wiley (1981)
15. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* 60, 259–268 (1992)
16. Ruiz, P., Mateos, J., Camps-Valls, G., Molina, R., Katsaggelos, A.K.: Bayesian active remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* (submitted, 2012)
17. Ruiz, P., Talens, J.V., Mateos, J., Molina, R., Katsaggelos, A.K.: Interactive classification oriented superresolution of multispectral images. In: *7th International Workshop Data Analysis in Astronomy ‘Livio Scarsi and Vito Di Gesù’ (DAA 2011)*, pp. 77–85 (2011)
18. Schölkopf, B., Smola, A.: *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press Series, Cambridge (2002)
19. Shawe-Taylor, J., Cristianini, N.: *Kernel Methods for Pattern Analysis*. Cambridge University Press (2004)
20. Tuia, D., Volpi, M., Copa, L., Kanevski, M., Muñoz-Marí, J.: A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal on Selected Topics in Signal Processing* 4, 606–617 (2011)
21. Vega, M., Mateos, J., Molina, R., Katsaggelos, A.K.: Super Resolution of Multispectral Images Using TV Image Models. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) *KES 2008, Part III. LNCS (LNAI)*, vol. 5179, pp. 408–415. Springer, Heidelberg (2008)

Interactive Image Retrieval Based on Relevance Feedback

Mauricio Villegas, Luis A. Leiva, and Roberto Paredes

Abstract. This chapter presents a prototype of a web image search engine that implements four approaches to improve the performance of interactive image retrieval systems. The first approach is classic relevance feedback, which relies on user feedback to provide better retrievals in an iterative process. It adopts a probabilistic model which leads to maximizing the relevance of the images retrieved. The second approach is based on user relevance feedback as well, but the attention is focused on combining several information sources to the retrieval mechanism. In particular, we propose a retrieval technique that combines both visual and textual features using dynamic late fusion. The third and fourth approaches are query refinement and tag cloud, both consisting of leveraging the information derived from the relevance feedback and the (textual) image annotations. In the former, a refinement of the initial textual query is suggested. In the latter, a tag cloud is given to provide an overall topic formation related to the user's image selection.

1 Introduction

From the beginning of the Internet, images have been a major and fast growing media. Allowing an effective search among such a huge amount of online images is a challenging task. Current approaches for retrieving relevant images have evolved from the text-based techniques used in classical information retrieval to the content-based image retrieval (CBIR) paradigm. CBIR helps to organize digital pictures by their visual content, and traditionally has involved a myriad of multidisciplinary fields such as computer vision, machine learning, human-computer interaction, database systems, statistics, and many more [3, 12, 28].

Mauricio Villegas · Luis A. Leiva · Roberto Paredes
ITI-PRHLT,
Universitat Politècnica de València, Spain
e-mail: {mvillegas, luileito}@iti.upv.es
rparedes@dsic.upv.es

Image retrieval has been investigated since the 80's and, in the 90's, CBIR became an active area of research. In CBIR, the objective is to find relevant images where the query is often described by an example image of the type of images the user is looking for. In practice, CBIR is still far away from being a solved problem. One way to increase retrieval performance is to capitalize on user feedback, i.e., a user issues a query with an example image and is then presented with a set of hopefully relevant images; from these images the user selects those images that are relevant and those which are not (possibly leaving some images unmarked) and then the retrieval system refines its results, hopefully leading to a better outcome after each interaction step.

The precision of CBIR systems has improved during the last years, but still it is not enough from a realistic and practical point of view. Users mostly prefer to use textual queries instead of submitting an image as an example of the desired results [13]. This leads to two well-known problems. On the one hand, it is not feasible to make the annotations of all images manually, and, on the other hand, it is not possible to translate some images (specially abstract concepts) into words—what is known as *the semantic gap* [19].

To this end, we developed a relevant image search engine (RISE) [1]. Two different goals were proposed before deploying the prototype. First, to design a base system that improves the state-of-the-art image retrieval systems by means of new image descriptors, new distance functions, and using textual queries. Second, to leverage the user knowledge to iteratively (and interactively) improve retrieval results, by means of relevance feedback and other techniques based on this kind of user intervention. This chapter presents the results and lessons learned while approaching these goals.

In Section 2 we describe the theoretical framework in which the RISE system lies. We propose a probabilistic model for relevance feedback, together with a multimodal variant to handle simultaneous signals (e.g., visual and textual image descriptors). In addition, we explore novel ways to enhance textual queries, by means of query refinement techniques. As we shall describe later, we explore two types of suggestion: a refined query and a tag cloud, each having a concrete utility for image retrieval tasks. Then, we thoroughly describe the prototype implementation in Section 3, including its architecture, design, user interface, and so on. Afterwards, we present a series of experiments in Section 4 that validate the techniques presented so far. Finally, we close the chapter with a summary and some conclusions in Section 5.

2 Methodology

2.1 Relevance Feedback

Relevance feedback has been studied in the field of image retrieval nearly as long as the field of information retrieval exists [17]. An overview of the early related

¹ <http://risenet.iti.upv.es/>

work on relevance feedback in image retrieval is given in the work of Zhou and Huang [31]. Most approaches use a set of images selected by the user as individual queries and combine the retrieved results. More recent approaches follow a query-instance-based approach [7] or use support vector machines to learn a two-class classifier [18].

In the the present work, the user interaction is handled using a probabilistic approach to model the relevance of candidate *image sets*. This leads to a significant boost in performance and also opens new ways to integrate consistency checks into the retrieval procedure. The approach most closely related to the one presented here is Bayesian browsing [25]. The formulation presented here, follows the concepts for interactive pattern recognition proposed in the work of Vidal et al. [26] and was first presented in the work of Paredes et al. [13]. The following description of the probabilistic interaction model is only a short overview. For further details, the reader should refer to the previously mentioned references.

2.1.1 Probabilistic Interaction Model

Let \mathcal{U} be the universal set of images and let $\mathcal{C} \subset \mathcal{U}$ be a fixed, finite *collection* of images. We assume that the user has in mind some relevant set of images $\mathcal{R} \subset \mathcal{U}$. This set is unknown and the system’s objective is to discover or retrieve N images of it, among the images in \mathcal{C} .

The interactive retrieval process starts with a query, q , given by the user, in which q could be either a textual query, or a *query image*, i.e. $q \in \mathcal{U}$, or possibly some other alternative. Using q , the system provides an initial set $\mathcal{X} \subset \mathcal{C}$, $|\mathcal{X}| = N$ of images, which would be obtained through textual retrieval if q is textual, or if q is an image, by retrieving “similar” images to q according to a suitable distance measure. These images are judged by the user who provides *feedback*, by selecting which images are relevant (and, implicitly, which are not relevant). Such feedback information is used by the system to obtain a new set of images \mathcal{X} , and the process is repeated until the user is satisfied, which means that all images \mathcal{X} are considered to be relevant.

At any step of this interactive process, the user feedback can be represented by two sets, $\mathcal{Q}^+ \subset \mathcal{R}$ and $\mathcal{Q}^- \subset \mathcal{C} - \mathcal{R}$, containing the images selected in previous interaction steps as relevant and non-relevant, respectively. In the current step, the (deterministic) feedback provided by the user, consists in marking as relevant some of the (previously unmarked) images which are added to \mathcal{Q}^+ , and the remaining presented images implicitly as non-relevant are added to \mathcal{Q}^- .

To optimize the user experience, we propose to maximize the probability that the images in \mathcal{X} are relevant according to $\mathcal{Q} = (\mathcal{Q}^+ \cup \mathcal{Q}^-)$, that is:

$$\hat{\mathcal{X}} = \arg \max_{\mathcal{X} \subset \mathcal{C}, |\mathcal{X}|=N} \Pr(\mathcal{X} | \mathcal{Q}), \quad (1)$$

where the query q is included in \mathcal{Q}^+ if q is an image, and if q is textual, then the collection of images \mathcal{C} can be filtered to include only images related to that textual query.

Applying the Bayes rule and dropping terms which do not depend on the optimization variable, \mathcal{X} , Eq. (1) becomes:

$$\hat{\mathcal{X}} = \arg \max_{\mathcal{X} \subset \mathcal{C}, |\mathcal{X}|=N} \Pr(\mathcal{Q} | \mathcal{X}) \cdot \Pr(\mathcal{X}). \quad (2)$$

For the first term of Eq. (2) we can use a model directly based on image distances:

$$\Pr(\mathcal{Q} | \mathcal{X}) \propto \prod_{\forall x \in \mathcal{X}} P(\mathcal{Q} | x), \quad (3)$$

where each term in the product (3) is a *smoothed* version of the classical class-conditional probability estimate based on nearest neighbors (5) using a suitable image distance $d(\cdot, \cdot)$:

$$P(\mathcal{Q} | x) = \frac{\sum_{\forall q \in \mathcal{Q}^+} d(q, x)^{-1}}{\sum_{\forall q \in \mathcal{Q}} d(q, x)^{-1}}. \quad (4)$$

Note that we use a product to combine probabilities in Eq. (3). This enables using a greedy search strategy to find approximate solutions to Eq. (2).

Now, if we assume that $\Pr(x)$ follows a uniform distribution, then it is constant in the maximization process and can be dropped, therefore the expression to maximize becomes:

$$\hat{\mathcal{X}} = \arg \max_{\mathcal{X} \subset \mathcal{C}, |\mathcal{X}|=N} \prod_{\forall x \in \mathcal{X}} P(\mathcal{Q} | x). \quad (5)$$

This defines a very simple procedure in which, to maximize this expression, we only have to choose those images x with maximum values of $P(\mathcal{Q} | x)$, which can be estimated using Eq. (4). This is known as the simplified Greedy Approximation Relevance Feedback algorithm (GARFs) (13).

2.2 Dynamic Visual-Textual Fusion

The use of several (complementary) sources of information is a fairly common approach for improving the performance of pattern recognition and information retrieval systems. In an image retrieval system, generally there are available many types of features extracted from the images, and also there are textual features, such as: metadata, annotations provided by users, or text surrounding the images from where they appear. Adequately using all this available information is a major goal in order to obtain the best performance possible.

Considering a general purpose image retrieval system, it can be observed that some queries will benefit more from visual information and others from textual information. For instance, if we think of images with clearly visible yellow “tigers”, then the visual features might be really important and therefore a visual retrieval

² We recall that only the notation $\Pr()$ stands for true probabilities; here we abuse the notation by letting $P()$ denote arbitrary functions used as models.

technique can perform best. However, for other queries a textual retrieval might be more appropriate. In general, it is expected that an adequate *fusion* between these two extremes can lead to a better performance. The main issue is that this fusion highly depends on the specific query and user's intent; and leaving this task to the user to specify the textual/visual combination is a heavy burden. Thus, an approach that improves retrieval accuracy of a search engine with less user involvement is valuable [10].

Fusion methods proposed in the literature can be mainly categorized as *early* or *late* fusion. The former combines the features and then a single retrieval system uses this combination. In the latter there are several retrieval systems, and the task is to combine all of the retrieval results.

For fusing visual and textual features, early fusion tries to discover the statistical dependencies between visual features and semantic concepts, by using unsupervised learning methods [14]. However, this is rather difficult mainly because image annotations often contain keywords that are not strongly associated with particular visual features or they only account for a small part of the image. Thus, since it is simpler and easier to integrate it in the relevance feedback interaction, late fusion has been the preferred strategy in the current work. The following subsections describe the proposed method for late fusion, and then an approach that takes advantage of the relevance feedback, in order to adjust the fusion depending on the query and the user.

2.2.1 Late Fusion

In late fusion, for each query q , there are K different retrieval methods working separately. This results in K ranked lists of documents (in this case a document being an image). The information of such K ranked lists is used to obtain a single ranked list. The final list is obtained by assigning a score to each document appearing in at least one of the K lists. A high score indicates that the document is more likely to be relevant to query q . Documents are sorted in decreasing order of their score and this is the final list of ranked documents, from which the top N are the ones retrieved.

In this work we consider a simple (yet very effective) score, based on a weighted linear combination of the document ranks through the different lists. The proposed score takes into account redundancy of documents and the individual performance of each retrieval method. Diversity and complementariness are brought to play by the heterogeneity of the considered independent retrieval methods (IRMs), while redundancy is considered through the use of several IRMs per modality. Given a document x and K lists $L_i, 1 \leq i \leq K$, a score $s_{\alpha}(x)$ is computed as follows:

$$s_{\alpha}(x) = |\{i : x \in L_i\}| \cdot \sum_{i=1}^K \frac{\alpha_i}{R(x, L_i)}, \quad (6)$$

where $R(x, L_i)$ is the rank position of document x in the ranked list L_i , and α_i , with $\sum_{k=1}^K \alpha_k = 1$, is the importance weighting for the i -th IRM, which can be determined using prior knowledge, such as in the proposal described in Section 2.2.2. The fused document ranking $R_\alpha(x)$ is obtained by sorting the documents by descending score $s_\alpha(x)$. Documents appearing in several lists at the top positions will receive a higher score, while documents appearing in few lists or appearing at the bottom positions most of the times will be scored low.

If only two IRMs (visual and textual) are considered, and the ranking of all documents is available for both IRMs, the following simpler linear combination can be used:

$$s_\alpha(x) = \frac{\alpha}{R_v(x)} + \frac{1-\alpha}{R_t(x)}, \quad (7)$$

where α is a single importance weight and $R_v(x) = R(x, L_v)$, $R_t(x) = R(x, L_t)$ are the visual and textual rankings, respectively.

2.2.2 Dynamic Linear Fusion

The main problem of the late fusion approach is how to obtain the importance weight α in Eq. (7). Clearly α should not be kept fixed for a given system, since it is known that in general some queries will perform better with visual information, or the other way around, and leaving this task to the user is too much burden. To deal with this dynamically variable weighting, we propose to take advantage of the relevance feedback information and solve an optimization problem. Two optimization criteria have been explored, specifically:

$$\hat{\alpha} = \arg \max_{\alpha} \sum_{x \in Q_i^+} \sum_{y \in Q_i^-} R_\alpha(y) - R_\alpha(x), \quad (8)$$

$$\hat{\alpha} = \arg \min_{\alpha} \sum_{x \in Q_i^+} R_\alpha(x), \quad (9)$$

where $R_\alpha()$ is the ranking obtained from the fusion Eq. (7) when fusing the visual and textual rankings of the previous interaction, and Q_i^+ and Q_i^- are the sets of relevant and non-relevant images, respectively, selected in the current interaction step i .

Intuitively speaking, Eq. (8) maximizes the area under the ROC curve, which globally ranks the relevant and non-relevant images far as possible. On the other hand, Eq. (9) only considers the relevant images trying to place them in the top positions.

At each interaction step, the system proposes a set of images which the user marks as relevant. Without further disturbing the user, the weighting α is updated on the basis of the user's intention, determined by the images marked so far in the current step.

2.3 Query Refinement

Nowadays in the popular Internet search engines it is almost a standard that when typing the query, a drop-down list of suggested query completions appears, giving the option to the user to select one and complete the query with much less effort. This drop-down list is generated by using the most popular queries among all of the users of the search engine. This technique works very well as long as the system has a very large amount of users and the query being searched is relatively common among other users. Since these assumptions are not always fulfilled, other approaches for query suggestion should be developed that are not so dependent on these factors.

In an image search engine that implements a relevance feedback approach to refine an initial search, query refinement suggestions can be derived by using the relevance information given by the user. The strategy would be that every time that the user changes the selection of relevant and non-relevant images, a new query or queries are suggested, which optionally the user can follow. In an image search engine, most of the area of the user interface is devoted to show the thumbnails of the images, and thus a drop-down list would hide parts of the images and interfere with the selection of images for the relevance feedback. So, in this context, a drop-down list is not appropriate. In the current version of the RISE prototype, only a single query suggestion is given, although there are other possibilities, such as giving several suggestions in the same line.

2.3.1 Suggestion Generation

The method for deriving the best query refinement to suggest is as follows. An initial set of N images are presented, and the user selects among these which are the ones considered relevant, leaving the rest implicitly selected as non-relevant. Each image has associated some weighted text, which for example can correspond to text extracted from the webpages where the image appeared. To suggest a new query to the user we propose an approach that essentially accounts for the (weighted) words that appear in the text associated to the relevant images but do not appear or are less important in the text associated to the non-relevant images. This selection of words can be accomplished in different ways but here we show the selection that gave us the most promising results.

Let $\{w_1, \dots, w_n\}$ be the words of the vocabulary, i.e. all of the different words that appear in the associated text of the N images being shown. We denote the set of relevant images as Q^+ and the set of non-relevant images as Q^- . Let \mathcal{W} be the set of words w_i that appear in all the relevant images. If this set is empty, then no suggestion is produced. If this set is not empty, we compute a score for each one of these words:

$$s(w_i) = \frac{\frac{1}{|Q^+|} \sum_{j \in Q^+} t_{ij}}{\frac{1}{|Q^+|} \sum_{j \in Q^+} t_{ij} + \frac{1}{|Q^-|} \sum_{k \in Q^-} t_{ik}}, \quad (10)$$



Fig. 1 Example of query refinement suggestions when searching for “marathon”. Images with a light green frame are the ones selected as relevant. The suggestions tend to be very specific when few images are selected (a, b). Conversely, when several images are selected the suggestions tend to be a common characteristic among them (c).

where $w_i \in \mathcal{W}$, and t_{ij}, t_{ik} are the scores of the word w_i in the relevant image j and non-relevant image k , respectively. See Section 3.4 for details on how these scores are obtained. The word relevance scores t_{i*} are assumed to be positive and add up to one for each image. For the words w_i that do not appear in any of the non-relevant images, then the score according to Eq. (10) always gives $s(w_i) = 1.0$, so in order to avoid ties in these cases, the value of $\frac{1}{|\mathcal{Q}^+|} \sum_{j \in \mathcal{Q}^+} t_{ij}$ is added to the score. Taking into account the aforementioned assumptions and the tie breaking, the scores are values between 0 and 2.

Finally at most the top 2 words $w_i \in \mathcal{W}$ with highest scores are suggested as a query refinement, taking care of not to include words from the original query. Figure 1 illustrates the basic functionality of the query refinement suggestion.

2.4 Tag Cloud

As a possible alternative to the query refinement presented in Section 2.3, the goal of the tag cloud is to provide the user with a weighted list of words, so that the most relevant topics of the selected set of images can be noticed visually. Furthermore, the tags in the cloud are also hyperlinks which refine the original text query by adding the respective word.

Tag clouds are visual presentations of a set of words, in which text attributes (such as size or color) represent the importance of some feature, e.g. word frequency. Depending on the context, tag clouds can support a wealth of different tasks, ranging from locating specific items in a list, to providing an overview and form a general

impression. Much as a table of contents or index can do for a book and a menu of categories can do for a website, they provide a means for users to get the “big picture” of the underlying set of content and a “gist” of what the book or site is about [15].

2.4.1 Tag Cloud Scoring

Similar to the query refinement approach, the tag cloud is generated from the weighted text associated to the set of images presented to the user, giving more importance to the words from the relevant images than the non-relevants. The associated text includes text near the image from the webpage that contains such image, text from the URL of the image, etc. These annotations are weighted depending on the position of the word in the page, the importance of the text in the DOM, etc. See Section 3.4 for details.

The scoring procedure is based on the one depicted in Eq. (10), again adding $\frac{1}{|Q^+|} \sum_{j \in Q^+} t_{ij}$ in the cases that $s(w_i) = 1.0$. There are two differences, though. First, the set of words \mathcal{W} includes all the words that appear in any of the relevant images, and second, added to the score is the number of relevant images E in which the word w_i appears, i.e. $E = |\{\forall j \in Q^+ : t_{ij} \neq 0\}|$. Therefore, tag scores are values between 0 and $|Q^+| + 2$.

This way, the tag cloud informs about the relevance of words for the images being selected. Since it is re-generated every time the users update their selection, the topic saliency is expected to pop up. Figure 2 illustrates this idea. In any case, similar to the query refinement approach, the words from the submitted query are not included in the list of tags.

The associated text is assumed to be normalized by image. However, the word weights can be somewhat improved by multiplying them by the Inverse Document Frequency (IDF) [9, 16, 30], which accounts for the relative importance of the word in the whole database of images, and implicitly reduces the impact of stopwords. This slows down the tag cloud generation, although not significantly since generally the vocabulary is not too large for the set of images that is presented to the user.

2.4.2 Tag Cloud Rendering

The above-mentioned ranking of words is used to feed a tag cloud generator, which uses the computed scores to render the tags according to the following set of CSS features: *font size*, *font weight* (or *bolding*), and *intensity*.

Concretely, each CSS feature f_i for a given w_i is computed by means of a linear interpolation:

$$f_i = \min \mathcal{F} + [s(w_i) - \min \mathcal{S}] \frac{\max \mathcal{F} - \min \mathcal{F}}{\max \mathcal{S} - \min \mathcal{S}} \quad (11)$$

where \mathcal{S} is the set of computed scores from \mathcal{W} and $\mathcal{F} = \{v_1, v_2\}$ denotes the bounded set of feature design values. For instance, to compute font sizes we could use $\mathcal{F} = \{80, 250\}$ to map sizes to percentage values or $\mathcal{F} = \{11, 36\}$ to map sizes to pixel values, respectively. Following the same rationale, to compute font weights



(a) Subset 1, related to the TV series.



(b) Subset 2, related to buildings.

Fig. 2 Some tag cloud examples for the query “house”. Images with a light green frame are the ones selected as relevant. As observed, tags (and their weights) are updated whenever the images marked as relevant change.

we could use $\mathcal{F} = \{100, 900\}$ where 900 denotes the maximum font weight (bold) according to the CSS standard³.

Tags are alphabetically sorted (although order sorting options are available), since it has been shown that alphabetized lists are the quickest to be scanned at a glance [8].

3 Prototype Implementation and Description

This section is devoted to describing the RISE system, together with the full steps to complete retrieval process; i.e. from the initial (textual) database query to having all retrieved images marked as relevant.

³ <http://www.w3.org/TR/CSS21/fonts.html#font-boldness>

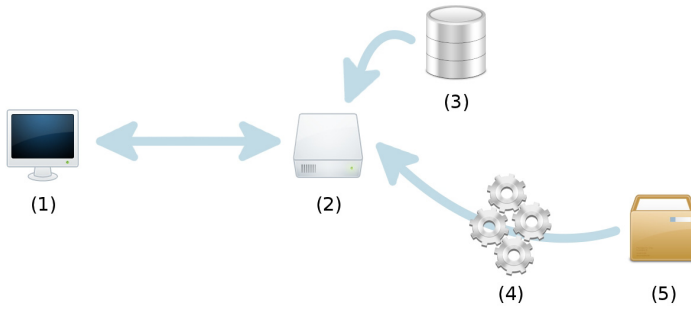


Fig. 3 RISE architecture. (1) user interface, (2) web server, (3) indexed database, (4) retrieval engines, and (5) image database.

We begin with an overview of the system architecture, followed by an illustration of the user interface. Next, we explain the image database, the image crawling, and the features used for later indexing. We end with a description of the hardware configuration.

3.1 System Architecture

As shown in Figure 3, the RISE system has 5 parts: (1) a user interface (UI), (2) a web server, (3) an indexed database, (4) a series of retrieval engines, and (5) a collection of crawled image files and their related metadata, named ‘image database’ from here onwards.

3.2 User Interface

The index page (Figure 4) is composed of a single text field to submit initial queries to the system. Once a query is submitted, the user is presented with a set of images (Figure 5). At this point, the user navigates to interactively retrieve a desired number of images (see Section 3.2.1 for more details).

In addition, the user can control some retrieval parameters such as the fusion amount between text- and visual-based engines, whether the interface should display suggestions for the current query, or if the user can inspect some image properties such as dimensions, size, etc.

3.2.1 Interactive Retrieval Process

As in other relevance feedback systems, the retrieval engine of RISE starts by receiving a textual query given by the user. This textual query is used to obtain and present to the user a set of N images that are relevant to the textual query. This initial set of images is obtained by a TF-IDF-like retrieval, using an inverted file index obtained from the weighted words extracted for images as described in Section 3.4.



Fig. 4 Index page

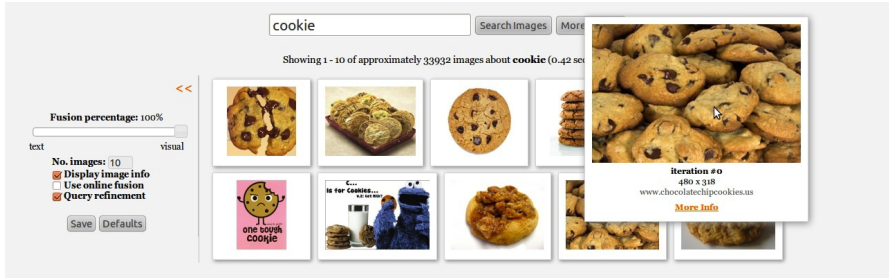


Fig. 5 Initial results, after querying the database

At this stage, the interaction itself starts, by letting the user select some of the images as being relevant. The unmarked images are considered as being non-relevant, in order to lower the burden on the user.

Using the information of the selected relevant and non-relevant images, the system provides several options to interact with the user. One possibility for the user is to click on a “Get Related Images” button, in which the classical relevance feedback approach is used to retrieve a new set of images. For selecting this new subset of images, the system uses the GARFs algorithm (see Section 2.1.1), with the Jensen-Shannon divergence as distance measure, and both visual and textual features

(Section 3.4) are used, fusing as described in Section 2.2.1. If the “Use online fusion” option is activated, the fusion parameter α is automatically adjusted starting from the second feedback interaction onwards. The optimization criteria used for this is Eq. (9) presented in Section 2.2.2. This was the preferred choice because it gave better performance both experimentally and practically.

The other options to interact with the system are either through the query refinement or the tag cloud. The generation of the query refinements and the tag cloud are described in Section 2.3 and Section 2.3 respectively. In both cases, the user has the option to click on the hyperlinks, which modifies the initial text query, thus presenting the user with a new set of images which (hopefully) is closer to what the user is searching for. Some illustrative examples are shown in figures 6 and 7 respectively. After this, again, the user can select more relevant images to refine further the results, using either the query refinement, the tag cloud or the relevance feedback, and this iterative procedure continues until the user is satisfied.

3.3 Web Server

The web server is ruled by a general purpose open-source LAMP configuration (Linux OS, Apache HTTP server, MySQL database, and PHP language). A series of PHP scripts are used to render the UI and connect it to both databases (images + metadata and MySQL).

3.4 Image Database

Among the objectives set out for the RISE prototype was that it had to work for a relatively large database of images and that the images should be of a general scope, so that queries about practically any topic could be submitted. In order to have a wide variety in a relatively small set of images (not billions), we opted to use the same crawling strategy as in [23], where the image URLs (and the corresponding URLs of the webpages that contain the images) are obtained by querying popular commercial image search engines. We selected Google, Bing and Yahoo, and queried them using words from the English dictionary that comes with the `spell` spelling checker. For each query word we kept at most the 200 first results for each search engine.

The next step in the crawling process was to download the images and corresponding webpages, and store a thumbnail of the images and a snapshot of the webpages. In the end, we obtained over 31 million of both images and webpages. In order to avoid duplicate images, several precautions were taken. First the URLs were normalized to prevent different versions of the same URL to be downloaded several times. However, there is also the possibility that the same image is found under different URLs. To account for this, the images were stored using a unique image identifier composed of: part of the MD5 checksum of an 864-bit image signature (in some aspects similar to the one presented in [29], which is independent to

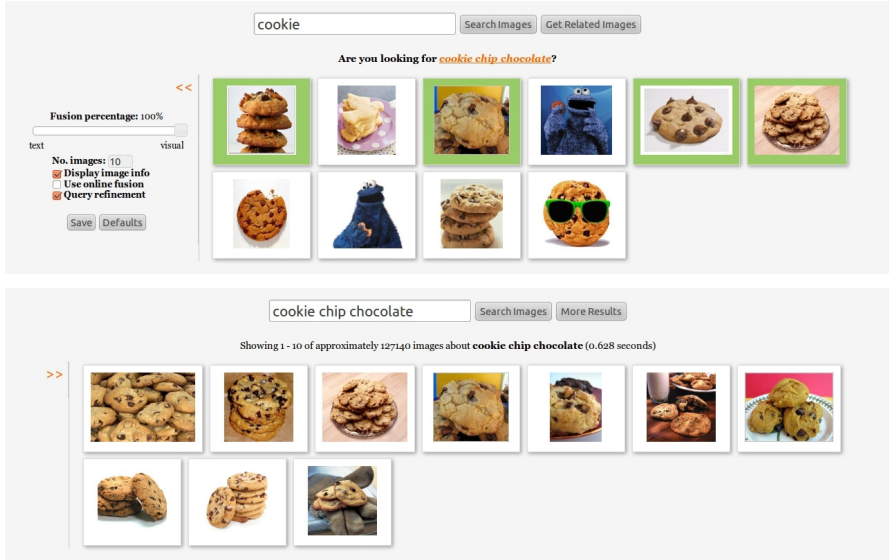


Fig. 6 Results after following a query refinement suggestion



Fig. 7 Results after following a tag from the tag cloud

simple image transformations) and, part of the MD5 checksum of the file. This scheme ensures storing exactly the same file only once and easily identifying duplicates or near duplicates (accounting for images in various formats, at different resolutions and with minor modifications such as watermarks). The final image identifiers are 16 base64url digits.

Even though the URLs were obtained from trustworthy search engines, inevitably there is a certain amount of problematic images that when indexing they tend to appear at the first positions of many queries. The problematic images we are referring to are for instance a message saying “image removed”, or dummy images some servers send specifically to web crawlers, or what is known as link farms. Removing this type of images is in itself a difficult problem, however we noticed that most of these tended to have many different URLs linking to them or are images that appeared in a large amount of webpages. So the approach to remove these images was simply not to include images that had more than K_u URLs linking to them or that appeared in more than K_w webpages. The values of K_u and K_w were set manually from a quick look at the images being considered for removal.

Another problem encountered when crawling the web is that an image could be still reachable, although the webpage where it appeared has been removed and does not supply the proper HTTP 404 code. In order to solve this issue, for the final set of images that were indexed, it was checked that at least one webpage contained each of the images. Any image not having a corresponding webpage was obviously not considered for inclusion. After this procedure, the resulting index contained almost 25 million images.

3.4.1 Textual Index and Textual Features

The RISE prototype starts by receiving an initial textual query for searching images. For this, the images were indexed using the corresponding webpages, which is briefly described in the following. Also this same webpage processing was used to extract the textual features required for the relevance feedback (see Section 2.1), the query refinement suggestions (see section Section 2.3) and the tag clouds (see section Section 2.4).

Since in the Internet it is very common that webpages have markup mistakes, the first step of processing was to convert the webpages to valid XML with UTF-8 encoding. Then, the script and style elements were removed and it was checked that the image appears in the webpage. Not all of the webpage text was considered, only the webpage title and all the terms that are closer than 600 in word distance to the image, not including the HTML tags and attributes. Each word was assigned a weight, defined as:

$$s(t_n) = \frac{1}{\sum_{\forall t \in \mathcal{T}} s(t)} \sum_{\forall t_{n,m} \in \mathcal{T}} F_{n,m} \text{sigm}(d_{n,m}), \quad (12)$$

where $t_{n,m}$ are each of the appearances of the term t_n in the document \mathcal{T} , $F_{n,m}$ is a factor depending on the DOM (e.g. title, alt attributes, etc.) similar to what is done

in [2], and $d_{n,m}$ is the word distance from $t_{n,m}$ to the image. The sigmoid function was centered at 35, had a slope of 0.15 and minimum and maximum values of 1 and 10 respectively. For each webpage, the word weights were normalized to add up to one. If an image appeared in more than one webpage, the resulting word weights are simply the average over all webpages, assuming that if a word does not appear in the document, then its weight is zero. Finally, for each image at most the 100 word-weight pairs with the highest weights were kept.

The resulting textual index corresponds to an inverted file over these textual features sorted by the word weights. In the index, the word positions within the webpages was also stored so that it could be possible to search for exact phrases and do proximity search. However, since the word weights depend on the distance to the image, the queries with and without proximity search tend to be very similar, thus the proximity search is currently not being used.

3.4.2 Visual Features

Color histograms are among the most basic approaches and widely used in image retrieval [6, 19, 21]. To show performance improvements in image retrieval, systems using only color histograms are often used as a baseline. The color space is partitioned and for each partition the pixels with a color within its range are counted, resulting in a representation of the relative frequencies of the occurring colors. The RGB color space has been used for the histograms, the images are divided into 3×3 regions, and for each region a 64 bin histogram is computed, leading to 576-dimensional feature vectors. Only minor differences were observed with other color spaces, which was also observed in [20]. Still the use of color histograms is only the basic technique. In the future, more powerful features can be used or even better would be to use a combination of several feature types [27].

3.5 Indexed Database

In the RISE prototype the image thumbnails and webpage snapshots are stored in an XFS filesystem, which is a mature filesystem that has a very good performance with millions of relatively small files. Also RISE uses a MySQL database with 3 relational tables, a configuration that has proven to be quite efficient to support our retrieval needs.

One table is used for the inverted file structure used for the initial textual query retrieval, see Table 1. The other two tables are used to store the textual and visual features, see tables 2 and 3, respectively. Since the relevance feedback requires to retrieve query dependent sets of visual and textual features, in order to have a fast response, the last two tables are stored in an SSD drive, which has a considerably faster seek time compared to standard hard drives.

Table 1 Table for word-images relation

Field	Type	Attributes	Description
word	varchar(128)	primary key	UTF-8 word
imgs	mediumblob	binary	list of weighted images
num	int(10)	unsigned	image count

Table 2 Table for image-words relation

Field	Type	Attributes	Description
img	char(16)	primary key	image identifier
words	mediumblob	binary	list of weighted words

Table 3 Table for image-features relation

Field	Type	Attributes	Description
img	char(16)	primary key	image identifier
feat	blob	binary	visual descriptor

3.6 Hardware Configuration

Everything is hosted on a single machine with 2 quad-core CPUs @ 2 GHz and 8 GB of RAM running Debian OS. The RISE prototype and the OS are both installed in a 130 GB SAS disk @ 15K rpm, which provides really good performance. As mentioned before, image and textual features are stored in a 120 GB SSD disk to speed up seek time and provide a fast performance of the relevance feedback. A batch of 6 SCSI disks of 2 TB each is attached as a single RAID unit to store image thumbnails and crawling information, and additionally there are two 2 TB hard drives to schedule regular backups.

