Advanced features for image representation: integrating relations, weights, depth, and time

Jean Martinet

HABILITATION À DIRIGER DES RECHERCHES

Discipline : Informatique

Advanced features for image representation: integrating relations, weights, depth, and time

Jean Martinet

Mémoire préparé au sein de l'équipe FOX de CRIStAL,
et soutenu publiquement le 15 décembre 2016

Jury :

Philippe Joly	Professeur	Université Toulouse 1 Capitole	Rapporteur
Georges Quénot	DR CNRS	Université Grenoble Alpes	Rapporteur
Nicu Sebe	Professeur	Università degli Studi di Trento, Italie	Rapporteur
Liming Chen	Professeur	Ecole Centrale de Lyon	Examinateur
Ichiro Ide	Professeur	University of Nagoya, Japon	Examinateur
Joemon Jose	Professeur	University of Glasgow, Royaume-Uni	Examinateur
Olivier Colot	Professeur	Université Lille 1 – Sciences et Technologies	Président
Chaabane Djeraba	Professeur	Université Lille 1 – Sciences et Technologies	Garant

Numéro d'ordre: 42290

Acknowledgements

This document describes my research activities during the last 15 years. Research is a collaborative process, and it is my pleasure to thank all the persons who contributed to this work.

First of all, I would like to warmly thank all jury members for accepting to evaluate my work, and for taking time to travel to Lille to attend the defence; I specifically want to thank the reviewers, Philippe Joly, Georges Quénot, and Nicu Sebe, for writing a report, in addition to carefully reading my habilitation thesis. I thank Olivier Colot for participating to the jury, and for taking on the role of the President of the jury.

I would like to thank Chaabane Djeraba for his support and guidance along the years, and for trusting me. I specially thank all students and colleagues involved in this work. I also thank all members (present and past) of FOX team, IRCICA, CRIStAL, and IUT A for sharing the everyday life at work.

I wish to thank Mohamed Daoudi, Ludovic Macaire, and Nicolas Anquetil for the useful proofreading.

Finally, I am happy to thank my family and especially my dear mother for being extremely supportive. I send a special thank to Fátima for her helpful dedication, Nilo for his encouraging smiles and songs, and Luna for urging me to finish, though she won the race when showing up a month in advance – exactly the time that I needed to finish writing up :-) But hey, life is life and there is always space to welcome a kid at home.

Abstract

Tremendous amounts of visual data are produced every day, such as user-generated images and videos from social media platforms, audiovisual archives, etc. It is important to be able to search and retrieve documents among such large collections. Our work in computer vision and multimedia information retrieval focuses on visual features for image representation. In particular, inside the entire processing chain ranging from visual data acquisition with sensors to the user interface that facilitates the interaction with the system, our research addresses the internal representation of visual data in the form of an index that serves as a reference for the system regarding the image contents.

In the general context of image representation, we describe in a first part some contributions related to the widely-used paradigm of "bags of visual words". We also discuss the general notion of relation, taken at several levels – the low level of visual words, the transversal level aiming for cross-modal annotation, and the high level of semantic objects. Finally, we focus on the definition of weighting models, that serve as visual counterparts to popular weighting schemes used for text.

Because of the specificity of persons and their faces compared to general objects, we focus in a second part on specific features and methods for person recognition. Two directions are developed to overcome some limitations of static 2D approaches based on face images, with the objective of improving systems' precision and robustness. One direction integrates depth in facial features, and the other takes advantage of temporal information in video streams. In both cases, dedicated features and strategies are investigated.

Keywords: Computer vision, Multimedia information retrieval, Image representation, Indexing, Visual features, Weighting scheme, Person recognition.

Résumé

D'immenses quantités de données visuelles sont générées tous les jours, telles que les images et vidéos produites par les utilisateurs des réseaux sociaux, les archives audiovisuelles, etc. Il est important de pouvoir chercher et retrouver des documents au sein de tels grands volumes de données. Notre travail en vision par ordinateur et recherche d'information multimédia porte sur les caractéristiques visuelles pour la représentation d'images. En particulier, dans la chaîne des traitements allant de l'acquisition des données visuelles via des capteurs jusqu'à l'interface utilisateur qui facilite l'interaction avec le système, notre recherche s'intéresse à la représentation interne des données visuelles sous la forme d'un index qui sert de référence pour le système concernant le contenu des images.

Dans le contexte général de la représentation d'images, nous décrivons dans une première partie quelques contributions liées au paradigme populaire des "sacs de mots visuels". Nous discutons également la notion générale de relation, prise à différents niveaux – le bas niveau des mots visuels, le niveau transverse qui vise l'annotation intermodale, et le haut niveau des objets sémantiques. Finalement, nous nous attachons à définir des modèles de pondération, qui servent de pendants visuels des schémas de pondération utilisés pour le texte.

En raison de la spécificité des personnes et visages en comparaison aux objets généraux, nous nous intéressons dans une seconde partie aux caractéristiques et méthodes spécifiques pour la reconnaissance de personnes. Deux directions sont développées pour pallier certaines limitations des approches 2D statiques basées sur des images de visages, avec l'objectif d'améliorer la précision et la robustesse des systèmes. L'une des directions intègre la profondeur dans les caractéristiques faciales, et l'autre exploite l'information temporelle dans les flux vidéo. Dans les deux cas, des caractéristiques et stratégies dédiées sont étudiées.

Mots-clés : Vision par ordinateur, Recherche d'information multimédia, Représentation d'images, Indexation, Caractéristiques visuelles, Schéma de pondération, Reconnaissance de personnes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

In a few decades, the research in multimedia Information Retrieval (IR) and computer vision has reached a very mature state, and yet there are numerous open problems that give the scientific community thrilling challenges. Visual data (images and video) arouse an increasing interest, due to the wide availability of such data, e.g. audiovisual archives, user-generated contents in social media platforms, surveillance, etc. Tremendous amounts of visual data are produced every day, and it is important to be able to search and retrieve documents among large colletions. Visual data gained a major societal importance. Many factors can explain this importance, such as the democratisation of digital cameras (nowadays small and cheap), the low cost of storage, and the generalized access to data networks. For a few examples, in France, there exists over 80 TV channels, meaning that 80 hours of video content are broadcasted every hour. The french INA [1] stores about 70 years of audiovisual archives, representing 300 years of continuous video playing. On YouTube, about 100 hours of video content are uploaded every hour.

The core problems in multimedia IR naturally come from those of text IR. The objective is to help users navigate, browse, or search through a multimedia collection. However, contrary to text documents, visual documents do not contain natural semantic evidence such as words or groups of words that can be used to represent their semantic content, and from which they can be retrieved. This fundamental difference between text and visual documents is often referred to as the *semantic gap*, and has been widely discussed over the last 15 years. This notion was defined by Smeulders *et al.* [SWS+00] as the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. This lack of coincidence originates from the difference between the way we perceive visual content and what the machine can extract.

1. Institut National de l'Audiovisuel: *National Audiovisual Institute*.

In the research presented here, I proposed solutions to overcome some of these problems. This document contains a summary of my research activities spanning from 2002 to 2016, and more specifically at Lille 1 University from 2008. After completing an MSc and a PhD in Computer Science at Grenoble 1 University in 2004, I was awarded a two-year postdoctoral fellowship by the JSPS [1] to visit the National Institute of Informatics in Tokyo (2005-2007). Then I joined Lille 1 University as an Assistant Professor in the FOX [2] research team, that is part of the IMAGE group in CRIStAL [3] laboratory (previously LIFL [4]).

1.2 Main contributions and overview

Multimedia IR is a research domain that is intrinsically multi-disciplinary, involving computer vision, pattern recognition, and image processing only for the visual aspects. My research focuses on image representations for objects and persons. Inside the entire processing chain ranging from data acquisition with visual sensors to the user interface that facilitates the interaction with the IR system, my work addresses the internal representation of visual data in the form of an *index* that serves as a reference for the system regarding the image contents.

A timeline of my research activities is given in Figure 1.1, showing the different contributions organized around the topics of image representations (upper part) and person recognition (middle part). For both topics, the figure shows a selection of related conference and journal papers. The lower part indicates my position across the time. These contributions were made possible by the help of many students and colleagues. Figure 1.1 also shows all students involved in the presented work. The contents of Chapter 2 and Chapter 3 describe our work related to image representations and person recognition, respectively. The description is organized by topic rather than chronologically.

Image representations (Chapter 2): We describe in Chapter 2 several contributions related to image representations (see the upper part of Figure 1.1). These contributions are in the continuity of my PhD and postdoctoral work, and the PhD of Mr Ismail El Sayad (October 2008 – July 2011), that I co-directed with Prof. Chaabane Djeraba. First, Section 2.1 describes several improvements around the widely-used paradigm of **bag of visual words**, that consists in representing images or video keyframes using a set of descriptors, that correspond to quantized low-level features. We present an enhanced bag-of-visual-words model that is based on the use of a new descriptor, the *Edge Context*, that improves the discriminative power by capturing the neighbourhood context of interest points. The Edge Context descriptor has been successfully used as a part of a larger image representation and retrieval model in Ismail's work. We also introduce two alternative methods to build a visual vocabulary, compared

1. Japan Society for the Promotion of Science, URL http://www.jsps.go.jp.
2. Fouille et indexation de dOcuments compleXes et multimedia.
3. Centre de Recherche en Informatique, Signal et Automatique de Lille.
4. Laboratoire d'Informatique Fondamentale de Lille.

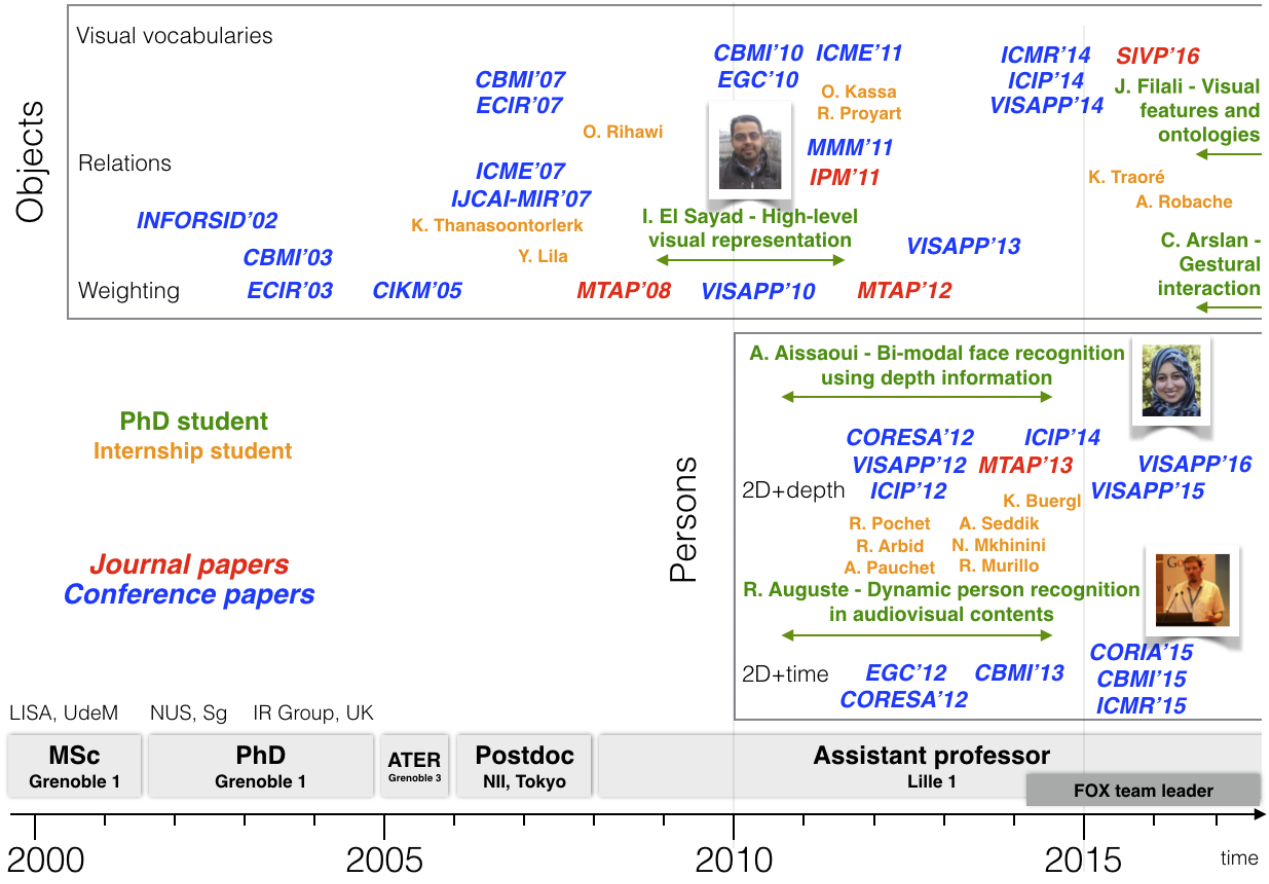


Figure 1.1 – Overview of the work presented in this document.

to the usual KMeans-based quantization of low-level features: (1) an *information-gain-based selection process* where words are iteratively filtered from a large set of randomly selected descriptors based on their information gain, and (2) a *split representation* dedicated to mobile image search, where a standard vocabulary is built on the server side; at query time, another query-specific, lightweight vocabulary is built on the client side (mobile device), that forms a compressed representation of the query image to be sent to the server for matching. Finally, a discussion is given about the basis behind adapting and applying text techniques to visual words.

Then Section 2.2 discusses the general notion of **relation** in image representations. In its general definition, a relation refers to the way in which two or more things are connected. In our work, this notion is taken at several levels: (1) *low level*: we consider relations between visual words in order to discover frequently co-occurring visual words patterns, and to create *visual phrases*, (2) *transversal level*: in cases where a multimedia document contains several modalities, we analyze the relations between visual words and other (textual) modalities in

order to allow cross-modal annotation, and (3) *high level*: we integrate in the representation model the relations between objects of an image to enrich the semantic representation.

Finally, Section 2.3 focuses on the issue of defining a **weighting scheme** for images. Indeed, since almost half a century, a number of weighting models for text have been designed and widely tested, giving more importance to relevant index terms. It is therefore a mature research domain with generally acknowledged good results. However, defining a weighting model for image is not trivial, because this rises the question of defining a notion of *importance* for image regions. This section discusses several attempts to define a weighting scheme of image parts at various levels of granularity, that aims to serve as an image counterpart to the *tf* formulation from the famous *tf-idf* weighting model [Sal71] used for text.

Person recognition (Chapter 3): While Chapter 2 addresses image representations for general objects, this chapter describes our work specifically related to person recognition (see the middle part of Figure 1.1). In image representations, persons – and more specifically their faces, are specific kinds of objects that need dedicated representations. Our objective is to address the limitations of *static 2D* approaches (i.e. those based on a face picture), that come from the appearance variations due to changes in lighting conditions, viewpoints, pose, and partial occlusions (glasses, hair, beard, etc.). We explored two directions for enhancing the accuracy and robustness of static 2D approaches of person recognition:

— the **use of depth** in face recognition (2D+depth),

— and the **use of time** (2D+time).

This work was done during the PhDs of Miss Amel Aissaoui (September 2010 – June 2014) and the PhD of Mr Rémi Auguste (November 2010 – July 2014).

In Amel's work (Section 3.1), we introduced a **2D-3D bimodal face recognition approach** that combines visual and depth features in order to provide higher recognition accuracy and robustness than monomodal 2D approaches. First, a *3D acquisition method* dedicated to faces is introduced. It is based on a stereoscopic reconstruction using an active shape model to take into account the topology of the face. Then, a novel descriptor named *Depth Local Binary Patterns* (DLBP) is defined in order to characterize the depth information. This descriptor extends to depth images the standard LBP originally designed for texture description. Finally, a *two-stage fusion strategy* is proposed, that combines the modalities using both early and late fusion strategies. The experiments conducted with different public datasets, as well as with a new dataset elaborated specifically for evaluation purposes, allowed to validate the three contributions introduced throughout this work. In particular, results showed a high quality for the data obtained using the reconstruction method, and also a gain in precision obtained by using the DLBP descriptor and the two-stage fusion strategy.

In Rémi's work (Section 3.2), we designed a **dynamic approach to person recognition** in video streams. This approach is dynamic as it benefits from the motion information contained in videos, whereas the static approaches are solely based on still images. The proposed approach is composed of two parts. In the first part, we extract *persontracks* (short video

sequences featuring a single person with no background) from the videos and cluster them with a re-identification method based on a new descriptor: *Space-Time Histograms* (STH), and its associated similarity measure. The original contribution in this work is the integration of temporal data into the descriptor. Experiments show that it provides a better estimation of the similarity between persontracks than other static descriptors. In the second part of the proposed person recognition approach, we investigate *various strategies* to assign an identity to a persontrack using its frames, and to propagate this identity to members of the same cluster, based on a standard facial recognition method. Both aspects of our contribution were evaluated using a corpus of real life TV shows broadcasted on BFMTV and LCP TV channels. The experimental results show that the proposed approach significantly improves the recognition accuracy thanks to the use of the temporal dimension both in the STH descriptor and in persontrack identification.

Chapter 4 gives a summary of all contributions presented in this document, and opens research perspectives. Finally, an extended curriculum vitæ is provided in Chapter 5.

Chapter 2

Image representations: visual vocabularies, relations, and weights

Since the early years of content-based image retrieval in the 1990's, there has been a great amount of work dedicated to image representations in indexing and retrieval systems. Indeed, a good representation is a needed ground for good IR system performances because the quality of search results highly depends on the quality of the index. This chapter deals with image representations, and it describes our contributions for enhancing such representations. The first part of this chapter, Section 2.1, is related to the widely-used paradigm of *bag of visual words* – or bag of features, or codebook. This popular approach for representing, searching or categorizing visual documents consists in describing images using a set of descriptors, that correspond to quantized low-level features such as SIFT [Low04] or SURF [BTVG06, BETVG08]. KMeans is the most widely used method for the quantization step. The clusters are used as a codebook where each descriptor is represented by the closest centroid – taken as a visual word, and images are described with visual words [1]. In some situations where e.g. the feature space dimensionality is large, the size of the dataset is large, or the system resources are limited, KMeans can be considered costly or even not feasible. Section 2.1 starts with a description of a new descriptor, the *Edge Context*, that is a part of our work during the PhD of Ismail El Sayad for defining an enhanced bag-of-visual-words model. This contribution was published at **CBMI'10** [**ESMUD10a**], and in more details in an **MTAP journal paper** [**ESMUD12**]. The section also explores two alternative ways to build a visual vocabulary:

1. The first method is an information-gain-based iterative selection process, in a joint work with my colleague Thierry Urruty (University of Poitiers, France) and co-authors. As an alternative to the usual KMeans-based quantization step, this approach consists in randomly selecting a subset of candidates from a large set, and then an iterative process filters out descriptors with the least *information gain* values until the desired number of features is reached, and considered as visual words forming the vocabulary. This work

1. In this chapter, we refer to this setting as the *standard* approach.

was published at **ICMR'14** [UGL+14].

2. The second contribution consists of a split representation dedicated to mobile image search, in a joint work with my colleagues José Mennesson (Lille 1 University) and Pierre Tirilly (Lille 1 University) during TWIRL project [1]. Despite recent technical advances, mobile devices (smartphones and tablets) still encounter limitations in memory, speed, energy and bandwidth, that represent bottlenecks from mobile image search systems. In the proposed approach, a reference visual vocabulary is built offline and kept on a server. At query time, a "disposable" lightweight vocabulary is built on-the-fly on the mobile device, using only the query image. Descriptors from this specific vocabulary are sent to the server to be matched to the reference vocabulary. The proposed method offers an acceptable tradeoff between the mobile device technical constraints and the search precision. This work was published at **ICIP'14** [MTM14].

Section 2.1 ends with a general discussion, published at **VISAPP'14** [Mar14], regarding the basis behind adapting and applying text techniques to visual words. Most visual words approaches are inspired from work in text and natural language processing, based on the implicit assumption that visual words can be handled the same way as text words. However, text words and visual words are intrinsically different in their origin, use, semantic interpretation, etc. More specifically, the discussion brings to light the fact that while visual word techniques implicitly rely on the same postulate as in text information retrieval, stating that the words distribution for a natural language globally follows Zipf's law – that is to say, such words appear in a corpus with a frequency inversely proportional to their rank, this postulate is not always true in the image domain.

The second part of the chapter, Section 2.2, discusses the general notion of *relation* in image representations. While in its general definition, a relation refers to the way in which two or more things are connected, in our work, this notion is taken at several levels: low, transversal, and high levels.

— **At a low level**, we consider the relations beween visual words, and more specifically, how their *occurrence patterns* can be used to define higher-level descriptors. The idea behind such descriptors originates from the concept of *phrases* in text indexing and retrieval. For instance, the phrases "dead end", "hot dog", or "white house" convey meanings that are different from the same words taken separately. Such phrases can automatically be discovered text corpus analysis with simple statistical tools, and by including them in the indexing vocabulary, the system can benefit from a finer representation. In a similar way, visual words that represent parts of real-world objects tend to co-occur in images in a close neighbourhood: they can be seen as *visual phrases*. The objective in this work is to build such visual phrases in order to both enrich and refine the image description, and thereby increase the system precision.

1. TWIRL project: June 2012 – October 2014, ITEA 2 Call 5 10029 – Twinning virtual World (on-line) Information with Real world (off-Line) data sources.

— **At a transversal level**, we analyse cross-modal relations in the context of image annotation. When multimedia documents contain several modalities (i.e. visual, text from optical character recognition or speech transcription), by allowing them to collaborate, the objective is to leverage textual information surrounding image to automatically extract annotations.
— **At a high level**, the relations beween *objects* in an image can be integrated in the description in order to allow the expression of semantic and spatial relations. For instance, "the dog is sitting in front of the door" is a finer description than just "dog" and "door" for an image.

The last part of the chapter, Section 2.3, focuses on the issue of defining a weighting scheme for images, which rises the question of defining a notion of *importance* for image regions. The purpose of a weighting scheme is to give emphasis to important terms, quantifying how well they semantically describe and discriminate documents in a collection. When it comes to images, the users' point of view is central for defining weights for image parts in the contexts of retrieval and similarity matching. Like in text retrieval, an image weighting scheme should give emphasis to important regions, quantifying how well they describe documents. The section explores several ways to define a weighting scheme for images.

2.1 Bag of visual words

The popular bag-of-visual-words approach for representing and searching visual documents consists in describing images using a set of descriptors, that correspond to quantized low-level features. In this section, we start by introducing a descriptor, the *Edge Context*, that is a part of an enhanced bag of visual words model. Then we describe two non-standard ways to build vocabularies. Finally, we discuss the very notion of *visual word*, and the application of text processing techniques (e.g. word filtering, weighting, etc.) to such words.

2.1.1 Discriminative descriptor for bag of visual words

During the PhD of Ismail El Sayad, we defined a novel descriptor for bag of visual words. The improvement of the standard bag-of-visual-words approach consists of enriching the existing SURF descriptor [BETVG08] with an *Edge Context*, that reflects the distribution of edge points around the detected interest points.

Enriching SURF with edge context

The motivation in this work is to refine the description of interest points by embedding a visual context of occurrence, thereby yielding a more discriminative feature. The originality lies in the use of edge points, to capture their distribution, yielding a more discriminative descriptor. This descriptor is inspired from the shape context descriptor proposed by Belongie

et al. [BMP02], with regard to extracting information from edge points distribution. In the proposed approach, the Fast-Hessian detector [BETVG08] (that was designed for SURF) is used to extract interest points, and the Canny edge detector is used to extract edge points. Both interest points and edge points are represented in a 5-dimensional colour-spatial feature space that consists of 3 (R, G, B) colour dimensions plus 2 (X, Y) position dimensions. The feature space, that includes both types of points, is modeled with a Gaussian Mixture Model (GMM). In an image with m interest/edge points, a total of m feature vectors: $f_1, ..., f_m$ are extracted, where $f_i \in \mathbb{R}^5$. In this representation, each point is assumed to belong to one of the n Gaussians of the model, and the Expectation-Maximization (EM) algorithm is used to iteratively estimate the Gaussians' parameter set [1]. The parameter set of Gaussian mixture is: $\theta = \{\mu_i, \Sigma_i, P_i\}$, $i = 1, ..., n$ where μ_i is the mean of the i$^{\text{th}}$ Gaussian cluster, Σ_i is the covariance of the i$^{\text{th}}$ Gaussian cluster, P_i is the prior probability of the i$^{\text{th}}$ Gaussian cluster. At each E-step, we can estimate the expected value of the log-likelihood function, with respect to the conditional distribution of the Gaussian g_i from which f_j comes, under the current estimate of the parameters θ_t at iteration t:

$$p(g_i|f_j, \theta_t) = \frac{p(f_j|g_i, \theta_t)p(g_i|\theta_t)}{p(f_j)} \tag{2.1}$$

$$p(f_j) = \sum_{k=1}^{n} p(f_j|g_k, \theta_t)p(g_k|\theta_t) \tag{2.2}$$

At each M-step, the parameter set θ of the n Gaussians is updated towards maximizing the log-likelihood:

$$Q(\theta) = \sum_{j=1}^{m}\sum_{i=1}^{n} p(g_i|f_j, \theta_t) ln(p(f_j|g_i, \theta_t)p(g_i|\theta_t)) \tag{2.3}$$

When the algorithm converges, the parameter sets of n Gaussians and the probabilities $p(g_i|f_j)$ are obtained. The most likely Gaussian cluster for each feature vector f_j is given by:

$$p_{f_j}^{max} = argmax_{g_i}(p(g_i|f_j)) \tag{2.4}$$

As an illustration in Figure 2.1, the vectors from an interest point in the 2D spatial image space are drawn towards all other edge points that are inside the same cluster in the 5-dimensional color-spatial feature space. The edge context descriptor for each interest point is represented as a histogram of 6 bins for the magnitude (distance from the interest point to the edge points) and 4 bins for the orientation angle. The resulting 24-dimensional descriptor shows several invariance proprieties:

— an invariance to translation, that is intrinsic to the edge context definition since the distribution of the edge points is measured with respect to a fixed interest point;

1. Note that this representation is also used to define a spatial weighting scheme, as described in Section 2.3.1.

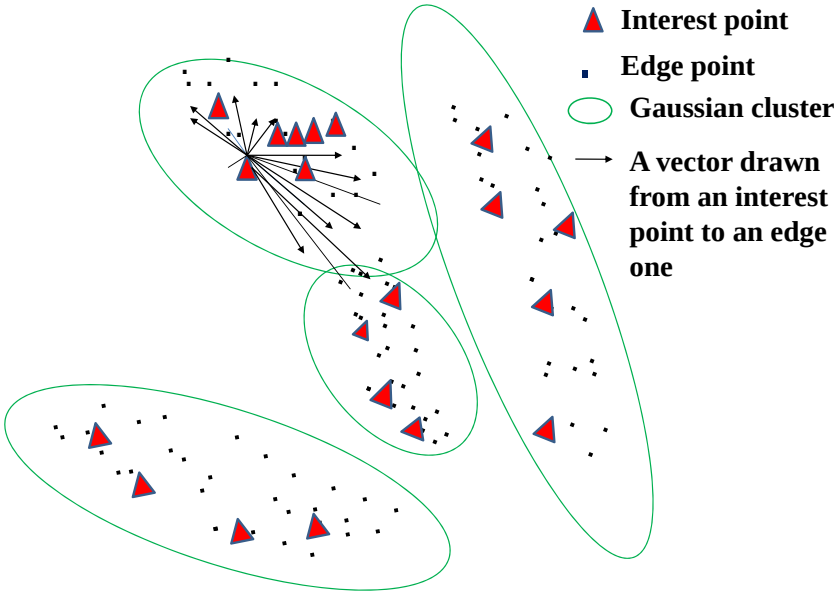


Figure 2.1 – Extraction of the edge context descriptor in the 2D spatial space, after clustering the points in the 5-dimensional color-spatial feature space with a GMM.

— an invariance to scale is achieved by normalizing the radial distance by a mean distance between the whole set of points inside the Gaussian;
— an invariance to rotation is achieved by measuring all angles relative to the orientation angle of each interest point.

After extracting the edge context descriptor, it is merged with SURF descriptor, and the final feature vector is composed of 88 dimensions: 64 from SURF + 24 from the edge context. The distribution of edge points enriches SURF descriptor with a local and discriminative information. Moreover, the distribution over relative positions yields a robust, compact, and highly discriminative descriptor. The set of features is then clustered with e.g. KMeans to form the visual vocabulary.

Evaluation of the proposed descriptor

We used the Caltech-101 dataset [FFFP07] to demonstrate the benefits of the proposed descriptor. Among the 8707 images of 101 classes in this dataset, we randomly selected 30 images from each class to build the visual vocabulary (i.e. 3030 images), and we randomly selected 10 other images from each class (1010 images) to build a test set. Query images are picked from this test set in the experiments. We used KMeans to construct the visual

vocabulary, with several values for K. For each value of K, we compare the accuracy of the proposed descriptor to the one of SURF taken alone.

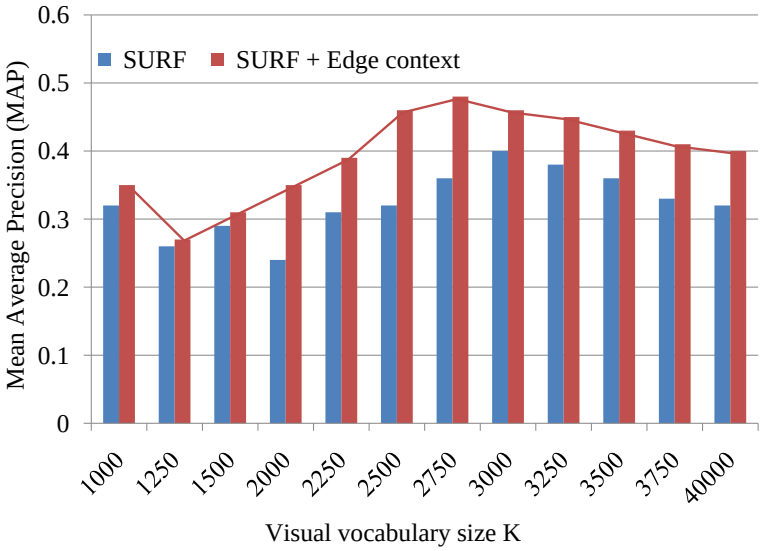


Figure 2.2 – Comparison of the precision obtained with the proposed descriptor and with SURF taken alone, with different vocabulary sizes.

Figure 2.2 shows the mean average precision (MAP) on all 101 object classes when the size of the visual vocabulary ranges from 1000 to 4000. We observe that we obtain the highest MAP value (0.483) with $K = 2750$ and with merging the local descriptors, which is over 30% of relative gain. More importantly, we see that for all different values of K, the proposed descriptor improves the accuracy over SURF.