4 Experiments and Results

We report here a series of experiments that illustrate the techniques discussed in Section 2. To begin, GARFs is compared against three baseline relevance feedback methods. Then, our dynamic linear fusion proposal to combine visual and textual features is compared to when each feature is used in isolation. Finally, we resort to two methods that aim to enhance the textual query that the user must submit to the indexed database. The first method consists in displaying a (refined) query suggestion, while the second method consists in displaying a tag cloud, where the user may pick one of the tags shown to refine the initial (textual) query.

4.1 Relevance Feedback Evaluation

Fortunately for this scenario there is adequate labeled corpora publicly available, hence, we can simulate the user interaction and compare different relevance feedback approaches [24]. Concretely, we simulated a user who wants to retrieve N relevant images. The retrieval engine would present N images which the user marks as relevant Q^+ and non-relevant Q^- (done automatically using the labels). Using this feedback, a new set of $N - |Q^+|$ images is retrieved using a relevance feedback algorithm. Again, the user can mark the new relevant images, and this iterative interaction continues until $|Q^+| = N$.

The number of images to be retrieved was set to $N = 20$, and the initial query was based on a single example image. Up to four feedback iteration steps were simulated, which including the initial query are I_1, \dots, I_5 . As a measure of performance, precision was estimated using the Leaving-One-Out procedure.

The GARFs method was compared to three baseline methods on the well-known Corel/Wang dataset. The baseline techniques were the following:

- **Simple Method:** Perform an exhaustive search of the next $N - |Q^+|$ images among the whole database, i.e., keeping the relevant images in each iteration.
- **Relevance Score:** Consider the best matching only among the positive and negative query images [7].
- **Rocchio Algorithm:** The best matching is the vector difference between the centroids of the relevant and non-relevant documents [17].

Images were represented by color histograms and Tamura texture histograms, as they are a very reasonable baseline in image retrieval [4]. The L_1 distance was used as dissimilarity measure. The features are the following:

- **Color Histograms:** The color space is partitioned in bins, each bin resulting in a representation of the relative frequencies of the occurring colors. We use the RGB color space for the histograms, and split each dimension into 8 bins leading to a $8^3 = 512$ -dimensional histogram.
- **Tamura Features:** From the features studied in [22], we use the ones that proved to be most significant: *coarseness*, *contrast*, and *directionality*. For each image pixel, we computed these features in a neighboring area, quantizing the values into 8 discrete bins, and creating thus a 512-dimensional joint histogram for each image.

The results are shown in Table 4. It is worth mentioning that, in relevance feedback, it is desirable to retrieve as much precise images as possible for the first feedback iterations. In this case, as observed, GARFs is the best performer, returning 94.5% of the relevant images in just 2 iterations.

Table 4 Precision (in %) for successive interaction steps

Method	I_1	I_2	I_3	I_4	I_5
Simple	73.6	83.2	88.0	91.0	92.9
Rocchio	73.6	92.7	97.3	99.2	99.8
Rel. Score	73.6	92.2	97.8	99.5	99.9
GARFs	73.6	94.5	98.9	99.9	99.9

4.2 Dynamic Linear Fusion Evaluation

To evaluate this approach, no public corpus is available. Hence, we manually labeled a subset of 21 queries with 200 images each from the RISE image database (Section 3.5). This way, like the previous experiments, the user feedback could be easily simulated in an automatic way. The reader may consult [24] for a brief description of each query, together with their respective images.

For consistency with the default UI (Figure 5), this time we simulated a user who wants to retrieve $N = 10$, which were shown at a time. So, in each iteration, the user would see 10 images and judge which were relevant or not.

The system used the α parameter (Section 2.2.1) to account for the fusion percentage between visual and textual retrieval engines. This way, $\alpha = 100\%$ is pure visual retrieval and $\alpha = 0\%$ is full textual retrieval. Visual features were comprised of color histograms (Section 3.4.2), while textual features were comprised of image annotations (Section 3.4.1). Results are shown in Figure 8.

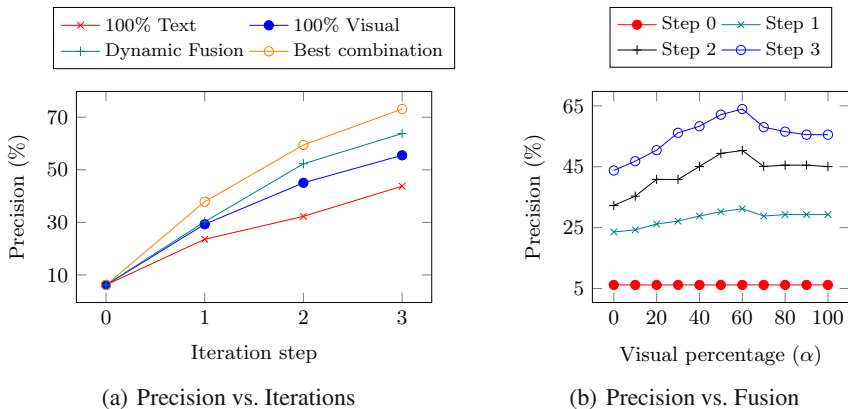


Fig. 8 Dynamic linear fusion results, for $N = 10$ images to be seen at a time. (a) Comparison of image retrieval techniques. (b) Precision as a function of α (visual percentage), for several feedback iteration steps.

4.2.1 Evaluation Discussion

Figure 8(a) shows the evolution of retrieval accuracy with the successive interaction steps for different retrieval strategies. As expected, both pure text and visual retrieval alone are worse performers. After one interaction step, the dynamic linear fusion approach performs better on average. The best fusion is just an upper bound, and therefore in practice it is unreachable.

It can be observed in Figure 8(b) that the system quickly gains accuracy with the progression of the user interaction steps. That is, the more information the system has about what the user considered relevant in previous steps, the better it can predict the best fusion parameter α for the current step. In the first step, there is a clearly ascendant slope towards the visual strategy, achieving high precision when full visual search is used. However, in the following iterations the best precision is not obtained on the extremes, which shows the importance of having a dynamic user/query-adaptative α to achieve always the best precision.

It is worth pointing out that this α parameter has no memory, and the user can change his mind while looking at a query or concept. For instance, the user can start looking for “apples”, choosing real eatable apples, and then change the objective starting to select as relevant images related to the company “Apple”. So, the system will automatically tune the value of α to fit the user’s retrieval needs. This allows much more flexibility to retrieve relevant results.

4.3 Query Refinement Evaluation

The image database used for the RISE prototype was built from real data gathered from the Internet with completely unsupervised annotations (see Section 3.4), so we have no ground truth, i.e., labeled samples. Furthermore, labeling a subset of the images in order to evaluate the query refinement suggestions is rather challenging. As commented in the previous section, the labeling would require to have a list of example queries, and for each query, several subsets of selected relevant images corresponding to different subclasses of the original query. Moreover, for each of these subsets we would require a list of possibly correct query refinements. Thus, in order to evaluate the proposed approach, we opted to conduct an informal field study. The procedure was simple: to measure the user’s subjectivity towards the query suggestion technique.

For the evaluation, we selected 81 of the 99 concepts from the ImageCLEF 2011 dataset⁴, and used these as the initial text search queries. The reason to remove 18 concepts was because they were related to specific image properties rather than high-level concepts, e.g., “Neutral Illumination”, “No Blur”, etc.

The evaluation task consisted of two stages [11]. First, users were presented with the first 10 ranked images for a given textual query, e.g., “cat”. Then the user would select a subset of the images which had a common concept or relation among them, e.g., “all are black cats”. If the system was able to derive a query refinement,

⁴ http://imageclef.org/system/files/concepts_2011.txt

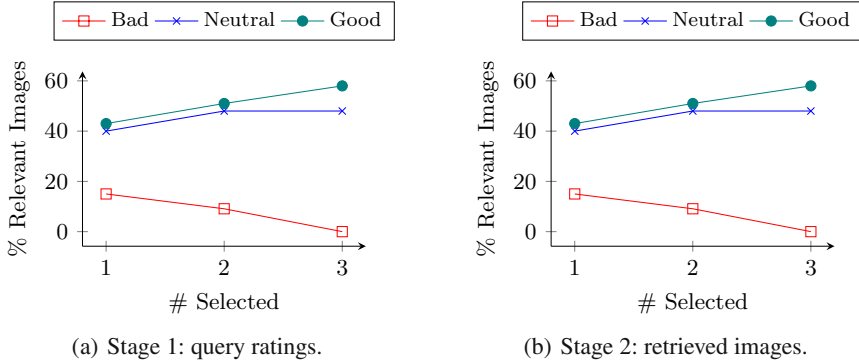


Fig. 9 Query refinement evaluation results. (a) Average rating of the suggested queries in relation to the number of initially selected images. (b) Percentage of images considered as relevant after following the query suggestions in relation to the number of initially selected images.

Table 5 Results for stage 1 of the query refinement evaluation, showing the ratings for the suggested query refinements of the system

# selected	# samples	Bad	Neutral	Good	NQ
1	194	35	42	106	11
2	74	11	13	29	21
3	24	2	4	6	12
>3	30	0	0	3	27
Overall	322	48	59	144	71

Table 6 Results for stage 2 of the query refinement evaluation, showing the mean (and the standard deviation) values of the number of retrieved relevant images after following the suggested queries

# selected	# ratings	Bad	Neutral	Good
1	183	1.5 (1.5)	4 (3)	4.3 (3)
2	53	0.9 (1.4)	4.8 (3.2)	5.1 (3.3)
3	12	0 (0)	4.8 (4.8)	5.8 (2.8)
>3	3	0 (0)	0 (0)	9.6 (0.5)
Overall	226	1.3 (1.5)	4.3 (3.2)	4.7 (3.2)

the UI would show it and let the user rate if the suggestion was either good, bad, or neutral. The number of times there was no query suggested (NQ) was also recorded (Table 5). In the second stage of the evaluation, users were presented with the images after following the query suggestion, and they had to mark all of the images

considered relevant to the concept they had in mind when selecting the images in the first stage of the evaluation. This two-stage process was repeated for all of the subsets of related images the user could identify. Three people from our department took part in the evaluation.

The results of the study are presented in Figure 9 and Tables 5 and 6, respectively.

4.3.1 Evaluation Discussion

Regarding the first stage of the evaluation, the first thing to note is that, as more images are selected, it is less probable that the system will suggest a query (see Table 5). This is understandable since it is less likely that there will be common terms to all selected images. Moreover, the terms associated to each image depends totally on the webpages in which the image appears, thus not all of the images will be well annotated. Nonetheless, most of the suggested queries were rated as being good, which indicates that this approach of deriving the suggestions based on the selected relevants can be quite useful.

Regarding the second stage of the evaluation, as expected, the query suggestions which were rated as good or neutral, retrieved more relevant images than the bad query suggestions (see Figure 9(b)). This is convenient since it is unlikely that a user will use a suggestion considered to be bad. A peculiar behavior that was also observed is that the performance tends to be better for suggestions that were derived using more selected relevant images. Then, overall, as more images are selected, it is less likely that the system will suggest a query; however if there is a suggestion, it tends to be a better one.

Another observation from the evaluation was that the quality of the query suggestions depends highly on the particular query. There are some queries where the images presented to the user clearly belong to different subgroups, which, if selected, most of the time a query will be suggested that relates to that subgroup. An example of a query that gives good suggestions was shown in Figure 10. In the figure it can be observed that when only one image is selected, the suggested queries tend to be very specific to that image. This is another possible factor why when fewer images were selected, the percentage of retrieved relevant images was lower: a very specific query might be very good, however this also means that there will be less images available of this type.

4.4 Tag Cloud Evaluation

Like in the query refinement evaluation (see Section 4.3) obtaining labeled data to be able to evaluate the tag cloud is rather difficult. Thus for evaluation, we conducted again an informal field study, using the same database of the RISE prototype, see Section 3.4. Fourteen users aged 31.42 (SD=5.34) were recruited via email advertising to participate in the evaluation. They were told to assess the relevance of the top scored tags suggested in the cloud for a series of queries (12 queries per person on average, 10 tags per query).

The list of queries was compiled by merging the two lists from ImageCLEF 2012: Photo Annotation and Retrieval. Concretely, the concepts from the ‘Scalable image annotation using general Web data’ subtask⁵ and the queries used in the ‘Visual concept detection, annotation, and retrieval using Flickr photos’ subtask⁶. The list comprised 164 search queries overall.

During the evaluation, each user had to follow their given list of queries and select a subset of images for different subtopics from the presented set of 10 images. Participants had no restrictions on subtopic selection, e.g., subtopics could have an arbitrary number of images, no minimum or maximum subtopics per query were imposed, etc.

Each time a relevant image was selected from the presented set, a list of tags was displayed in order of relevance (the most relevant tags at the beginning of the list, in a left-to-right order). The relevance of each tag was computed as described in Section 2.4.1. A checkbox was attached to each tag, so that users could mark whether the tag was considered relevant to the subtopic or not. Figure 10 shows the evaluation results.

It is worth pointing out that no tag cloud was displayed, but a list of tags sorted by relevance instead, since we wanted to avoid any possible visual bias in the study (see Section 4.4.1 for a discussion).

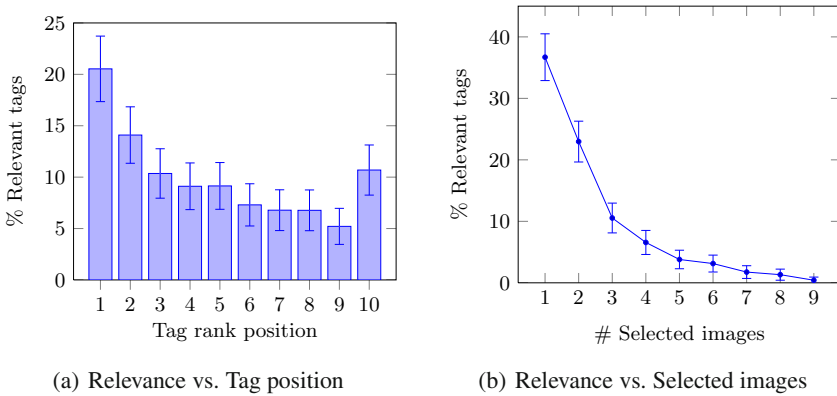


Fig. 10 Evaluation results of the tag cloud, with 95% confidence intervals

4.4.1 Evaluation Discussion

In Figure 10(a) each bar represents on average the percentage of relevant tags (normalized by the number of selected relevant tags) for each tag rank position. The relevance differences between the tags presented in rank 1 position and the tags presented in other positions are statistically significant.

⁵ <http://imageclef.org/2012/photo-flickr>

⁶ <http://imageclef.org/2012/photo-web>

As expected, tags in the first positions of the cloud tended to be perceived more often as relevant. Nonetheless, it was interesting to notice that, most of the time, tags in the last position (i.e., those with the least relevance score) were found to be more informative than those ranging from 4th to 8th positions.

A study by Bateman et al. [11] reported that tags with a larger number of characters tended to be selected less often. We looked whether this observation could be explained in our study. We computed the tag length ratio as the division of the average length of selected tags by the average length of all suggested tags and obtained 1.03 (SD=0.16), which means that selected tags were around the average tag length overall. Furthermore, only 10% of the time a user chose a tag that had more than a 1.1 of tag length ratio. This suggested that the length of a tag was not determinant to assess its relevance towards a particular query, but also that users did not choose neither shorter or longer tags.

Therefore, regarding the last tag being chosen as much informative as the 3rd one, it can be explained by the fact that users scanned the tag list rather than reading it carefully, and hence attention was more focused around the “edges” of the list. In fact, this behavior has been reported to some extent elsewhere [11, 8]. This effect could be mitigated by presenting to the evaluators the tags in random positions.

Figure 10(b) depicts the proportion of relevant tags according to the number of selected images. Similar conclusions to those in the works referenced in the previous paragraph were derived: 1) as more images are selected, the overview that provides the tag cloud about such a set of images tends to be very general; and 2) the quality of the tags depends highly on the particular query.

As observed, therefore, when a single image is selected, nearly 40% of the tags are considered as relevant, since the tag cloud is specifically tailored to the user selection. Then, this proportion decays dramatically as more images are selected. This suggests that when many images are selected, a new strategy for generating tag cloud should be devised. Nonetheless, on average, 21.87% (SD=9.9) of the presented tags were considered as relevant at any time.

Users reported that sometimes the tags were found to be really useful and beneficial to the current query, but also sometimes they were found to be meaningless. This fact is explained by the noise due to the image indexing procedure, which was completely unsupervised and therefore the cloud may contain irrelevant tags for a particular query.

All in all, our study indicates that the tag cloud approach supports its intended goal, i.e., impression formation about a particular set of relevant images. Furthermore, the tag cloud gives more options to the users to refine the initial text query. As such, we believe that the tag cloud has more potential than the query refinement strategy.

5 Conclusion

In this chapter, we have considered several approaches for relevance-based interactive image retrieval. Each approach is based on a probabilistic model to handle user interaction.

The first one, relevance feedback with late fusion, mixes visual and textual descriptors depending on an α parameter. It has been shown that fusion performs considerably better than other retrieval engines working separately. Typically, a single value of α must be known beforehand; however we have demonstrated that it can be automatically learned. Thus, the *dynamic linear fusion* scheme learns from the user intentions the type of query the user is searching for (either text- or visual-oriented) at each interaction step.

The second one, query refinement suggestion, is a query modification scheme to give the user precise information (i.e., a *continuation* of the textual query) using information derived from the selected images. An empirical evaluation has shown that this approach is specially useful when few images are selected. Overall, when a user follows a query suggestion, the number of retrieved images that are considered as relevant tend to increase regarding to not following such suggestion and using pure relevance feedback instead.

The third one, a tag cloud, is an extension of the query suggestion scheme. Its goal is to give a gist about the most relevant topics of the selected set of images, so that the user can form a general impression of the underlying contents of such image set. We observed that tags in the first positions of the cloud tended to be perceived more often as relevant, but also that the last tag was more informative than those ranging from central positions. All in all, the tag cloud gives more options to the users to refine the initial text query.

In sum, classical image retrieval strategies can be notably improved. This chapter has envisioned some ways to do so, and has demonstrated that they are both suitable and convenient. However, further research still remains to be done. It is our belief that the techniques presented here can push the boundaries of other relevance-based retrieval domains (e.g., video and multimedia contents).

References

1. Bateman, S., Gutwin, C., Nacenta, M.: Seeing things in the clouds: the effect of visual features on tag cloud selections. In: Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia (HT), pp. 193–202 (2008)
2. La Cascia, M., Sethi, S., Sclaroff, S.: Combining textual and visual cues for content-based image retrieval on the world wide web. In: Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL), pp. 24–28 (1998)
3. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys* 40(2), 1–60 (2008)
4. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: An experimental comparison. *Information Retrieval* 11(2), 77–107 (2008)
5. Duda, R., Hart, P.: *Pattern Recognition and Scene Analysis*. John Wiley, New York (1973)
6. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and effective querying by image content. *Journal of Intelligent Information Systems* 3(3/4), 231–262 (1994)
7. Giacinto, G., Rolli, F.: Instance-based relevance feedback for image retrieval. In: *Neural Information Processing Systems, NIPS* (2004)

8. Halvey, M.J., Keane, M.T.: An assessment of tag presentation techniques. In: Proceedings of the 16th International Conference on World Wide Web (WWW), pp. 1313–1314 (2007)
9. Hiemstra, D.: A probabilistic justification for using $tf \times idf$ term weighting in information retrieval. *International Journal of Digital Libraries* 3(1), 131–139 (2000)
10. Jin, H., Tao, W., Sun, A.: Vast: Automatically combining keywords and visual features for web image retrieval. In: International Conference on Advanced Communication Technology (ICACT), pp. 2188–2193 (2008)
11. Leiva, L.A., Villegas, M., Paredes, R.: Query refinement suggestion in multimodal interactive image retrieval. In: Proceedings of the 13th International Conference on Multimodal Interaction (ICMI), pp. 311–314 (2011)
12. Moran, S.: Automatic image tagging. Master's thesis, School of Informatics, University of Edinburgh (2009)
13. Paredes, R., Deselaers, T., Vidal, E.: A Probabilistic Model for User Relevance Feedback on Image Retrieval. In: Popescu-Belis, A., Stiefelhagen, R. (eds.) *MLMI 2008*. LNCS, vol. 5237, pp. 260–271. Springer, Heidelberg (2008)
14. Pham, T.-T., Maillot, N.E., Lim, J.-H., Chevallet, J.-P.: Latent semantic fusion model for image retrieval and annotation. In: Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM), pp. 439–444 (2007)
15. Rivadeneira, A.W., Gruen, D.M., Muller, M.J., Millen, D.R.: Getting our head in the clouds: toward evaluation studies of tagclouds. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), pp. 995–998 (2007)
16. Robertson, S.E., Sparck-Jones, K.: Relevance weighting of search terms. *Journal of the American Society for Information Sciences* 27(3), 129–146 (1976)
17. Rocchio, J.J.: Relevance feedback in information retrieval. In: Salton, G. (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323. Prentice-Hall (1971)
18. Setia, L., Ick, J., Burkhardt, H.: Svm-based relevance feedback in image retrieval using invariant feature histograms. In: IAPR Workshop on Machine Vision Applications (MVA), pp. 542–545 (2005)
19. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Trans. PAMI* 22(12), 1349–1380 (2000)
20. Smith, J.R., Chang, S.-F.: Tools and techniques for color image retrieval. In: *SPIE Storage and Retrieval for Image and Video Databases*, pp. 426–437 (1996)
21. Swain, M.J., Ballard, D.H.: Color indexing. *International Journal of Computer Vision* 7(1), 11–32 (1991)
22. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. *IEEE Transaction on Systems, Man, and Cybernetics* 8(6), 460–472 (1978)
23. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(11), 1958–1970 (2008)
24. Toselli, A.H., Vidal, E., Casacuberta, F. (eds.): *Multimodal Interactive Pattern Recognition and Applications*, 1st edn. Springer (2011)
25. Vasconcelos, N., Lippman, A.: Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing. In: *ICIP*, pp. 153–157 (1998)
26. Vidal, E., Rodríguez, L., Casacuberta, F., García-Varea, I.: Interactive pattern recognition. In: Proceedings of the 4th International Conference on Machine Learning for Multimodal Interaction (MLMI), pp. 60–71 (2008)
27. Villegas, M., Paredes, R.: Image-text dataset generation for image annotation and retrieval. In: *II Congreso Español de Recuperación de Información, CERI 2012*, pp. 115–120 (2012)

28. Wang, J.Z., Boujemaa, N., Del Bimbo, A., Geman, D., Hauptmann, A.G., Tešić, J.: Diversity in multimedia information retrieval research. In: Proc. MIR, pp. 5–12 (2006)
29. Chi Wong, H., Bern, M., Goldberg, D.: An image signature for any kind of image. In: Proc. of International Conference on Image Processing, pp. 409–412 (2002)
30. Yu, C.T., Salton, G.: Precision weighting—an effective automatic indexing method. *Journal of the ACM* 23(1), 76–88 (1976)
31. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. *Multimedia Systems* 8, 536–544 (2003)

An User-Driven Tool for Interactive Retrieval of Non Annotated Videos

M. Angeles Mendoza, Tomás Arnau, Isabel Gracia,
Filiberto Pla, and Nicolás Pérez de la Blanca

Abstract. A prototype to retrieve videos from non-annotated video databases is proposed. We focus on the problem of retrieving relevant videos from the audiovisual signal when the query is unknown for the system, since it is assumed that most of the available annotations are useless, as it is the case for most of the videos from common users in Internet. The approach presented is defined inside of the on-line learning paradigm where user and system collaborate to improve alternative rankings of the items dataset. The user guides the system in the semantic level and the system tries to adapt the low-level similarity distance between items according to the user preferences. The user interacts with the system until a prefixed number of relevant items is retrieved. The video database is represented as a dense graph where a semi-supervised algorithm is used to propagate the user feedback.

1 Introduction

This chapter describes the architecture of a video retrieval prototype for unconstrained video databases (consumer videos) [14]. In the era of images as information source, the low cost of the capture technology and the increasing capacity of storage media led to the fact that most of current personal electronic devices incorporate a video camera making of each person a potential image and video provider. The capacity of the human being to analyze and to understand events from images have boosted these media as the biggest resource for information exchange. In this

M. Angeles Mendoza · Nicolás Pérez de la Blanca
Department of Computer Science and A.I.,
University of Granada, Spain
e-mail: nines@decsai.ugr.es, nicolas@ugr.es

Tomás Arnau · Isabel Gracia · Filiberto Pla
Institute of New Imaging Technologies,
University Jaume I, Spain
e-mail: {tarnau, gracia, pla}@uji.es

situation, one of the most exciting technical problem is how to efficiently exploit this huge amount of information. In most of the cases the video file is annotated with some type of surrogate information, what is mainly truth for TV videos or equivalent signals. However this type of information is usually absent or useless when unconstrained videos are considered. Here we assume that the video signal is the only available information.

So far, the most popular retrieving approaches [33] have been focused on shots retrieval, that is, short pieces of video with homogeneous content. Each shot is summarized in a descriptor using combinations of low-level feature values and concepts detector output. Eventually, these descriptors are used to train a bank of classifiers representing the set of possible actions to test. This approach has shown to be very successful when used on the TV-signal and edited video where many samples of the same concept are available [34, 35, 29, 39]. However, recent results have shown that this approach does not scale adequately when the number of trained concepts increases [28]. The visual representation variability of a concept is so high that many different samples are necessary in order to classifiers can learn with high generalization power. In edited videos this is possible for a reasonable amount of concepts [29, 1], but for unconstrained ones it is not the case.

When the problem is to recognize the event represented in a piece of video, the so-called *event recognition* [32], a more challenged problem appears. Now new complexities appear from longer videos with multiples shots in a specific temporal order. In this case, current state of the art for concept combination approach is even more limited. A more efficient approach appears when the user collaborates with the system identifying the items to retrieve and therefore providing correlations between descriptors and semantic concepts.

The Relevance Feedback Approach (RFA) in image and video retrieval [40, 22] is the technique that allows pinpointing faster the desired targets of the users query. Different implementations of this idea have been suggested according to the complexity of the problem, the user feedback information and the searching process adaptation. Three main approaches can be identified [16]: query-shifting approach, feature re-weighting approach, and probability-based approach. Query-shifting methods, such as Multimedia Analysis and Retrieval system (MARS) [23, 3], move the query onto the feature domain and find similar images by a weighted combination of feature vectors of the relevant images. Feature re-weighting methods map the low-level features to the high-level concepts by feature transformations. Different classifiers as Support Vector Machine, Adaboost, etc (see e.g. [10]) has been used to achieve this goal. In contrast to query-shifting methods; which find the best center location in the fixed feature space, feature re-weighting methods transform the feature space around the relevant query images. The third category is probability-based methods, which calculate the probability of query images in each category.

Working with non annotated videos the query-shifting approach could be a good option but the semantic gap (see, [27]) makes difficult to improve the query from simple functions of the low-level features. Although the feature re-weighting

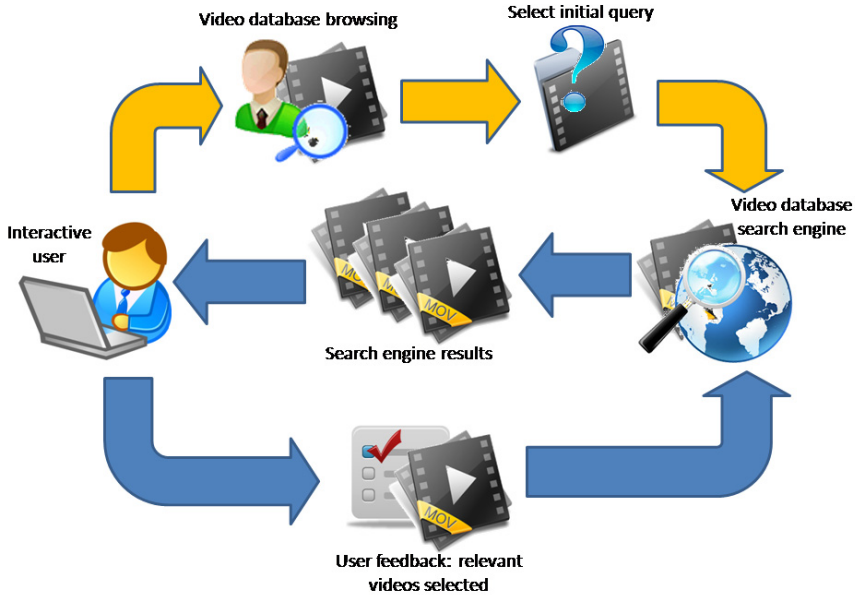


Fig. 1 Flowchart of the prototype interactive process

methods are a good option for non-interactive frameworks, in the interactive case the size of the dataset can be a drawback if we have to retrain the full system to incorporate the user feedback.

The RFA can be defined inside of the online learning paradigm [5] when, on each iteration, the system learns from the user feedback to update its internal representation for the user preferences. In our approach user and system develop a top-down dialog where the user guides the system in the highest level and the system try to adapt the similarity distance between items according to the user preferences. The user interacts with the system until a prefixed number of relevant items are at the top of the system output.

In summary, this chapter presents an online learning algorithm based on an adaptive graph model to retrieve non annotated videos from large databases. The simplified flowchart of the interaction process is summarized in Fig. 1. To start a searching session, the user first browses the video database through a graph clustering-based structure, in order to find an initial video semantically good enough to start the retrieving process. The system updates its internal representation producing a ranked list of the full dataset. The interactive loop is closed with the user feedback, that is a user re-labeling of the system output. This system-user learning loop is repeated until the cumulated system outputs contain a prefixed number of relevant items.

2 Related Works

Many efforts have been done in last years in order to improve the state of the art in video retrieval (e.g., [33, 29, 2]). Different approaches has been proposed to summarize pieces of videos in simpler representations defined by low-level features and object detector combination [30]. The most popular approaches assume that a piece of video can be summarized in a small set of relevant images named key-frames where each key-frame represent a short and almost stationary piece of video (usually 2 seconds). Although this summary eliminates any kind of motion information it has been very successful in many different applications; which point out the high spatio-temporal redundancy presents in the video signal. Different types of features representing the spatio-temporal information have been used: multiscale bank of filters and color histogram for holistic features, Harris-Laplace detector [24] and {HOG, SIFT} [7, 21] as interest region detector and descriptor respectively. However, in this approach the relevant temporal events must be estimated from the still images, and this is not always possible or simple. In [17] the Spatio Temporal Interest Point (STIP) detector was proposed to detect actions in video. In multimodal approaches the audio signal is usually characterized with its Mel Frequency Cepstral Coefficients (MFCC) although more sophisticated techniques can also be used [6]. In our experiment we have only used combinations of SIFT and MFCC features.

More recently, the video-shot retrieval from different query modalities has been the focus of attention. Query-by-text, query-by-concept or query-by-examples are the common query interfaces present in the current technology (e.g., [36, 8, 13, 39]). In these cases a concept lexicon is learned off-line to index the items. Experiments with different lexicon sizes have been carried out to evaluate the significance of the size in a concept combination modeling . In [28] is shown that large size lexicon does not improve the classification showing the high uncertainty of many of the learned concepts and the difficulty of characterizing the video semantic by concept combinations. Fusion techniques from multiple queries and strategies to automatically learn the most relevant concepts for each query has also been considered [28], but with unclear success; the classifier fusion architectures design remains an open issue.

Although some experimental results have shown better retrieval scores when several modalities are combined in the query, there is not a sound theoretical support to the fusion of modalities yet. In most cases the redundancy in the multimedia signal is so high that most of the queries can be solved using a single modality. For the complex cases the coding architectures connecting the low-level features to the semantic ones remains an open problem. In recent years some contributions in the machine learning field are been suggested (e.g., [18, 19]).

In contrast with the above approaches, our proposal follows an online learning framework where the user preference feedback is propagated to the dataset using a semi-supervised graph model. Each video is summarized off-line in a set of feature vectors from which a video descriptor using a bag of words model is built up. A dense graph model is used to represent the descriptors using an affinity measure to weight the graph arcs. We assume the affinity measure defines some relevant

correlation with the semantic meaning of the items. The retrieval process starts when the user after browsing the database clicks a small set of initial items, ranked or not, as representative of his/her interest. The system learns from this preferences and generate a ranking of all items by decreasing relevance. A partial list from the most relevant is shown to the user for a new re-rank. On each interaction the user provides a new ranking as feedback. By clicking on some of the items the user modify the system ranking shifting some of the items to the top of the ranking and others to the bottom. In our approach the feedback is a structured object defined by a full rank of the dataset, but the user does not provide a strict order but preferences on some items. In fact one specific feedback can be defined for many different rankings.

User interface is another issue that has been getting more relevance for video retrieval for the last years ([12]). Most of the work in the past in video retrieval was focused on retrieval engine algorithms, storing data and so on. Since the appearance of TREC Video track and TRECVID competition [26], more attention has been focused on the user interface aspects.

In [11] it is noted that image and video searching systems must cope with the fact that users can scan a large number of images and make a selection among them very quickly, as opposite to text retrieval. Thus, they pointed out the increasing popularity of the Rapid Serial Visual Presentation (RSVP) techniques; in which the series of images are shown to the user replacing existing ones in the screen at the same position. A variation of this approach is used in the Forkbrowser visualization technique for scanning video search results in the MediaMill system [8].

3 Prototype Functional Definition

As it has already been mentioned, we are concerned with unconstrained video retrieval through an interactive system. In this framework we focus on three main points with the following general functional requirements:

- *The modeling and representation of the video data collection.* A representation based on generic low level features, including static and motion features is used. The representation of videos in the low level feature space are assumed to represent content in such a way that two videos with similar semantic content are assumed to be near in the low level feature representation space chosen. To this end, high dimensional spaces usually help to facilitate to link videos with similar higher level semantic contents. The approach most currently used as low level representation for recognition tasks is the bag of words model. Different combinations of low level features to extract several types of bag of words representation will be used (see subsection 5.2).
- *The feedback coding.* Feedback will be a structured object where the user opinion on the system ranking output is coded (positive and negative video samples selected by user). This information will be the input for the next functional level in the user interaction loop (see Fig. 1)

- *An efficient mechanism to propagate the feedback information into the database items.* Eventually, the mechanism to propagate the feedback from the user must work in the feature space where videos are represented, although this mechanism must integrate the directions and flows related to the semantic relationships between contents of similar video samples. This is done by means of a semi-supervised learning algorithm, propagating the user feedback information, in such a way this propagation is done according to high level semantic relationships among video representations.

3.1 The Representation and Updating Model

Let $\mathcal{V} = \{v_i, i = 1, \dots, N\}$ be a video collection of N videos, where each video is represented by a finite vector descriptor v_i . Let $A = \{A(v_i, v_j) \geq 0, i, j = 1, \dots, N\}$ be a matrix defining an affinity measure between each two descriptors. Let $\mathcal{W} = \{W_{ij}, i, j = 1, \dots, N\}$ be a distance matrix between items where $W_{i,j} = \exp(-0.5 \cdot A(v_i, v_j)/\sigma)$ and $\sigma > 0$ represents the scale of the kernel. Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{W}\}$ be an undirected graph defined on the video collection \mathcal{V} , where \mathcal{E} denotes the arc set and \mathcal{W} denotes the weight matrix of the arcs. Let D be a diagonal matrix with elements $d_i = \sum_j W_{ij}$, then matrix $L = D - W$ defines the non-normalized Laplacian.

The goal in semi-supervised classification is to propagate the information from very few labeled items to the entire dataset in order to label the entire dataset. The propagation process is defined as a two class problem where the labeled items are updated iteratively. One class, class-1, is defined by the relevant items and the other class, class-0, by the non-relevant items. The goal is to label the rest of the dataset from the labeled items.

Let Y be an $(N \times 1)$ -vector with value 1 on the class-1 indexes, -1 on the class-0 indexes, and 0 in the rest. Let $f : \mathcal{V} \rightarrow [-1, 1]$, a mapping defining the class each video v_i belong to. Let $F = f(\mathcal{V})$ be the $(N \times 1)$ -vector with the f -values. Since $F^T L F = \frac{1}{2} \sum_{i,j} W_{ij} (f(i) - f(j))^2$, with value zero when f is constant, the Laplacian defines a smoothness penalty term on the F values. Thus, to estimate F a quadratic optimization problem can be established as follows (see [42]).

$$\min_F \{Q(F) = F^T L F + \mu \|F - Y\|^2\} \quad (1)$$

where the first term, $F^T L F$, imposes local smoothness and the second term, $\|F - Y\|^2$, defines the regularization to the fixed labels. The constant $\mu > 0$ measures the strength of the regularization term in the solution.

In (I) all points influence the regularization term with the same weight, but not all points have the same importance on the graph, therefore weighting each vertex independently must improve the retrieving process. We can rewrite (I) as

$$\min_F \{Q(F) = F^T L F + (F - Y)^T \Lambda (F - Y)\} \quad (2)$$

where Λ is a diagonal $(N \times N)$ -matrix with $\Lambda_{ii} = \mu \lambda_i$ if i belongs to the ground truth items and zero otherwise. The Λ values are calculated inside each class as the

relative measure of the importance of each point according to its connection weights. Similar weighting has been also suggested in [37]. That is,

$$\lambda_i = \frac{d_i}{\sum_k d_k} \text{ for } i, k \in \{\text{set of index of the same class}\} \quad (3)$$

The solution of (2) is obtained from the linear system $(L + \Lambda)F = \Lambda Y$. Inverting the dense matrix $(L + \Lambda)$ has in general complexity $O(N^3)$, which can be very inefficient in time and space for large N values. In [4, 25] has been suggested that a linear combination of a few eigenvectors of L can be used to obtain reasonable solutions reducing dramatically the dimension of the problem.

Let $\{\phi_i, i = 1, \dots, N\}$ and $\{\sigma_i, i = 1, \dots, N\}$ be the eigenvectors and eigenvalues respectively of the generalized problem $L\phi = \sigma D\phi$. Then any vector $F \in R^N$ can be written as $F = \sum_i \sigma_i \phi_i$ where the smoothness of each ϕ -vector is given by $\phi^T L \phi = \sigma_i$. Clearly the smoothness of F increases if the eigenvector of lower eigenvalues are only considered. Let U be the $(N \times k)$ -matrix of the k eigenvectors with lower eigenvalue and let α be a k -vector of coefficients. Let us approximate F as $F = U\alpha$, then substituting in (2) we obtain a minimization problem in α

$$\min_{\alpha} \{Q(\alpha) = \alpha^T (\Sigma + U^T \Lambda U) \alpha - 2\alpha^T U^T \Lambda Y\} \quad (4)$$

where $\Sigma = U^T L U$. By setting the derivative of (4) to be zero w.r.t. α , we have

$$(\Sigma + U^T \Lambda U) \alpha = U^T \Lambda Y \quad (5)$$

Now the dimension of this system only depends on the number of eigenvectors considered that drastically reduce the complexity of the problem. Nevertheless, the eigenvectors of L have to be computed. Different approaches to estimate the very large matrix eigenvectors have been suggested [41, 31, 15, 9]. In our case we use the approach suggested in [9].

In summary our model for each iteration is defined by a semi-supervised propagation technique on a dense graph where a regularization problem is solved combining a quadratic fitting term with a smoothness penalty term induced by the non-normalized Laplacian.

3.2 The Feedback

The user feedback is defined as a structured object from a re-ranking of the system output. The user by clicking on the relevant items groups the system output in three categories relevant, non-relevant and unknown, represented by the labels $\{1, -1, 0\}$ respectively. The non-relevant are those items seen by the user but no clicked as relevant. The rest of the items are labeled as unknown. This grouping cannot be considered neither strict nor free of mistakes. We assume the user does not spend too much effort in exploring a large list of videos in detail and therefore some mistakes can be introduced in the feedback. In the prototype interface this fact is emphasized since initially only a keyframe per video is shown. Furthermore, the user could not

have a sharp understanding of his/her wishes, only indicating high or low preference on the items. A label remains fixed but it can be changed by the user.

Let \mathcal{D} denote the dataset, that we assume fixed. Let \mathcal{R} and \mathcal{L} be the set of all possible ranking and labeling of the dataset items respectively. Let $\mathcal{Y}_t = \{(r_i, l_j) | r_i \in \mathcal{R}, l_j \in \mathcal{L}\}$, the set where each element is defined by a ranking and a label. The values $l_j \in \{-1, 0, 1\}$ define the user preferences on the items. All elements in $\mathbf{y}_0 \in \mathcal{Y}$ are labeled to 0 but those representing the query are labeled to 1. The process starts sending $\mathbf{y}_0 \in \mathcal{Y}$ to the system, then the system computes its answer $\mathbf{y}_t \in \mathcal{Y}$, and the user returns a feedback $\tilde{\mathbf{y}}_t \in \mathcal{Y}$, $t = 1, 2, \dots$.

In this model, learning from the feedback means to update the system labels $\{l_t\}$ and hence the $\{\Lambda\}$ weights. Let $\nabla\Lambda = \Lambda_i^{t+1} - \Lambda_i^t$ denote the λ_i variations from the iteration t to the $t + 1$. According to (2) we can write

$$\lambda_i^{t+1} = \lambda_i^t + f(\tilde{\mathbf{y}}_{t+1}, \tilde{\mathbf{y}}_t) \quad (6)$$

since $\nabla\Lambda : \mathcal{Y}\mathcal{X}\mathcal{Y} \rightarrow R$ is a function of the user feedbacks. In our case is possible to have an understanding on how the function f works. Each time the user adds a new item to one of the classes the relative weight of the points inside the class decrease, since the d_i value remains constant and the weight of the connections $\sum_j d_j$ inside the class increases. The consequence is a lower propagation weight from these point on the rest of the items, but compensated by a higher number of propagation focus. In the same way, when a point is removed from a class the rest of class items increases its propagation strength. According to this property this model provides a tradeoff between propagation strength and number of propagating focus. Semantic diversity in the retrieved items can be also part of the user preferences.

In this model the user feedback is the only mechanism to increase the semantic diversity but sometimes a large number of videos have to be examined. To improve this shortcoming some alternatives have recently been suggested. In [38] it is suggested to substitute the L -matrix in (1) by a new manifold where closeness is defined from a local regression model. However, in this model a linear system have to be computed on each dataset items, which makes it useless for large datasets or interactive systems. A possible alternative would be to use richer functions f_i making use of the local structure properties in \mathcal{G} .

3.3 The Algorithm

The summary of the proposed algorithm in pseudo-code is as follows:

```

THE INITIAL OBJECT
Input:  $Q$  - user query vector
1.0 Compute  $R = \{r_i | r_i = \text{distance}(Q, e_i), e_i \in \mathcal{D}\}$ 
2.0 Rank  $R$  in increasing order
3.0 Built  $\mathbf{y}_0$  from  $R$ 
3.0 Output  $\mathbf{y}_0$ 

```

RETRIEVAL ALGORITHM

Input: U = the $n \times k$ eigenvectors-matrix of the Laplacian.
Input: Σ = the matrix defined as $\Sigma = U^T L U$
Input: Λ = the weighting values on the items.
Input: \mathbf{y}_0 : the structured object with the initial items.
Set $t=0$
0.1 Compute Λ_0 applying (Eq.3) on \mathbf{y}_0
1.0 REPEAT
1.1 Solve $(\Sigma + U^T \Lambda_t U) \alpha = U^T \Lambda_t Y$
1.2 Compute $F_t = U \alpha$
1.3 Rank the F_t values in decreasing order
1.4 Compute \mathbf{y}_t from $\{j | F_t(j) > 0\}$ fixing $l_j = 0$
1.4 REPEAT
1.4.1 Show a panel of videos using the \mathbf{y}_t ranking as index
1.4.2 Update the current feedback $\tilde{\mathbf{y}}_{t+1}$ from the user interaction
1.5 UNTIL NO-MORE-PANELS
1.6 Save in the set \mathcal{S} the current relevant items
1.7 Compute Λ_{t+1} applying (Eq.3) on the feedback $\tilde{\mathbf{y}}_{t+1}$
1.8 Set $t=t+1$
1.9 UNTIL USER END
2.0 Output \mathcal{S}

3.4 User Interaction

As already established here we focus the interest on retrieval from the video signal. In this case, the user can provide a video query or could browse the dataset to identify some initial videos. In order to help in the browsing process a fixed hierarchical grouping structure is created from an agglomerative hierarchical clustering algorithm, where the selected grouping depends on the size of dataset. Panels of fixed size of random selected videos from each cluster are shown to the user until a relevant video is clicked. This video is considered the initial query. Next, on each retrieval iteration, the user clicks on relevant items to indicate its preference. We exploit the user clicks according to the following rules:

- i By default, all items showed to the user are initially considered as non-relevant.
- ii The user can click items with the mouse left button on each individual item to toggle the class of the item, either from relevant to non-relevant or vice-versa. Positive items are marked with a green framework (see Fig2).
- iii Combining the shift key when clicking an item as positive, all previous non-relevant items are marked and considered relevant up to the next relevant item backwards.

After each user interaction the number of propagating points increases. The number of the relevant items can change by addition or removing, but the number of



Fig. 2 User interaction: relevant items in green

non-relevant ones always increases with the feedback. According to the first rule any item showed to the user and not selected as relevant is assigned as non-relevant. The propagation from non-relevant items plays a very important role eliminating possible false positives. The algorithm stops when a prefixed relevant size or a maximum number of iterations is reached. The retrieving process can be canceled by the user saving the relevant items. A new retrieving process could be started later using the saved items as initial seeds.

An important issue is to account for the mistakes in the user interaction. Here we assume the user is right when click on relevant items but due to the initial partial information shown on interface's panels in combination with the assumption from the first rule, false negative can be introduced. The user can remove relevant items from the current list en each feedback but the non-relevant ones remains. In the experimental section we evaluate the consequences of this effect.

4 Prototype Design

4.1 Architecture

For the implementation of the prototype, two main components have been considered:

- The search engine.
- The graphical user interface (GUI), which provides facilities both for starting a query and for the further user interaction.

The search engine is a standalone program that initially loads the pre-computed database features and performs a video query as described in section 3.3. For the sake of efficiency, this program has been implemented using the C++ programming language and the OpenCV library [20]. Currently, it receives all query parameters and provides all query results through plain text files. Thus, it is very easy to test the search engine without using the GUI component. This also allows to simulate the user behavior in order to obtain quantitative measures of the obtained results.

The GUI component has been developed using a standard client-server architecture through the web. An application server, Tomcat v6 in this case, provides modules to run servlets, JavaServer Pages (JSP), and the JavaServer Standard Tag Library (JSTL). The client side is based on HTML5, Javascript and CSS3. It has

been mainly developed and tested using Firefox (version 10.0 or later) and Chrome (version 17.0 or later), because they better satisfy the current standards. However, almost any modern web browser satisfying these standards could be used to run the video retrieval prototype application. This component allows the user both to start a query and to refine the results obtained performing several iterations. The information provided by the user on the client side is sent to the server where it is conveniently transformed to the required format. Then, the application server runs the search engine, analyzes the results provided and sends them to the client where they are shown to the user.

For this very first version of the prototype, the search engine is run for every query/iteration. This means that its initialization (the loading of the pre-computed database features) is repeated every time. In a forthcoming version of the prototype, the search engine will be started as a system daemon, which will receive queries and will provide results through a communication channel (hard pipelines, sockets, etc.). Therefore, the loading of the database features will be done only once when the daemon is started.

Taking into account that the video database contains videos of different qualities (some of them are high definition videos) and in order to reduce the traffic of information between the client and the server sides, the GUI uses a different copy of the video database where all videos have been scaled to fit the video display area of the GUI. Also the *webm* standard format has been used to take advantage of the video playing facilities provided by the HTML5 standards.

4.2 *Prototype GUI*

In order to start a query, the user must provide at least one video file to the system, in either of the following ways:

- Directly, providing a video file.
The user can browse the local filesystem and select a video file that will be transferred to the server. Once in the server, the file will be preprocessed and the corresponding features will be extracted. These features will be used to start a first query over the database.
- Selecting a video file from the database.
In this case, the user can browse through the video database to choose some initial query videos related to the semantic concepts the user is looking for. As already mentioned a clustering process (see subsection 3.4) has been performed over the whole video database. Chunks of 32 items are formed from cluster representatives, randomly chosen. One item only appears in one chunk. The user is requesting new chunk to the system until he/she finds one or more videos to start the query.
- Selecting one of a series of prefixed concepts/categories.
A series of prefixed concepts/categories will be shown to the user, who can select one or more out of them. The interface will then show the user some samples of

the selected category(ies) and the user can select one or more video samples to start the initial query.

Once the initial query has been initialized, the server processes the query and provides a video ranked index. The interface uses the index to show the videos in pages of fixed size (32) allowing the user to move forward or backward over all available pages. Although the interface only shows one keyframe from each video, when the user hovers the mouse pointer over a keyframe, a video summary of keyframes is shown. Furthermore, if the user clicks on a keyframe, the whole video will be reproduced in the video display area.

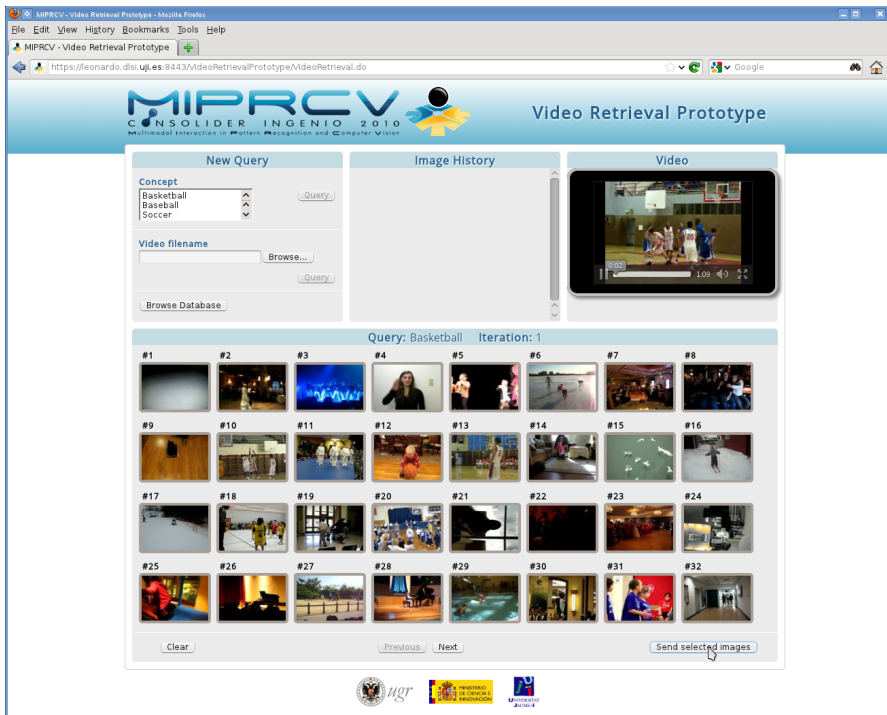


Fig. 3 Video retrieval web interface. Top part: query and video player area. Bottom part: retrieval results area.

Fig. 3 shows a screenshot of the prototype. At the top of the window, we can see the *New Query* area from where the user can start a new query by providing a local video file or by browsing the pre-computed clusters. At the top right part, we can find the panel where a specific video can be played. The central top section of the window contains the history of the selected videos in previous iterations, as shown in Fig 4. The bottom area shows the keyframes of the videos provided in the last iteration of the current query. Fig 5 shows an example of retrieved videos after two iterations.



Fig. 4 The "Image History" at the top center of the interface shows the selected videos by the user from previous iterations. The user can browse a video summary by hovering on a video key frame.

5 Database and Experiments

5.1 The CCV Database

The experimental results for testing the algorithm, has been done using the Columbia Consumer Video (CCV) Database –a Benchmark for Consumer Video Analysis [14]. The video collection is designed large enough and diverse as to run experiments with the different combinations of video descriptors. At the moment there are more than 9.000 videos into the CCV database. There are about 200 hours of Youtube videos over 20 semantic categories, and the average length of the videos is 80 seconds.

The database was collected with extra care to ensure relevance to consumer's interest and originality of video content without post-editing. Such videos typically have very little textual annotation and thus can benefit from the development of automatic content analysis techniques [14].

The dataset for the experiments consists of 8.351 videos, a subset of the full CCV, annotated in 20 classes: Basketball, Baseball, Soccer, IceSkating, Skiing, Swimming, Biking, Cat, Dog, Bird, Graduation, Birthday, WeddingReception,



Fig. 5 The retrieved videos in each iteration are shown in a ranked order of relevance (bottom area)

WeddingCeremony, WeddingDance, MusicPerformance, NonMusicPerformance, Parade, Beach, Playground, Other.

Fig. 6 shows the number of positive samples per category in CCV database. The used videos were a fewer less than the original database amount, due to there were some unavailable videos to download. In fact, the videos used for the experiments were 8435 instead of 9317 of the original dataset. Moreover, in the 8435 video set were 84 unlabeled videos. Thus, 8351 videos were used eventually.

5.2 Video Preprocessing

Different types of low-level descriptors are considered in this prototype to decode the static and motion information. The CCV database [14] provides a set of descriptors based on SIFT and MFCC features in order to perform evaluation experiments as a benchmark. In our experiments we have used these descriptors.

The descriptors provided in [14] have the following characteristics: a) SIFT: each video is summarized in a 5000 dimension histogram descriptor built after collapsing the SIFT descriptors in bag of words using two different vocabularies of 500 words and two spatial layouts; b) STIP: the STIP descriptors are collapsed in a bag of

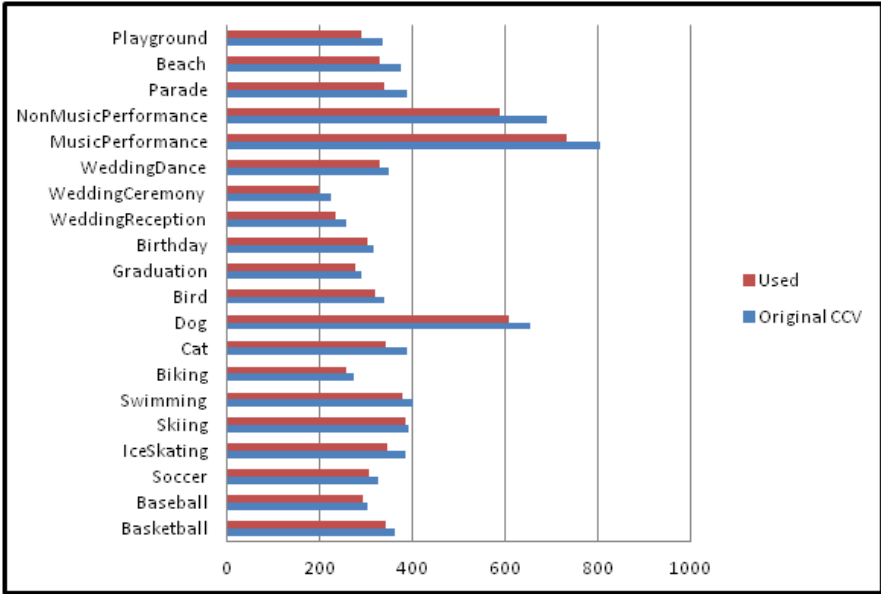


Fig. 6 Number of positive samples per category

words using a vocabulary of 5000 words; c) MFCC: in this case the vocabulary size is 4000 words. In all cases a PCA transformation have been applied in order to reduce the dimension to 512, keeping 98% of the variance. Eventually, all vectors are L2 normalized.

5.3 Experiments

Several experiments using different features have been conducted in order to evaluate the prototype. SIFT, MFCC and SIFT+MFCC features were evaluated. In all the experiments the improvement of the SIFT+MFCC combination is very small (<1%) compared with SIFT. Here only the SIFT results are shown. We start evaluating the propagation model for a large number of iterations. To evaluate the retrieval score we run experiments with a maximum of five feedback iterations with the goal of retrieving 20 new items. In the last group of experiments we evaluate the incidence of assuming a percentage of false negatives in the user decision. All experiments are simulated and the results shown are the average of five repetitions. In all cases the user only examines a panel with the first 32 items from the system output to provide the feedback. In all the experiments the λ_i -values are estimated using only the twenty-five values with higher d_i -values, that is $\lambda_i = \frac{d_i}{\sum_{j=1}^{25} d_j}$.

Fig. 7 and 8 show the results on all classes with ten iterations using the SIFT and MFCC descriptors. The graphics show a near lineal behavior in the number of retrieved items for most of the classes. However the retrieval ratio per iteration

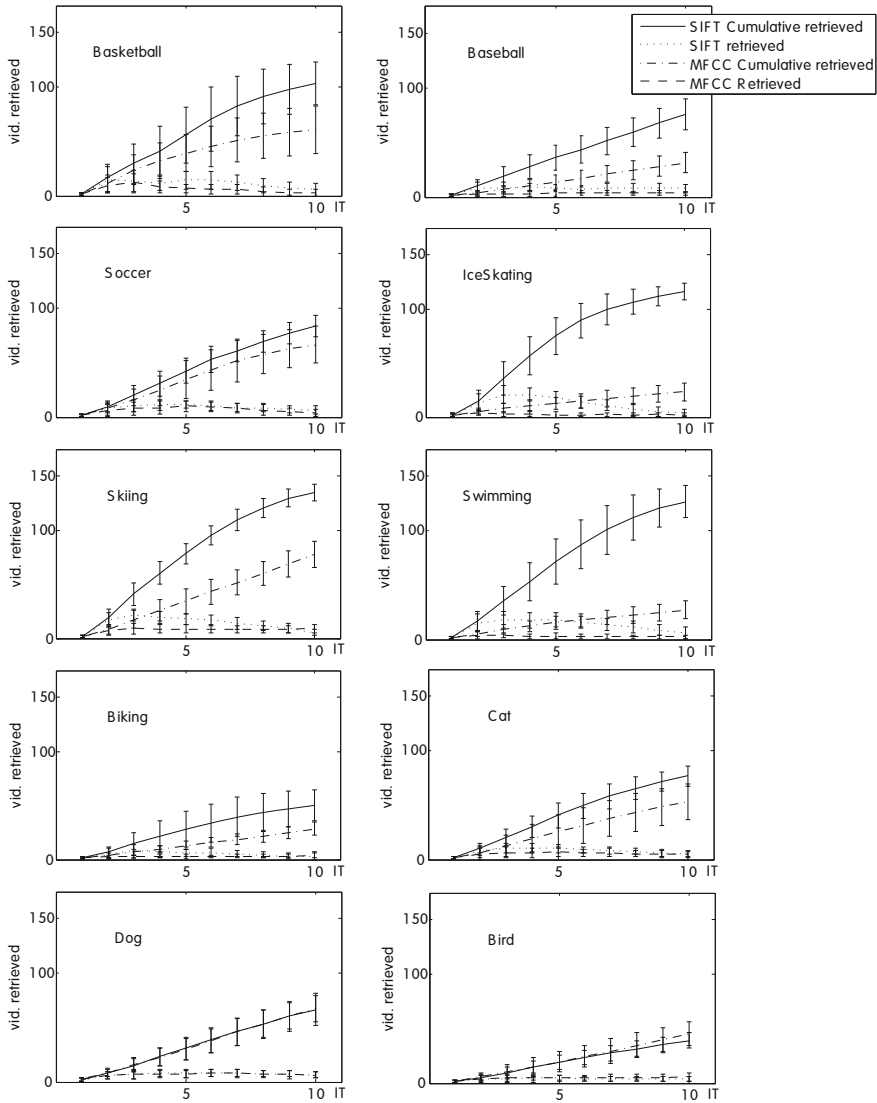


Fig. 7 Curves show the average number of retrieved items and cumulative values on each iteration. The vertical segment indicates the standard deviation of the estimation.

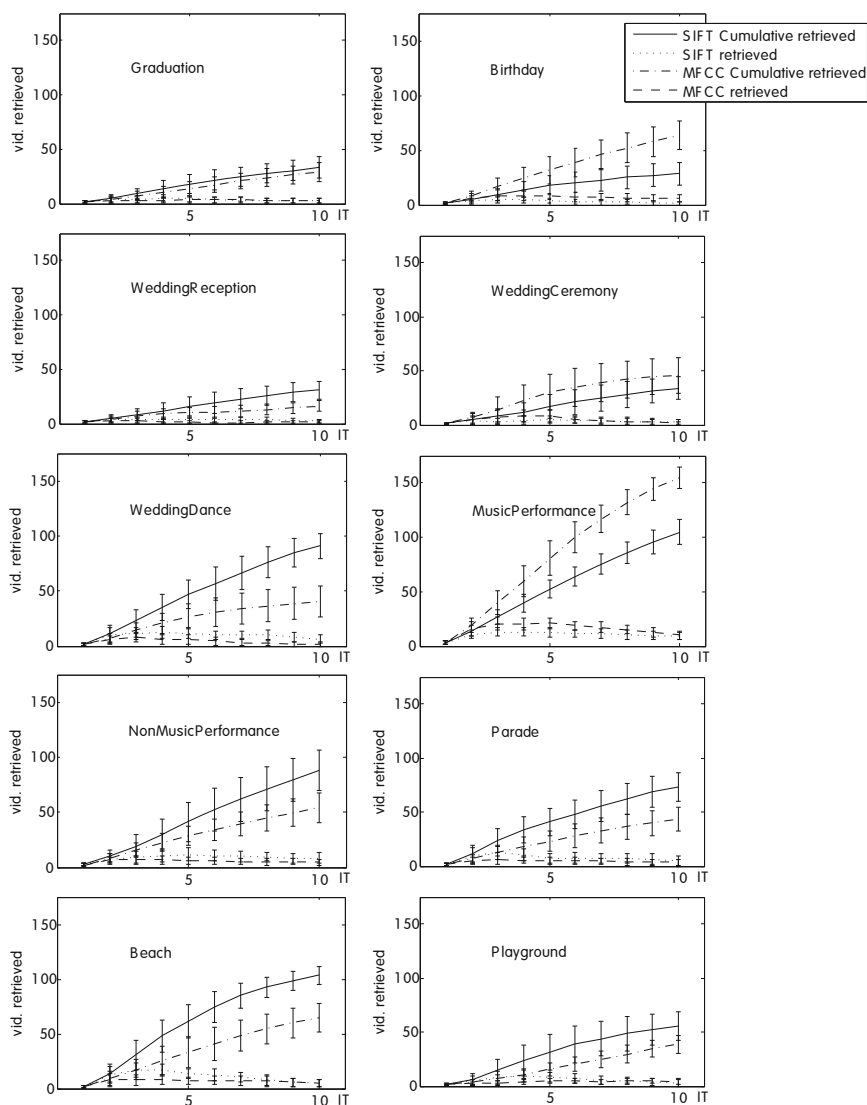


Fig. 8 Curves show the average number of retrieved items and cumulative values on each iteration. The vertical segment indicates the standard deviation of the estimation.

Table 1 Mean and standard deviation (std) of retrieved items on the fifth iteration

Class	1	2	3	4	5	6	7	8	9	10
Mean	39.1	13.4	34.1	12.8	34.4	15.3	12.8	25.5	29.9	19.4
Std	17.1	6.8	17.3	5.7	11.3	5.8	3.9	15.3	9.6	9.4
Class	11	12	13	14	15	16	17	18	19	20
Mean	13.6	31.7	9.9	30.3	26.7	80.7	28.2	23.0	33.0	15.6
Std	6.6	12.3	5.3	16.6	11.1	15.5	8.5	9.7	13.7	5.9

changes very much with respect to the class. Comparing the SIFT and MFCC curves it is observed that in general the SIFT descriptor is more discriminative, although some classes such as WeddingCeremony or Birth the MFCC descriptor works better than the SIFT one. The Table 1 shows the variation interval for the fifth iteration. It can be noted the high value of the standard deviation indicating a high variability in the results depending on the initial query.

Table 2 Retrieval score after five iterations on the 20 classes using SIFT features. Each row shows the percentages of labeled items (0%, 3%, 5%) respectively. 100% score means 20 retrieved items.

Class	1	2	3	4	5	6	7	8	9	10
0%	93.1	94.9	96.0	98.6	100.0	98.1	77.7	98.8	97.4	83.4
3%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.7
5%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.3
Class	11	12	13	14	15	16	17	18	19	20
0%	84.6	82.0	74.5	62.9	96.5	100.0	97.6	99.0	100.0	84.7
3%	98.8	99.8	96.8	97.4	100.0	100.0	100.0	100.0	100.0	100.0
5%	100.0	98.1	97.2	98.9	100.0	100.0	100.0	100.0	100.0	100.0

Table 2 and 3 show the retrieval results for SIFT and MFCC respectively assuming that the dataset contains a small percentage of labeled relevant items. To run these experiments we randomly select from each class forty videos to be used as video-queries. Experiments have been performed using descriptors from the SIFT, MFCC and SIFT+MFCC features. These numbers represent the average result of the forty queries on the five different partitions. From Fig. 6 it can be seen the number of labeled items per category for the fixed percentages (0%, 3%, 5%) is in general low, in most cases lower than ten. The SIFT results show that 15 out of the 20 classes saturate with 3% of labeled items. Of course the existence of labeled items increases very fast the number of retrieved items since all items sharing a label are automatically selected when the the user clicks on one of them. For the MFCC features the result is not so good as for the SIFT one, but it also reaches a high score when a 3% of labeled items is assumed.

Table 3 Retrieval score after five iterations on the 20 classes using MFCC features. Each row shows the percentages of labeled items (0%, 3%, 5%) respectively. 100% score means 20 retrieved items.

Class	1	2	3	4	5	6	7	8	9	10
0%	92.9	69.7	79.6	57.1	98.5	74.4	64.2	85.0	97.8	79.8
3%	100.0	95.0	100.0	95.8	100.0	97.8	92.0	98.3	100.0	98.1
5%	100.0	98.1	100.0	96.2	100.0	98.8	93.3	100.0	100.0	99.7
Class	11	12	13	14	15	16	17	18	19	20
0%	72.1	96.3	44.9	75.7	83.6	100.0	98.8	90.4	91.0	66.5
3%	96.1	100.0	90.9	98.9	99.6	100.0	100.0	99.6	100.0	98.4
5%	97.0	100.0	84.3	99.6	99.6	100.0	100.0	100.0	100.0	99.6

Table 4 Retrieval score in the first five iterations on the 20 classes using SIFT features. Each row shows the iteration score for the first five iterations (IT-*). 100% score means 20 retrieved items.

Class	1	2	3	4	5	6	7	8	9	10
IT-1	56.8	18.9	35.2	47.1	47.0	64.2	19.1	20.3	31.7	15.8
IT-2	78.8	48.2	74.0	85.8	84.2	91.2	50.3	59.5	67.8	34.6
IT-3	83.0	67.0	88.0	94.8	94.6	95.3	70.3	80.1	83.7	55.6
IT-4	85.4	82.2	94.0	96.5	98.7	97.7	78.0	89.3	92.7	72.5
IT-5	87.2	91.6	97.4	97.9	99.9	98.7	83.2	96.0	96.7	83.7
Class	11	12	13	14	15	16	17	18	19	20
IT-1	21.4	26.9	34.9	16.0	43.2	25.9	26.8	51.4	50.8	29.5
IT-2	44.4	52.9	67.9	37.5	85.8	68.2	59.2	79.5	81.0	54.9
IT-3	64.7	70.2	80.4	54.0	95.2	90.3	78.4	90.5	91.4	72.2
IT-4	76.9	77.8	84.7	65.2	97.1	97.5	91.2	95.7	96.7	82.7
IT-5	85.1	82.8	87.3	73.9	98.2	99.5	96.7	99.0	99.0	90.0

Table 5 Retrieval score in the first five iterations on the 20 classes using MFCC features. Each row shows the iteration score for the first five iterations (IT-*). 100% score means 20 retrieved items.