This descriptor is a part of the PhD of Ismail El Sayad, that include several contributions forming an image indexing and retrieval model. In this model, the descriptors are quantized to form a vocabulary tree, and visual words are filtered using a multi-layer pLSA [LRH09], and weighted as described in Section 2.3.1. In the part of his work, the contribution is essentially a new descriptor, that was published at **CBMI'10** [**ESMUD10a**], and in more details in an **MTAP journal paper** [**ESMUD12**]. In the next two sections, we introduce alternative ways to build a visual vocabulary.

2.1.2 Iterative vocabulary selection based on information gain

The motivation is here to propose a simpler, faster, and more robust alternative to the standard KMeans-based quantization step step used in bag-of-visual-words approaches. We designed a method for building a visual vocabulary, that consists in iteratively selecting visual words from a random set of features, using the *tf-idf* scheme and saliency maps.

Words selection and filtering

Instead of using a clustering algorithm, the proposed method randomly selects descriptors among a large set of candidate visual words, and then iteratively filters out words in order to retain only the *best* words in the final vocabulary:

1. In the first step, a subset of candidate descriptors are randomly selected from a large and heterogeneous set (typically the whole set of descriptors from all images in a dataset) to form the initial vocabulary of visual words. All remaining descriptors are assigned to their closest visual word.
2. The second step iteratively identifies descriptors that have the highest information gain values in the candidate set: visual words with low information gain values are then discarded, and "orphan" descriptors (that were previously assigned to discarded words) are re-assigned to the closest remaining words.

The initial vocabulary is purposely large, and the process iterates until the desired vocabulary size is reached. The information gain formulation IG for this work combines two sources: the *tf-idf* weighting scheme [vR79a] and Itti's saliency maps [IKN98]:

$$IG_w = \frac{n_{wD}}{n_D} \log \frac{N}{n_w} + \frac{\sum Sal_{wD}}{n_{wD}} \tag{2.5}$$

where IG_w is the information gain value of the visual word w, n_{wD} is the frequency (i.e. the number of occurrences) of w in the dataset D, n_D is the total number of descriptors in the dataset, N is the number of images in the dataset, n_w is the number of images containing the word w, and $\sum Sal_{wD}$ is the accumulated saliency score for all the keypoints assigned to word w. Note that the standard *tf* for text is defined document-wise, and the formulation $\frac{n_{wD}}{n_D}$ in Equation 2.5 is defined collection-wise. During each iteration, an amount of words (defined by a fixed ratio of the current vocabulary size) with lowest IG_w value are discarded. This formulation means that the only words to be kept in the vocabulary have:

— either a high *tf-idf* value, i.e. large number of occurrences in the collection, only in a limited number of images,
— or a high saliency score, i.e. high saliency values for keypoints attached to the word,
— or both a high *tf-idf* value and a high saliency score.

Since the vocabulary is built from a random selection of words, it is natural to expect that several runs would produce different results. Although the experimental results proved little variations across different runs, we included in the method a stabilisation step to guarantee more stable results, that consists in combining several vocabularies. This idea is to generate k vocabularies as described above, and to combine them in a new set of words to be filtered again with the same iterative process. Experiments show that in addition to offering more stability, this process also further improves the results, since only the "best" words from k vocabularies remain in the final vocabulary. Note that the proposed method is feature-independant, and therefore it can be used with a wide variety of low-level features.

Evaluation of the iterative process

Two datasets were used in the experiments to validate our proposal: the University of Kentucky Benchmark (UKB) by Nistér and Stewénius [NS06] that contains 10,200 images, and the PASCAL Visual Object Classes challenge 2012 (VOC2012) [EVGW+] that contains 11,530 images. Since it is important in the proposed approach to pick random words from a large and heterogeneous feature set in order to give the selection process the best chance to end up with a high-quality vocabulary, another bigger dataset, MIRFLICKR-25000 dataset [HL08], was used to build the vocabulary. The main findings of the evaluation results are summarized below.

— The first interesting result, as depicted in Figure 2.3, is that when using a mere random vocabulary (i.e. step 1 only, without the *IG*-based selection process), with a vocabulary size of 2000 words, the results are very similar to those obtained with a standard KMeans-based vocabulary of equal size on UKB images. Furthermore, this is true for several types of features we tried: Color Moments, Color Moments Invariants, SIFT, OpponentSIFT [vdSGS10], and SURF. Figure 2.3 shows the results for Color Moments Invariants (CMI), OpponentSIFT (OppSift), and SIFT. However, obtaining high results with a random vocabulary is mainly explained by the specificity of UKB retrieval dataset, where one query image is used to retrieve all 4 images of a given object under different viewpoints. Therefore, when the vocabulary size is large enough, the visual words selected with the random approach show sufficient diversity to be able to precisely select only the target object, without matching with other objects.

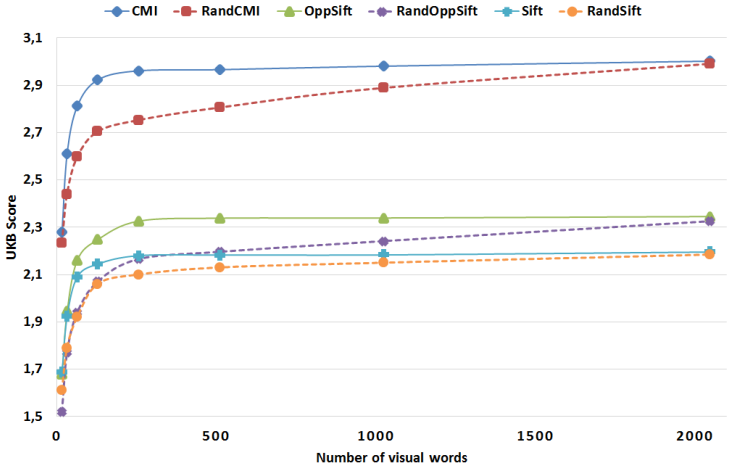


Figure 2.3 – KMeans-based vocabulary versus random selection of words.

— Another noticeable result is that the *IG*-based selection process improves the performances, and best scores with UKB are reached with a very small vocabulary, e.g. 200-300 words, as shown in Figure 2.4; these scores are higher than the scores obtained

with standard vocabularies of any size. This confirms that the iterative selection process helps selecting most useful visual words to represent the image content. With a vocabulary containing as little as 150 words, the iterative selection achieves higher scores than the KMeans-based vocabulary. Besides, despite the random nature of the method, we observed a high stability in the results after several runs.

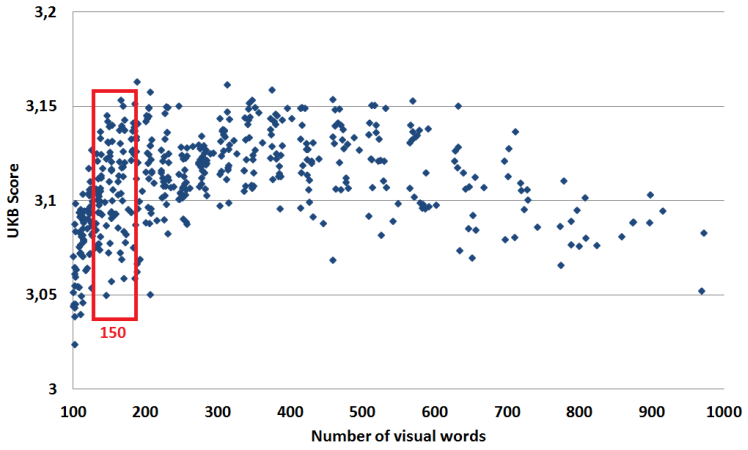


Figure 2.4 – Iterative selection scores; the initial vocabulary size is 2048.

— We found that it is important that the initial vocabulary size be large enough (e.g. over 2000 for UKB, that we see as a diversity threshold) to achieve high scores. However, above the diversity threshold, using larger initial vocabularies brings no improvement, as observed in Figure 2.5.

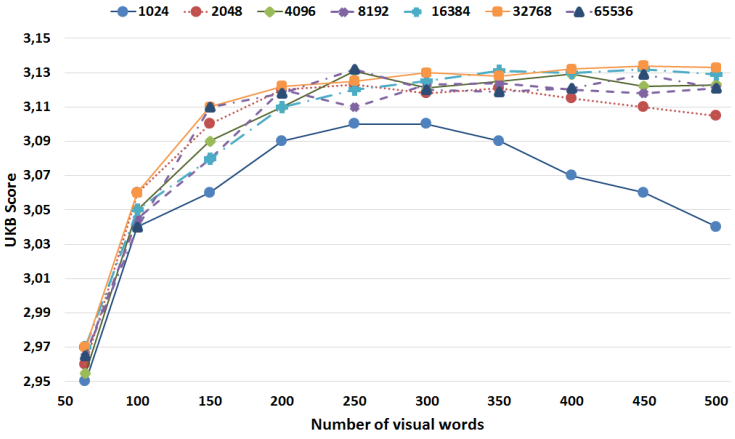


Figure 2.5 – Mean scores for several initial vocabulary sizes, from 1024 to 65536.

— Finally, the combination of several vocabularies improves the selection of the *"best"* words, and increases the scores. Empirically, $k = 3$ proved to be optimal, and $k \geq 4$ gives identical results and yet requires more processing. This improvement (about +7%) is observed with most low-level features on UKB, and the scores are higher than Fisher Vectors [PD07] or Vector of Locally Aggregated Descriptors (VLAD) [JDSP10]. Indeed, our highest score is 3.23 (with 256 visual words selected from 4096, using CMI and combining $k = 3$ vocabularies), while Jégou *et al.* [JDSP10] reported a score of 3.17 for VLAD, and 3.09 for Fisher Vectors, and 2.99 for a standard KMeans-based vocabulary (also with 256 visual words, using CMI). We also made experiments with Pascal VOC2012: the results are very similar to (however not higher than) those obtained with a standard vocabulary. Indeed, Pascal VOC2012 is a classification dataset with cluttered highly images, that is less suitable for retrieval. Nevertheless, the vocabulary construction process in the proposed approach is much lighter, and therefore the processing time is much shorter.

In further research [LUG+16], we also compared 3 other information gain models in addition to *tf-idf*: *tfc* [SB88] (the term frequency component is a normalized version of *tf-idf* that includes the differences in documents' length), *Okapi bm25* of Roberstson *et al.* [RWH00] (the weight is based on a probabilistic framework) and an *entropy* formulation by Rojas López *et al.* [LJSP07] (the weight is based on the distribution of a term in a single document as well as in the whole dataset). We observed that except *Okapi bm25*, all information gain formulations give slightly better results than the standard vocabulary (about +2% for UKB and +4% for Holidays dataset [JDS08]). The results also confirm the low performance with Pascal VOC2012 for all information gain formulations. Besides, we showed that the iterative vocabulary selection can be successfully applied in the context of visual phrases, bringing about 3% increase in the UKB score.

In addition, it should be highlighted that this method is robust since it yields high scores using a vocabulary that was built on a third party dataset, which is a good indication of versatile vocabulary. More details on the iterative vocabulary selection can be found in our **ICMR'14 paper [UGL+14]** and **SIVP journal paper [LUG+16]**. The next section introduces another alternative to standard vocabularies, in a mobile context.

2.1.3 Split representation for mobile image search

This work was done in the context of image search on mobile devices in a scenario where image descriptors are extracted on the mobile device and transferred to a remote server for the retrieval task. The increasing popularity of mobile devices leads to a growing need to adapt image representation and retrieval methods to the constraints of such devices. Indeed, despite the huge technology advances, mobile devices are rather still limited in memory, speed, energy and bandwidth. To deal with these constraints, three scenarios for mobile image search were proposed by Girod *et al.* in [GCGR11].

1. **Server-side processing:** The first one consists in transferring a compressed version of

the query image to a remote server that is in charge of extracting descriptors, retrieving the most similar images and returning results' thumbnails to the mobile device. However, highly compressed images tend to contain visual artefacts that make difficult the detection of regions of interest.

2. **Client-side processing:** The second scenario consists in performing the whole retrieval task on the mobile device. It requires the whole database index to be stored on the device. Because the available memory is limited, the size of the database is restricted, even when using memory-efficient indexing methods. Moreover, the retrieval process is likely to require more computational power than the device can provide.
3. **Hybrid method (shared processing):** The last strategy consists in extracting the descriptors on the device, and to transfer them to the server for the retrieval task – possibly after a descriptor selection/compression step. For example, the Compressed Histogram of Gradients (CHoG) descriptor [CTC+09] was designed by Chandrasekhar *et al.* to follow this strategy.

The approach proposed in this work also adopts the third strategy. We propose to limit the bandwidth use by reducing the amount of data required to describe images. During TWIRL project, José Mennesson, Pierre Tirilly and I proposed to leverage the repetitiveness (or *burstiness* [JDS09]) of visual elements in images, and build a lightweight representation of images. By taking advantage of this property, our approach represents groups of local descriptors as single representative descriptors, called *elementary blocks*, that are transmitted to the server. Assuming that images are composed of a set of elementary blocks, they can be represented using only a few well-chosen features extracted on the mobile device, and sent to the server for matching. The results are then sent back to the mobile device. This scenario is illustrated in Figure 2.6. Compared to the entire client-side processing and the entire server-side processing, this strategy is preferred to provide a good trade-off between hardware constraints (memory, speed, energy, and bandwidth) and retrieval effectiveness.

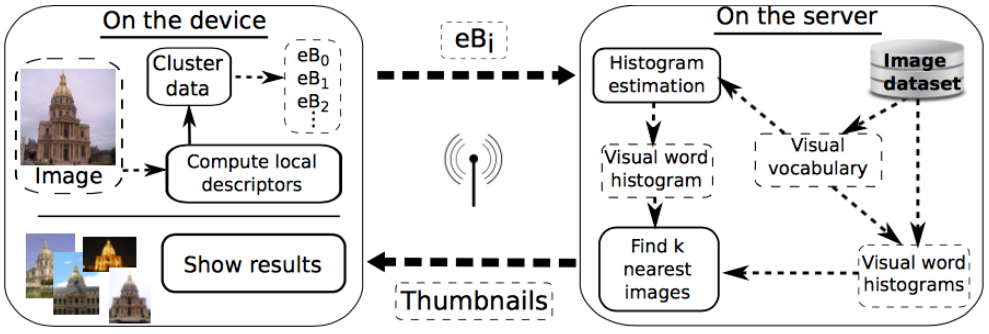


Figure 2.6 – Illustration of the mobile image search scenario.

Elementary blocks

The key idea of the presented approach consists in having two different vocabularies:
— one reference visual vocabulary (visual words), built only once, kept on the server side;
— one local vocabulary (elementary blocks) that is built at query time on the device using a single image.

From a given query image, local descriptors (e.g. SIFT) are extracted, and a quantization process determines the representative features eB_i – with a record of the cluster sizes $o(eB_i)$, that is to say the number of descriptors assigned to each cluster center eB_i. Figure 2.7 shows an example of elementary blocks built using KMeans in a query image of *Les Invalides* from the Paris dataset [PCI+08] using SIFT descriptors and $K = 20$, i.e. 20 elementary blocks. Figure 2.7-(a) shows the keypoints locations; each keypoint color corresponds to a given block. Figure 2.7-(b) presents samples of visual elements belonging to each block. Some structures of the building emerge, such as pieces of roof, columns, balustrades, etc.