Class	1	2	3	4	5	6	7	8	9	10
IT-1	38.5	10.8	19.9	16.1	31.5	19.4	7.0	21.6	26.8	14.9
IT-2	62.8	25.7	54.3	35.1	65.0	41.2	17.2	47.3	54.6	32.2
IT-3	79.3	39.8	71.7	47.1	82.7	56.5	27.3	63.1	76.5	49.3
IT-4	88.2	52.4	82.7	54.5	92.2	66.8	39.0	75.8	89.6	64.7
IT-5	91.7	65.1	89.6	60.3	96.2	74.8	49.4	84.8	96.7	75.7
Class	11	12	13	14	15	16	17	18	19	20
IT-1	17.6	29.8	26.9	21.0	30.3	52.8	32.6	13.7	33.6	14.2
IT-2	33.5	54.9	49.5	38.7	55.1	84.2	64.5	36.6	64.5	28.7
IT-3	45.7	74.8	65.1	52.5	73.8	96.3	82.8	52.4	80.7	43.1
IT-4	55.0	90.7	71.8	64.0	82.0	99.3	91.5	65.9	89.5	56.6
IT-5	63.8	96.5	76.5	71.5	87.3	99.8	96.8	75.3	93.8	68.1

Table 4 and 5 show the evolution of the retrieval results along the five iterations for the SIFT and MFCC features. In this experiment no labeled items in the dataset are assumed, all relevant items have to be clicked by the user. In this case the per class score on the fifth iteration is a bit lower than the one showed in the Table 2 and 3. Table 6 shows the average percentage of retrieved items per iteration. The most relevant fact from this table is the increasing in retrieval after the first feedback. As shown in the Fig. 7 and 8 the improving rate is almost linear with the iterations.

Table 6 Average rate of retrieved items on each iteration on the goal of 20 items in five iterations

ITER-1	ITER-2	ITER-3	ITER-4	ITER-5
1.4	9.0	5.6	2.8	1.2

Table 7 Retrieval score in the first five iterations on the 20 classes using SIFT features assuming a 50% false negatives on the user decision. Each row shows the iterations 1-5. 100% score means 20 retrieved items.

Class	1	2	3	4	5	6	7	8	9	10
IT-1	7.2	6.4	7.0	7.8	7.8	7.5	6.3	7.5	10.7	7.8
IT-2	58.8	35.9	35.8	57.4	65.1	59.4	20.8	33.0	35.5	24.9
IT-3	80.3	62.8	63.9	86.8	91.7	86.7	39.6	60.5	61.1	40.9
IT-4	89.5	80.4	82.0	95.0	97.9	94.1	56.1	79.8	81.2	57.4
IT-5	93.4	90.8	92.4	97.9	99.0	96.2	69.6	92.0	91.2	71.5
Class	11	12	13	14	15	16	17	18	19	20
IT-1	7.3	7.2	6.3	5.2	7.2	12.5	10.9	7.8	6.8	6.8
IT-2	21.8	19.9	17.9	17.4	40.4	50.2	37.3	44.1	48.3	24.4
IT-3	37.7	34.2	30.7	31.0	72.3	84.2	60.5	73.2	82.8	45.3
IT-4	52.6	47.5	43.6	44.6	90.7	96.4	81.9	87.9	95.6	65.1
IT-5	66.1	58.2	55.5	57.8	95.5	99.2	93.2	95.0	98.7	78.5

Table 7 and 8 show the results after assuming a 50% of false negatives on the user feedback. This has been implemented selecting with probability 0.5 the relevant items present in the ranked system output. If we compare these tables with the Table 4 and 5 it can be noted that, as it was expected, the false negative have a strong incidence. The Table 9 shows the retrieved averages using SIFT descriptors. The row N-FN shows the results when false negatives are not considered, that is, the same result shown in the Table 4 and 5. The row FN50-P shows the case in which the false negatives have been taken into account as recovered items by the system but they are labeled as non-relevant in the feedback. The third row shows the results when the false negatives are considered as non-relevant. The results from the MFCC descriptors present a similar behaviour but with lower scores.

Table 8 Retrieval score in the first five iterations on the 20 classes using MFCC features assuming a 50% false negatives on the user decision. Each row shows the iterations 1-5. 100% score means 20 retrieved items.

Class	1	2	3	4	5	6	7	8	9	10
IT-1	6.5	6.8	6.5	5.9	7.5	10.7	10.2	7.3	7.1	6.7
IT-2	20.2	26.5	16.8	24.7	29.7	62.1	36.4	26.0	33.2	16.7
IT-3	32.2	52.9	26.7	47.4	51.7	89.3	63.6	47.0	57.1	28.7
IT-4	42.6	73.4	33.7	61.5	66.0	96.7	81.7	63.2	72.0	41.8
IT-5	52.5	86.5	38.9	68.5	74.2	99.6	90.5	75.0	82.4	55.9
Class	11	12	13	14	15	16	17	18	19	20
IT-1	7.1	7.0	7.3	7.5	8.0	7.7	7.2	7.2	10.5	7.4
IT-2	36.4	17.8	30.8	18.8	38.5	23.6	16.6	27.4	31.7	23.4
IT-3	60.3	29.6	54.2	31.6	68.7	38.2	26.4	50.2	53.9	39.7
IT-4	73.2	42.5	70.5	41.3	85.4	51.1	37.6	65.3	74.5	55.9
IT-5	81.0	54.2	81.2	48.1	93.6	60.6	49.3	74.8	88.5	68.2

Table 9 Retrieved average in percentage on 20 items in the fifth iteration using SIFT. N-FN indicates that non false negative are considered. FN50-P indicates that false negative are partially considered (see text). FN50 indicates that false negative are consider in full.

	ITER-1	ITER-2	ITER-3	ITER-4	ITER-5
N-FN	0.34	0.65	0.79	0.87	0.92
FN50-P	0.09	0.47	0.73	0.85	0.91
FN50	0.07	0.27	0.47	0.61	0.71

6 Discussion and Conclusion

An interactive prototype for non annotated video retrieval in a large dataset has been proposed. The system representation model is an adaptive dense graph, where the weights are updated after each user feedback. The prototype engine has been implemented in C/C++ with the help of the OpenCV library [20]. The interface has been developed also using a mix of several web technologies. A simple retrieving protocol has been used for testing: to retrieve 20 new relevant items in five interactions where the feedback is only defined from 32 items. The experimental results show a high average retrieving score although there is not homogeneity among classes. A group of classes shows lower general score in all the experiments.

The decreasing in retrieval rate with respect to the iteration number seen in Fig. 7 and 8 and the large value of the standard deviation in all cases shows the high influence of the initial items. This is partially justified since we always start from only one item and the semantic gap precludes descriptor homogeneity inside the classes. This result show the needing of better updating functions for the λ_i weights and the use of better smoothing functions.

The comparison between rows N-FN and FN50-P indicates that deleting a large percentage of relevant items in the feedback has a low influence in the results of the next iteration. This could be explained assuming the similarity distance from each

relevant item to its class is small. The influence of the false negative on the retrieval score points out the needing of an interface precluding as much as possible the loss of relevant items.

In summary, this prototype presents a good relevance feedback representation model on the difficult problem of retrieving non-annotated videos from the CCV database. The retrieval scores are encouraging for the used testing protocol. Of course some shortcoming have been identified and some properties need much more experimentation. Future work is directed to an in deep evaluation on a larger database in order to analyze and obtain a better understanding of the propagation model for further improvements in the retrieval rate.

References

1. Lscom lexicon definitions and annotations version 1.0. Technical Report 217-2006-3, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report
2. ACM Grand Challenge (2010), <http://comminfo.rutgers.edu/conferences/mmchallenge/>
3. Multimedia Analysis and Retrieval System
4. Chapelle, O., Scholkopf, B., Zien, A.: *Semi-Supervised Learning*. MIT Press (2006)
5. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large Scale Online Learning of Image Similarity through Ranking. In: Araujo, H., Mendonça, A.M., Pinho, A.J., Torres, M.I. (eds.) *IbPRIA 2009*. LNCS, vol. 5524, pp. 11–14. Springer, Heidelberg (2009)
6. Cotton, C., Ellis, D.: Audio fingerprinting to identify multiple videos of an event. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing* (2010)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection, vol. 1, pp. 886–893. *IEEE Computer Society*, Washington, DC (2005)
8. de Rooij, O., Worring, M.: Browsing video along multiple threads. *IEEE Transactions on Multimedia* 12(2), 121–130 (2010)
9. Fergus, R., Weiss, Y., Torralba, A.: Semi-supervised learning in gigantic image collections. In: *NIPS* (2009)
10. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining Inference and prediction*. Springer (2009)
11. Hauptmann, A.G., Lin, W.H., Yan, R., Yang, J., Chen, M.Y.: Extreme video retrieval: joint maximization of human and computer performance. In: *14th Annual ACM International Conference on Multimedia*, pp. 385–394. ACM Press, New York (2006)
12. Hearst, M.A.: *Search user interfaces*. Cambridge University Press (2009)
13. MARVEL IBM, <http://mp7.watson.ibm.com/imars/demos/>
14. Jiang, Y.-G., Ye, G., Chang, S.-F., Ellis, D., Loui, A.C.: Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: *Proceedings of ACM International Conference on Multimedia Retrieval (ICMR)*, Oral Session (2011)
15. Kumar, S., Mohri, M., Talwalkar, A.: Sampling techniques for the nystrom method. In: *AISTATS* (2009)
16. Kushki, A., Androustos, P., Plataniotis, K.N., Venetsanopoulos, A.N.: Query feedback for interactive image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 14(5), 644–655 (2004)
17. Laptev, I., Pérez, P.: Retrieving actions in movies, pp. 1–8 (October 2007)

18. Le, Q.V., Ranzato, M.A., Mong, R., Devin, M., Chen, K., Corrado, G.S., Dean, J., Ng, A.Y.: Building high-level features using large scale unsupervised learning. In: Twenty-Ninth International Conference on Machine Learning (2012)
19. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Twenty-Sixth International Conference on Machine Learning (2009)
20. The OpenCV Library, <http://opencv.willowgarage.com/wiki/>
21. Lowe, D.C.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
22. Müller, H., Clough, P., Deselaers, T., Caputo, B. (eds.): *ImageCLEF, Experimental Evaluation in Visual Information Retrieval*. The Information Retrieval Series, vol. 32. Springer (2010)
23. Rui, Y., Huang, T.S., Mehrotra, S.: Content-based image retrieval with relevance feedback in mars. In: *Proc. IEEE Int. Conf. on Image Proc.*, pp. 815–818 (1997)
24. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. *IJCV* 37(2), 151–172 (2000)
25. Schoelkopf, B., Smola, A.: *Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press (2002)
26. Smeaton, A.F., Over, P., Kraaij, W.: Trecvid: evaluating the effectiveness of information retrieval tasks on digital video. In: *12th Annual ACM International Conference on Multimedia*, pp. 652–655. ACM, New York (2004)
27. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12), 1349–1380 (2000)
28. Snoek, C.G.M., Worring, M.: Concept-based video retrieval. *Foundations and Trends in Information Retrieval* 4(2), 215–322 (2009)
29. Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.-M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: *Proceedings of ACM Multimedia, Santa Barbara, USA*, pp. 421–430 (2006)
30. Snoek, C.G.M., Everts, I., van Gemert, J.C., Geusebroek, J.M., Huurnink, B., Koelma, D.C., van Liempt, M., de Rooij, O., van de Sande, K.E.A., Smeulders, A.W.M., et al.: The mediamill trecvid 2007 semantic video search engine. In: *5th TRECVID Workshop (2007)*
31. Talwalkar, A., Kumar, S., Rowley, H.: Large-scale manifold learning. In: *CVPR (2008)*
32. TRECVID Multimedia Event Detection Track, <http://nist.gov/itl/iad/mig/med.cfm>
33. Trecvid, <http://trecvid.nist.gov/>
34. Trecvid-2004, <http://www-nlpir.nist.gov/projects/tv2004/tv2004.html>
35. Trecvid-2005, <http://www-nlpir.nist.gov/projects/tv2005/tv2005.html>
36. Wactlar, H.D., Kanade, T., Smith, M.A., Stevens, S.M.: Intelligent access to digital video: Informedia project. *IEEE Computer* (1996)
37. Wang, J., Jebara, T., Chang, S.-F.: Graph transduction via alternating minimization. In: *Proceedings of the 25th International Conference on Machine Learning (2008)*
38. Yang, Y., Nie, F., Xu, D., Luo, J., Zhuang, Y., Pan, Y.: A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4), 723–742 (2012)
39. Zheng, Y.-T., Neo, S.-Y., Chen, X., Chua, T.-S.: Visiongo: towards true interactivity. In: *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR 2009*, pp. 51:1. ACM, New York (2009)

40. Zhou, X.S., Huang, T.S.: Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst.* 8(6), 536–544 (2003)
41. Zhu, X., Lafferty, J.: Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. In: *ICML (2005)*
42. Zhu, X.: Semi-supervised learning literature survey. Technical Report Computer Sciences TR 1530, University of Wisconsin – Madison (Last modified on July 19, 2008)

Exploiting Multimodal Interaction Techniques for Video-Surveillance

Marc Castelló, Jordi González, Ariel Amato, Pau Baiget,
Carles Fernández, Josep M. Gonfaus, Ramón A. Mollineda,
Marco Pedersoli, Nicolás Pérez de la Blanca, and F. Xavier Roca

Abstract. In this paper we present an example of a video surveillance application that exploits Multimodal Interactive (MI) technologies. The main objective of the so-called VID-Hum prototype was to develop a cognitive artificial system for both the detection and description of a particular set of human behaviours arising from real-world events. The main procedure of the prototype described in this chapter entails: (i) *adaptation*, since the system adapts itself to the most common behaviours (qualitative data) inferred from tracking (quantitative data) thus being able to recognize abnormal behaviors; (ii) *feedback*, since an advanced interface based on Natural Language understanding allows end-users the communication with the prototype by means of conceptual sentences; and (iii) *multimodality*, since a virtual avatar has been designed to describe what is happening in the scene, based on those textual interpretations generated by the prototype. Thus, the MI methodology has provided an adequate framework for all these cooperating processes.

1 Introduction

The main objectives of a Video Surveillance (VS) system are typically set to achieve detection of anomalies and/or recognition of a predefined set of agent behaviors

Marc Castelló · Jordi González · Ariel Amato · Pau Baiget · Carles Fernández ·
Josep M. Gonfaus · Marco Pedersoli · F. Xavier Roca
Centre de Visió per Computador, Dept. Ciències de la Computació, Universitat Autònoma de
Barcelona, Barcelona, Spain
e-mail: {mcastello, poal, aamato, pbaiget, perno},
 {gonfaus, marcopedede, xavir}@cvc.uab.es

Ramón A. Mollineda
Instituto de Nuevas Tecnologías de la Imagen, Universitat Jaume I, Castelló, Spain
e-mail: ramon.mollineda@lsi.uji.es

Nicolás Pérez de la Blanca
Dpto. Ciencias de la Computación e I.A., ETSI Informática y de Telecomunicación,
Universidad de Granada, Granada, Spain
e-mail: nicolas@ugr.es

arising from real-world events and in real-time [1]. One strategy for embedding *cognition* in VS systems is to convey such recognized events to end-users based on Natural-Language texts. This skill is justified by the fact that we can demonstrate that we understand what is going on in a scene if we are able to describe it using Natural Language [2].

This communication capability is considered an important step towards developing a *Cognitive Video Surveillance* (CVS) system [4]. For example, a CVS system exploiting language capabilities would require to implement the following functionalities [6]: (i) *adaptation*, i.e. the system adapts its recording conditions for best keeping track of agents and for best inferring contextual knowledge; (ii) *feedback*, i.e. an advanced interface based on Natural-Language understanding supports conceptual feedback where input texts define the end-user's queries and commands; and (iii) *multimodality*, i.e. inferred conceptual predicates are converted into Natural-Language sentences for allowing the system its communication with end-users, for example by means of a virtual avatar.

An envisaged prototype focused on the aforementioned cognitive aspects requires the implementation of several modules like, at the very least, to detect motion while keeping track of agents over time; to infer high-level, conceptual descriptions based on agent trajectories combined with context (i.e. interacting objects and knowledge about the scene [3]); and to communicate those inferred interpretations to human operators using Natural-Language texts (and preferably in multiple languages).

This chapter introduces a CVS system as defined before. In essence, the prototype we have developed consists of the following steps:

- Those events detected in image data-streams are obtained from a system composed of static or active cameras, see Fig. 1(i), for example following the architecture presented in [7].
- Since common background subtraction detection and filter-based tracking allows the system to estimate the trajectories of moving agents within the scene, these trajectories constitute the basis of knowledge for further inference, as in [4].
- Semantic models defined by experts using logic formalism are used to infer interactions and behaviors. The use of ontologies is very helpful for guiding the top-down modelling of the expert database, while providing a framework which centralizes the multiple types of knowledge involved in CVS (e.g. visual, conceptual and linguistic).
- The generation of Natural-Language is usually based on linguistic models which convey those semantic concepts inferred from behavioral models.
- Understanding textual queries from end-users is the basis for feedback, since these queries can modify the functionality of the system by incorporating new knowledge (e.g. refining a description or giving a better estimation of *why* a specific behavior is happening) or by modifying the acquisition conditions.
- Lastly, an advanced, multimodal interface is designed to communicate the results to end-users using texts [5]. These Natural-Language sentences are reported by means of a virtual avatar who *explains* us using speech what is happening in the scene, see Fig. 1(ii).

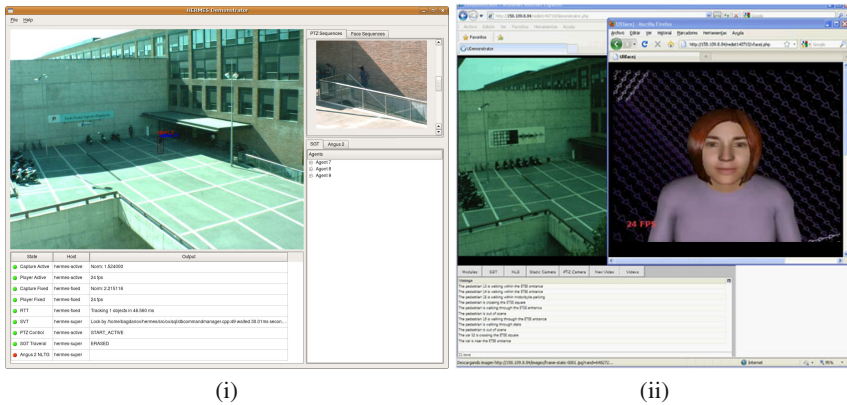


Fig. 1 Examples of advanced interfaces exploiting Natural-Language, which support conceptual feedback plus a communication with human operators using virtual avatars

In the sequel, we will describe the methodology used to implement the capabilities of adaptation, feedback and multimodality in a surveillance prototype.

2 Methodology and Theoretical Background

2.1 Adaptation: Anomaly Detection Based on Trajectory Analysis

Nowadays, the detection of anomalies in video sequences is considered a hot topic in video understanding research [6]. This issue is caused not by the difficulty of implementing an anomaly detector, but because it is unclear which is the best definition of anomaly. On the one hand, the concept of anomaly is usually related in video-surveillance to suspicious or dangerous behaviors, i.e. those for which an alarm should be fired when detected.

From a statistic point of view, normal behavior occurs more frequently than anomalous behavior. This is the main assumption of all works performed in anomaly detection research [12, 11, 4]. Since an anomaly is a deviation from what is considered normal, these two concepts, normality and anomaly, are complementary. In the video-surveillance domain, standard learning procedures use observations extracted by a motion tracking algorithm over a continuous recording to build a model of the scenario that will determine somehow the normality or abnormality of new observations: trajectories, understood as the series of positions of an object over time, from entering to exiting a scene, are considered by most authors as the most useful information to embed the behavior of moving objects [15].

Extensive work has been done on behavior understanding based on trajectory analysis: Makris and Ellis [4] considered spatial extensions of trajectories to construct path models, which were updated when new trajectories were matched. A similar approach was previously used in [14]. Piciarelli and Foresti [9] presented



Fig. 2 Main steps for anomaly detection

an online modeling algorithm to obtain a hierarchy of typical paths. Hu et al. [11] obtained motion patterns by spatially and temporally clustering trajectories using fuzzy c-means. More Recently, Basharat et al. [12] modeled a probability density function at every pixel location by means of Gaussian Mixture Models (GMM), considering not only spatial coordinates but also object sizes.

Inspired in these works, the procedure depicted in Fig. 2 is applied for each pair $(s, e) \in S \times E$. Let $T_{s,e} = \{t_1, \dots, t_N\}$ be the set of trajectories starting at position s and ending at e . Each trajectory t_i is represented by a sequence $r(t_i)$ of K equally spaced control points sampled from the tracked points $P_{s,e} = \{p_{i_1}, \dots, p_{i_K}\}$.

The input of the algorithm consists of the set of trajectories starting in an entry point $s \in S$ and ending in an exit point $e \in E$. These trajectories are represented in a $K \times N$ matrix, where N is the number of trajectories and K is the number of control points that have been sampled for each trajectory. Each trajectory is associated to one of the routes R_c that form the trajectory $P_{s,e}$. This is represented by points (k, c) , which contains the list of k -th points of all the trajectories being currently associated to the route c .

Subsequently, the number of Gaussian components that best represent each route, see 3 route examples in Fig. 3. The list of trajectories for each route is modeled using a one-component and a two-component GMMs. After applying the algorithm to each pair of entry and exit areas, M contains the set of normal paths that trajectories should pass in the future. Thus, a trajectory deviating from M will be considered as an anomaly.

Thus the prototype is able to differentiate among three kinds of anomaly with respect to the current trajectory over the scenario, see Fig. 4:

- Soft Anomaly (SA): Some parts of a trajectory t_a are classified as SA if they follow a modeled path, but there are sudden changes in speed or orientation that differ from the learnt routes from s to e .
- Intermediate Anomaly (IA): A trajectory t_b is classified as an IA if the most probable path $P_{s,e}$ changes from one learnt route to another.
- Hard Anomaly (HA): A trajectory t_c is classified as a HA if it has performed a completely unobserved path. This can be caused because the trajectory started from an entry point e/E or because the probability of any path beginning in the actual entry point is too low for the whole scene.

This description of anomalies represents different degrees of deviation between a new observation and the learnt model. Indeed, SA can be considered as a route

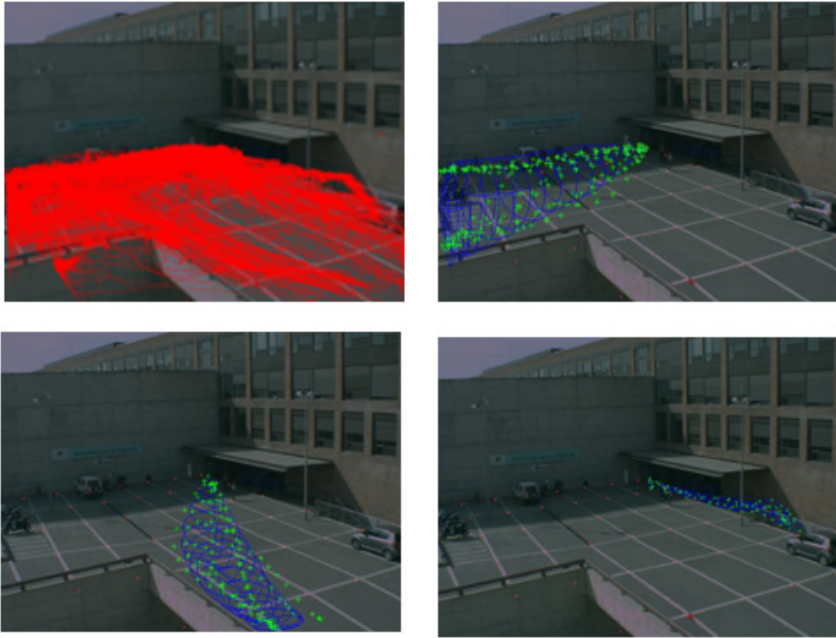


Fig. 3 Learned trajectories and the probabilistic paths of three routes

deviation inside a path. IA detects path deviations inside the same model M . Finally, HA represents a complete deviation from M .

Summarizing, Fig. 5 shows the modular schema of our anomaly detector within the Multimodal Interactive paradigm developed during the MIPRCV project.

2.2 Feedback: Interaction Based on Natural-Language Generation and Understanding

Natural Language Generation (NLG) and Natural Language Understanding (NLU) are both subfields of Natural Language Processing (NLP), which in turn can be seen as subfields of both computer science and cognitive science [23, 6]. NLG focuses on computer systems which can automatically produce understandable texts in a natural human language, and NLU studies computer systems which understand these languages, see Fig. 6(i). Both NLG and NLU are concerned with computational models of language and its use.

In general terms, the two processes have the same end points but opposite directions, see Fig. 6(ii). Nevertheless, the internal operations of these processes hold several differences in character. NLG has been often considered as a process of choice, whereas, NLU has been best characterized as one of hypothesis management. In NLG, we have several means available, and must choose the most suitable

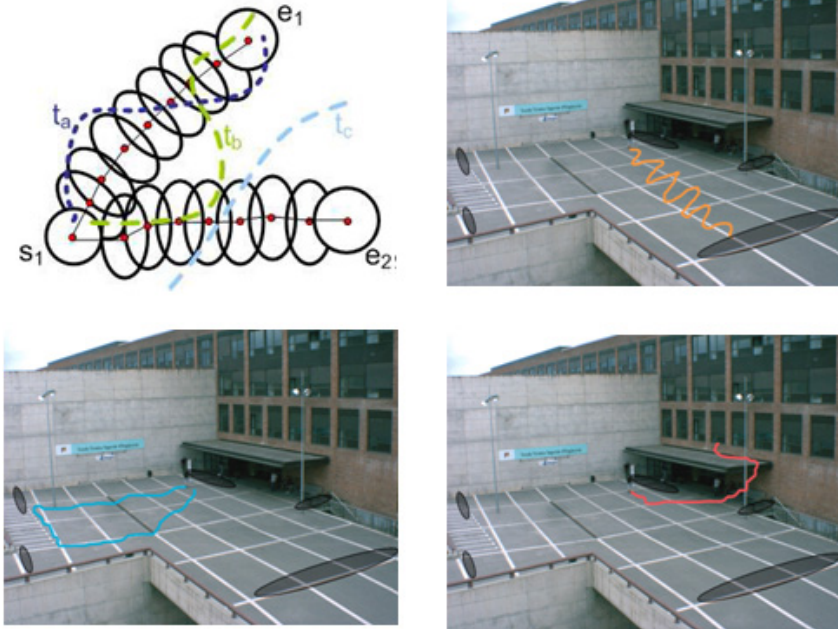


Fig. 4 Examples of the different types of anomalies detected in our prototype

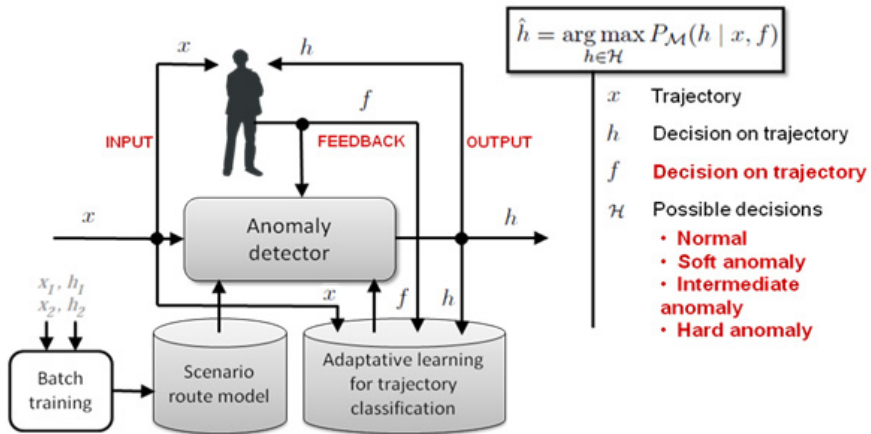


Fig. 5 Adaptive model for anomaly detection

one to achieve some desired end. In NLU, we must select the most appropriate interpretation out of a multiple set of them, given some input.

Therefore, the two strategies which have been considered to design the NL interface are different for the two NLG vs. NLU processes. In NLG we control the set of situations which need to be expressed, and can define one correct form of

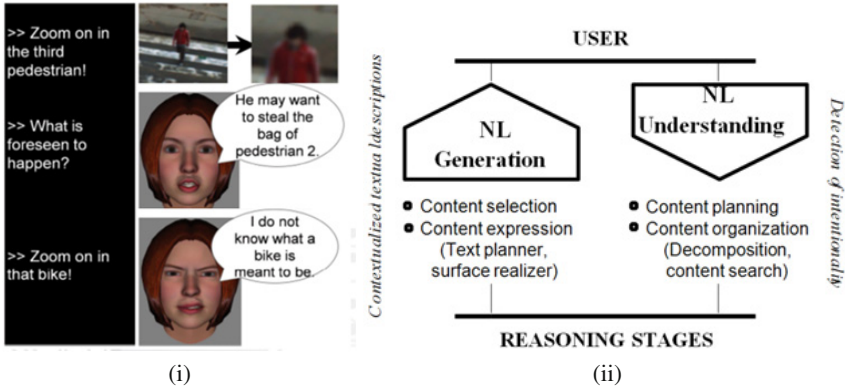


Fig. 6 (i) Examples of interfaces exploiting Natural-Language and (ii) NLG vs. NLU processes

expressing that information in a clear and natural way for each language considered. On the other hand, the NLU process provides us with an open number of possible user queries which need to be interpreted; we need to somehow restrict to a set of intentions which we assume the user can show.

From a general perspective, some characteristics have been considered for NLG:

- We describe situations contained in the implemented behavioral model. In our case, the situations are those ones defined in a domain ontology and used for behavioral analysis in Situation Graph Trees [25].
- Following the cognitive situatedness/embeddedness property, outputs have been restricted to interpretations of the possible situations uniquely for the defined domain [24]. These interpretations will be expressed linguistically by native speakers of each language.
- Such linguistic utterances are built and adapted into the system using rule based parsing techniques and functional grammars [20].
- The linguistic model is based on the Prototype Theory from cognitive linguistics, in which elements are categorized using sets of semantic features [22]. As explained in some of the following chapters, this approach entails a series of advantages, like the lack of rigidity to formalize linguistic properties, or the interoperability of linguistic knowledge at different stages.

There are several operations fulfilled by the NLG module to generate textual descriptions in NL from high-level semantic predicates [19, 21]. Basically, this module uses two kinds of knowledge sources, ontological/situational and linguistic; three kinds of grammar, ranging from semantic to morphological considerations; and an onomasticon to handle the history of instantiations during a discourse.

On the other hand, NLU has usually been regarded as a process of hypothesis management that decides for the most probable interpretation of a linguistic input. Following this idea, the NLU module links plotline sentences to their most accurate interpretations in the domain of interest, in form of high-level predicates referring to

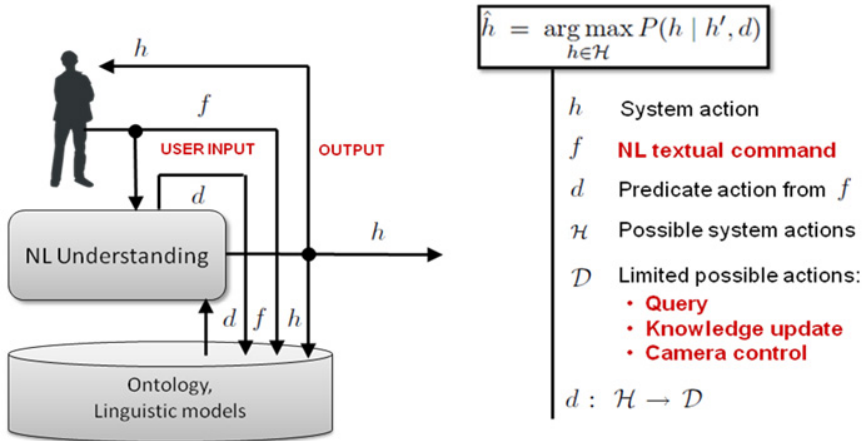


Fig. 7 Feedback model for Natural Language interaction

known concepts or instances within the scene. In essence, we detect all valid queries which apply to predefined goals in the domain of interest, which can be generally classified as questions, commands, or information updates. An ontology is used to restrict the semantic domain of validity of these queries.

Once a proper formatting has been applied, an input sentence is analyzed through a sequence of 3 processes: first, a morphological parser tags each input word with linguistic features depending on the context of apparition. Secondly, a syntactic/semantic parser recursively builds a dependency tree for the tagged sentence. Finally, the resulting dependency tree, already having ontological references, is assigned to the most related high-level predicate found [25]. The morphological and syntactical processes convert the NL sentence into a tree representation, which is finally converted into a goal predicate measuring the distances from the tree to the predicates stored in the ontology. Subsequently, each plotline predicate produced by the NLU module instantiates a high-level event, which must be converted into a list of explicit spatiotemporal actions which are conveyed to the end-user.

Summarizing, Fig. 7 shows the modular schema of our Natural Language interaction within the MI paradigm developed during the MIPRCV project.

2.3 Multimodality: Animation of Virtual Avatars for Communication with End-Users

XfacePlayer is an application designed by the Cognitive and Communication Technologies (TCC) division of ITC-irst and modified for the animation of 3D talking heads through either the MPEG-4 Facial Animation standard or key frame interpolation¹. XfacePlayer relies on Microsoft's Speech API 5.1 Text-To-Speech engine

¹ <http://xface.fbk.eu/index.htm>

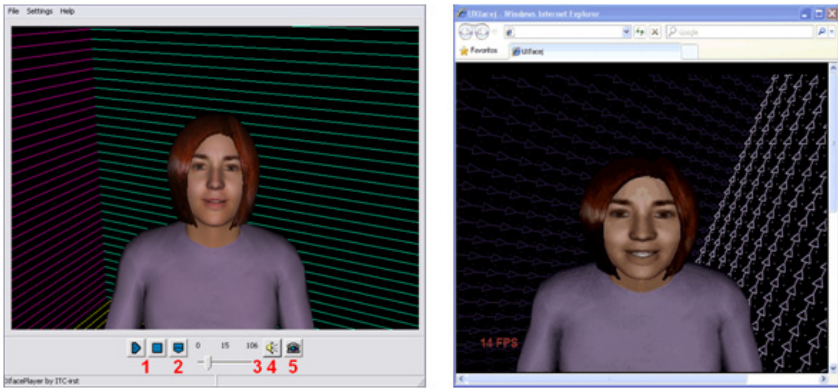


Fig. 8 Multimodal communication using virtual avatars: XfacePlayer interface adapted to the xMontse model(left), and its web version (right)

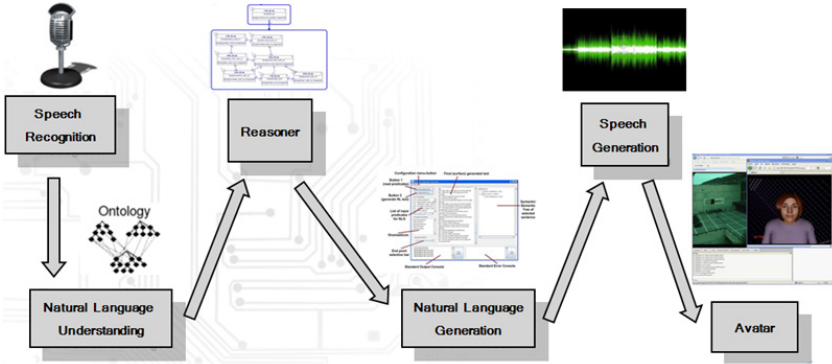


Fig. 9 Multimodal interaction complete process from voice commands or queries to the response using an avatar and a speech generation module

and NeoSpeech’s Kate16 voice for speech synthesis in order to generate the voice for the model. The version used in the VID-Hum prototype is written in java and works in a web browser, see Fig. 8.

From the multimodal point of view, the system depicted in Fig. 9 is a complete set of modules that goes from recognizing the human speech to reply simulating an intelligent virtual human answer and/or doing some action. Voice recognition module implements a signal (voice) to text transformation. Speech synthesis (TTS), along with the avatar animation, implements a text-to-signal transformation. The reasoner is designed to fit in the middle, to provide the text-to-text transformation that appears to produce an intelligent reply to the human input, and could execute some instructions like to zoom in a pedestrian. The Avatar allows any user to see a virtual head pronouncing any given text while displaying facial gestures and emotions.

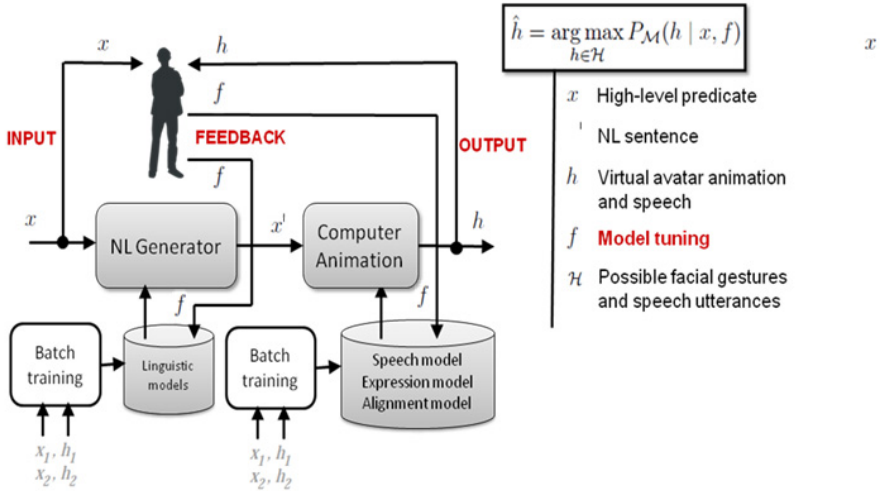


Fig. 10 Multimodal model for human-computer communication

Summarizing, Fig. 10 shows the modular schema of our multimodal model within the MI paradigm developed during the MIPRCV project.

3 The VID-HUM Demonstrator

The prototype platform has been installed on top of the CVC building in Barcelona. The cameras survey a scene in front of the engineering building on the campus of the Universitat Autònoma de Barcelona, see Fig. 11 for an example scene view. The hardware platform for the VID-Hum prototype consists of one fixed camera and a camera mounted in a pan/tilt platform and fitted with a zoom lens.

The hardware platform for the prototype consists of one fixed camera, a camera mounted in a pan/tilt platform and fitted with a zoom lens, three dedicated servers to provide raw computational power and a fast 1 Gb Ethernet switch. The hardware integration architecture is illustrated in Fig. 12. The main components of the hardware infrastructure are Cameras, PTZ platform, Compute Servers and Network Infrastructure, as described next.

Two Pulnix TMC-1405GE cameras are used for the VID-Hum prototype. These cameras are GigE-compatible and deliver very high-resolution (1392x1040) imagery at high framerate (30fps). Each camera is connected by a dedicated, 1000base-TX Ethernet connection to ensure constant, high-framerate streaming from both cameras. The connection between Servolens zoom optics and the active pc is done by an RS-232 serial communication, at 9600 baud of speed. One of the Pulnix cameras is mounted in a Directed Perception PTU-D100 pan/tilt platform that allows 350° pan and 180° tilt surveillance. It also allows a wide range of pan and tilt speeds (0.0075°/sec to over 120°/sec), and is fully sealed for outdoor operations.



Fig. 11 The scene is surveyed from atop the CVC building using one fixed (right one in the image) and one active camera (left)

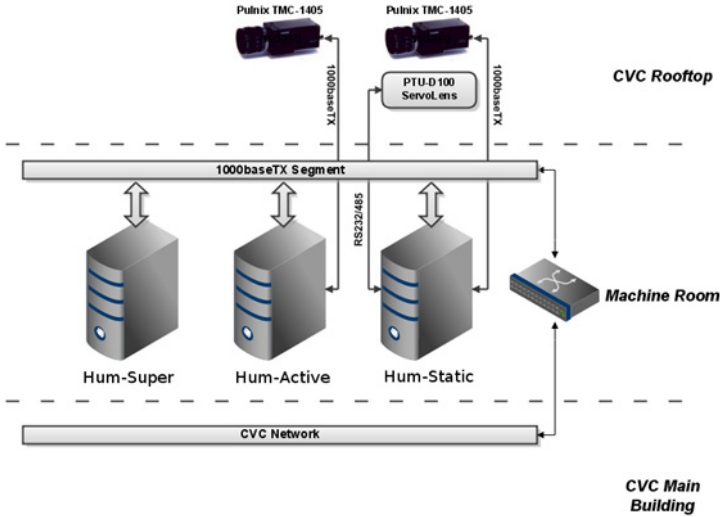


Fig. 12 The VID-Hum prototype hardware infrastructure

Three dedicated Dell Poweredge T100 servers are used. Two of these are directly connected to the Pulnix cameras and are primarily dedicated to video acquisition. The third server is used for components not requiring direct access to the cameras, such as the supervisor tracker and SGT reasoning subsystems described below. These three machines are referred to as Hum-Super, Hum-Static, and Hum-Active to emphasize their roles in the prototype platform. Every computer has 3,2GB of RAM of up to 8GB possible if was necessary, one Intel Dual Core Xeon E3120 at

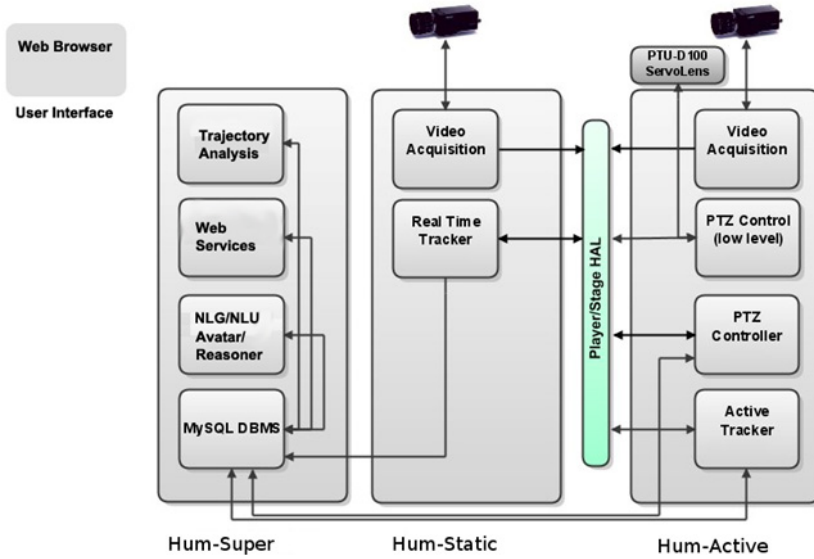


Fig. 13 The VID-Hum prototype software architecture

3.16GHz with two processors, 300GB of Hard Disk and Ubuntu 8,10 (Intrepid) OS with kernel 2,6,27,11-generic.

Fig. 13 shows the modular software architecture designed for the VID-Hum prototype platform. There are two components in the prototype architecture that greatly aided integration: the MySQL database server used to communicate between medium-to-high-level components and the use of the Player/Stage system to provide a layer of hardware abstraction between the VID-Hum prototype and the low-level hardware of the cameras and the PTZ platform.

The major software components in the prototype architecture are: Video Acquisition, PTZ Controller (low level), Real Time Tracker, Avatar, Trajectory Analysis, PTZ Controller, The Reasoner, NLG, NLTU Player/Stage Hardware Abstraction Layer, and MySQL DBMS. A Real Time Tracker (RTT) is one of the fundamental components in the prototype platform. The tracker software communicates through a standardized interface, which allows the parallel development of the prototype as well as the tracking method itself. The RTT tracks multiple moving targets in the fixed camera view and writes its observations into a table on the MySQL server.

The main features and components of the prototype interface are:

Administration and Monitoring of Components: Given the number and diversity of components in the prototype, administration and monitoring of these distributed components is one of the main functions of the user interface. The interface also allows individual components or subsets of components to be tested in isolation from the others.

Visualization of Video Streams: The largest panel in the graphical interface, in the upper left of Figure 14, displays video streaming from one or both

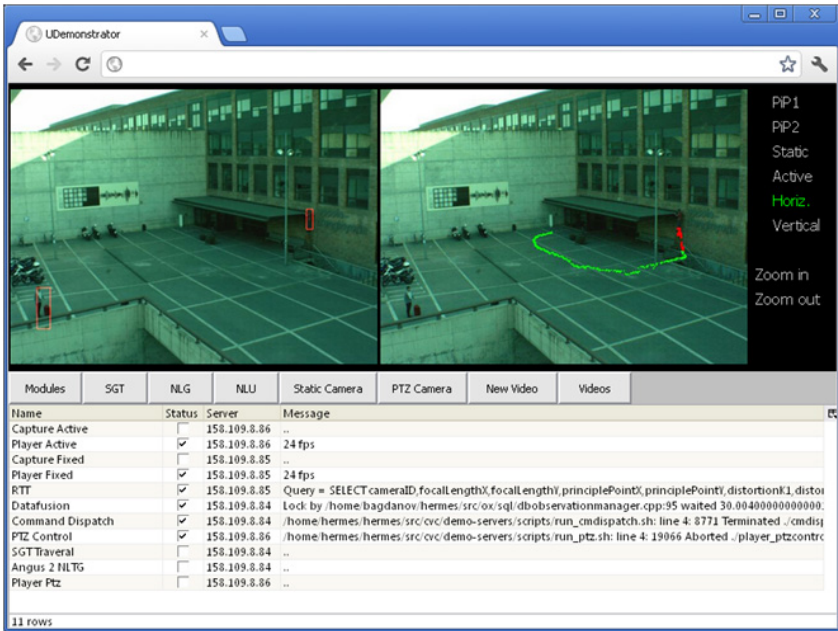


Fig. 14 The GUI for the real-time active surveillance prototype

prototype cameras. Streaming video can be viewed from each camera individually, or simultaneously in split-screen mode.

Visualization of Tracker Telemetry: When the tracker is active, the observations emanating from it are displayed in real-time, overlaid on the streaming video from the fixed camera.

Trajectory Analysis: The module tracks people and then studies if the trajectories are normal or not, without any deterministic information of the scene, the trajectory is classified and its result is showed in real time as an overlay to the camera stream.

Visualization of Reasoner Inferences: When the reasoner component is active, inferences coming from it are displayed in the SGT tab. The time-stamp and predicate from each inference are displayed in a list.

Visualization of NLG: The NLG tab allows to see the texts generated from Natural Language Text Generation.

Command Window for NLU: The NLU tab allows to enter commands or queries in natural language and see its results.

Virtual Avatar: The Avatar allows any user to see a virtual head pronouncing any given text while displaying emotions and other facial gestures in a 3D environment. When the avatar is selected a new window appears and it tells us what the natural language module is generating.

Summarizing, the VID-Hum prototype platform consists of many distributed software components. So a web graphical user interface was designed and implemented to organize, administer and visualize the operation of these diverse components. Fig. 14 illustrates the VID-Hum prototype GUI. The interface allows visualization of both camera streams individually or in split-screen mode which simultaneously streams scaled versions of both camera streams.

4 Discussion

In this chapter we have detailed a proper architecture for a multimodal video surveillance application. Although the results are presented using a single outdoor scenario, the modular design of the architecture makes our prototype suitable to be installed in other scenarios. That means, the low-level tasks should be adapted depending on the characteristics of the new scene, like the motion detection and tracking procedures plus the design of a 3D calibrated, conceptual scene model. For example, in our current implementation, the tracking system is unable to cope well with crowds and multiple occlusions: therefore, the tracking strategy should be substituted if applied to scenarios with more complex pedestrian behaviours. On the other hand, the 3 types of anomalies presented in this chapter are statistically learnt from tracked routes, so the basic procedure for detecting these anomalies would adapt well to other scenes if tracking works well: in our particular implementation, the quantitative results and details of the tracking system using in VID-Hum can be found in [20].

The proof-of-concept in this chapter is that Natural-Language and Virtual Avatars are appropriate when incorporating multimodality and feedback in surveillance prototypes. So, regarding the high-level tasks of the prototype (NLU and NLG, for example), these modules are independent from the scene: in these cases, since a MySQL database server is used to communicate between tracking-to-understanding components, modularity is granted. In addition to a proper quantitative evaluation of the NL modules implemented in this chapter (which can be found in [5] showing how accurate and useful a NL framework can be in surveillance scenarios), we have shown that there are several advantages when converting numerical data obtained from tracking to the conceptual data used by NL: dealing with abstract symbols, generating complex queries, assisting low-level motion tasks in a top-down fashion, describing high-level events, reasoning about the logic terms instantiated from raw pixels, predicting plausible future events, etc. Basically, the use of NL makes possible that the final output of the system is a text which not only describes, in human-readable terms, what is happening but also explains the interactions which take place while inferring why a particular behaviour is being detected.

Lastly, the use of virtual avatars (merging visual and acoustic features) for surveillance purposes has several advantages, as demonstrated in [25]: we consider the use of virtual avatars as an intelligent multimodal affective interface since it combines speaking features like prosody, gaze and spoken words together with visual cues like head nods and lips motion. The aim in the VID-Hum prototype is

not only to perceive the virtual avatar as an intentional agent, but also in the future as a means to evaluate the virtual agents behaviours based on the knowledge (and experience) of the human agent.

5 Conclusions

The ability to communicate is innate in a natural cognitive system. There exist several ways to reach this goal artificially, although Natural Language is usually taken as a primary choice, being a flexible, unconstrained, and economical tool that is also intrinsic to end-users. In the MIPRCV project, we have developed linguistic modules to close the communication loop between the system and external users within the surveillance domain.

The main procedure of the prototype described in this chapter entails: (i) *adaptation*, since the system adapts itself to the most common behaviours (qualitative data) inferred from tracking (quantitative data) thus being able to recognize abnormal behaviors; (ii) *feedback*, since an advanced interface based on Natural Language understanding allows end-users the communication with the prototype by means of conceptual sentences; and (iii) *multimodality*, since a virtual avatar has been designed to describe, using synthetic speech, what is happening in the scene, based on those textual interpretations generated by the prototype. Summarizing, MI provided an adequate framework for all these cooperating processes.

In the implementation of the prototype, instrumental to the success of the integration effort was the adoption of a MySQL interface for communication between nearly all components of the prototype platform. This, along with the adoption of the Player/Stage hardware abstraction layer for routing video and PTZ commands, greatly eased integration efforts. In addition to easing integration, the database of observations preserves a complete picture of everything that happens in the scene of surveillance, including inferences and PTZ camera actions. This, combined with recorded video, allows detailed after-the-fact inspection of the results of cooperative detection and recognition of human actions.

Acknowledgements. The authors acknowledge the support of the Spanish Research Programs Consolider-Ingenio 2010: MIPRCV CSD200700018; Avanza I+D ViCoMo (TSI-020400-2009-133) and DiCoMa (TSI-020400-2011-55); along with the Spanish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02.

References

1. Fusier, F., Valentin, V., Brémond, F., Thonnat, M., Borg, M., Thirde, D., Ferryman, J.: Video understanding for complex activity recognition. *Machine Vision and Applications* 18(3), 167–188 (2007)
2. Arens, M., Gerber, R., Nagel, H.-H.: Conceptual representations between video signals and natural language descriptions. *Image and Vision Computing* 26(1), 53–66 (2008)

3. Dee, H.M., Fraile, R., Hogg, D.C., Cohn, A.G.: Modelling Scenes Using the Activity within Them. In: Freksa, C., Newcombe, N.S., Gärdenfors, P., Wöflf, S. (eds.) *Spatial Cognition VI. LNCS (LNAI)*, vol. 5248, pp. 394–408. Springer, Heidelberg (2008)
4. Makris, D., Ellis, T., Black, J.: Intelligent Visual Surveillance: Towards Cognitive Vision Systems. *The Open Cybernetics and Systemics Journal* 2, 219–229 (2008)
5. Fernández, C., Baiget, P., Roca, F.X., González, J.: Determining the Best Suited Semantic Events for Cognitive Surveillance. *Expert Systems with Applications* 38(4), 4068–4079 (2011)
6. González, J., Rowe, D., Varona, J., Roca, F.X.: Understanding Dynamic Scenes based on Human Sequence Evaluation. *Image and Vision Computing* 27(10), 1433–1444 (2009)
7. Bellotto, N., Sommerlade, E., Benfold, B., Bibby, C., Reid, I., Roth, D., Van Gool, L., Fernández, C., González, J.: A Distributed Camera System for Multi-Resolution Surveillance. In: 3rd ACM/IEEE International Conference on Distributed Smart Cameras (2009)
8. Makris, D., Ellis, T.: Learning semantic scene models from observing activity in visual surveillance. *IEEE Trans. on Systems Man and Cybernetics-Part B* 35(3), 397–408 (2005)
9. Piciarelli, C., Foresti, G.L.: Online trajectory clustering for anomalous events detection. *Pattern Recognition Letters* 27(15), 1835–1842 (2006)
10. Johnson, N., Hogg, D.C.: Learning the distribution of object trajectories for event recognition. In: *British Machine Vision Conference*, pp. 583–592 (1995)
11. Hu, W., Xiao, X., Fu, Z., Xie, D.: A system for learning statistical motion patterns. *IEEE Trans. on PAMI* 28(9), 1450–1464 (2006)
12. Basharat, A., Gritai, A., Shah, M.: Learning object motion patterns for anomaly detection and improved object detection. In: *IEEE Conference on CVPR* (2008)
13. Hu, W., Xie, D., Tan, T.: A Hierarchical Self-Organizing Approach for Learning the Patterns of Motion Trajectories. *IEEE Trans. on Neural Networks* 15(1), 135–144 (2004)
14. McKenna, S., Nait-Charif, H.: Summarizing Contextual Activity and Detecting Unusual Inactivity in Supportive Home Environment. *Pattern Analysis and Applications Journal* 7(4), 386–401 (2004)
15. Zhang, Z., Huang, K., Tan, T., Wang, L.: Trajectory Series Analysis based Event Rule Induction for Visual Surveillance. In: *IEEE Conference on CVPR* (2007)
16. Yao, B., Wang, L., Zhu, S.: Learning a Scene Contextual Model for Tracking and Abnormality Detection. In: *IEEE Conference on CVPR Workshops* (2008)
17. Morris, B., Trivedi, M.: Learning trajectory patterns by clustering: Experimental studies and comparative evaluation. In: *IEEE Conference on CVPR* (2009)
18. Bremond, F., Thonnat, M., Zuniga, M.: Video understanding framework for automatic behavior recognition. *Behavior Research Methods* 38(3), 416–426 (2006)
19. Arens, M., Nagel, H.-H.: Behavioral Knowledge Representation for the Understanding and Creation of Video Sequences. In: Günter, A., Kruse, R., Neumann, B. (eds.) *KI 2003. LNCS (LNAI)*, vol. 2821, pp. 149–163. Springer, Heidelberg (2003)
20. Fernández, C., Baiget, P., Roca, F.X., González, J.: Interpretation of Complex Situations in a Semantic-Based Surveillance Framework. *Signal Processing: Image Communication* 23(7), 554–569 (2008)
21. Gerber, R., Nagel, H.-H.: (Mis-?)-Using DRT for Generation of Natural Language Text from Image Sequences. In: Burkhardt, H., Neumann, B. (eds.) *ECCV 1998. LNCS*, vol. 1407, pp. 255–270. Springer, Heidelberg (1998)
22. Lakoff, G.: *Women, fire, and dangerous things*. University of Chicago Press (1987)

23. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press (2000)
24. Wilson, R.A., Keil, F.C. (eds.): The MIT Encyclopedia of the Cognitive Sciences. Bradford Books (2001)
25. Fernández, C., Baiget, P., Roca, F.X., González, J.: Augmenting video surveillance footage with virtual agents for incremental event evaluation. *Pattern Recognition Letters* 32(6), 878–889 (2011)

Interactive Video Surveillance for Perimeter Control

Javier Ortells, Henry Anaya-Sánchez, Raúl Martín-Félez,
and Ramón A. Mollineda

Abstract. This chapter presents an interactive video-surveillance solution for assisting human operators in the control of movements across a multi-region scenario (perimeter control). It has been conceived as a multi-camera system to detect anomalous trajectory events, such as entering or leaving a region or changing the walking speed, by means of a dynamic collection of decision rules. They relate spatio-temporal patterns and event categories (anomalous, unknown, normal), and are used to assess and classify trajectory events. The interactive paradigm has been adopted as a natural framework to progressively learn and update rules, particularly at early stages of the system operation. The approach of continuously improving system knowledge from user feedback conducts to adaptive, reliable and increasingly automatic systems in a relatively short period of time.

1 Introduction

Video surveillance has become part of our everyday lives. The interest on remote visual systems for security purposes has grown significantly in the last years, particularly for those public spaces where a great number of people pass. Some examples are supermarkets, airports, underground stations, stadiums, shopping centers, etc. Traditionally, these systems have required a huge amount of human supervision both in real-time and in a posterior video analysis to state the truth. However, with the current proliferation of cameras, this exhaustive search has turned into unfeasible.

A major breakthrough of video-surveillance systems is coming about with the development of video analytic techniques. They can be roughly defined as autonomous understanding of events occurring in a scene monitored by multiple video cameras [16]. However, most present-day video-surveillance systems are far from being

Javier Ortells · Henry Anaya-Sánchez · Raúl Martín-Félez · Ramón A. Mollineda
Institute of New Imaging Technologies, Universitat Jaume I of Castelló
Av. Sos Baynat s/n, 12071, Castelló de la Plana, Spain
e-mail: {jortells, henry.anaya, martinr, mollineda}@uji.es

fully autonomous, particularly when video analytics should deal with complex and crowded scenes. Furthermore, autonomy is not always an end purpose for some task solutions, which have an intrinsic interactive nature. While some of them can be substantially benefited from human feedback, such as speech transcription and machine translation [18], others require the human involvement in a critical decision making process, for example, in computer-assisted medical image diagnosis.

The interactive paradigm in computer vision applications allows users to progressively enhance priors and constraints by adding their knowledge within an operational loop of the system. This approach has several benefits on learning. Firstly, it exploits human know-how in a natural way (through feedback) for model refinement. Secondly, it leads to solutions that better fit the user demands. Finally, it favors a dynamic adjustment of models when the user needs or the context changes.

A typical video-surveillance application is an intelligent system designed for monitoring one or more regions of a scene, in order to detect and track particular objects or situations according to predefined safety rules. When this kind of system is deployed in simple scenarios, for example, the garden of a house, it is easy to customize the behavior of the system through a few basic and steady rules [8]. However, in more complex and changing environments such as large commercial areas, big factories or public squares, it could be unrealistic to try to foretell every particular situation that requires a specific management. Thus, a prior and hard-to-change collection of safety rules does not appear to be the most suitable strategy. Under such shifting conditions, an interactive framework for supporting the on-line management of unseen events can be a natural alternative to keep a set of meaningful rules updated.

This manuscript proposes an interactive solution to the general problem previously described. A video-surveillance application for interactively controlling movements across a multi-region scenario (perimeter control) is here presented. It is a two-camera prototype to detect anomalous actions on a region model defined in a real-world scene. The analysis of trajectory data is performed by using a collection of dynamic spatio-temporal rules. A suitable graphical user interface allows on-the-fly rule management via emerging windows when a detected event is unknown (rule definition) or when it is classified as anomalous (alarm and event refinement). This feedback mechanism has been adopted for progressively building and keeping updated an effective series of rules.

This chapter is organized as follows. Section 2 summarizes the related state-of-the-art works. The interactive learning approach is explained in Sect. 3. A description of the implemented system is given in Sect. 4, including scope, system architecture, main software modules and resources needed. Finally, Sect. 5 gathers the conclusions of the chapter.

2 Related Work

One of the main applied research areas of video analytic methods focuses on visual event detection, tracking and classification, looking for anomalies in order to alert human operators. A broad range of video-surveillance applications can be

envisioned from this general operational context, such as access and perimeter control, human identification at a distance, detection of anomalous behaviors of people or vehicles, etc. An inspiring survey about methods and applications for video surveillance of people and vehicle is given in [10]. It presents a general framework of visual surveillance in dynamic scenes from multiple cameras, and discusses the principles and methods involved in the main parts of the framework. That paper also sheds light upon a number of key research problems and directions, providing a consistent basis for the forthcoming developments. However, although interactive functions are suggested as a key direction for surveillance applications, no discussion is provided about the importance or convenience of using human feedback for managing critical decisions with higher confidence in some tasks.

A complementary study of the state-of-the-art of visual surveillance systems was provided afterwards in [17]. This work focuses in distributed automated surveillance systems, and it reviews a number of real commercial applications. One of them, DETER [15], consists of two main modules, one for detection, recognition and tracking of objects, and a second one for threat assessment and alarm management. The lack of a feedback loop in the alarm module to improve its performance is highlighted in [17]. In spite of this observation, no other claim toward human interaction in surveillance systems appears in this survey.

Focusing on specific surveillance solutions, it is compulsory to mention two early prominent proposals [5, 9]. As a part of the Video Surveillance and Monitoring (VSAM) project (1997-1999), an ambitious initiative of the DARPA, the Robotics Institute at Carnegie Mellon University (CMU) and the Sarnoff Corporation developed a system for autonomous video surveillance and monitoring [5]. This system included robust methods for detecting and tracking moving objects by multiple cooperative video sensors, and for classifying these objects into semantic categories such as human, human group, car and truck. Besides, people activity was also classified as walking or running. The system GUI had a module for managing Regions Of Interest (ROIs), that allowed the interactive creation of a polygonal ROI from a collection of boundary points. The user could also link object types (e.g. human, vehicles) to ROIs, providing the system with a mechanism for triggering events like “enter”, “pass through”, “stop in”, etc. The detected activities from all sensors were transmitted to a central unit, where they were displayed and analyzed.

In [9], a real-time visual surveillance system (W^4) for detecting and tracking people in an outdoor environment is described. From gray-scale and infrared video imagery, W^4 is able to locate and track people and their parts (head, hands, feet, torso) using models of people’s appearance. W^4 can also detect and segment objects that are being carried by people, so interactions among people and objects can be described, such as depositing, removing and exchanging objects. However, despite of the relevance of the above two cited papers, none of them documents any human feedback function for model improvement.

A methodological work is presented in [7], where a number of critical issues related to video-surveillance requirements are addressed. For instance, it describes several low-level image and video processing techniques such as change detection for fixed and mobile cameras, background updating, multi-camera view registration, etc.

This paper provides valuable guidelines and resources to new researchers in video surveillance, to design low-level vision modules. High-level video analysis functions for scene understanding are beyond the scope of this paper.

Recent progress in embedded sensor technology is encouraging the development of distributed sensor networks. In particular, smart cameras comprise sensing, processing and communication functions, all built in within the same device. Two papers about distributed networks of embedded smart cameras are [3, 4]. In [3], a distributed network of smart cameras for surveillance purposes is introduced, along with the low-level video processing algorithms developed for particular nodes. Some of the low-level routines included are segmentation, labeling, tracking and classification of detected objects. On the other hand, in [4], a traffic surveillance application is proposed. A scalable smart camera hardware architecture is designed as an open technology to develop distributed intelligent video-surveillance systems. Each smart camera is assigned to a particular region, and some motion vectors are defined to represent the spatial relationship among the cameras. These vectors allow to check whether an object is moving in the correct direction. A multi-camera object-tracking application is implemented, where a tracking agent migrates from one camera to another in no more than one second. This parameter imposes conditions on both maximum vehicle speed and minimum camera distance.

In the literature reviewed, taking advantage of the user's feedback for a continuous improvement of the models is not a target. We believe that this practice can help to build adaptive, reliable and efficient systems in a short period of time, which is especially useful when critical decisions must be made. For example, in the case of distributed smart camera networks, the inter-node spatial relationships could be modeled by an interactive approach, in which the network would propose spatial hypotheses to a human operator for their refinement and validation. These validated hypotheses (knowledge) could help to infer new and more accurate hypotheses, and so on. This is the cornerstone of interactive learning: each piece of human feedback information must contribute to increase the system robustness and, usually, to decrease the human-machine interaction effort.

3 Interactive Learning Strategy

As discussed previously, the singularity of the proposed visual surveillance solution is the interactive strategy to learn and update safety rules as context conditions change. Object trajectories are monitored, and those events relevant to the system end-purpose are detected and classified by a collection of rules, which can be interactively defined or modified either on-line or off-line.

From a conceptual perspective, the implemented prototype embodies the following two main interactive learning requirements:

- System knowledge increase, by allowing on-the-fly creation of new rules to model unseen trajectory events.
- System knowledge adapting, by allowing on-the-fly updating of existing rules, when they are triggered, to model new user needs or a changing environment.

A simple rule-based approach has been implemented to manage these requirements. It is able to encapsulate human knowledge through a collection of “if then” declarative statements, which is an easy way for human beings to understand the know-how description. Considering our goals, this approach has two main strengths: 1) it makes quick decisions in a repeatable form (deterministic procedure), and 2) it admits an agile adaptation when a new piece of knowledge is added or an existing one is updated. These useful qualities are due to the absence of hard constraints on the order of evaluating the rules. Generally, a conflict-resolution policy is needed to decide which rule to fire if some of them match with a given problem state. The simplest approach is possibly to define a rule order for choosing the first triggered rule. Thus, when a new rule is added, an ordered insertion is enough to adapt the collection of rules for dealing with an extended problem scope.

An alternative would have been a rule-based decision tree, a structure that can be induced from a set of rules [2]. Its main merit is that it can potentially organize rules in a concise and efficient way to take the best decision readily. However, when the collection of rules changes, the decision tree must be completely re-built in order to keep it consistent. The complexity of this process depends on the number of rules, the attributes involved in rule premises, the intersection level between rules, and the heuristic criteria used to create the tree. In the case of an interactive solution that starts with no knowledge of the task to be solved, which is also expected to change frequently, a hard-to-adapt model is not certainly the most suitable choice.

The video-surveillance system here introduced should monitor, describe and classify (as anomalous, unseen or normal) the trajectory points of moving object across a scene view. Given a trajectory point description, a rule-based reasoning process tests the condition of each decision rule against the trajectory point data (known fact), regarding a predetermined evaluation order. This process searches for a first matching rule to produce an output. On success, the trajectory event is classified as anomalous or normal, according to the action associated to the triggered rule. Otherwise (when no rule is applicable), it is considered as unknown.

The interactive strategy adopted allows a progressive learning and updating of decision rules, particularly at early stages of the system operation. This approach leads to the following knowledge life cycle:

Starting Point: The system starts with no knowledge about the surveillance task to be solved. That is, there are no rules at the beginning, so every trajectory event is initially classified as unseen.

Learning a New Rule: When an unseen trajectory event is detected, an emerging GUI asks the user to infer and define a decision rule from this particular event¹.

The new rule is then inserted in an orderly manner into the rule set, in such a way that rules which involve more specific conditions are prioritized. To this end, an algorithm computes the generality/specificity relation between the new rule and each existing rule, using two criteria (in the order they appear): 1) the higher the number of attributes in a rule condition is, the more specific the rule is, and 2)

¹ It is also possible to create rules using the standard GUI, out of the main system operation loop.

when the conditions of two rules are expressed in terms of the same attributes, an analysis about their intersection is performed. Three possible results are considered from this analysis: a) if there is no intersection between their attribute values, both rules are considered as not related; b) if the attribute values that satisfy one rule are subsets of the corresponding attribute values of the other rule, the former is considered more specific than the latter (the first one can be deemed as an exception of the second one); and c) when other types of intersections are found out, the system warns the user about a possible contradiction between both rules, and no insertion is carried out. Note that the binary relation defined, “more specific than”, is antisymmetric and transitive, what avoids critical ambiguities in deciding where to insert. When no generality/specificity relation exists between two rules, their order of appearance determines the rule order.

Updating a Rule: When an alarm rule is triggered, a related emerging GUI allows the user to update the rule to fit a new requirement of the task. For example, the user might modify its parameters or deactivate the rule, *i.e.* to transform the alarm into a normality rule.

This procedure of continuously improving system knowledge from user feedback should conduct to an adaptive, reliable and increasingly automatic solution.

4 Prototype Description

In this section, we provide details about the functional scope, system architecture and the main application modules of the proposed video-surveillance solution.