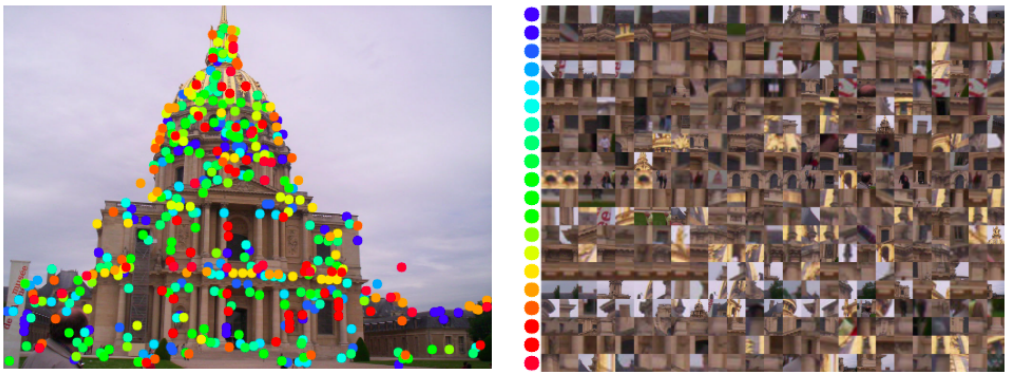


Figure 2.7 – Elementary blocks of *Les Invalides*. (a) SIFT keypoints are represented by colored dots; each color corresponds to a cluster. (b) Each row contains 25 sample patches from a given cluster. Best seen in color.

Elementary block assignment

Once the elementary blocks are extracted from the query image, they are sent to the server. It is necessary to assign them to the reference visual words, in order to allow matching them with the dataset images. Of course, there is no guarantee for an eB_i to perfectly correspond to a single visual word. A soft assignment function is used to assign an elementary block eB_i to several reference words C_j according to their distances to the descriptor, using a weighting function. Euclidean distances between each eB_i and all C_j (denoted $D(eB_i, C_j) = ||eB_i - C_j||^2$) are computed, then normalized by rescaling them between 0 and 1 based on their minimum and maximum values. A radial weighting function $w(eB_i, C_j)$ defines a weight, inspired from

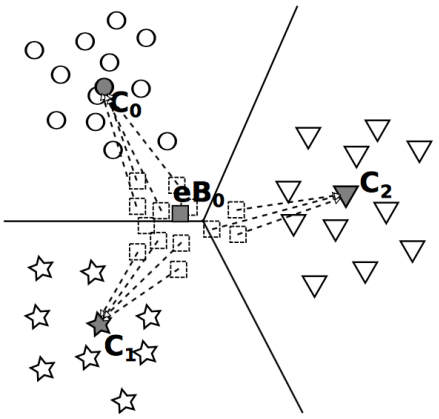


Figure 2.8 – Assignment of elementary block eB_0 to visual words C_0, C_1, and C_2.

soft assignment methods [PCI+08, GCC+11], that is assigned to each (eB_i, C_j) pair:

$$w(eB_i, C_j) = e^{-\frac{D(eB_i, C_j)}{2\sigma^2}} = e^{-\kappa \times D(eB_i, C_j)} \tag{2.6}$$

where $\sigma \in \mathbb{R}^+$ is a free parameter controlling the slope of the exponential function and $\kappa = \frac{1}{2\sigma^2}$. The parameter κ controls the softness of the assignment: high κ values will result in a steep slope for the radial function, and therefore a hard assignment to only close visual words; on the contrary, low κ values will give a moderate slope, so the assignment will be soft, and will consider visual words in a large neighbourhood. As illustrated in Figure 2.8, descriptors corresponding to eB_0 can be assigned to C_0, C_1 or C_2. In order to estimate the distribution of all $o(eB_i)$ descriptors over the visual vocabulary, these weights are divided by the sum of these quantities over the vocabulary:

$$w_n(eB_i, C_j) = \frac{w(eB_i, C_j)}{\sum_{j=0}^{N_C} w(eB_i, C_j)} \tag{2.7}$$

where w_n is the normalized weighting function and N_C is the size of the visual vocabulary. The final number of occurrences $o(C_j)$ of the visual word C_j is estimated with:

$$o(C_j) = \sum_{i=0}^{N_{eB}} w_n(eB_i, C_j) \times o(eB_i) \tag{2.8}$$

where N_{eB} is the total number of elementary blocks. This assignment process is illustrated Figure 2.9.

Evaluation of the split representation

We conducted experiments varying the vocabulary size N_C and the number of elementary blocks N_{eB}, using three datasets: Paris [PCI+08], Oxford [PCI+07], and Holidays [JDS08].

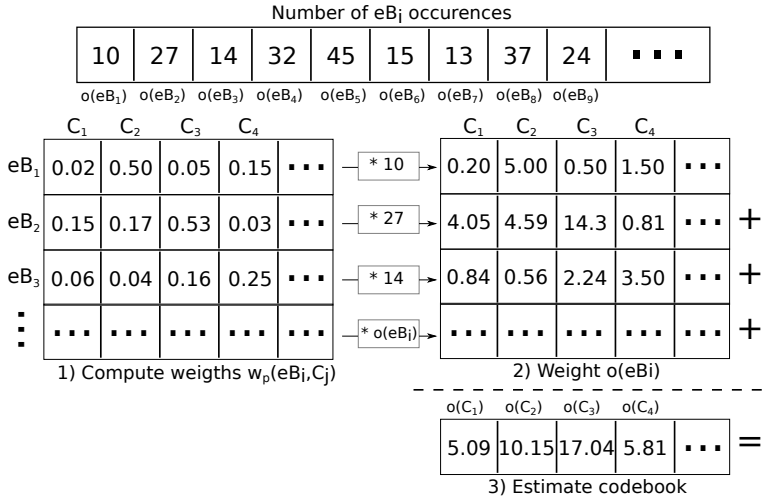


Figure 2.9 – Assignment of eB_i to C_j to estimate the associated visual word histogram.

We used a vocabulary size $N_C \in [5.000, 50.000]$, and the number of elementary blocks N_{eB} was set to a fixed ratio of the number of descriptors N_D extracted from the query image (with typically $N_D < 2.000$): $N_{eB} = \alpha \times N_D$, with $\alpha \in]0, 1]$. Since the number of keypoints per image is limited, the running time of KMeans for block computation is acceptable (< 1 second). Table 2.1 reports running times of Kmeans for various values of K (i.e. N_{eB}) and various numbers of descriptors (N_D), with a desktop computer [1].

N_D	500	1,000	2,000
$\alpha = 0.3$	0.05 s.	0.20 s.	0.58 s.
$\alpha = 0.5$	0.07 s.	0.21 s.	0.77 s.
$\alpha = 0.7$	0.08 s.	0.26 s.	0.87 s.

Table 2.1 – Running time (in seconds) of KMeans with $N_{eB} = \alpha \times N_D$.

The main experimental results are summarized in Table 2.2. They show that with $\alpha = 0,3$ (i.e. when using only one third of the descriptors extracted from the query image), we can achieve results similar to the standard vocabulary approach. For Paris and Oxford datasets, the mean Average Precision (mAP) values are even higher with the proposed approach than with the standard vocabulary in most settings. For Holidays dataset, the proposed approach did not achieve the same results as the standard approach. The main reason is that in this dataset, contrary to the two other datasets, query images require various query-specific settings for N_{eB} (number of elementary blocks) to reach similar results. Therefore, a query-specific setting of α would be more suited than using a fixed value for all queries.

1. We have not tried yet to run the system on a mobile device. We used a desktop computer with Intel core 2 CPU 2.66GHz×2, 2GB RAM.

N_C	5,000			10,000			50,000		
Database	Paris	Oxford	Holidays	Paris	Oxford	Holidays	Paris	Oxford	Holidays
Baseline	36.83	33.48	**48.41**	36.18	33.33	**44.85**	50.40	42.89	**60.10**
Random 30%	33.74	29.71	42.78	34.98	32.43	43.64	44.17	36.39	53.34
Random 50%	35.60	31.78	44.26	**36.31**	32.79	44.25	47.91	40.14	57.28
Random 70%	36.59	32.21	46.93	35.93	33.12	44.50	49.16	40.42	59.40
Ours ($\alpha = 0.3$)	**37.20**	31.25	43.16	**37.88**	**34.49**	43.93	46.86	36.75	49.99
Ours ($\alpha = 0.5$)	**37.25**	**33.69**	46.12	**37.52**	**33.98**	44.76	49.82	40.52	55.83
Ours ($\alpha = 0.7$)	**37.16**	**33.97**	46.95	**36.79**	**33.78**	44.30	**50.66**	**42.90**	58.99

Table 2.2 – mAP (in %) on Paris, Oxford and Holidays datasets for a standard approach (baseline), a vocabulary with randomly sampled descriptors, and our method using different values of N_{eB}, N_C. Results above the baseline are shown in **bold**.

We also found in the experiments that the optimal value of κ depends on the vocabulary size: smaller κ values are needed with larger vocabularies. Indeed, a high number of visual words implies a sparse space. This means that we get a larger number of histogram bins, and consequently a reduced chance to find matching eB_i in similar images. Therefore, it is necessary to distribute the query image descriptors over more reference words in order to maximize the chances of matching. In this case, such a soft assignment brings a solution to a sparse feature space.

Besides, for Paris and Oxford datasets, higher values of N_C will require higher values of α to outperform the standard approach. In other words, with a large vocabulary, a large number of elementary blocks is needed. Once again, when the space becomes sparse, a higher number of descriptors will increase the chance of matching for similar images.

Finally, Table 2.2 also shows the results of a random sampling among query image descriptors. Random sampling generally gives lower results, justifying the need to carefully select elementary blocks e.g. with the proposed approach. It is intersting to notice that here, even though the matching process is different, the behaviour with the random sampling is similar to this of the approach presented in the previous Section 2.1.2 (see in particular Figure 2.3): in both cases, the random sampling of descriptors yields results slightly lower than the baseline, and such results increase with the number of sampled descriptors until the baseline level reached.

In summary, the presented method shows that using only one third of the query image descriptors, we can achieve results comparable to the bag-of-visual-words approach for Paris and Oxford datasets. Coming back to the objective of the method, it is an aceptable trade-off between hardware constraints and retrieval effectiveness. For example, with $N_D = 1000$ and $\alpha = 0.3$, the average clustering time in our experiments was 0.20 sec. In this setting, the size of data to be transmitted to the server is 1000×512 B = 500 KB (512 bytes is the size of 1 SIFT descriptor) for the standard approach, and 500 KB $\times$ 0.3 = 150 KB for the proposed approach. For a comparison, the first strategy described earlier in this section would require to send the entire image (say 1 MB) to the server. More details on these experiments and results can be found in our **ICIP'14 paper** [MTM14].

2.1.4 Textual vs. visual words

As a last part of Section 2.1 related to visual vocabularies, we discuss here the very notion of *visual words*, and the application of text processing techniques (e.g. word filtering, weighting, etc.) to such words. Most of existing approaches for visual words are inspired from the workwork in text indexing, based on the implicit assumption that visual words can be handled the same way as text words. However, to the best of our knowledge, no work ever casted doubt on this assumption for visual words.

Zipf's law and Luhn's model

A central aspect of text IR approaches is that they rely on important characteristics of the words distribution in a natural language. Zipf's law links word frequencies in a language to their ranks, when they are taken in decreasing frequency order [Zip32]. This law stipulates that words occurrences follow a distribution model given by:

$$P_n = \frac{1}{n^a} \tag{2.9}$$

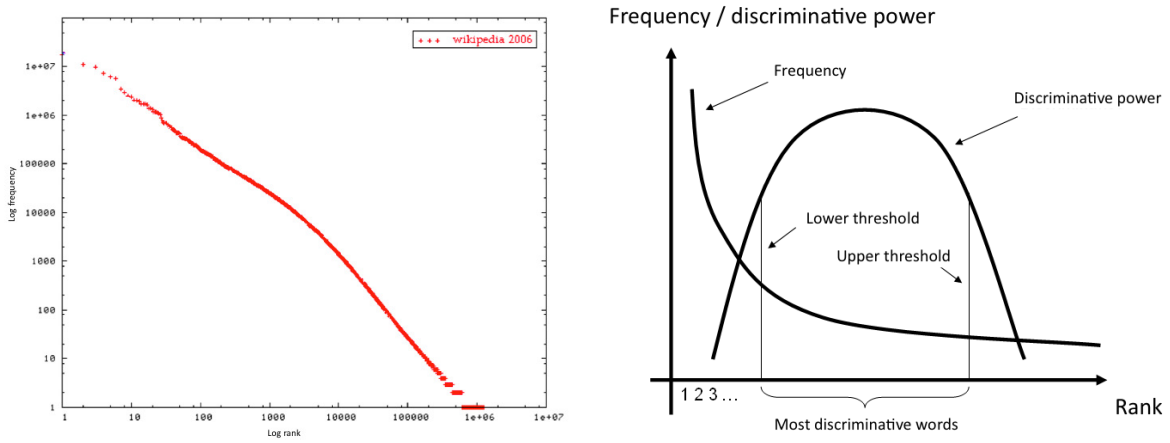
where P_n is the occurrence probability of the word at rank n, and a is a value close to 1. For instance, in large text collections in english language, the term *"the"* is generally the most frequent, representing about 7% of the total word count. The second most frequent is *"of"*, with 3,5% of the total, etc. As an illustration, Figure 2.10-(a) shows the word distribution from Wikipedia entries in November 2006 [1] in a bi-logarithmic plot, which reveals the logarithmic distribution.

This logarithmic distribution, interpreted with Shannon's information theory [Sha48], laid down the theoretical foundations of word filtering and weighting schemes. In particular, early selection techniques for significant terms were mainly grounded on hypotheses from Luhn's model [Luh58], and this model originates from Zipf's law. This model indicates a relation between the rank of a word and its *discriminative power* (or resolving power), that is to say its capacity of identifying relevant documents (notion of recall) combined with its capacity of distinguishing non-relevant documents (notion of precision). This relation is illustrated Figure 2.10-(b): less discriminative words are those with a low rank (very frequent), and also those with a high rank (very rare). The most discriminative words are those located in-between, and therefore these terms should be selected to create the indexing vocabulary.

Relations between words distribution and search performance

Interestingly, we found in our experiments that the visual words distribution highly depends on the clustering method used to build the vocabulary, and less on the descriptor. We used two image datasets: Caltech-101 [FFFP07] and Pascal VOC2012 [EVGW+]. In this work,

1. Source : Wikipedia, URL : http://en.wikipedia.org/wiki/File:Wikipedia-n-zipf.png.



(a) (b)

Figure 2.10 – (a) Word distribution from Wikipedia entries in November 2006 in a log-log plot (Source: Wikipedia). (b) Zipf's law and Luhn's model.

we compared the distributions of words from vocabularies built using SIFT and SURF, and with two different clustering methods: KMeans and Self-Organising Maps (SOM). The combination yields four vocabularies for each dataset: KMSIFT, KMSURF, SOMSIFT, SOMSURF. While the words distributions are highly similar with SIFT and SURF, the use of KMeans yields a different vocabulary from SOM, and therefore the word distributions are different. In particular, the KMeans-based vocabularies prove a rather flat distribution compared to the SOM-based vocabularies, indicating that the words occurrences are more evenly distributed, that is to say clusters are more identical in size. We observe that the words distributions for the SOM-based vocabularies are closer to the word distribution of a natural language (we considered the english language), as illustrated in Figure 2.11.

Besides, the results of a search task using the two image collections proved higher precision results with SOM, independently from the choice of descriptors, indicating that visual word distributions that are closer to natural languages tend to yield higher result when using standard text-like techniques. The complete study and experiments details can be found in our **VISAPP'14 paper** [Mar14]. This discussion about the notion of visual vocabulary concludes Section 2.1, and we focus on relations in the next section.

2.2 Relations between descriptors

The notion of **relation** discussed here takes on several shapes. First, we focus on the notion of relation at a low level, by analyzing the occurrence patterns of visual words in large datasets. The objective in this work is to statistically detect regularities, namely frequently occurring patterns, and to use them to build an intermediate image representation, between the low

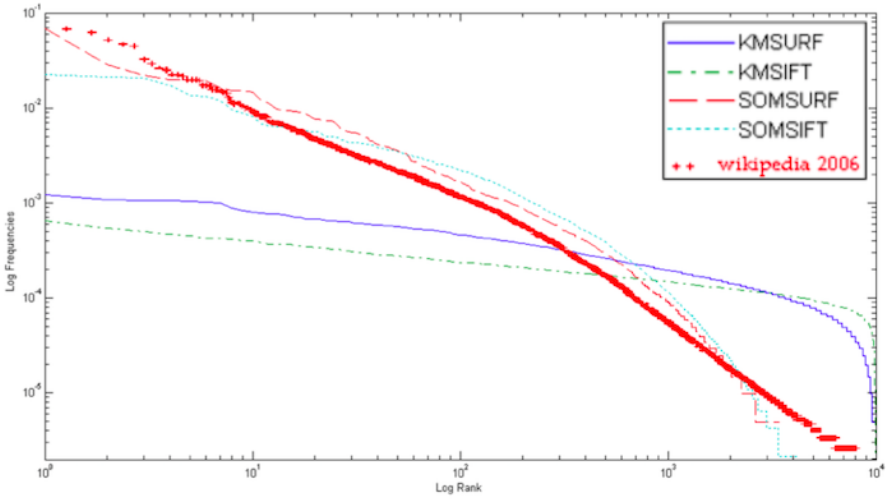


Figure 2.11 – Word distributions for Caltech-101. A similar tendency is observed with Pascal VOC2012.

level of pixels and the high level of objects. The aim of building such mid-level descriptors is to help bridge the semantic gap. This part of the work is published at **CBMI'07** [MS07c] (early work during my postdoc at NII in Tokyo), **EGC'10** [ESMUD10b] (use of visual words and visual phrases built with association rules), **VISAPP'10** [ESMU+10] (inclusion of the Edge Context descriptor and a spatial weighting scheme), **MMM'11** [ESMUD11] and **ICME'11** [ESMU+11] (extension of the work with a semantic model based on visual and high latent concepts), and finally an **MTAP journal paper** [EMUD12] (detailed description of the entire approach plus extended experiments), as contributions from the PhD of Ismail El Sayad.

When multimedia documents contain several modalities, we can apply the same kind of analysis of occurrence patterns in an inter-modal manner. The objective is to detect and use links between the image and related text (surrounding description, caption, optical character recognition or speech transcription, etc.) to generate an annotation. Our work related to cross-modal analysis were published at **ECIR'07** [MS07b] and **IJCAI-MIR'07** [MS07a] (description of the information-theoretic-based model), and **ICME'07** [SYM+07] (application of cross-modal analysis to the automatic quizz generation from TV news).

Finally, at a higher level, it is desirable for an image representation and retrieval model to be able to express the relations between objects in an image (e.g. semantic relations, spacial relations, etc.) We designed a model that combines the vector space model of IR and the knowledge representation formalism of conceptual graphs. This work is part of my PhD research, and it was published at **INFORSID'02** [MCM02], **ECIR'03** [MOCM03], and **CBMI'03** [MCMO03].

2.2.1 Visual phrases taken as mid-level descriptors

At the syntactic level, we see a correspondence between a text document and an image, where a text document is a particular arrangement of letters in a 1D space, and an image is a particular arrangement of pixels in a 2D space. In Figure 2.12, a view of the syntactic granularity of an image and a text document is shown, with *analogies* between their constituent elements. In this analogy, pixels correspond to letters, image patches to words, and group of patches to phrases. In text processing, beyond the use of (key) words in the indexing vocabu-

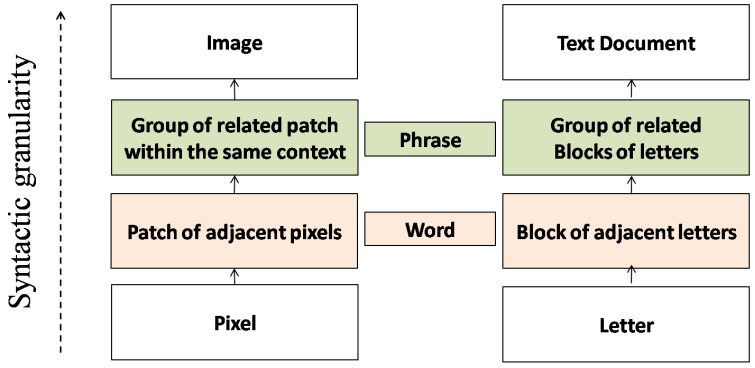


Figure 2.12 – Syntactic granularity in an image and a text document.

lary, it is beneficial to use *phrases* to refine the representation. Phrases were used e.g. to discover trends in text databases [LAS97], and for text representation and retrieval [Had03, AWH15], like Google's US patent on phrase-based indexing in an IR system [Pat09]. The objective of the proposed approach is to discover **frequently occurring patterns of features** in the dataset – called visual phrases. Such features are likely to correspond to image regions that are parts of real word objects (like buildings, cars, or faces) that often occur together in specific geometrical configurations. In the proposed approach, we use association rules to analyze the visual words occurrences together with their spacial configuration.

Association rules for mining visual words patterns

Association rules are popular in sales transactions analysis [AIS93], especially for market basket analysis. Questions such as *"if a customer purchases product A, how likely is (s)he to purchase product B?"* and *"What products will a customer buy if (s)he buys products C and D?"* are answered by association-finding algorithms. Association rules are used in this approach to spot frequently occurring local patterns of visual words. Given a vocabulary of visual words $V = \{w_1, w_2, \cdots, w_m\}$, we define a **local context** $d \subseteq V$ as a set of visual words occurring together locally in image parts. A local context can be implemented e.g. with a window sliding over images, or by considering, for a given visual word w in an image, all words located around w at a distance under a threshold. A local context $d \in D$, that we see as a basket of items in transaction analysis, contains a set of visual words, and is part of the set D of all local

contexts in the dataset. Association rules are mined from D, in order to discover implications of the form $X \Rightarrow Y$, where $X \subset V$, $Y \subset V$, and $X \cap Y = \emptyset$. The rule $X \Rightarrow Y$ holds in the set D with confidence c if $c\%$ of contexts in D that contain X also contains Y.

$$\text{confidence}(X \Rightarrow Y) = \frac{|d \in D : X \cup Y \subseteq d|}{|d \in D : X \subseteq d|}$$

The support of a rule is defined to be the fraction of all considered contexts that satisfy the union of objects in the antecedent and consequent of the rule. Hence, the rule $X \Rightarrow Y$ has support s in the set D if $s\%$ of contexts in D contain $X \cup Y$.

$$\text{support}(X \Rightarrow Y) = \frac{|d \in D : X \cup Y \subseteq d|}{|d \in D|}$$

The support quantifies how frequently a rule is applicable (its condition of application), while the confidence is a measure of the certainty of the association in this context. Given the set D, the problem of mining association rules is to discover all *strong* rules, i.e. that have support and confidence greater than some pre-defined minimum support and confidence values. Although a number of algorithms are proposed improving various aspects of association rule mining, Apriori [AS94] remains the most commonly used algorithm.

After mining D, new descriptors are generated when both the rule $X \Rightarrow Y$ and the symmetric rule $Y \Rightarrow X$ are strong, indicating a high correlation between the descriptors. The newly created descriptor is a phrase p, that is made up of the union of visual words involved in the rule: $p = X \cup Y$. The reason why we consider both rules is that considering only $X \Rightarrow Y$ is not enough to create a new phrase, since the support of Y is not taken into account. In Figure 2.13, for example, the upper representation shows a high confidence rule, and the lower representation shows a lower confidence rule: confidence$(Y \Rightarrow X)$ is low. In this example, only the support of Y is changed between the two versions. In other words, if confidence$(X \Rightarrow Y)$ is high (above the confidence threshold), it indicates that most occurrences of X take place in the context of Y. However, we need to take a look at the other occurrences of Y before deciding to merge them. Indeed, if *most occurrences* of Y happened out of the context of X, the loss of information yielded by the fusion would be more important than the gain in the representation.

Illustration of visual phrases

As an illustration of the proposed approach, we applied the approach to keyframes of TRECVID 2004 video dataset. A visual vocabulary is generated by quantizing color and texture features extracted from square image patches in a grid. Local contexts are defined with a sliding window, and association rules are mined on the set of local contexts extracted from the image dataset. Figure 2.14 shows sample images containing a phrase composed of visual words $\{2, 6\}$. This phrase was generated automatically by one of the strongest rules. We can see that this combination of visual words (with color and texture corresponding to

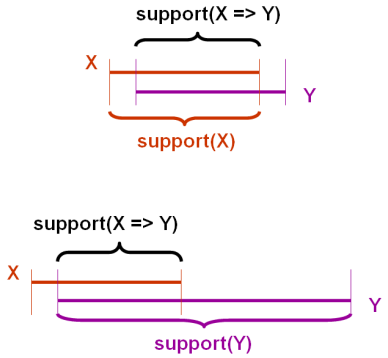


Figure 2.13 – Example of representation of two rules. The upper representation shows a high confidence rule, and the lower representation shows a lower confidence rule. Only the support of Y is changed between the two cases.

different parts of a face: mainly the mouth, the nose, the eye and the beginning of the hair) tend to be correlated in the data. By using the proposed method, this combination of visual words is handled as one single descriptor. More details about these results can be found in our **CBMI'07 paper** [MS07c].

This approach was implemented in a system where visual words are produced in a generative probabilistic framework inspired from the multi-layer pLSA by Lienhart *et al.* [LRH09]. In this system, Semantic Visual Words (SVWs) are generated by filtering classical visual words according to a relevance criterion, and Semantic Visual Phrases (SVPs) are generated by mining association rules as described above. In this approach, words and phrases produce distinct spaces, and the matching results are combined with classifier vote. Figure 2.15 shows a comparison of the classification precision for SVWs, SVPs, and their combination SVWs+SVPs. The results for the classical visual words are also displayed. For clarity purposes, the 101 classes are arranged from left to right in the figure with respect to the ascending order of results from the proposed approach. When considering only SVPs, the performance is slightly better than the setting in which only SVWs are used. We also observe that the SVWs representation is better than the setting using classical visual words over the 101 except in 5 categories, revealing the positive impact of the words filtering. Finally, the combination of SVWs and SVPs yields better results than all other representations taken separately.

Besides, we compared the retrieval results with two other phrase-based approaches:

— Yuan *et al.* [YWY07] proposed a higher-level lexicon, i.e. a visual phrase lexicon, where a visual phrase is a meaningful spatially co-occurring pattern of classical visual words. This higher-level lexicon is much less ambiguous than the lower-level one and likely to describe an object or part of the object.

— Zheng *et al.* [ZG08] proposed a visual-phrase-based approach to search objects in images. A visual phrase is defined in their approach as a pair of adjacent local image patches with a high co-occurrence count.

Figure 2.14 – Example of a visual phrase corresponding to faces, made of strongly correlated visual words. Best seen in color.

After constructing the representations for these two approaches proposed by Yuan *et al.* and Zheng *et al.*, respectively, the images are indexed with the visual words and phrases. Figure 2.16 plots the experimental results (MAP values) for the 3 systems. Here again, the 101 classes are arranged from left to right with respect to the ascending order of results from the proposed approach. We can see that the proposed approach gives higher MAP values for most classes, excepts for the 7 classes pointed with an arrow. The results given Figure 2.15 and Figure 2.16, published at **MMM'11** [ESMUD11], demonstrate the usefulness and the increase of matching precision of including in the index phrases built automatically by analyzing relations between visual words.

2.2.2 Cross-modal relations for image annotation

The idea of analyzing the relations and interactions beween visual words can be applied when dealing with several modalities. Indeed, a natural approach to automatically annotate multimedia documents is to exploit the information redundancy across modalities with an inter-modal analysis [SW05]. Here, a modality refers to an application-dependant abstraction of the raw signal, which is different from its meaning in human-computer interaction. Note that a single medium may contain several modalities. Cross-modal (also called multi-modal

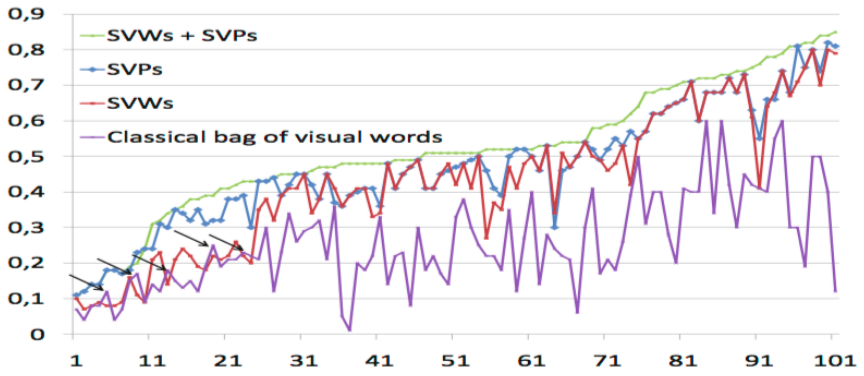


Figure 2.15 – Comparison of the classification precision for SVWs, SVPs, and their combination, and also the classical visual words.

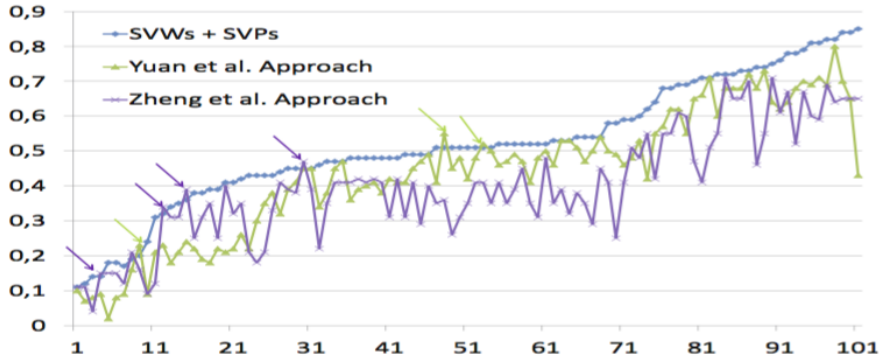


Figure 2.16 – Comparison of the classification precision for the proposed and two other state-of-the-art approaches.

or inter-modal) analysis refers to the process of combining several modalities in a synergistic (collaborative) way, in order to automatically extract semantics. In particular, when some of the modalities are textual, cross-modal analysis can be used to automatically annotate other modalities. For instance, in image and video annotation, the visual modality may contain the visual part of image and video documents, and textual modality contains any textual resource that can be used as a description of the content of the document (e.g. document name, surrounding text from a web page containing the document, open and closed captions, etc.) Applications of cross-modal analysis include automatic document – or image region – annotation, organisation of a document collection, and text illustration (finding visual documents related to a given text in order to illustrate it). Figure 2.17 shows a symbolic example of multimedia stream containing two modalities, and objects are represented in orange boxes with their relations. The relations can possibly exist between objects *inside the modality* (intra-modal relations, as discussed in the previous Section 2.2.1 about visual phrases) or *from other*

modalities, within the space-time window or even *between neighbour windows* (only the time dimension of the window is represented here).