4.1 Functional Scope

The implemented prototype has been devised as a multi-camera system for perimeter control over a region map defined for a real-world scene. A multi-camera object tracking process allows trajectory monitoring and assessment through a collection of rules, which can be interactively defined either on-line or off-line.

Apart from the interactive learning capacities (see Sect. 3), the prototype provides a number of user functions for camera configuration, scene region definition, region projection over distinct scene views, rule definition, alarm management, real-time display of the monitored scene views, video recording for off-line analysis, among others. These end-user functions are supported by some other operations such as multi-camera acquisition, view-dependent object detection, view-dependent object tracking, object matching between different camera views, trajectory-based event detection, rule-based event classification and real-time information display.

Next sections provide details of this operational context.

4.2 System Architecture

Figure 1 illustrates the high-level system architecture. Five main modules can be identified in the diagram:

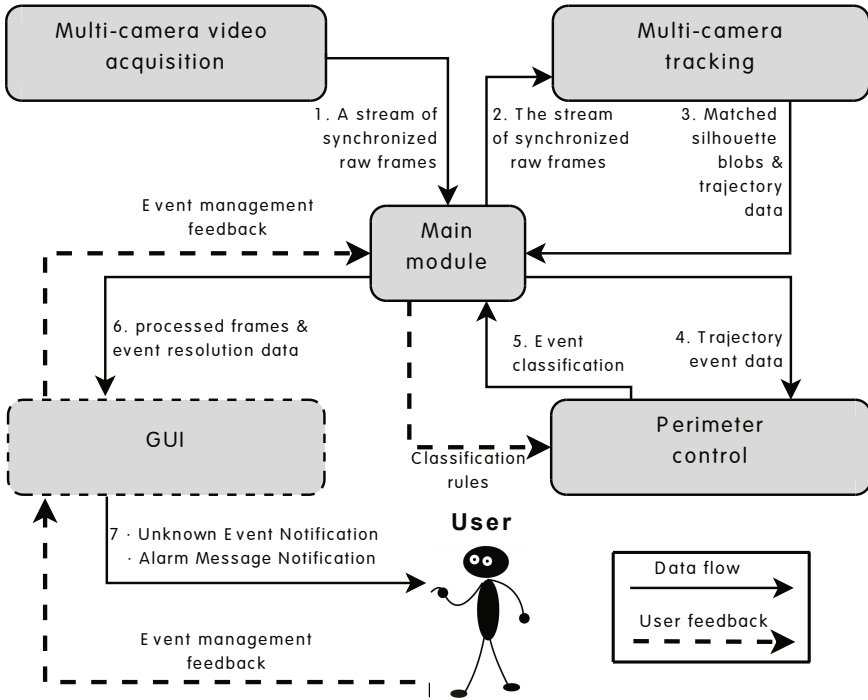


Fig. 1 High-level architecture of the video-surveillance prototype for perimeter control

Multi-camera Video Acquisition: It consists of multi-port firewire cards where one or more video cameras are plugged into, low-level software drivers to gain access to the cards, and high-level software components to make easier the interaction between the system and the cameras. The output of this module is a stream of synchronized raw frames for each camera, which feeds the tracking module.

Multi-camera Tracking: The inputs to this module are the streams of synchronized raw frames coming from different cameras and provided by the Video acquisition module. Object tracking is separately performed on each stream by a particular tracking thread. Then, a collection of synchronized tracking results (blobs) from all the tracking threads is given to the matching algorithm. It assesses all possible correspondences between blobs tracked on different cameras, looking for multiple views of the same object/subject. When two blobs are matched, they are tagged with a same code, thus uniquely identifying the corresponding real-world object/subject in the system. Figure 2 illustrates this process. The two blobs of a same subject are linked by a common colour of their bounding box edges, and also by an identical numeric label located on the upper edge of the bounding box. The output of this module is a collection of matched blobs, which model the 1 : n relationships between scene real objects and their camera projections.



Fig. 2 An example of two-camera object tracking. The two silhouettes of a same subject share a common colour in their bounding box edges and an identical numeric label.

Perimeter Control: The input to this module is trajectory data of tracked objects. Given a multi-region model defined on a real-world scene and projected onto the different camera views (view region maps), a collection of rules classify trajectory points obtained from tracking objects across the regions. The use of multiple cameras covering different viewpoints of a scene should lead to a more accurate detection of anomalous movements.

GUI: In human-machine interaction, the graphical user interface (GUI) is a crucial part of the system. On the one hand, the GUI shows a real-time view of each camera with the bounding boxes of the detected objects overprinted on them. On the other hand, it allows the user to provide feedback to the system for performance improvement, what mostly occurs as a part of a work session. An intuitive, simple and easy-to-use GUI has been designed and implemented. In Sect. 4.2.3, the most important functions provided by the GUI are described.

Main Module: This module coordinates all data flows across the system. As more than one video stream must be concurrently managed, a multi-threaded design is set to allow a synchronized video acquisition, tracking and analysis. It also keeps the main data structures updated, including the collection of decision rules (see Sect. 3) and the estimated trajectory data obtained from blob positions such as the movement speed, crossed regions, and so on.

In the following subsections, the three most relevant architecture modules to the end-purpose of the system are described in detail.

4.2.1 Multi-camera Tracking

As stated above, the term *multi-camera tracking* is used to refer the process of assigning a unique label to multiple camera projections of a same real object/subject. In this paper, this function consists of two main steps: view-dependent foreground segmentation, and a matching algorithm to find the proper correspondences between blobs detected on different cameras. These two steps are explained below.

Foreground Segmentation – In the proposed system, background modeling is combined with post-processing techniques to retrieve the foreground from each video frame. The background modeling relies on the adaptive approach introduced in [11]. The aim is to make the system robust enough to deal with dynamic changes that can appear in the scene views (e.g. global illumination changes produced from day-night transitions or long-term background updates corresponding to events such as parking a car in front of a building).

In the background model of a scene view, each color pixel i is represented by three non-stationary Gaussian distributions $\left\{N_c^{(i)}\left(\mu_c^{(i)}(t), \sigma_c^{(i)}(t)\right)\right\}$, $c \in \{R, G, B\}$, where each distribution models a RGB-component of the pixel at time t . The methodology for segmenting the foreground proceeds as follows.

Initially, the learning of the background model of the scene view comprises a sequence of video frames given as training. Then, to retrieve the foreground of an incoming frame, its pixels are classified into either background or foreground. A pixel i is classified into the foreground class if at least one of its three measured RGB values are out of a confidence region of their corresponding Gaussian distributions. Otherwise, the pixel is considered as background, and its Gaussian representation is updated using the equations (1) and (2):

$$\mu_c^{(i)}(t+1) = \alpha \mu_c^{(i)}(t) + (1 - \alpha) z_c^{(i)}(t) \quad (1)$$

$$\begin{aligned} \left(\sigma_c^{(i)}(t+1)\right)^2 &= \alpha \left(\left(\sigma_c^{(i)}(t)\right)^2 + \left(\mu_c^{(i)}(t+1) - \mu_c^{(i)}(t)\right)^2 \right) + \\ &+ (1 - \alpha) \left(z_c^{(i)}(t) - \mu_c^{(i)}(t+1) \right)^2 \end{aligned} \quad (2)$$

where $z_c^{(i)}(t)$ is the observed value for the RGB-component c of pixel i in the frame, and the parameter α is the adaptation rate ($0 < \alpha < 1$). Finally, the foreground is represented by a binary pixel map, according to the classification of the pixels.

Additionally, post-processing techniques are applied on the binary pixel map since it frequently contains noise due to motion of small background objects or shadows. This post-process includes morphological operations and shadow filtering.

Matching Algorithm – In this second step, the multi-camera tracking process is completed by combining the single-camera tracking corresponding to each scene view with a matching procedure operating on 2D-scene regions.

Broadly, the single-camera tracking is based on both 1) an appearance model for each individual in the scene view and 2) a blob overlapping criterion to estimate the regions occupied by a tracked individual through the video stream frames.

In the matching procedure 2D-scene regions are modeled as Gaussian distributions, each one being parameterized by both a mean vector and a covariance matrix. That is, a given region is represented by the maximum likelihood estimated Gaussian distribution that describes the locations of its pixels in the view, and the distance between regions is calculated in terms of the Bhattacharyya distance.

From these settings, the matching procedure is build upon a stochastic model of pairs of regions that represents the correspondence between 2D-regions in the different views according to the real-world 3D-region they represent. This correspondence model between regions in the different scene views is represented by a finite set of categories $\mathcal{C} = \{C_1, \dots, C_k\}$; where each category C_i ($1 \leq i \leq k$) is centered at a pair of 2D-regions $\langle u_i, v_i \rangle$ (in different scene views) with a high likelihood of representing the same real-world region. Each category C_i is also provided with deviation parameters σ_{i_1} and σ_{i_2} .

In this way, given a pair of 2D-regions $\langle u, v \rangle$ in different views, the matching procedure estimates a measure of the confidence this pair of regions is generated from the model \mathcal{C} as showed in Eq. 3:

$$p(\langle u, v \rangle | \mathcal{C}) = \frac{1}{k} \sum_{i=1}^k \frac{1}{\sigma_{i_1} \sigma_{i_2}} \mathcal{K} \left(\frac{d(u, u_i)}{\sigma_{i_1}}, \frac{d(v, v_i)}{\sigma_{i_2}} \right) \quad (3)$$

where d represents the distance function between regions and \mathcal{K} is the Gaussian of the Eq. 4:

$$\mathcal{K}(x, y) = e^{-\frac{1}{2}(x^2 + y^2)} \quad (4)$$

From the estimated value $p(\langle u, v \rangle | \mathcal{C})$, the matching procedure makes a decision on whether the regions $\langle u, v \rangle$ represents the same real-world scene or not by relying on an empirical threshold. In the system, the set of categories \mathcal{C} is learned from a synchronized training sequence of video frames from the scene views.

4.2.2 Perimeter Control

The aim of the Perimeter control module is to evaluate and classify trajectory events using a collection of rules. This section describes how this module works based on two important data structures: the *view region map* and the *rule*.

The term *view region map* is used to refer the 2D projection of a 3D region model onto a particular camera view. It provides spatial information to the trajectory evaluation process. Each view region map is coded by a logic matrix whose size matches the view resolution. Thereby each camera pixel corresponds to a matrix cell where the identifier of the pixel region is stored. This data structure makes easy to check which region a pixel belongs to.

On the other hand, a *rule* can be formally defined as a condition that can be formulated in terms of one or more of the following context parameters:

Time Range: This parameter can adopt three different formats, depending of the nature of the time period used:

- i If the rule is intended to control events within the same time period every day, the starting and finishing times are required.

- ii If the rule is intended to control weekly events within a time period between two week days (e.g. from every Friday at 22:00 h. to next Monday at 6:00 h.), these two week days and the starting and finishing times must be provided.
- iii If the rule is intended to control events within a time period between two calendar days (e.g. from July 31st at 22:00 h. to September 1st at 6:00 h.), these two calendar days and the starting and finishing times must be provided.

Current Region: It identifies the region where the object is physically located.

Previous Region: This parameter identifies the region, if any, on which the object was before entering into the current region.

Movement Speed: This parameter sets a maximum speed limit, in such a way that higher speeds are considered anomalous.

In addition, the rule must be defined/labeled as either normality or alarm, which will determine the *action* taken by the system. The fulfillment of a normality rule does not cause any visible result. On the contrary, when an alarm rule is triggered, a warning message alerts the operator, who could also interact with the system to modify the involved rule.

4.2.3 GUI Capabilities

This section provides an overview of the main functional capabilities of the GUI of the implemented video-surveillance prototype.

Camera Configuration: At the beginning of a work session, the user can choose which cameras will be used among those plugged into the multi-port firewire cards. Besides, the user can configure the cameras by choosing values for some operating parameters such as the resolution and the sampling frequency.

Region Definition and Region Map Outlining: Regions are defined in a real-world scene, but they are manually drawn (projected) over each camera view. This user function supports region definition by a meaningful name and a specific color, and the creation of view region maps. A drawing tool to outline piecewise linear closed contours on fixed images was implemented. It is used to draw the view region map on some key video frame with suitable background information of each camera view. The tool also supports the visual superposition of a region map on the video frames as a semi-transparent layer when the system is running.

Interactive Event Management:

- *Alarm notification.* When an alarm rule is fulfilled, a pop-up window is triggered (see Fig. 3). It allows the user to deactivate the rule, *i.e.* to transform the alarm into a normality rule, or to modify its parameters.
- *On-line rule definition.* When a trajectory event is classified as unknown, a new window emerges (see Fig. 4) describing the current state and allowing the user to transform it into an alarm or a normality rule by the GUI shown in Fig. 5. Some rule parameters are filled in with the collected information. It is also possible to ignore an unknown event when normality and alarm criteria are not clear (see the Ignore button in Fig. 4). When this decision is made, a short-term normality rule for a parameterized period of time is automatically



Fig. 3 An alarm rule is satisfied when a subject enters to the grass area



Fig. 4 An example of an unknown state detection that requires user interaction

created. That means, for a limited amount of time, those situations similar to the one ignored will be considered as normal. When this time ends, the provisional rule is removed and similar events will be unknown again.

- *Off-line rule definition.* A surveillance rule can also be defined at any time from the application menu. Figure 5 illustrates an example of the definition of a new alarm rule using time, region and speed parameters.

4.3 Hardware and Software Resources

The implemented prototype has been the result of using a number of specific hardware and software resources. This section summarizes the most relevant ones.

Hardware Resources:

- Two firewire 'B' AVT Stingray F-080TM cameras. The prototype was also tested using other AVT cameras such as Guppy F-036TM (Firewire 'A').
- A firewire 'B' card with two independent buses to connect Stingray cameras. The firewire 'A' cameras were plugged into two other suitable cards.

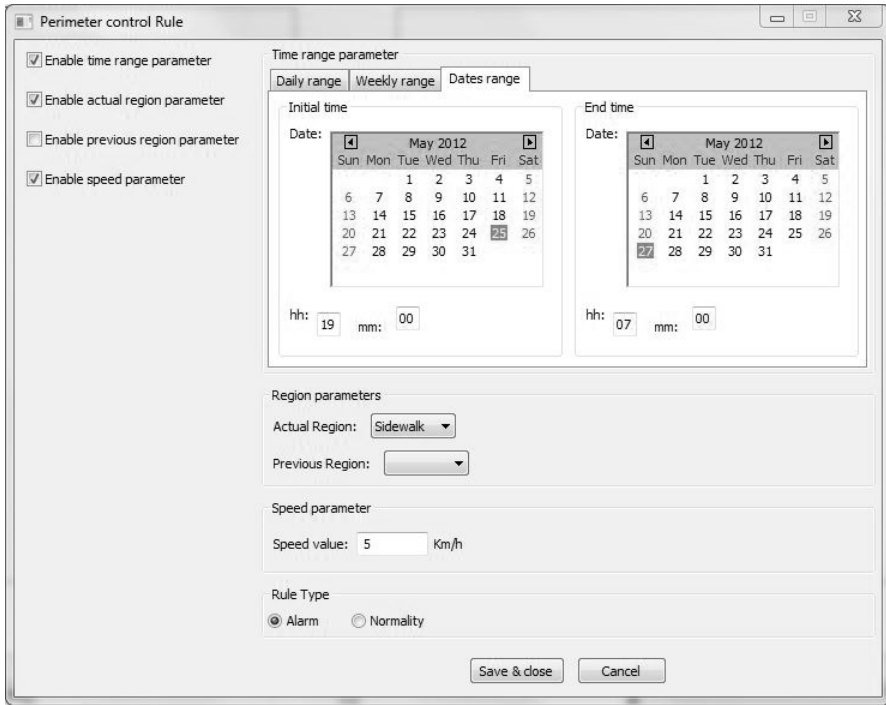


Fig. 5 Example of defining a new alarm rule using time, region and speed parameters

- A Dell XPS 730XTM desktop computer to develop and test the prototype. Its main features are: Intel CoreTM i7 920 (2.67GHz) CPU, nVidia GeForce GTX 285TM GPU, 6GB RAM, Windows 7 ProfessionalTM 64 bits.

Software Resources:

- Microsoft Visual Studio 9TM [12] was used as development environment.
- AVT Universal Package, which is a programming API to easily operate with AVT (Allied Vision TechnologiesTM [11]) cameras.
- wxWidgets cross-platform GUI library [19], which is a C++ library that supports the development of graphic interfaces. It also offers useful methods and data structures for general purpose programming.
- OpenCV library [13] provides some image processing functions that were used with the video frames.
- CvBloblib library [6] was used to perform some tracking and perimeter control tasks. Some pieces of its source code were modified to fit our purposes.

5 Conclusions and Future Work

This work proposes an interactive video-surveillance solution for assisting human operators in the control of movements on a multi-region scenario (perimeter

control). Trajectory monitoring is performed by a multi-camera tracking algorithm, while trajectory event assessment is supported by a dynamic collection of spatio-temporal rules. An interactive approach has been implemented for allowing human operators to progressively extend, adapt and refine the system knowledge (rules) by providing feedback within an operational loop of the system. When an anomalous or an unknown event is detected, suitable emerging user interfaces allow on-the-fly rule management for defining or updating rules. This kind of interaction leads to increasingly automatic operating modes, because the more feedback the user provides, the less future interactions are expected to be demanded by the system and more reliable decisions can be autonomously made.

Future developments could involve the integration of non-intrusive biometric functions such as those based on face, gait or even on a combination of them under a multi-modal biometric system. Other related applications might be re-identification of people over different cameras do not sharing the same scene, detection of abandoned luggage, and so on. Furthermore, due to the system capability of storing trajectory data, other parameters such as the covered path distance, the scene entry and exit points, and the exposure time can be straightforwardly estimated, which might be useful for instance in a control access application.

Acknowledgements. The work documented in this paper has been partially funded by the five-year Multimodal Interaction in Pattern Recognition and Computer Vision (MIPRCV) research project (2007-2012) with code CSD2007-00018. Other projects that have also contributed to support this work are TIN2009-14205-C04-04 from the Spanish Ministry of Innovation and Science, P1-1B2009-04 from Fundació Bancaixa, and PREDOC/2008/04 grant from Universitat Jaume I.

References

1. Allied Vision TechnologiesTM website, <http://www.alliedvisiontec.com>
2. Abdelhalim, A., Traore, I.: Converting Declarative Rules into Decision Trees. In: Proc. of the World Congress on Engineering and Computer Science (WCECS), San Francisco, USA (2009)
3. Benet, G., Sim, J.E., Andreu-Garcia, G., Rosell, J., Sanchez, J.: Embedded low-level video processing for surveillance purposes. In: Proc. of 3rd Conference on Human System Interactions (HSI), Rzeszow, Poland, pp. 779–786 (2010)
4. Bramberger, M., Doblander, A., Maier, A., Rinner, B.: Distributed embedded smart cameras for surveillance applications. *Computer* 39(2), 68–75 (2006)
5. Collins, R.T., Lipton, A.J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N., Hasegawa, O., Burt, P., Wixson, L.: A System for Video Surveillance and Monitoring. Technical report CMU-RI-TR-00-12, Carnegie Mellon University, Pittsburgh, and The Sarnoff Corporation, Princeton, NJ (2000)
6. cvBlobslib library website, <http://opencv.willowgarage.com/wiki/cvBlobsLib>
7. Foresti, G.L., Micheloni, C., Snidaro, L., Remagnino, P., Ellis, T.: Active video-based surveillance system: the low-level image and video processing techniques needed for implementation. *Signal Processing Magazine* 22(2), 25–37 (2005)

8. Frejlichowski, D., Forczmański, P., Nowosielski, A., Gościńska, K., Hofman, R.: SmartMonitor: An Approach to Simple, Intelligent and Affordable Visual Surveillance System. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (eds.) IC-CVG 2012. LNCS, vol. 7594, pp. 726–734. Springer, Heidelberg (2012)
9. Haritaoglu, I., Harwood, D., Davis, L.S.: W⁴: Real-Time Surveillance of People and Their Activities. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(8), 809–830 (2000)
10. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man, and Cybernetics - Part C* 34(3), 334–352 (2004)
11. McKenna, S.J., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking Groups of People. *Computer Vision and Image Understanding* 80(1), 42–56 (2000)
12. Microsoft Visual Studio 9™ (2008),
<http://www.microsoft.com/visualstudio/en-gb/products/2008-editions>
13. OpenCV (Open Source Computer Vision) library website,
<http://opencv.willowgarage.com>
14. Ortells, J., Anaya-Sánchez, H., Mollineda, R.A.: A demo of an interactive video-surveillance prototype for perimeter control (2012),
http://miprcv.iti.upv.es/index.php?option=com_%5fcontent&task=view&id=220&Itemid=205
15. Paulidis, I., Morellas, V.: Two examples of indoor and outdoor surveillance systems. In: Remagnino, P., Jones, G.A., Paragios, N., Regazzoni, C.S. (eds.) *Video-based Surveillance Systems*, pp. 39–51. Kluwer Academic Publishers, Boston (2002)
16. Regazzoni, C.S., Cavallaro, A., Wu, Y., Konrad, J., Hampapur, A.: Video Analytics for Surveillance: Theory and Practice. *IEEE Signal Processing Magazine* 27(5), 16–17 (2010)
17. Valera, M., Velastin, S.A.: Intelligent distributed surveillance systems: a review. *IEE Proc. - Vision, Image and Signal Processing* 152(2), 192–204 (2005)
18. Vidal, E., Rodríguez, L., Casacuberta, F., García-Varea, I.: Interactive Pattern Recognition. In: Popescu-Belis, A., Renals, S., Bourlard, H. (eds.) *MLMI 2007*. LNCS, vol. 4892, pp. 60–71. Springer, Heidelberg (2008)
19. wxWidgets cross-platform GUI library website, <http://www.wxwidgets.org>

Interactive Training of Human Detectors

David Vázquez, Antonio M. López, Daniel Ponsa, and David Gerónimo

Abstract. Image based human detection remains as a challenging problem. Most promising detectors rely on classifiers trained with labelled samples. However, labelling is a manual labor intensive step. To overcome this problem we propose to collect images of pedestrians from a virtual city, *i.e.*, with automatic labels, and train a pedestrian detector with them. The resulting detector performs correctly when such virtual-world data are similar to testing one, *i.e.*, real-world pedestrians in urban areas. When testing data is acquired in different conditions than training ones, *e.g.*, human detection in personal photo albums, dataset shift appears. In previous work, we treat this problem as one of domain adaptation and solve it with an active learning procedure. In this work, we focus on the same problem but evaluate a different set of faster to compute features, *i.e.*, Haar, EOH and their combination. In particular, we train a classifier with virtual-world data, using such features and Real AdaBoost as learning machine. This classifier is applied to real-world training images. Then, a human oracle interactively corrects the wrong detections, *i.e.*, few miss detections are manually annotated and some false ones are pointed out too. A low amount of manual annotation is fixed as restriction. Real- and virtual-world difficult samples are combined within what we call *cool world* and we retrain the classifier with this data. Our experiments show that this adapted classifier is equivalent to the one trained with only real-world data but requiring 90% less manual annotations.

1 Introduction

Image based human detection is of paramount interest due to its potential applications in fields such as advanced driving assistance, video surveillance and media

David Vázquez · Antonio M. López · Daniel Ponsa · David Gerónimo

Computer Vision Center (CVC) and the Computer Science Dept.

at the Autonomous University of Barcelona (UAB)

e-mail: {david.vazquez, antonio, daniel, dgeronimo}@cvc.uab.es

analysis. However, by reading some recent surveys of the field [8, 11, 7] we see that even detecting non-occluded standing humans remains challenging. This is not surprising due to the great variety of backgrounds (scenarios, illumination) in which humans are present, as well as their intra-class variability (pose, clothes, occlusion). Nowadays, the most relevant baseline human detector relies on a (holistic) human classifier that uses the so-called histograms of oriented gradients (HOG) as features, and the support vector machines (SVMs) as learning algorithm [5, 4]. New methods have been developed on top of this baseline in order to take into account relative pose of human parts [9], to handle occlusions [27], to take advantage of color [26], etc.

One important aspect of a human detector is its computational cost. HOG features are very effective but expensive to compute. Some works tried to speed up its computation by using integral histograms [18] or specific hardware [20], being the former the most promising one. Haar features combined with AdaBoost [23] were one of the first proposals for *pedestrian*¹ detection, and also made use of integral features. It was extended with Edge Orientation Histograms (EOH) features too [10]. More recently [6] presented a detector based on different integral features, such as color and gradient orientations, which is one of the best performing ones in the state of the art. This work was extended by [2] proposing the fastest pedestrian detector to the date, running at more than 100 frames per second.

One can deduce that the most promising human detectors rely on classifiers developed by following the discriminative paradigm, *i.e.*, trained with labelled samples, being integral features and AdaBoost key ingredients. However, labelling is a manual labor intensive step, especially in cases such as human detection in which labelling objects (humans) means to provide at least bounding boxes. Note that this is more costly for a *human labeller* than just answering to *yes/no*-questions like *is there any human in this image?* (*i.e.*, without specifying *where* in the affirmative cases). In addition, it is well accepted that having sufficient variability in the labelled samples is decisive to train classifiers able to generalize properly [3]. However, traditional (passive) manual labelling does not evaluate the degree of variability achieved by the labelled samples. A common approach is to assume that the larger the set of labelled samples the higher the variability. However, just subjectively adding more examples does not guarantee a higher variability, *e.g.*, it can happen that we are just adding human samples too similar to the ones we already collected.

In order to obtain good samples to train as well as significantly reducing human labelling effort, in [16] we used a video game to collect city images with automatically labelled pedestrians (Fig. 1). By using such virtual-world data we trained a pedestrian classifier that was used within a pedestrian detector operating in real-world images. We employed HOG as pedestrian descriptor and linear SVM as learning machine. The results provided by the virtual-world based pedestrian detector were equivalent to a counterpart detector which pedestrian classifier was trained on real-world images.

¹ We use the term *pedestrian* to refer to a human as a traffic participant.

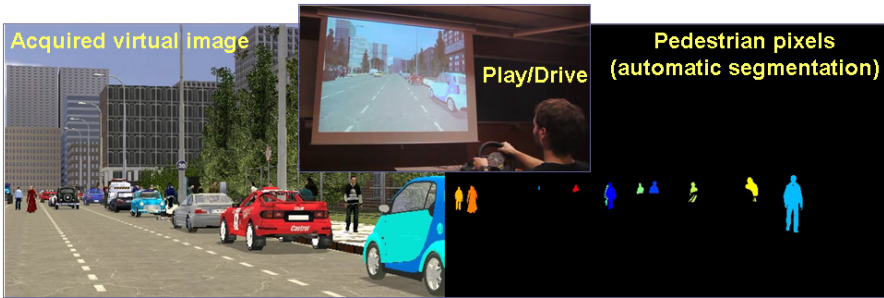


Fig. 1 Virtual-world images with pixel-level groundtruth of pedestrians

In [22, 21] we applied the same procedure for detecting humans in more general images, for instance, in holiday photos. In this case the performance shown by the virtual-world based pedestrian detector was far worse from the one obtained by its real-world counterpart. In fact, we illustrated how the problem remains the same when training and testing with real-world data coming from different domains. In other words, we were suffering dataset shift, but not because the training data was from virtual-world but just because it came from a domain different than the one where the pedestrian detector operated, *e.g.*, training with urban sequences and testing in landscape or indoor scenarios. Accordingly, we casted the problem in a domain adaptation framework based on active learning, *i.e.*, we followed a semi-supervised domain adaptation approach (Fig. 2). Such a framework allowed us to obtain the desired performance by combining our virtual-world data with just a few real-world one (25%) actively labelled.

In [16, 22, 21] we focused on HOG/linear-SVM. In this chapter we extend our study to Haar, EOH and Haar with EOH descriptors, employing AdaBoost as learning algorithm. This is an important setting since, as we mentioned before, it can lead to fast pedestrian detectors thanks to the use of integral images and decision cascades. We will see that Haar/EOH/HaarEOH with Adaboost also presents dataset shift. Fortunately, as for HOG/linear-SVM, we will show how our semi-supervised domain adaptation proposal provides the desired results in this case. We restrict more the number of allowed real-world pedestrian annotations, *i.e.*, those done manually. In [16, 22, 21] we allowed the 25% of the virtual-world pedestrians; here we have reduced it to 10%. User interaction comes in the form of a human oracle who actively annotates some difficult samples from real-world images.

The rest of the chapter is organized as follows. Sect. 2 presents the HaarEOH AdaBoost human detection method. In Sect. 3 we show the details of the proposed semi-supervised domain adaptation algorithm. In Sect. 4 we present and discuss the obtained results. Finally, Sect. 5 summarizes our conclusions.

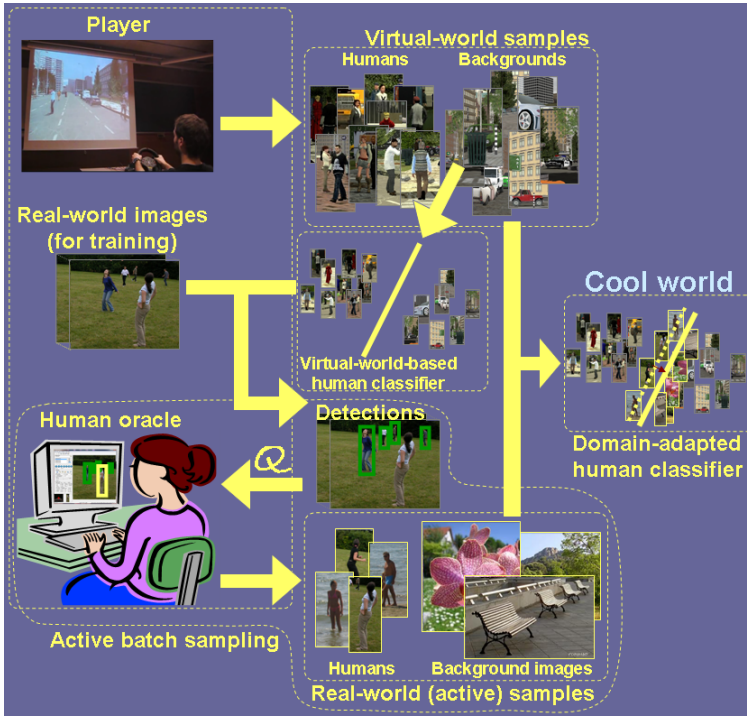


Fig. 2 Our proposal in a nutshell: domain adaptation based on active learning

2 HaarEOH-Based Pedestrian Detection

2.1 Detection Architecture

A pedestrian detector, *scans* an image with a window determining if it contains a pedestrian (*positive*) or not (*negative*) by using a learnt pedestrian classifier which comes from a *learning machine* process. The classifier gets the features computed over each window as input and its class, *i.e.*, positive or negative, as output. Since multiple positive windows can be detected for a single pedestrian, we must *select* a representative one, *i.e.*, the window *detecting* the pedestrian. Let us briefly review the features, learning machine, scanning, and selection.

2.1.1 Features

We use two different features that can be computed using the so-called integral image which has been demonstrated that speeds up object detection [25] and that recently is attracting much interest [6, 2].

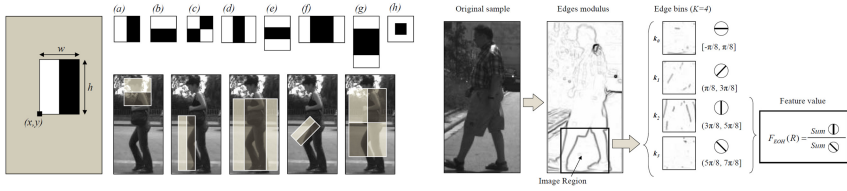


Fig. 3 *Left: Haar filters.* Example of a filter with parameters (x, y, w, h) with basic forms of the Extended Haar set and examples of filters that give high response in regions containing pedestrians. *Right: EOH features.* The feature is defined as the relation between two orientations of a region. In this case, vertical orientations are dominant with respect to the diagonal orientations (k_3), so the feature will have a high value.

Haar filters were introduced in [15] to detect pedestrians using a static camera. A single feature of this set is defined as the difference of illumination between two areas (white and black, see Fig. 3 left). The sum of the pixel values of a given region, $E(R)$, can be efficiently computed by only four accesses to the integral image. The feature value is:

$$Feature_{Haar}(R, f) = \frac{E(R_{white})E(R_{black})}{wh\sqrt{(E(R))^2 - E^2(R) + \varepsilon}},$$

where R_{white} and R_{black} are the white and the black rectangles of the filter f , and w , h refer to the width and height of rectangle R . $E^2(R)$ is the sum of the square of the pixels of a given region. Note that the denominator is a contrast normalization factor that depends on the region size and standard deviation. Originally, [15] presents three basic filters, namely (a)(b)(c) in Fig. 3. Posteriorly, [24] add filters (d) and (e) to the previous set in order to achieve face detection and for pedestrian detection using a static camera in [25]. This set is referred in this chapter as Simple Haar set. In our work, we use filters from (a) to (h), coming to use the Extended Haar set described in [14] to detect faces.

EOH features were proposed in [13] for face detection and used for pedestrian detection in [10] as well. These features are based on gradient information, which not only maintains invariance to global illumination changes, but is also able to extract shape information difficult to capture by Haar filters. This feature extracts similar information to the HOG feature, which is the standard pedestrian detection descriptor, but EOH can be easily computed by using integral images. Features are computed as follows (see Fig. 3). First, the image derivative is computed with a Sobel mask to get the edge orientation. Then, the derivative image pixels are classified according to its edge orientation into K (in our case $K = 6$) images corresponding to K orientation bins. Therefore, a pixel in bin $k_n \in K$ contains its gradient magnitude

if its orientation is inside k_n range, otherwise is null. Integral images are now used to store the accumulation image of each of the edge bins. Finally, the feature value is defined as the ratio between two orientations, k_i and k_j , of region R :

$$Feature_{EOH}(k_i, k_j, R) = \frac{E_{k_i}(R) + \varepsilon}{E_{k_j}(R) + \varepsilon}.$$

If the feature value is greater than a given threshold, then the orientation k_i is dominant to orientation k_j at the region R , which can be exploited as a weak hypothesis too. The small value ε is added to the factors for smoothing purposes.