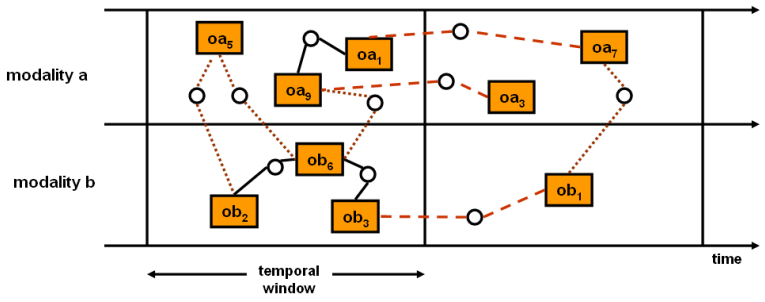


Figure 2.17 – Example of multimedia stream containing two modalities. Objects are represented with orange boxes inside the modalities, and relations between them are shown (plain line: intra-modal, same temporal window ; dashed line: intra-modal, neighbour windows ; dotted: cross-modal, same window).

By using a simple annotation model based on entropy and mutual information between modalities, we automatically extracted meaningful information from multimodal datasets containing a textual modality. The results on one image dataset and one video dataset proved higher average prediction precisions than two other annotation models: the co-occurrence model and the Naive Bayes model. The details of this model can be found in our **ECIR'07 paper** [MS07b]. Another entertaining application of cross-modal analysis is the automatic generation of quizzes from news video archives. In this application, video archives from NHK News 7 channels are used to extract images and related statements to automatically generate entertainment quizzes, as described in our **ICME'07 paper** [SYM+07].

While in this section, we focused on cross-modal relations, in the next section, we move on to relations between objects (high-level features, concepts) in a image, and we introduce a variation of the famous vector space model of information retrieval to allow expressing such relations.

2.2.3 Integrating term relations into the vector space model

Semantic-based approaches of image indexing and retrieval integrate a semantic interpretation of the image content. Simple keywords can be used to describe the main elements in an image, e.g. "bike" or "face". However, some information contained in the image cannot be expressed or modeled by keywords themselves [OP98b], such as the semantic and spatial relations between objects, or object attributes. This section describes a proposal to integrate **relations** in the image index, by combining the standard Vector Space Model of IR [Sal71, SM83] with the knowledge representation formalism of Conceptual Graphs introduced by Sowa [Sow84]. Our proposal, published in an **IPM journal paper** [MCM11], is based on the use of *star*

graphs, defined as elementary graphs made up of a single relation attached to one or more concepts representing objects in images. We show that using star graphs as index terms in the vector space model combines the efficiency, flexibility, and speed of the vector space model, and the expression power and precision of conceptual graphs.

Vector space model

The Vector Space Model of IR is a widely used model that represents documents and queries as *n-dimensional* vectors in a space where each of the n terms (or keywords) represents one dimension [Sal71]. The index of a document d_j is the vector:

$$\vec{d_j} = (w_{1,j}, w_{2,j}, \cdots, w_{n,j})$$

where $w_{i,j} \in [0,1]$ is the weight of the term t_i in the document d_j. A query is represented by a vector $\vec{q} = (w_{1,q}, w_{2,q}, \cdots, w_{n,q})$, where $w_{i,q} \in [0,1]$ is the weight of the term t_i in the query q. The weight of a term denotes its importance in the document and its ability to discriminate among documents in the collection, in order to distinguish relevant documents from irrelevant ones.

Conceptual graphs

This knowledge-representation formalism proposed by Sowa [Sow84] was used as a formal framework for image indexing and retrieval. A conceptual graph (CG) is a bipartite oriented graph, composed with concepts and relations. A concept has a *type* and a unique *referent* (label) that identifies the instance. The concept types are organized in a lattice $< T_c, \leq_c >$ where T_c is the set of concept types and $\leq_c$ is the *IsA* (generic/specific) relation between concept types (e.g. "Horse" $\leq_c$ "Animal"). Similarly, the conceptual relations are organized in a lattice $< T_r, \leq_r >$ (e.g. "standing" $\leq_r$ "position"). The advantage of representing concepts in a lattice is that the *IsA* relation between concepts is integrated into the model, grounded in the first order logic interpretation of conceptual graphs subsumption, because the generic/specific relation of concepts corresponds to a hyponym/hypernym relation of subsumption for the labels.

The CG formalism was used for image representation by Mechkour [Mec95] and for image retrieval by Ounis and Paşca [OP98b]. In such approaches, the similarity between a document and a query is determined using the *graph projection operator* that consists of a graph isomorphism. The first drawback of this operator is its exponential complexity [CM92, MC96]. This point impacts the matching processing time negatively, as pointed out in [Mec95]: in their experiments, using a dataset containing 650 images, the processing time was close to 2 minutes for a query, which does not match the usability constraints of an end-user system [Nie94, BKB00]. Ounis and Paşca proposed to shift a part of the data processing to the indexing time (i.e. it is done offline), in order to make the query processing algorithm polynomial. However, the indexing time exponentially grows with the size of the collection. As a second

drawback, a major problem of projection-based matching is that the projection does not allow to *rank* the relevant documents in decreasing order of relevance, but only to organize documents into three relevance classes:

1. total matching,
2. partial matching,
3. no matching at all.

Therefore, it is not possible to rank documents according to their estimated relevance. The objective of combining VSM and CG is to design an IR model able to both integrate relations in the document description, and to provide a flexible way to rank the relevant documents in decreasing order of relevance.

Star-graphs as image descriptors

Our proposal consists in using *star graphs* as image descriptors. A star graph is an elementary conceptual graph made up of one relation attached to one or several concepts. Star graphs are more expressive than simple keywords, and less complex (by construction) than large conceptual graphs. Hence, we see them as a satisfying trade-off between the lack of expressivity of keywords and the processing complexity of conceptual graphs.

Figure 2.18 shows an example of star graphs, taken from the description of the left image in Figure 2.19. When considering only the concepts found in the star graphs, the four labels *"Jean"*, *"Matthieu"*, *"Boat"*, and *"Sky"* would apply to both images in Figure 2.19. However, the star graphs only apply to the left image.

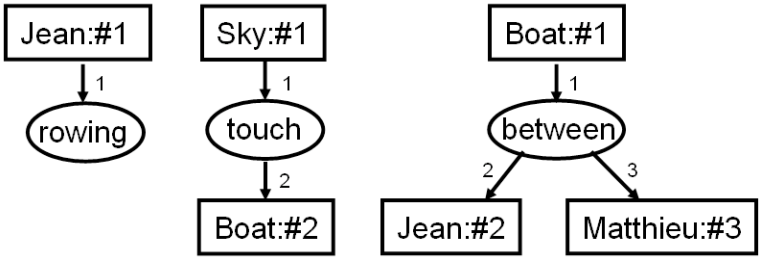


Figure 2.18 – Star graphs of unary, binary and ternary relations (from left to right).

The set of star graphs extracted from the collection of conceptual graphs has a lattice structure, composed with sub-lattices of relations sharing the same arity (number of concepts), as shown in Figure 2.20. Its partial order relation $\leq_{sg}$ is the graph projection [Sow84]. In the proposed model, star graphs serve as index terms, and documents and queries are represented by vectors of **star graphs**.

Figure 2.19 – Example of two images likely to be labeled with *"Jean"*, *"Matthieu"*, *"Boat"*, and *"Sky"*. While these four labels do not make it possible to distinguish between these images, the star graphs given Figure 2.18 help differentiating them since they describe only the left image.

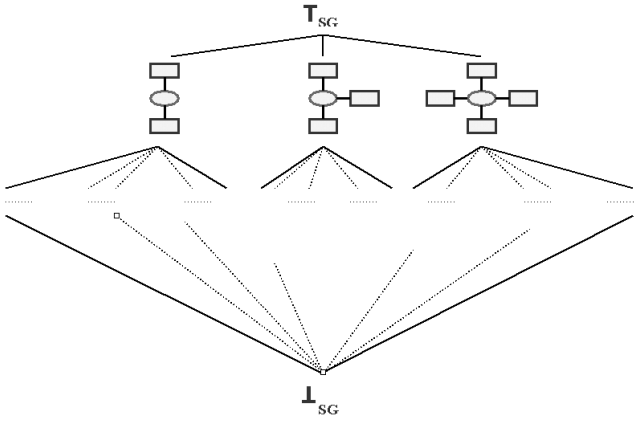


Figure 2.20 – General outline of the star graph lattice.

Automatic document expansion

When a document is indexed by a given index term, it is also implicitly described by all index terms that are *generic* to the given index term. For instance, a document that is originally described by the term *left(Boat, PineTree)* is also implicitly described by the term *left(Boat, Tree)*, assuming that the concept type lattice contains the information $PineTree \leq_c Tree$. This automatic document expansion, which is an important part of the indexing process in our model, consists in adding certain related index terms to the index of the original document. In the above example, the automatic document expansion process would add the second term, if not already there, to the index of the document.

The operation of *adding* an extra term to the index simply means changing its weight from 0 to a non-zero value. If the term that is added already belongs to the index, its weight will

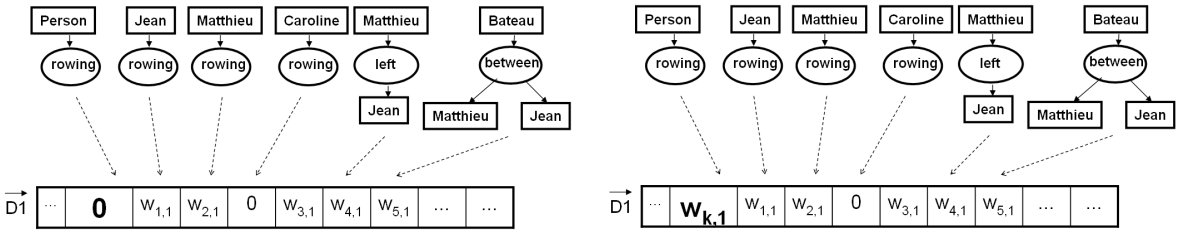


Figure 2.21 – Vector of the document before (left) and after (right) the document expansion.

still be increased. Figure 2.21 shows the index of a document before (left) and after (right) the automatic document expansion. Dimensions corresponding to terms that are generic to the original index terms are assigned non-zero values, e.g. the weight of the term *rowing(Person)* is 0 before the document expansion, and becomes $w_{k,1} > 0$ after the document expansion, assuming that *Jean*, *Matthieu* and *Caroline* are specific concept types of *Person*.

The justification behind the operation of document expansion as part of the model originates from the first order logic interpretation of conceptual graphs subsumption. Indeed, a projection of a graph g into a graph h exists if and only if $h \Rightarrow g$ [Sow84, CM92]. Since our model grounds on a lattice to represent the indexing vocabulary, the set of star graphs is not flat, and it is necessary to guarantee the consistency of the vector matching with regard to the logical implication. Hence, this pre-processing step ensures a non-null term intersection between indexes containing star graphs that can be compared using $\leq_{sg}$.

Benefits of including relations

We evaluated the benefits of using star graphs in the indexing vocabulary, compared to bare concepts; we also compared the differences in precision between a star graph vector space system and a conceptual graph total projection system. We used two image collections. The first one, which we will refer to as *Coll-1*, is composed of about 800 personal photographs. The images were manually segmented and indexed by defining polygons corresponding to real-world objects, and assigning a semantic label to them. This collection contains a wide range of holidays photographs. A set of 24 queries was designed for this collection, containing an equal number of queries with a single keyword (i.e. one unique concept, like *"boat"*), several keywords, one relation (i.e. one unique star graph, like *"water stream under leafed tree"*), and several relations.

The second collection (*Coll-2*), contains about 2300 images that were indexed automatically [Lim01]. The automatic indexing process is based on an artificial neural network, that is trained to classify 20×20 image blocks into categories such as rock, building, water, etc. Regions (or blobs) are formed by aggregating spatially connected blocks sharing the same label. For comparison purposes, we took the same query set as in the experiments presented in [ML02], in which authors used a set of 24 queries. We implemented the following systems:

1. **VSM-C:** The first system is a vector space system based on concepts only, which cor-

responds to an implementation of the standard VSM for text. Star graphs are not used in this system.

2. **VSM-SG:** The second system is a vector space system based on star graphs.
3. **DIESKAU** [1] **[ML02]:** The third system, is a CG-based system that uses the total graph projection operation to select relevant images for a query (two-class binary relevance).

Note that both VSM-C and VSM-SG can be run in the *original space* or in the *extended space*, where the operation of expansion adds some new dimensions to the vector space and also new terms to the indexes.

We compared the results of VSM-C and VSM-SG in terms of MAP and recall-precision curves to measure the increase in precision brought about by integrating relations into the VSM. The results given in Figure 2.22 show the MAP value difference between the concept-based system and the star-graph-based system for both collections in the extended space, i.e. after the document expansion. The MAP values for the graph projection (DIESKAU) are also displayed for comparison. These results indicate that for Coll-1 (Figure 2.22-above), relations significantly increase the VSM precision in the extended space, from 0.461 to 0.585, that is a 27% relative gain. This can also be seen in the recall-precision curves given in Figure 2.24-left by comparing *BEST EXTENDED STAR CONCEPTS* and *BEST EXTENDED GRAPH*.

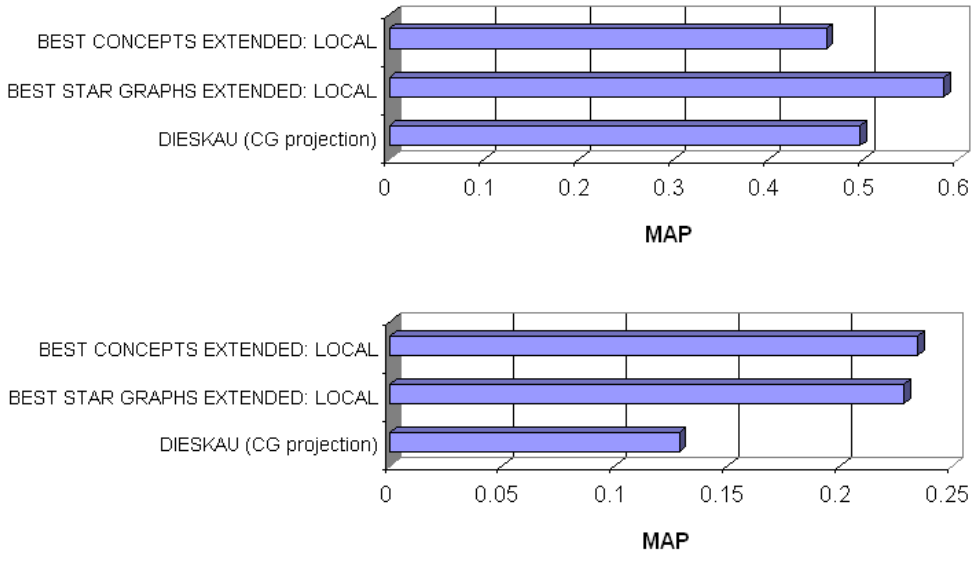


Figure 2.22 – From concepts to star graphs: impact of integrating the relations. Above: Coll-1, below: Coll-2.

For Coll-2 (Figure 2.22-below), the results show very similar MAP values (note the scale of the chart): 0.234 for VSM-C and 0.227 for VSM-SG. The recall-precision curves given

1. Digital Image rEtrieval System based on Knowledge and imAge featUres

in Figure 2.24-right (compare *BEST EXTENDED STAR GRAPH* and *BEST EXTENDED CONCEPTS*) confirm the similar results. This similarity is due to the two following reasons:

— the reasonably small size of graphs in Coll-2 documents and queries (on average 3.01 concepts/document and 1.58 concepts/query),

— the large proportion of non-relational query in this collection (half of the queries do not involve any relation).