2.1.2 Learning Machine

The feature descriptors we use are of high dimensionality. Accordingly, we need a machine learning algorithm able to work in such spaces. Boosting algorithms are the most suitable ones as they automatically select a subset of the features that best characterize the problem to learn. From the different boosting proposals, we use real AdaBoost [19], more specifically, we use the implementation of [17]. The key idea is to build a strong classifier by combining a set of weak classifiers.

In an iterative manner, real AdaBoost chooses the weak classifiers that best classify the training set. In the algorithm, each sample has a weight depending on prior classifications; this value is increased in case it has been misclassified by previous rules. Hence, at each iteration, the algorithm focuses its efforts on previously misclassified samples. Finally, the strong classifier is composed of n weak classifiers, where n is defined by the user and usually is much lower than the total number of features of the samples. In our case we collect a total of 22,848 features per window for the Haar descriptor and 42,840 for the EOH one. We restrict the strong classifier to select the same amount of weak classifiers as the dimension of the HOG descriptor we used in [22, 21], which is 3780.

In fact, the learning process could continue until constructing a cascade of strong classifiers (as in [24] for face detection), where the first layer discards clear non-pedestrians, the second layer would discard less clear non-pedestrians and so on, being pedestrians those windows that are not rejected at any layer. This cascade procedure speeds up detection as most of the windows are rejected at the early stages of the cascade. However, in this chapter we are more interested in showing that the pedestrian detector based on Haar and EOH can be learnt using virtual-world samples and domain adaptation techniques. Therefore, we only present results based on training a single layer with the real AdaBoost algorithm.

2.1.3 Scanning

As scanning procedure we apply the pyramidal sliding window [4]. It consists in constructing a pyramid of scaled images, for the range of scales in which we want to

detect the pedestrians. The bottom of the pyramid (higher resolution) is the original image, while the top is limited by the size of the smaller pedestrian to detect. At the pyramid level $i \in \{0, 1, \dots\}$, the image size is $\lceil d_x/s_p^i \rceil \times \lceil d_y/s_p^i \rceil$, being $d_x \times d_y$ the dimension of the original image ($i = 0$), and s_p a provided parameter. We down-sample the image using bilinear interpolation with anti-aliasing as in [9] for building the lower resolution levels of the pyramid. Then, a canonical window (CW) of fixed size scans each pyramid level according to strides s_x and s_y , in x and y axes, respectively. We set $\langle s_x, s_y, s_p \rangle := \langle 8, 8, 1.2 \rangle$ like in [4] as it is a good tradeoff between processing time and final detection performance. This scanning procedure is not the standard for pedestrian detectors based on descriptors as Haar or EOH. Haar and EOH features are usually scaled themselves instead of using an image pyramid. However, we have experimentally seen that, in general, the pyramid with anti-aliasing boosts the performance of the pedestrian detectors based on self-scaled descriptors.

2.1.4 Selection

Detection over multiple scales and different positions usually yields several detections which frequently refer to a single object. In order to obtain a unique detection per object (pedestrian), we apply the non-maximum-suppression approach proposed in [12].

2.2 Datasets: Real- and Virtual-World Samples

To perform our experiments we use the generic real-world dataset INRIA [5, 4] and our virtual-world one [16]. The widespread INRIA dataset for human detection contains color images of different resolution (320×240 pix, 1280×960 pix, etc.) with persons photographed in different scenarios (urban, nature, indoor). INRIA data includes a set of training images with the bounding box annotation of 1,208 humans (that can be vertically mirrored to obtain 2,416 positive samples). In addition, 1,218 human-free images are provided for training. For testing, INRIA includes a dataset consisting of 563 annotated humans in 288 images, and 453 human-free images.

The virtual-world dataset [16] is generated with Half Life 2 videogame by city driving. It is composed of color images of 640×480 pix. From the provided virtual-world data we mimic INRIA settings for fair comparison. Thus, we use 1,208 virtual-world humans that are vertically mirrored to obtain 2,416 ones, as well as 1,218 human-free virtual-world images. Virtual-world data is only used for training, *i.e.*, for being domain adapted to the training data of INRIA.

2.3 Training with Virtual- vs Real-World Samples

As we mentioned in Sect. 2.1.1 in this chapter we use Haar, EOH and HaarEOH features and real AdaBoost learning machine for training human/pedestrian



Fig. 4 Top: virtual pedestrians and city scenarios. Bottom: INRIA photographs with humans and diversified scenarios as city, countryside, beach, etc. Humans appear also in such scenarios. Domain adaptation by batch active learning (Fig. 2) will bring together virtual-world samples and difficult real ones to learn real-world human classifiers.

classifiers. Accordingly, we train the INRIA human classifier using the INRIA training set and the virtual-world pedestrian one using virtual-world training data. During training, *bootstrapping* is used, *i.e.*, enriching the respective negative training sets with hard negative samples and then re-training. Hard negatives are collected from the corresponding negative training images by applying the initially learnt classifier. The process is iterated until very few new negatives are incorporated. In practice, these particular training sets saturate with a single step.

2.4 Testing with Real-World Images: Dataset Shift

In order to evaluate the performance of the pedestrian detectors we follow the procedure proposed in [7] for this purpose. This means that we use performance curves of *miss rate vs false positives per image*. We focus on the range $FPPI=[0.1, 1]$ of such curves, where we provide the *average miss rate* (AMR) by averaging its values taken at steps of 0.01. Accordingly, such an AMR is a sort of expected miss rate when having one false positive per five images.

Fig. 5 plots the performance of human detectors based on INRIA and virtual-world training data, applied to INRIA testing set. Comparing the performance of the HOG/linear-SVM and the HaarEOH/real-AdaBoost we realize that is almost the same. Moreover, we show the results of three different pedestrian detectors based on different sets of features: Haar/real-AdaBoost, EOH/real-AdaBoost and HaarEOH/real-AdaBoost. The pedestrian detectors trained on the INRIA dataset clearly outperforms their counterparts trained on the virtual-world one. The gap of performance is over 10 points. We argue, as in [22, 21] that this gap is due to dataset shift. In the next section we will explain how to solve this problem by using a semi-supervised domain adaptation technique.

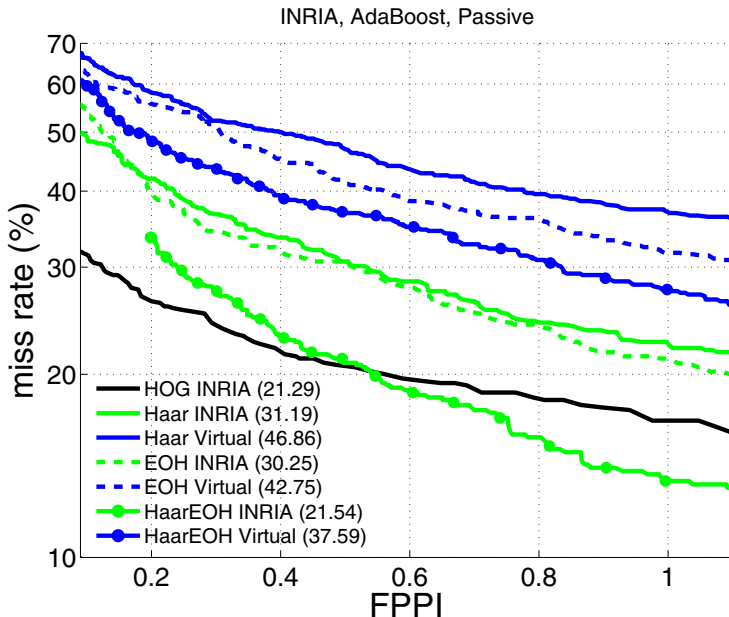


Fig. 5 Per-image evaluation of different pedestrian detectors. The notation *Feature DB* means that the corresponding classifier was learnt using *Feature*, and *DB* training data. In all cases the number inside the parenthesis indicates the average miss rate (AMR) in percentage, for the plotted FPPI range.

3 Semi-supervised Domain Adaptation

The dataset shift problem can be solved with a domain adaptation technique. In this case we use an active learning procedure similar to the one we employed in [22, 21] but using other kind of sample selection methods to reduce the amount of needed supervision during the annotation of the active samples. Additionally, we reduce the amount of allowed actively collected samples from the 25% in [22, 21] to 10%, which is a much more restrictive scenario.

Let us start by introducing some notation and concepts. We denote by \mathcal{D}_s and \mathcal{D}_t two domains from which we observe samples. We refer to \mathcal{D}_s as the *source* domain, while \mathcal{D}_t is the *target* domain. Our problem is that given a sample $x_t \in \mathcal{D}_t$, we want to know if $x_t \in w_t$, using w_t to denote the samples in \mathcal{D}_t with a particular property in which we are interested in. We want to face this problem by learning a classifier \mathcal{C} able to answer if $x_t \in w_t$. To learn \mathcal{C} we want to follow a discriminative paradigm, *i.e.*, learning from labelled samples. If $x_t \in \mathcal{D}_t$, its corresponding label ℓ_{x_t} equals +1 if $x_t \in w_t$ and -1 otherwise. It turns out that we have very few labelled samples drawn from \mathcal{D}_t to learn a reliable classifier. However, we have enough labelled samples drawn from \mathcal{D}_s . This scenario is called semi-supervised. If the distributions



Fig. 6 Labelling tool. For each displayed image, the human oracle (Fig. 2) performs as follows: (1) if there are not humans, it marks the image as *human-free*; (2) if there are humans, some of them have been detected by the previous classifier (green bounding box), but others may not (not framed). The undetected humans must be manually framed by the human oracle (yellow bounding box).

of the samples in \mathcal{D}_s and \mathcal{D}_t are uncorrelated, then the task would be impossible. However, if they have a sufficient correlation, then we can cast the problem as one of *domain adaptation* [1]. More specifically, we can use the large amount of labelled data from \mathcal{D}_s and a low amount of labelled data from \mathcal{D}_t to learn a \mathcal{C} with chances of succeeding in the task of classifying unseen samples from \mathcal{D}_t . Roughly speaking, our \mathcal{D}_s is the set of image windows cropped from virtual-world images, and our \mathcal{D}_t the set of image windows cropped from the real-world images in which we want to detect humans. A sample x_t is just an image window, w_t is the property of imaging a human (*human class*), and \mathcal{C} a human classifier.

Since we can collect in a cheap way as many samples as we need from our virtual cities, the setting for \mathcal{D}_s holds. However, we assume that we start with no labelled samples from \mathcal{D}_t . As we have seen in Sect. 2.4, a pedestrian classifier trained on virtual-world samples does not perform as good as we expect when applied to some real-world images. However, the obtained performance allows us to assume that there is sufficient correlation between \mathcal{D}_s and \mathcal{D}_t , to the eyes of the features and base learning machine we use. Of course, as we deduce also from results in Fig. 5, \mathcal{D}_s and \mathcal{D}_t are not equal at all. In our case, \mathcal{D}_t is more general (*i.e.*, human detection is more general than pedestrian detection) because more types of scenarios are faced (\mathcal{D}_s is urban like).

Therefore, our problem is reduced to obtaining some labelled samples from \mathcal{D}_t , in a cheap way. Our proposal consists in an extension of the *active learning* procedure proposed in [22, 21] using a *human oracle* to label *difficult samples* and to confirm *easy samples* as right classified. All these samples are coming from \mathcal{D}_t . Usually, the difficult samples are defined as those falling in the ambiguity region of the base classifier at hand, *i.e.*, the area close to the decision boundary. However, in these cases, \mathcal{D}_s and \mathcal{D}_t follow the same distribution and the aim is to label as few samples as possible but being meaningful. Our case, however, is different. Let us say that \mathcal{C}_s has been learnt from \mathcal{D}_s and that $x_t \in \mathcal{D}_t \wedge x_t \in w_t$. If $\mathcal{C}_s(x_t)$ is a negative value, large in magnitude, it turns out that from the viewpoint of \mathcal{D}_s , x_t is far from being in w_t , from imaging a human in our case. In our domain adaptation proposal, we do not consider such x_t as an outlier. On the contrary, these are the informative samples for adapting the domains, *i.e.*, the samples that must label the human oracle. As an extension we also include easy samples, *i.e.*, those ones falling out of the ambiguity region.

Accordingly, a given collection of real-world images it is processed using \mathcal{C}_s to detect pedestrians. Detections are kept. By detections we consider those image windows x_t for which $|\mathcal{C}_s(x_t)| \geq th$. For our real AdaBoost, $|\mathcal{C}_s(x_t)| \geq 1$. Then, it is started a working session in which such images and detections are presented to the human oracle. The responsibility of the oracle is to say if a given image contains no humans (*yes/no*-question), to label missed humans with a rectangular bounding box (Fig. 6) and to confirm the correct detections. Once the whole sequence is processed by the oracle, a new classifier is trained using the labelled samples that were used to build \mathcal{C}_s (virtual-world ones) as well as the new collected difficult samples (real-world ones). We call *cool world*² to the joint space of virtual- and real-world samples. This type of active learning is termed as *batch mode*, because a set of images is processed before re-training. We think that a noticeable fact is to use virtual- and real-world samples to train a human classifier. This kind of process can be iterated. The overall approach is summarized in Fig. 2

4 Experimental Results

Plots of Fig. 5 show the effect of the dataset shift on the performance. To solve this problem we employed the semi-supervised domain adaptation technique proposed in section 3. Fig. 7 and 8 show these experiments in three different plots: Haar, EOH and HaarEOH. The plots compare the performance of the passive trained classifiers shown in Fig. 5 with the active ones explained in section 3. Four experiments are tested following the active learning procedure. Each experiment relies on a different manner of collecting the samples. As a reference we show two baselines that differ in the data used to train their classifiers: *INRIA 10%* one uses a 10% of the INRIA training dataset and the *Rand* one uses the virtual-world training set plus a 10% of the INRIA training dataset. *Act* refers to the active learning experiments explained

² We use the *cool world* term as a tribute to the 1992 movie with that title. In this movie, there is a real world and a cool world, the latter shared by real humans and cartoons.

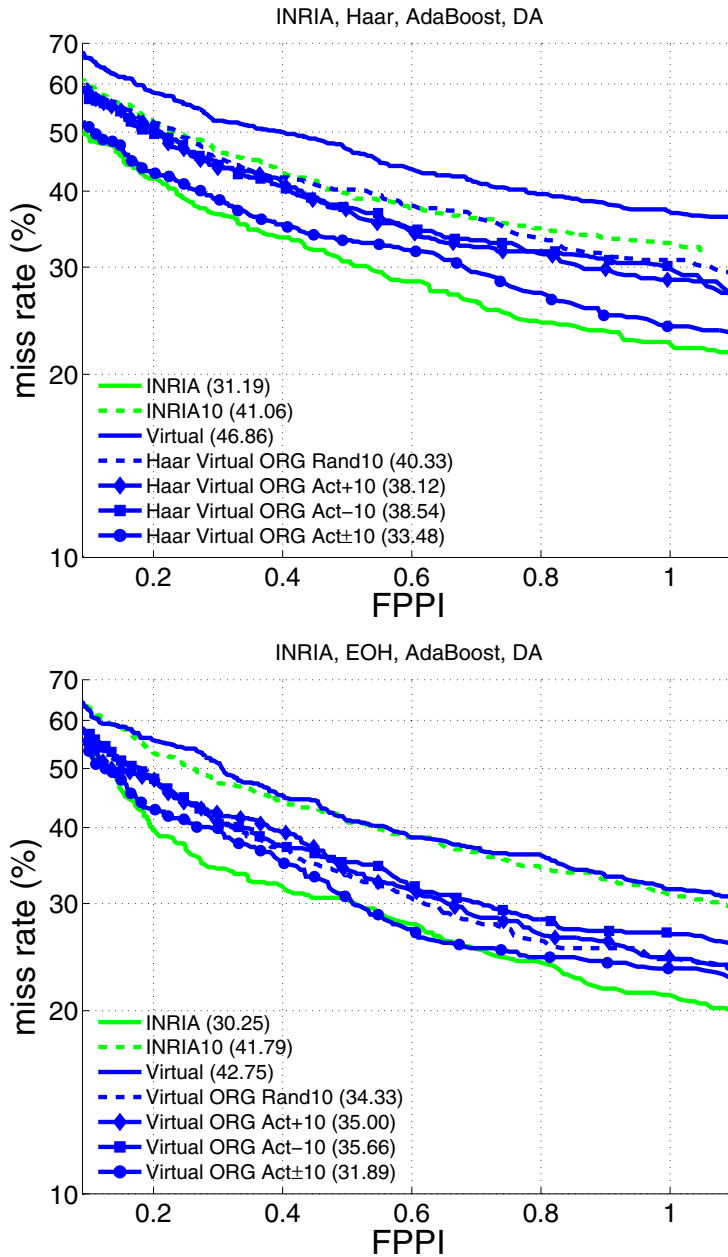


Fig. 7 Semi-supervised domain adaptation experimental results

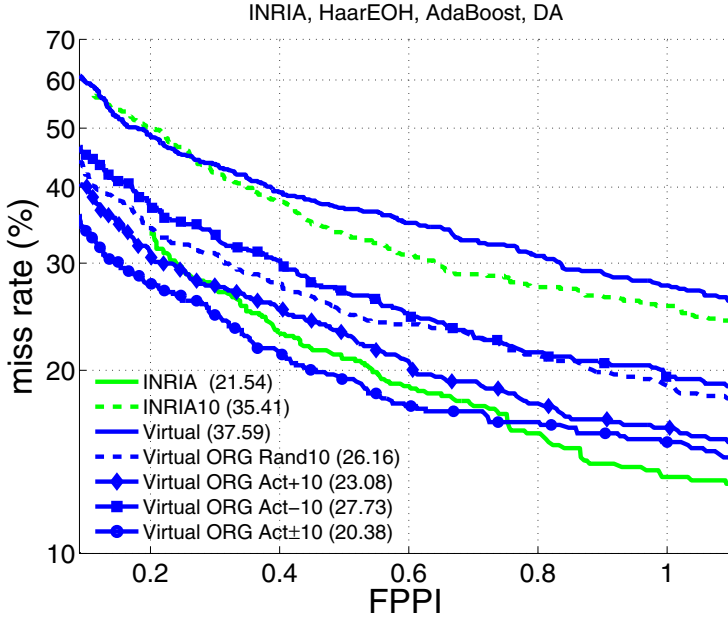


Fig. 8 Semi-supervised domain adaptation experimental results

in section 3. They differ on the oracle annotation procedure. In *Act+* the oracle labels the difficult samples, *i.e.*, the not detected humans, by framing them with a bounding box. In *Act-* the oracle annotates the easy samples, *i.e.*, the correctly detected humans, by just eliminating false positives. Note that *Act-* requires much less annotation effort. Accordingly, *Act±* refers to the combined annotation effort of *Act+* and *Act-*.

From the experiments we can draw the following observations:

- Reducing the training data of INRIA to the 10% decreases the performance of any trained detector by more than 10 points of AMR.
- All detectors benefit from adding a 10% of random INRIA data to the virtual-world set. This benefit varies from 6 to 10 points of AMR.
- Almost all of the tested *Act* experiments outperform the *Rand* and *INRIA 10%* baselines. Note that the trivial procedure of adding random data performs well in other contexts and it is usually difficult to outperform. Our proposed active learning procedures clearly outperform the random ones.
- For Haar and EOH *Act+* and *Act-* perform equally but requiring *Act-* less annotation effort. However, for HaarEOH *Act+* performs better.
- In all the cases *Act±* outperforms the baselines and the other *Act* experiments, even slightly outperforming the INRIA trained pedestrian detector for the most important case, the HaarEOH.

5 Conclusion

In this chapter we have addressed a core problem in the field of human detection, namely, the acquisition of good samples to train at low cost. In order to collect most of the human and background samples we rely on players/drivers of a videogame, *i.e.*, we automatically collect labelled samples while enjoying a game. With them we learn a virtual-world based pedestrian classifier that must work as a human classifier in images depicting the real world. In INRIA images, the virtual-world trained pedestrian classifier cannot reach the performance of a classifier learnt using data manually labelled for training in such dataset. In order to keep the advantage of the cost-free labelling in virtual-worlds, we have cast the problem of transforming the virtual-world based pedestrian classifier into a human classifier for real-world images of general scenarios as a domain adaptation problem. To perform the adaptation, we have used a batch active learning technique that, with just a few manually labelled humans from the real-world images, is able to reach the same performance than a human classifier entirely trained from a much large amount of manually labelled data. Ultimately, our human classifier has been trained by using HaarEOH features that can be computed fast using integral images. We observe that, in a way, we have adopted a multimodal approach from two view points: (1) using two different types of raw data (virtual and real), and (2) collecting the data by playing in the one hand and by working on the other. Besides, user interaction is key as we need a human oracle to annotate real-world samples. Finally, we would like to mention that our proposal can be extended in the future in several ways, *e.g.*, detecting other targets and incorporating spatio-temporal features.

Acknowledgements. This work was supported by the Spanish MICINN, in particular by projects TRA2011-29454-C03-01, TIN2011-29494-C03-02 and Consolider Ingenio 2010: MIPRCV (CSD200700018).

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. *Machine Learning* 79(1), 151–175 (2009)
2. Benenson, R., Mathias, M., Timofte, R., Van Gool, L.: Pedestrian detection at 100 frames per second. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, Providence, RI, USA (2012)
3. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer (2006)
4. Dalal, N.: *Finding People in Images and Videos*. PhD thesis, Institut National Polytechnique de Grenoble, Advisors: Cordelia Schmid and William J. Triggs (2006)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA (2005)
6. Dollár, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features. In: *British Machine Vision Conference*, London, UK (2009)

7. Dollár, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: an evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 34(4), 743–761 (2012)
8. Enzweiler, M., Gavrila, D.M.: Monocular pedestrian detection: survey and experiments. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31(12), 2179–2195 (2009)
9. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA (2008)
10. Gerónimo, D., Sappa, A.D., López, A.M., Ponsa, D.: Pedestrian detection using adaboost learning of features and vehicle pitch estimation. In: *IASTED Int. Conference on Visualization, Imaging and Image Processing*, Palma de Mallorca, Spain (2006)
11. Gerónimo, D., López, A.M., Sappa, A.D., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32(7), 1239–1258 (2010)
12. Laptev, I.: Improving object detection with boosted histograms. *Image and Vision Computing*, 27(5), 535–544 (2009)
13. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: the importance of good features. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, Washington, DC, USA (2004)
14. Lienhart, R., Maydt, J.: An extended set of Haar-like features for rapid object detection. In: *IEEE Int. Conf. on Image Processing*, Rochester, NY, USA (2002)
15. Sinha, P., Osuna, E., Oren, M., Papageorgiou, C., Poggio, T.: Pedestrian detection using wavelet templates. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, San Juan, PR, USA (1997)
16. Marin, J., Vázquez, D., Gerónimo, D., López, A.M.: Learning appearance in virtual scenarios for pedestrian detection. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA (2010)
17. Ponsa, D., López, A.: Cascade of Classifiers for Vehicle Detection. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) *ACIVS 2007. LNCS*, vol. 4678, pp. 980–989. Springer, Heidelberg (2007)
18. Yeh, M., Zhu, Q., Avidan, S., Cheng, K.: Fast human detection using a cascade of histograms of oriented gradients. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, San Diego, CA, USA (2005)
19. Schapire, R.E., Singer, Y.: Improved boosting using confidence-rated predictions. *Machine Learning* 37(3), 297–336 (1999)
20. Sudowe, P., Leibe, B.: Efficient Use of Geometric Constraints for Sliding-Window Object Detection in Video. In: Crowley, J.L., Draper, B.A., Thonnat, M. (eds.) *ICVS 2011. LNCS*, vol. 6962, pp. 11–20. Springer, Heidelberg (2011)
21. Vázquez, D., López, A.M., Ponsa, D., Marin, J.: Cool world: domain adaptation of virtual and real worlds for human detection using active learning. In: *Advances in Neural Information Processing Systems. Domain Adaptation Workshop: Theory and Application*, Granada, Spain (2011)
22. Vázquez, D., López, A.M., Ponsa, D., Marin, J.: Virtual worlds and active learning for human detection. In: *ACM International Conference on Multimodal Interaction*, Alicante, Spain (2011)
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, HI, USA (2001)
24. Viola, P., Jones, M.: Robust real-time face detection. *Int. Journal on Computer Vision* 57(2), 137–154 (2004)

25. Viola, P., Jones, M., Snow, D.: Detecting pedestrians using patterns of motion and appearance. *Int. Journal on Computer Vision* 63(2), 153–161 (2005)
26. Walk, S., Majer, N., Schindler, K., Schiele, B.: New features and insights for pedestrian detection. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA (2010)
27. Wang, X., Han, T.X., Yan, S.: An HOG-LBP human detector with partial occlusion handling. In: *Int. Conf. on Computer Vision*, Kyoto, Japan (2009)

Robot Interactive Learning through Human Assistance

Gonzalo Ferrer, Anaís Garrell, Michael Villamizar, Iván Huerta, and Alberto Sanfeliu

Abstract. This chapter presents some real-life examples using the interactive multimodal framework; in this work, the robot is capable of learning through human assistance. The basic idea is to use the human feedback to improve the learning behavior of the robot when it deals with human beings. We show two different prototypes that have been developed for the following topics: interactive motion learning for robot companion; and on-line face learning using robot vision. On the one hand, the objective of the first prototype is to learn how a robot has to approach to a pedestrian who is going to a destination, minimizing the disturbances to the expected person's path. On the other hand, the objectives of the second prototype are twofold, first, the robot invites a person to approach the robot to initiate a dialogue, and second, the robot learns the face of the person that is invited for a dialogue. The two prototypes have been tested in real-life conditions and the results are very promising.

1 Introduction

Humans live interacting with other people and perform tasks in individual and collective ways everyday. Robotic researchers are interested in designing robots that can interact with people in the same way that humans do. In order to reach this goal, robots must learn from the interaction with humans and learn humans skills used in everyday life to acquire robot social behaviors that can then be used in a wide range of real-world scenarios: domestic tasks, shopping, assistance, guidance, entertainment, surveillance, rescue or industrial shop-floor.

There are many examples where these interactions occur, but some of them are very basic and people do not realize the extreme difficulty that entails executing such tasks for a robot. For example, the navigation in crowded environments,

Gonzalo Ferrer · Anaís Garrell · Michael Villamizar · Iván Huerta · Alberto Sanfeliu
Institut de Robòtica i Informàtica Industrial CSIC-UPC

e-mail: {gferrer, agarrell, mvillami, ihuerta, sanfeliu}@iri.upc.edu

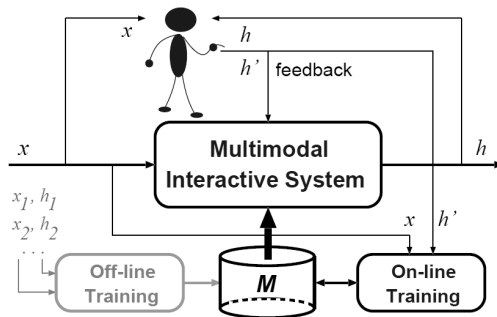


Fig. 1 General multimodal interactive framework

such as crossing streets or shopping malls, or the social engagement to initiate a conversation, are simple examples where this interaction occurs. In the last years important academic and private research efforts have been carried out in this field. Examples can be seen in automatic exploration sites [32], evacuation of people in emergency situations [4], crafting robots that operate as team members [29], therapists [7], robotic services [24] or robot guiding [16, 14].

In this chapter, we will present some examples where the robots learn from the interaction with humans using the general multimodal interaction framework. We will show how the general multimodal system is used in two specific tasks namely: interactive motion learning for robot companion; and on-line face learning using robot vision.

The general idea of the multimodal interactive framework used in the present work is depicted in Fig. 1. As it can be seen, the model can be learned off-line or on-line, and the human -the oracle- uses the information coming from inputs and the outputs to train again the system in order to improve the model. We will see in the two examples how this framework is used.

We have developed two prototypes where the interaction occurs and it is used to improve the systems. The first prototype is “interactive motion learning for robot companion”. The objective is to learn how a robot has to approach to a pedestrian who is going to a destination, minimizing the disturbances to the expected person’s path. In this prototype, the robot has to detect the person’s path, forecast where the person is going to move and approach to the target while taking into account the person intentionality.

The second prototype, “online face learning using robot vision”, has two main objectives. On the one hand, the robot seeks the interaction proactively, the objective is to invite a person to approach the robot to initiate a dialogue. The robot has to take into account the person behavior (reactions) to convince the person to approach the robot. The robot uses a perception system to know the person position and orientation and uses a dialogue and robot motions to invite the person to approach.

On the other hand, the robot learns people's faces. The system learns the face of the person by means of a sequence of images that the robot vision system captures while the person is in front of the robot. The robot only asks the person when the captured face image is very different with respect to the learned face model. If the person agrees with the new face image, the robot uses this image as a positive image to improve the face classifier. In case that the person rejects that face image, the robot uses the image as a negative image to also improve the face classifier. The on-line face learning is done in real-time and is robust to varying environment conditions such as lighting changes. Moreover, it is robust to different people independently of the aspect and gender.

Throughout the two prototypes, the multimodal interactive system improves the accuracy and robustness of the prototypes thanks to the use of a human in the loop. The human plays the role of a teacher with the robots, that is, it evaluates and corrects the results of the robots' tasks in changing environment conditions and human behaviors. The system has been tested in real-life situations and the tests show the improvements of using this framework with respect to using classical non-interactive approaches in several robot tasks.

The remainder of the chapter is organized as follows. In section 2, the interactive motion learning for robot companion approach towards humans is explained. Section 3 describes how the robot performs his active behavior and the online face learning using robot vision to detect and identify the people. Finally, the last section briefly reviews the topics discussed in the different sections of this chapter and establishes the final concluding remarks of this work.

2 Interactive Motion Learning for Robot Companion

Navigation in crowded urban environments, such as crossing streets or shopping malls, is an easy task for humans. However, it is extremely difficult for a robot due to the high environment uncertainties and the variability of the human behavior. The uncertainties associated to the problem can be partially overcome using the multimodal interaction (MI) framework, shown in Fig. 1, where the human can teach specific issues of the robot companion approach.

The aim of this prototype is to show how a robot can learn to accompany a person and navigate safely and naturally in urban settings, minimizing the disturbances to the expected person's paths in two different situations: when crossing the path of a person and when approaching a person to guide him/her to a destination. We are considering for this prototype that we know the urban map, the obstacles and that the robot guides one person. The person can move in any direction, but the goal of the person is to arrive to a given destination, and the robot must accompany the person minimizing the disturbances to his(her) trajectory. Due that the person can change anytime his(her) trajectory, the robot must track the person and anticipate to his(her) path using a human motion predictor. In summary the system has to take into account the following requirements:

- The robot has to track the person path, while handling occlusions and crossings.
- The human motion predictor must infer the person motion intentionality (goal), forecasting the path required to get there.
- The robot has to use its navigation model and a human motion predictor to take into account the person's motion intentionality.

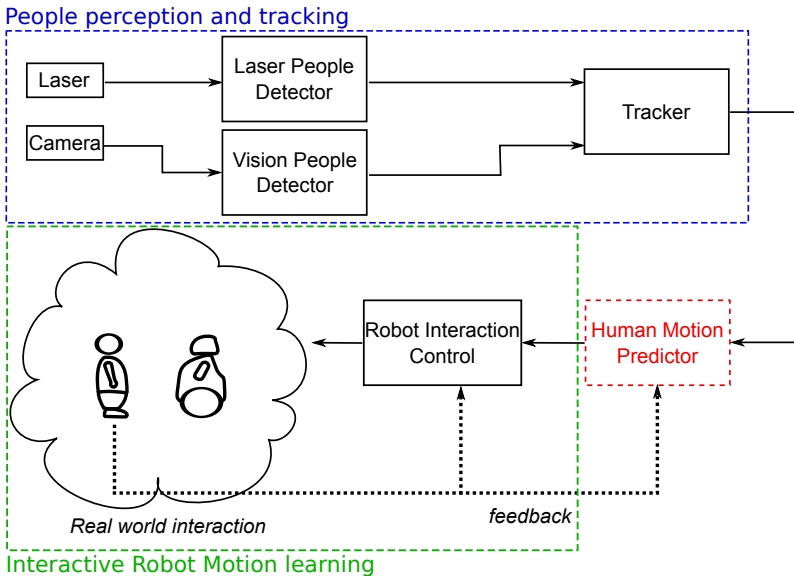


Fig. 2 Interactive Motion Learning: Schematic prototype of the interactive motion learning for robot companion.

The prototype scheme is depicted in Fig. 2. We can realize that it shares some issues of the general multimodal interaction framework shown in Fig. 1. The input to the system is the robot motion and the person path, which are obtained through the robot odometry and the robot laser/vision person tracker. The output of the system is the robot motion approaching or guiding the person. The human in the loop provides the multimodal interaction and he/she can modify the robot motion behavior in different ways. We have used in this prototype the on-line feedback of the person by using a subjective measure of comfortableness of the target being approached or guided. This measure allows to learn some parameters of the robot motion.

2.1 People Detection and Tracking

People detection is needed to track person motion and to extract the learning parameters for comfortable robot navigation in urban sites. Our tracker combines the information of a laser detector, based on [2] and a vision detector based on the Histogram of Oriented Gradient [6]. The people tracker uses the ideas of the work

of [1, 25] with some variations, for example instead of using a Kalman filter, we use a particle filter.

The information of both detectors, the laser and the camera, is fused to obtain a robust detection of the people. The output of this fusion is used as the tracker input.

2.2 Human Motion Prediction and the Social-Force Model Applied to Robot Companion

As we have commented, we need a human motion predictor to know where the person will be after some period of time and a navigation model that allows to navigate safely in the urban area, and that can learn the best parameters to accompany a person.

There are several human motion predictors in the literature. The work of Bennewitz [3] learn the different human motion paths using clustering techniques. The work of Foka [11] uses a geometric model to find the best trajectory from the person position and the destination. The work of Ferrer [10] uses a geometric model but using the present and the previous person path to infer the destination. We have used in this prototype a new model, a Bayesian human motion predictor that calculates the person posteriori probabilities to reach all destinations in the scene. The path to the destination that obtains the highest probability is used as the trajectory that will follow the person, that is the human motion prediction model.

With respect to the robot navigation model, there exists in the literature a high number of models, but they are oriented to the navigation of a robot in a static environment or when the moving objects are not humans. When there are humans in the robot trajectory or when the robot must accompany persons, then there are few works that deal with this issue. The best well known model is based on “social forces” and it has become important for human robot interaction studies. The social-force model was proposed by Helbing [20] to explain the human to human “virtual” forces that appear when two or more humans have motion interactions, that means one person guides another one, both persons follow the same trajectory to collide, one person wants to transverse a group of people, etc. The Helbing’s approach treats each person as a particle abiding the laws of Newtonian mechanics, more specifically, there are several forces in the motion interaction between humans, for example the dragging force that appears when a person follows another one, or the push force that happens when a person is approaching another person without stopping. An extension of Helbing’s work that takes into account the time of collision has been proposed by Zanlungo [37]. We have extended this social-force model to the relations between robots and humans [15] and applying it for guiding people in urban areas with two or more robots.

In this prototype we use the social-force model including additional forces for accompany a person to a destination. The aim is to obtain the force that the robot must apply at each instant i , F_i . This force F_i governs the trajectory to the

destination goal p_i and it is computed as the summation of the attractive force to go to the goal f_i^{goal} and the robot interaction force F_i^{int} to the static and dynamic objects or persons.

$$\mathbf{F}_i = \mathbf{f}_i^{goal} + \mathbf{F}_i^{int} \quad (1)$$

Let us go to describe each one of these forces. Assuming that a pedestrian tries to adapt his or her velocity within a *relaxation time* k_i^{-1} , the attractive force to go to the goal, \mathbf{f}_i^{goal} , is given by:

$$\mathbf{f}_i^{goal} = k_i(\mathbf{v}_i^0 - \mathbf{v}_i) \quad (2)$$

The relaxation time is the interval of time needed to reach a desired velocity and a desired direction.

The interaction force F_i^{int} is the summation of all the repulsive forces, $\mathbf{f}_{i,q}^{int}$, that interact with the robot coming from static (obstacles) and dynamic objects (people, cars, ...). This force prevents humans from crashing with static obstacles o , humans (or dynamic objects) p_i or the robot r . These person-robot interaction forces are modeled as:

$$\mathbf{f}_{i,q}^{int} = A_q e^{(d_q - d_{i,q})/B_q} \frac{\mathbf{d}_{i,q}}{d_{i,q}} \quad (3)$$

where $q \in P \cup O \cup \{r\}$ is either a person (or any moving object), a static object of the environment or the robot. A_q and B_q denote respectively the strength and range of interaction force, d_q is the sum of the radii of a pedestrian and an entity and $\mathbf{d}_{i,q} \equiv \mathbf{r}_i - \mathbf{r}_q$.

The parameters of the previous equation are obtained in a two-step optimization: first we optimize the parameters of the model forces describing the expected human trajectories under no external constraints and consequently we obtain the k parameter and second, we optimize the parameters of the force interaction model under the presence of a moving robot, taken into account that these are the only external force altering the outcome of the described trajectory, obtaining $\{A, B, d\}$. All optimizations are carried out using genetic optimization algorithms [17].

The robot force \mathbf{F}_i is the result of applying all the forces that are needed for the robot navigation. By computing this force at each instant i , we obtain a robot trajectory that can be seen as a reactive navigation system. When we incorporate the human motion prediction to the computation of this force, then the behavior of the system is more than reactive, then we can improve the robot motion because it is anticipating the human motion. This is specially important for guiding or approaching people, because the robot anticipates his(her) motion trajectory.

In this prototype, we have gone a step further, we have incorporated a multimodal interaction approach to modify the robot forces (and indirectly its velocity and trajectory) to improve the comfortableness of the person when moving to a destination and a robot perturbs his(her) trajectory. For our experiments, the person that is approached and guided by the robot, has a video-game controller (a wii device)

to modify the parameters that control the robot forces (we will explain these parameters in the next section). We will call this person, person-controller. The person-controller through a video-game controller dynamically modifies the robot forces meanwhile tries to perform a determined trajectory aiming to a given destination. In our experiments, first the robot has to approach the person-controller and then the robot accompanies it to the destination. In the first part of the experiment, the robot is far away from the personal space of the person-controller and he/she can modify its trajectory (or the robot velocity or trajectory using the wii device) if he/she feels that the robot can collide with him/her). In the second part of the experiment, when the robot is near the personal space of the person-controller, he/she can control the robot velocity or trajectory if he/she feels that the robot is moving too fast or too slow.

2.3 *Interactive Robot Motion Learning*

We will explain in this section how the human can modify the robot forces using the subjective measure of comfortableness, and how we learn these parameters. As we have commented previously the person-controller uses a wii device to send the on-line feedback to the robot.

The robot motion is based on the social forces commented in the previous sections, and the robot autonomously moves to the destination goal, first looks for the person and then accompanies him/her to the destination goal. While the robot accompanies a person, interaction takes place continuously, through the social forces and also using the human feedback of comfortableness, to learn different robot approaching behaviors. There are few articles regarding this topic. The work of Fox [12] or more recently the work of Fraichard [13] analyzes dynamical obstacle avoidance strategies for robot navigation; the work of Kanda [23] uses prediction strategies in social robots in a train station; and the works of Chung [5] or Henry [21] deal with robot control design.

In our system, the on-line feedback is a subjective measure, which varies some parameters of the system by weighting the contribution of all the active forces. The forces that we have considered are:

- Force to the target destination: we infer the target destination by using the intentionality prediction described at section 2.2 and thus the robot aims to the most expectable target destination.
- Force aiming to the person: either the current person position as well the expected motion prediction are known.
- Force of interaction: that is a repulsive force due to the relative position and velocity between the robot and the target.

The combination of these three forces determines the behavior of the robot while the robot is approaching the person. In contrast to the social-force model, two different goals are taken into account. First, a force makes the robot to approach to the

predicted destination $f_{r,dest}^{goal}$. Furthermore, the robot must approach the person who must accompany, hence, a second goal pushes the robot to move closer to the person p_i , $f_{r,i}^{goal}$, which are analogous to eq. 2

$$F^r = \alpha f_{r,dest}^{goal} + \beta f_{r,i}^{goal} + \gamma F_{r,i}^{int} \quad (4)$$

The most interesting part of the system so far, resides in the fact that the approach proposed does not require static targets, the robot is able to navigate near to moving persons.

Although we want to obtain a general approaching rule, it highly varies from person to person in addition to the highly noisy environment in which we are working. Accordingly, we propose the use of an $erfcf(x)$ function to measure the contribution of the human feedback provided $\{\alpha, \beta, \gamma\}$. By using this function we guarantee a slow change in the contribution of these parameters near its constraints. While iteratively repeating the robot physical approach, the provided feedback refines the weights of the force parameters and we can infer a basic interactive behavior where the person feels comfortable under the presence of the robot.

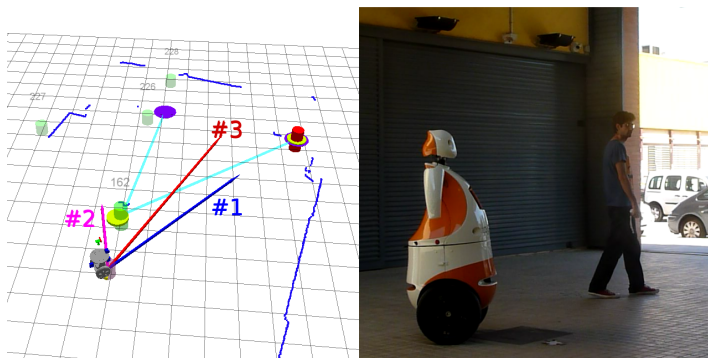


Fig. 3 Illustration of the experiment. On the left is depicted the robot interface, in which the social forces can be appreciated, centered on the robotic platform. On the right hand side of the picture appears the real scene.

As can be seen in Fig. 3, we have reproduced the experiment under controlled conditions. The left figure shows the robot motion and after a few approaches to the target, the robot captures the behavior of the person, by heading towards the most expectable destination of the target. The attractive force to the target destination is plotted as the #1 arrow, and the force approaching the person is plotted as the #2 arrow. The interaction force represents the repulsion generated by the target towards the robot. This force is important to reach the state where the robot does not approach too close to the target, as this behavior will most likely produce repulsion. The result of all the weighted forces is represented as the #3 arrow.

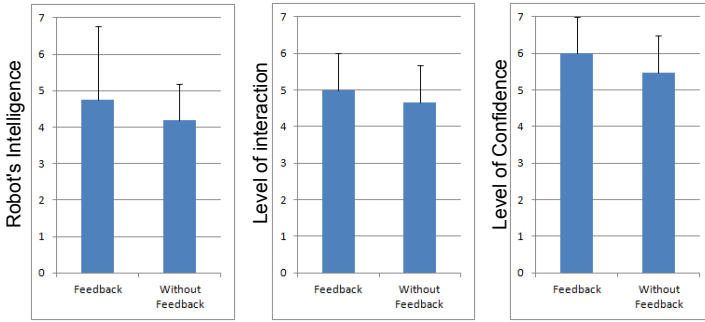


Fig. 4 People’s perception of the use of the interaction (remote control). **Left:** Robot’s Intelligence. **Center:** Level of interaction. **Right:** Level of confidence.

2.4 Experimental Results

In order to validate the usefulness of our contributions to the robot companion subject, that is, making use of human motion prediction and a human feedback as a measure of comfortableness, we have made a set of experiments combining these characteristics and evaluating the overall performance of each combination:

- With feedback
- Without feedback

The measurement of the performance of the overall system is a simple rating on a Likert scale between 1 to 7. For the evaluation score, ANOVA measurements are conducted. It is necessary to study if the use of the remote control enhances the interaction between the robot and a person.

In order to analyze if the use of the remote control enhances the interaction between the robot and a person, three different scores are examined: “Robot’s Intelligence”, “Level of interaction” and “Level of confidence”, plotted in Fig. 4. To summarize, the multimodal feedback under the shape of a wii remote controller improves the subjective performance, according to the poll, nevertheless, the improvement is marginal.

3 Autonomous Mobile Robot Seeking Interaction for Human-Assisted Learning

In the last years, great efforts have been carried out by researchers around the world with the aim of creating robots capable of initiate and keep dynamic and coherent conversations with humans [27]. If robots are able to start a conversation, they create an active engagement with people which can be used to seek assistance from them. This engagement is particular convenient to improve some robot skills. For example, a human can act as a teacher to guide and correct the robot’s behavior or

its response. This active interaction leads to improve the robot capabilities using the human knowledge.

In this section, we present a multi-modal framework where robot and human interact actively to compute an on-line and discriminative face detector. To achieve this objective, the proposed framework consists of two main components or steps. The first one corresponds to create the engagement between the robot and a human, whereas the second step refers to the computation of the on-line face detector once the engagement and the dialogue are established.

More specifically, during the first step, the robot seeks and approaches to a human in order to initiate the conversation or interaction. This is done using its sensors and approaching algorithms. Once the conversation is initialized, a coherent dialogue is conducted during the second step to compute and refine the face detector using the human assistance. This results in a robust and discriminative face detector that is computed on the fly and is assisted in difficult circumstances.

The proposed framework is described in the following. Sec. 3.1 shows the proactively seeking interaction between the robot and humans (first step), and Sec. 3.2 describes the on-line face detector and the procedure used to assist the classifier using human-robot interactions (second step).

3.1 *Robot's Proactively Seeking Interaction*

Recently, social robots have begun to move from laboratories to real environments to perform daily life activities [30, 31, 35]. To this end, the robots must be able to interact with people in a natural way. Recent studies have shown robots which are able to encourage people to begin interaction [8, 19], but using a strategy based on people approaching to the robot in order to establish the interaction and dialogue. Contrary, we present, in this section, a method where the robot is proactive and approaches to people to initiate the interaction and establish the engagement. This is exemplified in Fig. 5.

This proactive way of creating engagements between people and robots enables numerous applications such as guiding robots, tourism robots, or robots focused in approaching people for providing information about a specific urban area. On the other hand, this engagement can be also useful to assist the robot and improve its skills. For example, using the human help, the robot can improve its vision skills. Therefore, it can detect objects and faces in a more robust and discriminative manner. The human can assist the robot to validate or correct the robot responses when it has uncertainty about its predictions. In this way, the robot capabilities are improved along with the number of human interventions. This particular application is addressed in Sec. 3.2.

To seek the interaction with humans, the robot has a people detector that allows to localize and identify humans in its neighbourhood. Once the person is localized, the robot approaches and invites the human to initiate and participate in the interaction. The robot is also able to respond according to human reactions. For instance, if the robot invites a person to approach, and he ignores it, the robot will return to insist.



Fig. 5 Robot approaching. The TIBI robot approaches to a human to start the interaction.

However, if human does not approach, the robot will search for another volunteer. Furthermore, if a person shows interest in the robot, it will start the interaction process with this person.

The active robot’s behavior is performed developing a finite state machine. This state machine allows robot to react depending on people’s behavior. The robot is able to decide if humans are interested in starting the interaction by tracking people positions only.

The robot’s behavior is based on the conceptual framework known as “proxemics” presented by Hall [18], which studied human perception and the use of the space. This work proposed a basic classification of distances between individuals:

- Intimate distance: the presence of other person is unmistakable, close friends or lovers (0-45cm).
- Personal distance: comfortable spacing, friends (45cm-1.22m).
- Social distance: limited involvement, non-friends interaction (1.22m-3m).
- Public distance: outside circle of involvement, public speaking (>3m).

Based on these proxemics, Michalowski et al. [26] classified the space around a robot to distinguish human’s levels of engagement while interacting or moving around a robot. In the present work, our robot tries to maintain a social distance through voice messages and movements.

In Table 1 some sample phrases uttered by the robot are presented. Allowing the robot to acquire the proactive behavior, the number of interactions between the robot and people increases, so, as it will be explained in section 3.2 humans are able to assist the robot in the the computation of an on-line method for face recognition.

Table 1 Robot’s utterances. Some utterances used during the human-robot interaction to keep an active and coherent conversation.

Invitation to create an engagement	Hey, how are you? I’m Tibi. I’m trying to learn to detect faces, will you help me?
	Hi, I am Tibi, I’d like to learn how to recognize different objects, can you be my teacher?
Invitation to continue the interaction	I only want to talk to you, can you stay in front of me?
	Please, don’t go. It will take just two
	Let me explain you the purpose of the experiment, and then, you can decide if you want to stay.
Invitation to start the engagement	Thanks for your patience. Let’s start the demonstration.
	Now we are ready to start. I’m so happy you’ll help me.

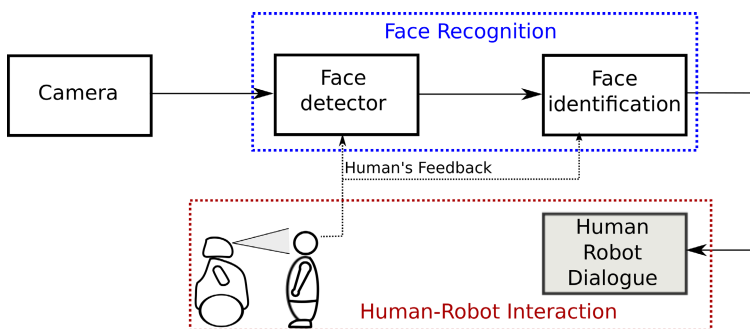


Fig. 6 On-line face learning. The proposed approach consists, mainly, of a face recognition module and a human-robot interaction module. The first module is in charge of detecting and identifying faces, whereas the second one establishes a dialog with a human. The synergically combination of both modules allows to compute a robust and efficient classifier for recognizing faces using a mobile robot.

3.2 On-Line Face Learning Approach

In order to detect and identify faces in images, we use an on-line and discriminative classifier. Particularly, this classifier is based on on-line random ferns [22, 33], which can be progressively learned using its own hypotheses as new training samples. To avoid feeding the classifier with false positive samples, the robot will ask for the human assistance when dealing with uncertain hypotheses. This particular combination of human and robot skills allows to compute a discriminative and robust face classifier that outperforms a completely off-line random ferns [28], both in terms of recognition rate and number of false positives.

Following, the main components of the proposed approach are described in detail. Fig. 6 sketches these constituents and the overview scheme. The synergically combination of a face recognition system with a human-robot interaction module gives the proposed approach: *on-line face learning*.

Human-Robot Interaction. The on-line classifier is learned and assisted using the mobile robot and its interaction with a human. To this end, the robot is equipped with devices such as a keyboard and a screen that enable a dynamic and efficient interaction with the human. The interaction is carried out by formulating a set of concise questions (Fig. 7(Left)), that expect for a ‘yes’ or ‘not’ answer. In addition, the robot has been programmed with behaviors that avoid having large latency times, specially when the human does not know exactly how to proceed. Strategies for approaching the person in a safe and social manner, or attracting people’s attention have been designed for this purpose [9, 36].

Greeting	Nice to meet you Can you teach me to detect faces/objects?
Assistance	Is your face inside the rectangle? I’m not sure if I see you, am I?
No detection	I can’t see you, move a little bit. Can you stand in front of me?
Farewell	Thank you for your help, nice to meet you I hope I see you soon.

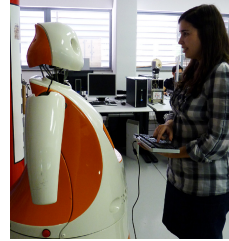


Fig. 7 Human-Robot Interaction. **Left:** Sample phrases uttered by the robot to allow the human assistance. **Right:** The interaction is carried out using diverse devices such as keyboard or touchscreen.

On-Line Face Classifier. The on-line classifier consists of a random ferns classifier [28] that, in contrast to its original formulation, is learned, updated and improved on the fly [33]. This yields a robust and discriminative classifier which is continuously adapted to changing scene conditions and copes with different face gestures and appearance.

Random Ferns (RFs) are random and simple binary features computed from pixel intensities [28]. More formally, each Fern F_t is a set of m binary features $\{f_1^t, f_2^t, \dots, f_m^t\}$, whose outputs are Boolean values comparing two pixel intensities over an image I . Each feature can be expressed as:

$$f(x) = \begin{cases} 1 & I(\mathbf{x}_a) > I(\mathbf{x}_b) \\ 0 & I(\mathbf{x}_a) \leq I(\mathbf{x}_b) \end{cases}, \quad (5)$$

where \mathbf{x}_a and \mathbf{x}_b are the pixel coordinates. These coordinates are defined at random during the learning stage. The Fern output is represented by the combination of their Boolean feature outputs. For instance, the output z_t of a Fern F_t made of $m = 3$ features, with outputs $\{0, 1, 0\}$, is $(010)_2 = 2$.

On-line Random Ferns (ORFs) are Random Ferns which are continuously updated and refined using their own detection hypotheses or predictions. Initially, the parameters of this classifier are set using the first frame. To this end, the opencv face detector is used to find a face candidate with which to start the on-line learning procedure. Subsequently, several random affine deformations are applied to this

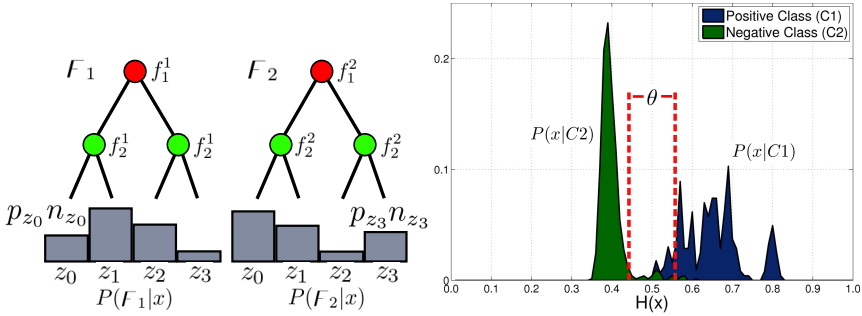


Fig. 8 On-line Random Ferns. **Left:** Ferns probabilities. **Right:** Human-assistance criterion.

training face sample in order to enlarge the initial training set, and initialize the RFs. In addition, the classifier is computed sharing a small set of RFs with the aim of increasing its efficiency, both for the training and detection stages [34].

As shown in Fig. 8(Left), during the on-line training, the number of positive p_z and negative n_z samples falling within each output of each Fern is accumulated. Then, given a sample bounding box centered at x and a Fern F_t , the probability that x belongs to the positive class is approximated by $P(F_t = z|x) = p_z/(p_z + n_z)$, where z is the Fern output [22, 33]. The average of all Fern probabilities gives the response of the on-line classifier:

$$H(x) = \frac{1}{k} \sum_{t=1}^k P(F_t|x), \quad (6)$$

where $\frac{1}{k}$ is a normalization factor. If the classifier confidence $H(x)$ is above 0.5, the sample x will be assigned to the positive (face) class. Otherwise, it will be assigned to the negative (background) class.

The classifier is updated every frame using its own hypotheses or predictions. In particular, the classifier selects the hypothesis (bounding box) with the highest confidence as the new face location. Using this hypothesis as reference, nearby hypotheses are considered as new positive samples, while hypotheses which are far away are considered as new false positive samples. These positive and false positive samples are then evaluated for all the Ferns to update the aforementioned p_z and n_z parameters, see Fig. 8(Left).

Human Assistance. ORFs are continuously updated using their own detection predictions. However, in difficult situations in which the classifier is not confident about its response, the human assistance will be required. The degree of confidence is determined by the response $H(x)$. Ideally, if $H(x) > 0.5$ the sample should be classified as a positive. Yet, as shown in Fig. 8(Right), a range of values θ (centered on $H(x) = 0.5$) is defined for which the system is not truly confident about the classifier response. Note that the width of θ represents a trade off between the frequency of required human interventions, and the recognition rates. A concise evaluation of this parameter is performed in the experimental section.

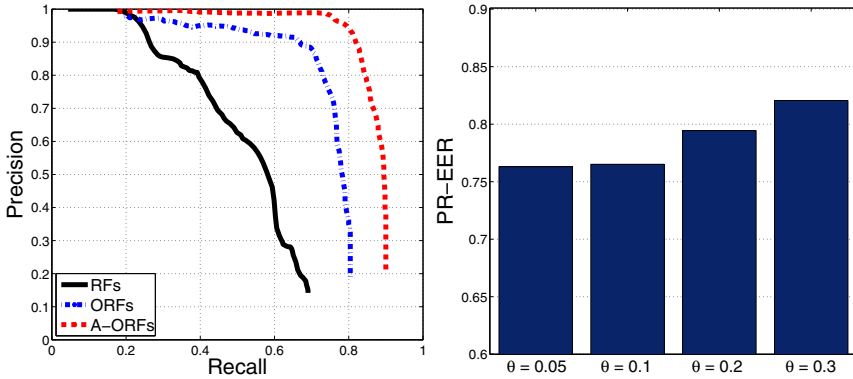


Fig. 9 Face Recognition Rates. **Left:** Precision-Recall curves for different detection approaches. **Right:** Recognition rates in terms of human assistance.

3.3 Experiments

The on-line face learning method is evaluated on a face dataset acquired using a mobile robot. This face dataset has 12 sequences of 6 different persons (2 sequences per person). Each face classifier is learned using an image sequence and tested in the second one. The dataset is quite challenging as faces appear under partial occlusions, 3D rotations and at different scales. Also, fast motions and face gestures disturb the learning method [33].

More precisely, the learning/recognition method is evaluated using three different strategies for building the classifier. First, an offline Random Ferns approach (RFs) is considered. This classifier is learned using just the first frame of the training sequence and is not updated anymore. The second approach considers an ORFs methodology without human intervention. Finally, the proposed human-assisted approach which is denoted by A-ORFs. Remind that the human assistance is only required during the learning stage. During the test, all classifiers remain constant, with no further updating or assistance.

Fig. 9(Left) shows the Precision-Recall curves of the three methodologies, and Fig. 3.3(Left) depicts the Equal Error Rates (EER). Both graphs show that the A-ORFs consistently outperform the other two approaches. This was in fact expected, as the A-ORFs significantly reduce the risk of drifting, for which both the RFs and ORFs are very sensitive, especially when dealing with large variations of the learning sequence.

What is remarkable about the proposed approach is that its higher performance can be achieved with very little human effort. This is shown both in the last 4 rows of the table in Fig. 3.3(Left) and in Fig. 9(Right), where it is seen how the amount of human assistance influences the detection rates. Observe that with just assisting in a 4% of the training frames, the detection rate with respect to ORFs increases a 2%. This improvement grows to an 8% when the human assists on a 25% of the frames.

Method	θ	PR-EER	Human Assistance
RFs	—	55.81	—
ORFs	—	74.79	—
A-ORFs	0.05	76.31	$4.66\% \pm 0.46$
A-ORFs	0.1	76.51	$9.54\% \pm 0.87$
A-ORFs	0.2	79.44	$16.25\% \pm 1.09$
A-ORFs	0.3	82.06	$25.72\% \pm 1.65$

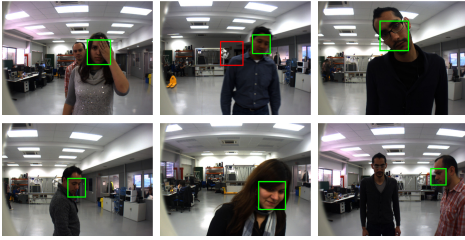


Fig. 10 Recognition Results. **Left:** Face recognition rates for different learning approaches: off-line Random Ferns (RFs), On-line Random Ferns (ORFs) and On-line Human-Assisted Random Ferns (A-ORFs). **Right:** Face detection examples given by the proposed human-assisted method.

Finally, Fig. 3.3 (Right) shows a few sample frames of the detection results, once the classifier learning is saturated (i.e., when no further human intervention is required). The on-line face classifier is able to handle large occlusions, scalings and rotations, at about 5 fps.

4 Conclusions

In this chapter we have presented two different ways of robot learning using the interaction with humans. Furthermore, we have described two different prototypes: interactive motion learning for robot companion; and mobile robot proactively seeking interaction plus human-assisted learning.

We have presented a complete interactive motion learning for robot companion, the “interactive motion learning for robot companion” prototype, in three stages. The first initial design, the perception module, has been implemented and tested extensively in indoor environments. The implementation of the second design, where an external agent moves the robot, was a key step in order to obtain a human intentionality predictor and a motion predictor. A database has been collected of the robot approach to a walking human and the data was used to calculate the model parameters of the intrinsic forces and the interaction forces. For the final stage, we have implemented a multimodal feedback system, where a behavior inference of the weighting parameters of the contributing forces is implemented on-line. All this stages went through intensive real experimentation in outdoor scenarios, by far more challenging scenarios. The results are measured using a poll and its results give information regarding the success of the system.

In the “online face learning using robot vision” prototype the human-robot interaction is performed in a very dynamic and efficient manner. Robot’s proactive behavior has advantages in comparison with passive conducts. Firstly, invitation service, a robot offers information and invites people to interact with it. And, secondly, this behavior increases the number of interactions, and therefore, people can assist the robot to improve its skills continuously. Furthermore, we have realized

that using the interactive multimodal framework, we are able to handle large occlusions, scaling and rotations in different environment and with diverse number of people.

Acknowledgements. This research was conducted at the Institut de Robòtica i Informàtica Industrial (CSIC- UPC). It was partially supported by the CICYT project RobTaskCoop (DPI2010-17112), the MIPRCV Ingenio Consolider 2010 (CSD2007-018) and European project CONET ((INFSO-ICT-224053).

References

1. Arras, K.O., Grzonka, S., Luber, M., Burgard, W.: Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. In: IEEE International Conference on Robotics and Automation, pp. 1710–1715 (May 2008)
2. Arras, K.O., Mozos, O.M., Burgard, W.: Using boosted features for the detection of people in 2d range data. In: IEEE International Conference on Robotics and Automation, pp. 3402–3407. IEEE (April 2007)
3. Bennewitz, M., Burgard, W., Cielniak, G., Thrun, S.: Learning motion patterns of people for compliant robot motion. *The International Journal of Robotics Research* 24(1), 31 (2005)
4. Casper, J., Murphy, R.R.: Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 33(3), 367–385 (2003)
5. Chung, S.Y., Huang, H.: A Mobile Robot That Understands Pedestrian Spatial Behaviors. *Learning*, 5861–5866 (2010)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE CVPR 2005 (June 2005)
7. Dautenhahn, K.: Robots as social actors: Aurora and the case of autism. In: *The Third International Cognitive Technology Conference*, pp. 359–374 (1999)
8. Dautenhahn, K., Walters, M., Woods, S., Koay, K.L., Nehaniv, C.L., Sisbot, A., Alami, R., Siméon, T.: How i serve you?: a robot companion approaching a seated person in a helping context. In: *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*, pp. 172–179. ACM (2006)
9. Feil-Seifer, D., Mataric, M.J.: Defining socially assistive robotics. In: *Proc. of the International Conference on Robotics and Automation*, pp. 465–468 (2005)
10. Ferrer, G., Sanfeliu, A.: Comparative analysis of human motion trajectory prediction using minimum variance curvature. In: *Proceedings of the 6th International Conference on Human-Robot Interaction, Lausanne, Switzerland*, pp. 135–136 (2011)
11. Foka, A.F., Trahanias, P.E.: Probabilistic Autonomous Robot Navigation in Dynamic Environments with Human Motion Prediction. *International Journal of Social Robotics* 2(1), 79–94 (2010)
12. Fox, D., Burgard, W., Thrun, S.: The dynamic window approach to collision avoidance. *IEEE Robotics; Automation Magazine* 4(1), 23–33 (1997)
13. Fraichard, T., Kuffner, J.J.: Guaranteeing motion safety for robots. *Autonomous Robots* (January/February 2012)
14. Garrell, A., Sanfeliu, A.: Local optimization of cooperative robot movements for guiding and regrouping people in a guiding mission. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE (2010)

15. Garrell, A., Sanfeliu, A.: Cooperative social robots to accompany groups of people. *The International Journal on Robotics Research* (2012) (published online September 4, 2012), doi: 10.1177/0278364912459278
16. Garrell, A., Sanfeliu, A., Moreno-Noguer, F.: Discrete time motion model for guiding people in urban areas using multiple robots. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 486–491. IEEE (2009)
17. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison-Wesley (1988)
18. Hall, E.T.: *The hidden dimension, man's use of space in public and private*. The Bodley Head Ltd., Great Britain (1966)
19. Hayashi, K., Sakamoto, D., Kanda, T., Shiomi, M., Koizumi, S., Ishiguro, H., Ogasawara, T., Hagita, N.: Humanoid robots as a passive-social medium: a field experiment at a train station. In: *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 137–144 (2007)
20. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. In: *Physical Review. E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, vol. 51, pp. 4282–4286 (May 1995)
21. Henry, P., Vollmer, C., Ferris, B.: *Learning to navigate through crowded environments*. Robotics and Automation (2010)
22. Kalal, Z., Matas, J., Mikolajczyk, K.: P-n learning: Bootstrapping binary classifiers by structural constraints. In: *Computer Vision and Pattern Recognition* (2010)
23. Kanda, T., Glas, D.F., Shiomi, M., Ishiguro, H., Hagita, N.: Who will be the customer?: a social robot that anticipates people's behavior from their trajectories. In: *Proceedings of the 10th International Conference on Ubiquitous Computing*, pp. 380–389. ACM (2008)
24. Kawamura, K., Pack, R.T., Bishay, M., Iskarous, M.: Design philosophy for service robots. *Robotics and Autonomous Systems* 18(1-2), 109–116 (1996)
25. Luber, M., Diego Tipaldi, G., Arras, K.O.: Place-dependent people tracking. *The International Journal of Robotics Research* 30(3), 280 (2011)
26. Michalowski, M.P., Sabanovic, S., Simmons, R.: A spatial model of engagement for a social robot, pp. 762–767 (2006)
27. Morales, Y., Satake, S., Huq, R., Glas, D., Kanda, T., Hagita, N.: How do people walk side-by-side? using a computational model of human behavior for a social robot. In: *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 301–308 (2012)
28. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast keypoint recognition using random ferns. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 448–461 (2010)
29. Scholtz, J.C.: Human-robot interactions: Creating synergistic cyber froces. In: *Multi-Robot Systems: from Swarms to Intelligent Automata: Proceedings from the NRL Workshop on Multi-Robot Systems*, p. 177 (2002)
30. Siegwart, R., Arras, K.O., Bouabdallah, S., Burnier, D., Froidevaux, G., Greppin, X., Jensen, B., Lorotte, A., Mayor, L., Meisser, M.: Robox at expo. 02: A large-scale installation of personal robots. *Robotics and Autonomous Systems* 42(3), 203–222 (2003)
31. Tasaki, T., Matsumoto, S., Ohba, H., Toda, M., Komatani, K., Ogata, T., Okuno, H.G.: Dynamic communication of humanoid robot with multiple people based on interaction distance. In: *ROMAN, 13th IEEE International Workshop on Robot and Human Interactive Communication*, pp. 71–76. IEEE (2004)
32. Trevai, C., Fukazawa, Y., Ota, J., Yuasa, H., Arai, T., Asama, H.: Cooperative exploration of mobile robots using reaction-diffusion equation on a graph. In: *ICRA* (2003)
33. Villamizar, M., Garrell, A., Sanfeliu, A., Moreno-Noguer, F.: Online human-assisted learning using random ferns. In: *International Conference on Pattern Recognition, Tsukuba, Japan* (2012)

34. Villamizar, M., Moreno-Noguer, F., Andrade-Cetto, J., Sanfeliu, A.: Shared random ferns for efficient detection of multiple categories. In: International Conference on Pattern Recognition (2010)
35. Wada, K., Shibata, T., Saito, T., Tanie, K.: Analysis of factors that bring mental effects to elderly people in robot assisted activity, vol. 2, pp. 1710–1715 (2002)
36. Wilkes, D.M., Pack, R.T., Alford, A., Kawamura, K.: Hudl, a design philosophy for socially intelligent service robots. In: American Association for Artificial Intelligence Conference (1997)
37. Zanlungo, F., Ikeda, T., Kanda, T.: Social force model with explicit collision prediction. EPL (Europhysics Letters) 93(6), 68005 (2011)