However, if we select among all queries the subset of 12 relational queries, we observe in Figure 2.23 a noticeable increase in precision: 9% from 0.254 for VSM-C to 0.278 for VSM-SG. With this result, we also notice that the MAP for VSM-SG remains comparable to that of VSM-C when using all queries, meaning that the star graph representation is still valid for non-relational queries. The MAP value for DIESKAU with the subset of relational queries shows a minor increase in precision (0.136), compared to the MAP value for all queries (0.129).

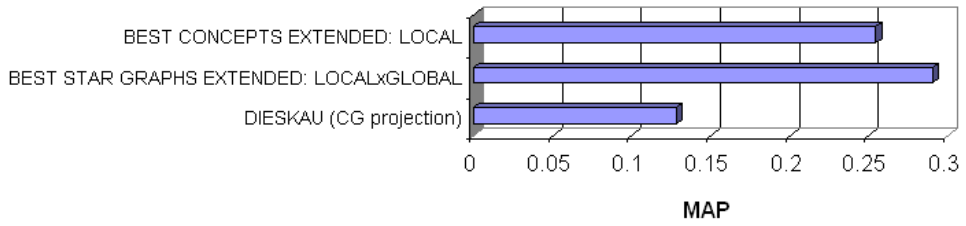


Figure 2.23 – Result for the subset of 12 relational queries of Coll-2.

Comparison to graph projection

We further discuss here the differences in search precision between the star graph approach (VSM-CG) and the conceptual graph approach (DIESKAU). Figure 2.24 shows the recall-precision curves from these systems for both collections. The curve for Coll-1 shows that for small recall values, the graph projection system (DIESKAU) gives a higher precision than the star-graph-based system. Indeed, DIESKAU is precision-oriented since it uses whole conceptual graphs as document indexes, and not just small parts like star graphs. Also, because the total graph projection is used to select relevant images, the system finds fully matching documents accurately. The star graph system, on the contrary, does not keep the **joint information** between star graphs. For instance, the joint information on $[Tree : \#1]$ (type $Tree$, referent $\#1$) in the star graphs:

$$[Tree : \#1] \rightarrow (left) \rightarrow [Building : \#2]$$

and

$$[Tree : \#1] \rightarrow (above) \rightarrow [Car : \#3]$$

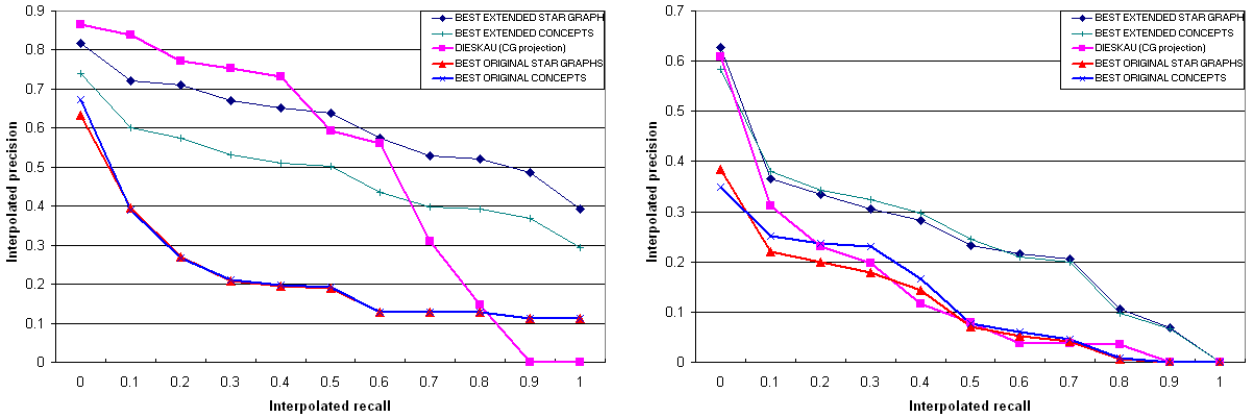


Figure 2.24 – Recall-precision curves for best-performing settings of each system (Left: Coll-1, Right: Coll-2).

is lost in the vector space representation. Indeed, these two star graphs are part of the index, which could describe both *"a tree located to the left of a building and above a car"* (the original description contained in the whole graph with the joint information) and *"a tree located to the left of a building, another tree located above a car"*. Hence, documents containing both index terms are retrieved by the star graph system regardless of this joint information. Therefore, the star graph system is not able in this situation to distinguish fully relevant documents from partially relevant documents, which is a limitation of the model.

For recall values over 0.5, DIESKAU shows a drop in precision, and the precision of the star-graph-based system remains reasonably high. This drop is a consequence of the rigidity of the projection system: while both the concept-based system and the star-graph-based system retrieve *partially* relevant documents, the total projection-based system misses these documents. In the example given above, DIESKAU would miss documents containing only one of these star graphs, while the vector space systems would retrieve them. From this point of view, the vector space system is more flexible than the graph projection system. This result confirms that the graph projection system is *precision-oriented*, while the vector space system is *recall-oriented*. This lack of flexibility is more explicit for Coll-2, where the curves show that DIESKAU's results are *always* lower than the star graph system.

Finally, when we compare the curves in the *ORIGINAL* space and the *EXTENDED* space given in Figure 2.24, we see the large increase in precision brought by the operation of expansion for both collections – with a larger difference in precision between VSM-C and VSM-SG for Coll-1 than for Coll-2. This confirms the usefulness of the expansion process.

In this section related to *relations*, we discussed this notion at low, cross-modal, and high levels. In the next section, we move to the question of weighting image parts according to their importance.

2.3 Weighting models for image parts

Because all index terms in a document do not describe it equally, it is useful to assign them numerical weights according to their importance. Popular weighting schemes for text retrieval are based on variations of *tf-idf* (*term frequency* $\times$ *inverse document frequency*) [Sal71, SWY75, SM83, SB88, BYRN99]. The *tf* value measures the local importance of a term in a document, and is usually defined according to the (normalized) number of occurrences of the term in the document. The *idf* value measures the discriminance of the term (also called *resolving power* [vR79b]), i.e. the ability of the term to distinguish between the documents of the collection.

Notion of *importance* in images

While *tf-idf* works well text, when it comes to images, it is less intuitive to design a weighting scheme. Indeed, what is *important* in an image? We attempted to address this (apparently naive) question from several sides. Our objective is to assign a *level of interest* or *semantic importance* to image objects (IOs), from a user perceptual point of view. For this purpose, we discuss the *aboutness* of a document considering a given region, that is to say the *importance* of the region regarding the document. Our definition of level of interest is related to the notion of document content in a much similar way to the standard notion of local term importance in the context of text retrieval. A weight is assigned to a region in order to evaluate to what extent the image containing the region is similar to another image containing a related region, and also to evaluate how relevant the document is, regarding a user query about the region content.

Considering an image I, and io an image object of I, io is labeled by a term t. The **importance** of an io regarding t is directly related to the **aboutness** of I considering t. For example, one will consider that an IO representing a boat (hence $t =$ "boat") in a given image is important if most users would consider that this image is **about** a "boat". In our opinion, **aboutness** is a notion mainly local to a given document. It might not be confused with **relevance**, a notion that usually compares a particular document to the whole corpus. The notion of importance presented here is local to images and thus cannot be directly assimilated to relevance evaluation. A relation between these two notions is somewhat illustrated by the *tf-idf* weighting model: given a term t of document D, *tf* of t stands for the **aboutness** of D considering t, while *tf-idf* is an estimate of the **relevance** of D considering t. The notion of occurrence – a valuable weighting element of index terms for textual documents – is much less intuitive in the case of multimedia documents. Images are two-dimensional data, and in this context one may postulate that the relevance of an image containing e.g. some boats is not only related to the fact that it represents one or more distinct boats. Other two-dimensional perceptive factors are possibly more important in that matter, such as for example the overall scene organisation of boats relatively to the background in the scene.

This section describes our contributions to design weighting methods for image, that are

based on several sources information:

1. **A spatial weighting scheme** for the visual words introduced in Section 2.1.1, that takes into account the spatial distribution of words over the image, as described in our **CBMI'10 paper** [ESMUD10a].
2. **A gaze-based weighting scheme** for regions, that is based on users' perception, published at **VISAPP'13** [Mar13].
3. **A geometrical weighting model** for image objects, that is based on geometrical properties of image objects. This model was first introduced at **CIKM'05** [MCM05], and an extended generalized version can be found in our **MTAP journal paper** [MSCM08].
4. **A weighting model** for star graphs described in Section 2.2.3, inspired from *tf-idf*, published at **INFORSID'02** [MCM02], **ECIR'03** [MOCM03], and **CBMI'03** [MCMO03].

The four models follow an increasing *granularity* (referring to the scale of items to be weighted): visual words, image regions, image objects, and star graphs.

2.3.1 Spatial weighting scheme for visual words

While popular weighting schemes for text are based on word occurrences, the bare number of visual words occurrences is not always representative of the importance. As an example, if a given visual word w is typical among *"forest images"*, an image containing 100 occurrences of w is not necessarily more **about** a *"forest"* image than an image containing only 25 occurrences of w. However, a classification system trained with the first image can be mislead by the high count and is likely to misclassify the second image as a *"non-forest"* image. We propose a weighting scheme that weights the visual words according to the spatial constitution of an image content rather than just the number of occurrences. In this section, we introduce a weighting model that uses the same GMM representation of points in the 5D color-spatial feature space as defined in Section 2.1.1. The model is inspired and adapted from [CHS09].

Formulation

The proposed spatial weighting is designed locally to an image for a visual word w. It is based on two terms: the first term denotes the average importance of the word w in the Gaussian cluster g_i, and the second term estimates how spread out is the distribution of w over the clusters. Suppose that in an image, there are local descriptors $\{f_1, f_2, \ldots, f_{n_i}\}$ obtained from a set of n_i interest points that belong to a given Gaussian cluster g_i, and that are assigned to a visual word w. The sum of the probabilities of the point occurrences indicates the contribution of visual word w to the Gaussian cluster g_i. Therefore, the weighted term frequency $tf_w^{g_i}$ of the visual word w with respect to g_i is defined as follows:

$$tf_w^{g_i} = \sum_{j=1}^{n_i} P(g_i|f_j) \tag{2.10}$$

The average weighted term frequency tf_w of the visual word w in an image where w occurs in n_w Gaussian clusters is defined as follows:

$$tf_w = \frac{1}{n_w} \sum_{i=1}^{n_w} tf_w^{g_i} \tag{2.11}$$

The weighted inverse Gaussian frequency of w with respect to an image with n clusters is defined as follows:

$$igf_w = ln \frac{n}{n_w} \tag{2.12}$$

Note that while the standard *tf-idf* for text is defined collection-wise, our tf_w and igf_w are defined document-wise. Finally, the spatial weight W_w of the visual word w is defined by the following formula:

$$W_w = tf_w \times igf_w \tag{2.13}$$

Evaluation of the spatial weighting scheme

The proposed spacial weighting method was evaluated using Caltech-101 dataset. We compare the proposed approach to the original bag-of-visual-words approach of Sivic and Zisserman [SZ03]. The 101 classes are ordered with respect to the ascending order of their average retrieval precision with the proposed weighting scheme in order to get a clear representation, as displayed in Figure 2.25. We can see that the spatial weighting scheme globally outperforms the standard approach except for 3 image classes out of 101 one in the used data set. The three classes are *dolphin*, *pizza*, and *stegosaurus*. The MAP of the standard approach is 0.302, which is below the proposed approach (0.483). This large increase (almost +60%) on a data set containing 101 classes validates the good performance of the proposed weighting scheme. More details about this approach and the experiments can be found in our **CBMI'10 paper** [**ESMUD10a**].

2.3.2 Gaze-based region importance

The possibility of taking advantage of the information conveyed in gaze opened many research directions, namely in image compression – where users' gaze can be used to set variable compression ratios at different places in an image, image understanding – where gaze data helps the segmentation process [RKS+10], in marketing – for detecting products of interest for customers, civil security – for detecting drowsiness or lack of concentration of persons operating machinery such as motor vehicles or air traffic control systems, and in human-computer interaction. In the latter for instance, users' gaze can serve as a complementary input device to traditional ones such as a keyboard and mouse, namely for disabled users.

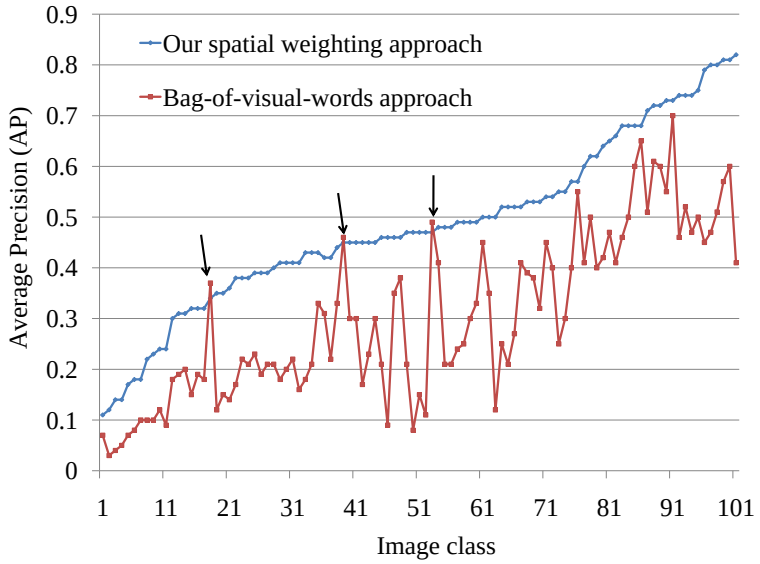


Figure 2.25 – Comparison between the spatial weighting approach performance with the original bag of visual words.

An alternative method for weighting image parts is to rely on users' perception of important regions, by taking advantage of users' gaze, based on the hypothesis that most *salient* areas correspond to the most important image parts. While the weighting model in the previous section was dedicated to visual words, we focus in this section on image regions. When enough gaze data is available, it can be used for selecting important regions and determining their importance. In this work, we propose to use gaze information as an input for designing a gaze-driven region selection and weighting method, as described in our **VISAPP'13 paper** [Mar13]. This weighting model is to be included in an image search system. Originally, this model was designed and used during the ANR [1]-funded project ANAFIX [2] to assess the quality of advertising videos, as illustrated in our study published in [MLLD09].

Visual attention, fixation, and saccades

The proposed weighting method is based on visual attention. Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things. For images and videos, the visual attention is at the core of the visual perception, because it drives the gaze to salient points in the scene. During visual perception, human eyes move and successively fixate the most informative parts of the image [Yar67].

1. Agence Nationale de la Recherche.
2. ANAlyse des FIXations et orientations du regard dans le cadre de l'affichage dynamique de films et images publicitaires, partners: Ouest Audiovisuel, LIFL (Lille 1 University), URECA (Lille 3 University). ANR-06-RIAM-0026, Dec 2006 – Mar 2009.

Gaze trackers provide the horizontal and vertical coordinates of the point of regard relative to the display device. Thus, it is easy to obtain a sequence of points corresponding to sampled positions of the eye direction. These points correspond to triplets of the form (x, y, t) and reflect the scan path. The scan path consists of a sequence of eye fixations (locations where the gaze is kept still, yielding regions with an important density of sample points), separated by eye saccades (fast movement, yielding large spaces with only few isolated points).

After combining gaze data from several persons by merging all their fixations into a single set, a clustering process allows reducing the spatial characteristics of fixations into a limited subset of clusters K_i, which define the important regions, as illustrated in Figure 2.26. The clustering can be achieved by unsupervised techniques, such as KMeans, because the spatial distribution of points from all scan paths is unknown, and their number is finite. Once a set of important regions is identified for an image and the corresponding objects are labeled, the labels can be given a weight according to their estimated importance in the image, with the final objective of integrating the weights into an image search system.

Region importance

We defined an importance measure for clusters, that can be interpreted in the same way as a *tf* measure of a word in a text document. Inspired from [NCS06], this measure is based on the following clusters' characteristics:

— cardinal – number of points in the cluster,
— surface – surface of its convex hull,
— variance – variance of gaze points from the cluster center,
— time-weighted visit count – number of points weighted by the visit time (gives higher values for first visited clusters), and
— revisit count – number of saccade revisits to the cluster during the scan path.

The region importance is calculated as a combination of the five criteria, and can be used to select important regions (by discarding regions with no gaze points, or regions whose importance value is below a threshold) and to weight such regions with the estimated importance value. Of course, since it is costly to gather user gaze data, a straightforward solution to do without such data is to use visual attention prediction models, such as Itti's model [IKN98, IK99].

2.3.3 Geometry-based weighting model for image objects

This section describes a higher-level weighting model that is designed for Image Objects (IOs), that consist of abstractions of physical objects represented in images. They are specific instances of Media Objects, defined in our **MTAP journal paper** [**MSCM08**] as semantic entities (from text, image, video, audio) from a multimedia document, inside given spatial-temporal bounds. More precisely, an Image Object is semantic visual entity defined as a 2D projection of one or several physical objects from the real world, made of an aggregation of non necessarily connected regions. Therefore, the granularity is higher than in the previous section

Figure 2.26 – Example of images displaying gaze points (yellow), and showing the processed clusters as superimposed disks (pink), with their estimated importance denoted by the size of the disks. Best seen in color.

that considers arbitrary image regions. A formal definition of IOs is given in our **IPM journal paper** [MCM11]. Image Objects provide a flexible model for describing image content, since an IO may correspond to several non-connected regions. Hence, they provide a solution for handling groups of objects in images. For instance, they can be used to describe a group of boats in a harbor, a group of trees in a forest, or a group of people in a crowd, with a certain level of granularity. This flexibility also makes it possible to take into account the common situation in which objects are partially occluded (e.g. a car behind a tree).

The weighting model discussed here is based on hypotheses that were experimentally validated. In a first step to design the proposed model, we identified several geometrical criteria related to the semantic importance of image objects from a user point of view. According to the identified criteria, we made corresponding hypotheses, which were validated with users. The valid hypotheses drove the definition of an importance model for image objects. We describe below the criteria and hypotheses about the overall scene organisation related to the importance of IOs.

Hypotheses statment and validation

Basically, an IO is a set of pixels whose basic low-level and geometrical characteristics (color, texture, area, position, etc.) can be easily computed. We propose to investigate the following criteria:

— **Size:** The importance of an IO varies in the same way as its size S.

— **Position:** The importance of an IO is maximal when its position P is at the center of the scene, and decreases when its distance from the scene center increases.

— **Fragmentation:** The fragmentation refers to situations where the IO corresponds to several non-connected image regions, either because of occlusion or when the IO includes

several instances of a given physical object [1]. The importance of an IO is maximal when it is not fragmented – i.e. the IO corresponds to a single image region, and decreases when its fragmentation F increases.

— **Homogeneity:** The importance of an IO varies in the same way as the homogeneity H of its embedding image. Contrary to the other criteria, this criterion independant from a given IO under consideration, and is defined relatively to the entire image.

Of course, we do not claim that these criteria are the only ones related to an effective definition of the importance of an IO. We propose here a step for future evaluations of geometrical features of image objects on their importance. For semantic image similarity matching, defining and validating such criteria is a first important step toward a well-founded weighting model for image objects.

For the purpose of validating the hypotheses by confronting them to actual human perception of images, we designed an image test set, containing several types and configurations of image objects. Relevant configurations were determined according to the 4 hypotheses, considering 2 qualitative absolute values for each of them:

— **size**: big/small,
— **position**: center/lateral,
— **fragmentation**: aggregated/fragmented,
— **homogeneity**: homogeneous/heterogeneous.

Considering all different combinations of these values, there are $2^4 = 16$ configurations as shown in Figure 2.27, where the IO is the disk – or group of disks.

Figure 2.28 shows an example of two photos of a *boat* in different configurations: on the left, the boat is small, non-fragmented, in the center of a homogeneous image; on the right, the boat is small, non-fragmented, on the side of an heterogeneous image. In order to prevent the results from being biased by the type of physical object, and also in order to check whether the hypotheses hold across different types, we used 3 object categories: boats, birds, and faces (children faces in general, and one specific face). The evaluation was carried out with 30 participants (14 men and 16 women) aged from 24 to 50. Participants are shown sequences of image pairs displaying a target object in different configurations, and they are given the specific task, for each image pair, to select the image out of both that is the most representative of (i.e. *about*) the object.

A statistical analysis of the collected data shown that among the 4 considered hypotheses, only those about **size**, **position**, and **homogeneity** are valid; the hypothesis about **fragmentation** is not valid. Moreover, the combination of valid criteria makes them stronger (for example, a big object is even more important when located in the center), and their combination with the fragmentation only has very little impact. The details of the analysis can be found in our **MTAP journal paper** [MSCM08]. A weighting model was designed according

1. For example: several *"trees"*, several *"persons"*, etc. Note that the question of annotating several trees with e.g. *"wood"* or *"forest"*, and several persons with *"group"* or *"crowd"* is left to the indexing vocabulary, and is out of the scope of the proposed weighting model.

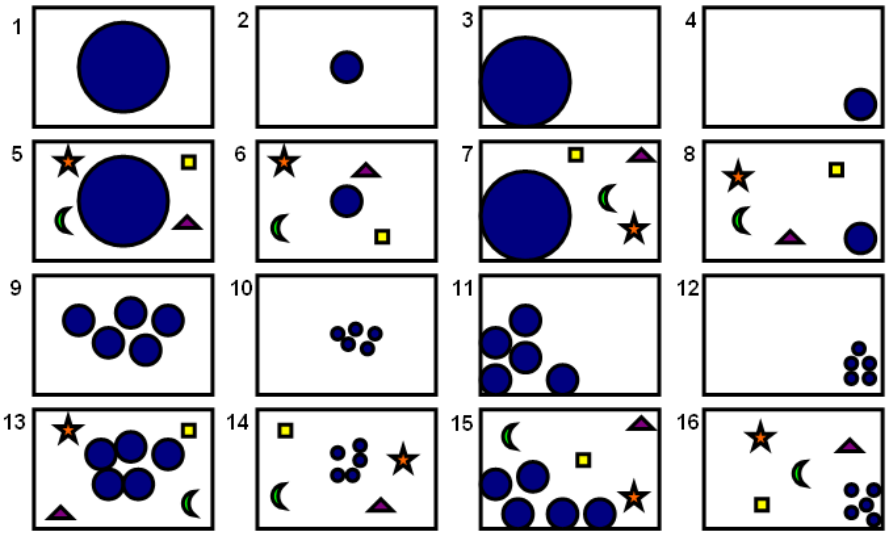


Figure 2.27 – 16 typical configurations for an IO represented by the disk.

Figure 2.28 – Example of photos showing configurations 2 (left) and 8 (right) of Figure 2.27.

to the findings of the study, in order to reflect the users' preferences about the importance of objects.

Evaluation of the geometry-based weighting model

In this experiment, we evaluate the retrieval accuracy of the weighting model designed based on the valid hypotheses. This model gives more importance to big objects located in the center of homogeneous images (not cluttered) images. Using a manually-indexed dataset containing about 800 images [1] and a set of 20 image queries, we assess the ability of our model to retrieve and rank images according to the importance of objects. The model is implemented

1. This dataset is the same as the one used in Section 2.2.3, when integrating relations into the vector space model.

in a vector space system in which documents are represented with vectors in a label space, where each coordinate denotes the importance of the corresponding label. We compare 3 weighting settings:

1. a Boolean weighting scheme,
2. our proposed weighting model (referred to as *Importance*), and
3. the *Importance* model combined with the standard *idf* component.

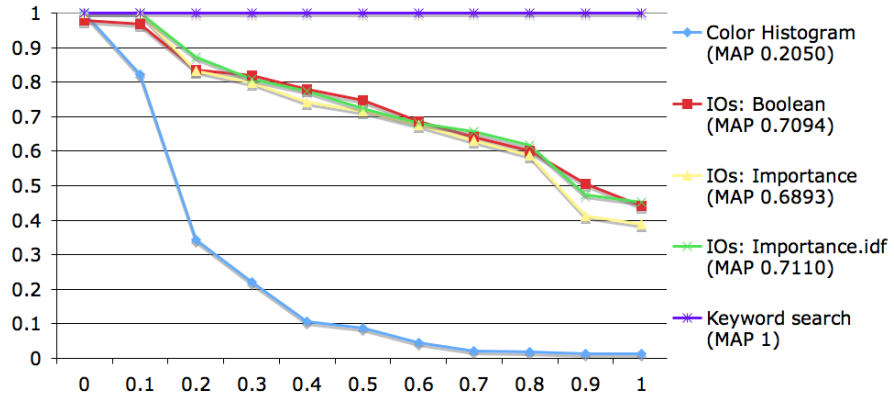


Figure 2.29 – Recall-precision graph for the global color histogram matching, the IO matching with a Boolean weighting scheme, and the IO matching with the proposed weighting scheme – with and without the *idf* component. The graph also displays the result of a text-based keyword search.

We also include the retrieval results of a simple color histogram search system. The results are given in Figure 2.29. The result of the color histogram matching is very high for a low recall values because the first result is the query image itself, and then the following results are often some photographs of the same scene, taken from a slightly different angle. Then the precision drops drastically and is very low from about 20% of recall. This expectable result is explained by the fact that visually similar images (with similar color histograms) are not necessarily semantically similar. The color histogram system can only retrieve results that are visually similar, even though their semantic content might be very different from the query.

When comparing the 3 weighting settings for the vector space system in Figure 2.29, we notice that they yield quite similar performances (the differences are not statistically significant), with mean average precisions of 0.7094, 0.6893, and 0.7110 for *Boolean*, *Importance*, and *Importance* $\times$ *idf* settings respectively.

This can be explained by the fact that, since the precision is very high for the 3 systems, the results contain a large part of relevant documents, and therefore the differences consist mainly of **minor ranking variations among many relevant documents and few non-relevant documents**, which can neither be captured nor differentiated correctly with a recall-precision plotting. In order to highlight this point, we included in Figure 2.29 the expectable result of a

text query-based search, where queries consist of single keywords corresponding to the label of the target objects. Since the images were carefully indexed manually, the result for this search contains *all relevant images*, and *only relevant images*, yielding a MAP value of 1. Moreover, this result remains the same whatever weighting scheme is used. The reason for this is that the standard recall-precision measures evaluate systems based on a ground truth about the *binary relevance of documents*. For this reason, it is necessary to use another metric to evaluate the quality of the ranking **among relevant images**, that would take into account not only the *binary relevance of documents*, but also their rank in the ground truth, when available.

We compare our system ranking of the images in decreasing order of importance to an assessors' ranking in decreasing order of aboutness, with the same set of 20 queries in a text format, consisting of one single keyword corresponding to the label of the target object (similarly to the text query-based search reported above). The system ranking for each query is defined as the sequence of relevant images sorted in decreasing order of importance values for the keyword. The ideal user ranking was generated by a group of 4 assessors [1] – 2 women and 2 men aged from 23 to 25 who have a good knowledge of the collection – and who carefully ranked relevant images for each query (on average, 6 relevant images were ranked for each query). An average ranking is generated by merging the individual rankings of the 4 assessors, as suggested by Ounis and Paşca [OP98a], which is used as a ground truth to which the system ranking is compared. The comparison to the ideal ranking indicates how close to the users' perception of aboutness our weighting model is. This comparison is based on the following divergence function, inspired from [OP98a] that gives low divergence values to similar rankings, while penalizing more system ranking errors at the top ranked images:

$$div(U,S) = \frac{1}{mdv_n} \sum_{i=1}^{n} \frac{rk(U,i) - rk(S,i)}{rk(U,i)}$$

where U and S are the user ranking and the system ranking of n images respectively, $rk(U,i)$ and $rk(S,i)$ are the ranks of image i in the user ranking and in the system ranking respectively; mdv_n is the maximum divergence value for n items [2]. This divergence function gives the value 0 when $U = S$, and it gives the value 1 when U and S are in reverse order. For instance, if we obtain u=[5,8,1,4,6,3,2,7] and s=[5,8,3,2,4,6,1,7] for a given query, meaning that the image 1 was ranked at the 3^{rd} position by the assessors, while it is ranked at the 7^{th} position by the system, then the divergence value for u and s is:

$$div(u,s) = \frac{1}{76.15} \cdot (\frac{16}{7} + \frac{9}{4} + \frac{9}{3} + \frac{1}{5} + \frac{0}{1} + \frac{1}{6} + \frac{0}{8} + \frac{0}{2})$$

This value corresponds to a ratio of the ranking quality of a system over the worst ranking, which is the reverse order as compared to the ideal ranking. Note that a Boolean system not implementing any weighting scheme is likely to provide an arbitrary (random) ranking,

1. These 4 assessors are different from the ones who participated in the hypotheses validation.
2. The maximum divergence value is reached when the two rankings are in a reverse order.

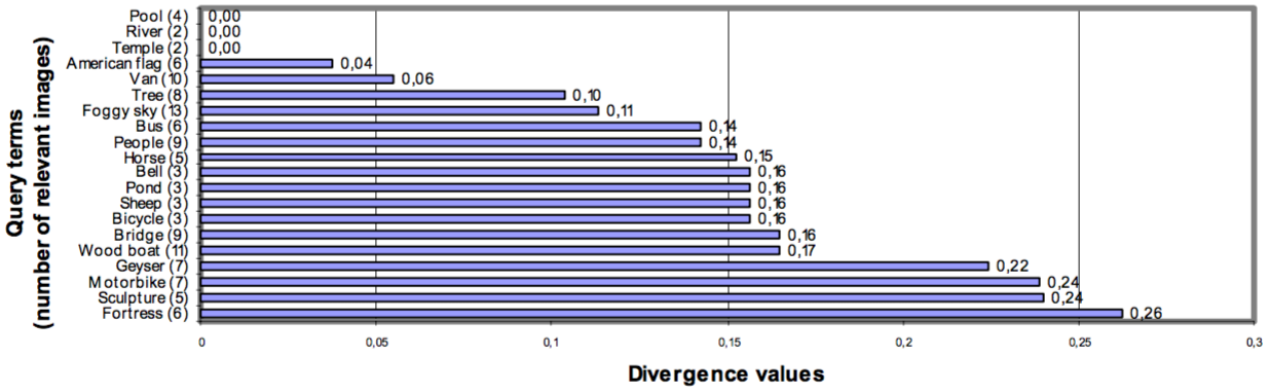


Figure 2.30 – Divergence values for all query terms.

possibly related to the order in which relevant images are found by the algorithm in a list, and the order may simply depend on the image file names in the system. Hence, this experiment provides a comparison between our weighting system and a Boolean system not implementing any weighting scheme, that would return a reverse order compared to the ideal ranking.

We use this divergence function in addition to the classical recall-precision measures because we are specifically interested in the order of relevant images, rather than in the usual binary relevance of images. Indeed, as discussed earlier, recall-precision measures do not make it possible to differentiate two systems retrieving all relevant documents and only relevant documents, **in different orders**. For instance, if the result of one system (that is to say the sorted list of images) is s=[5,8,3,2,4,6,1,7], and the result of another system is s'=[5,8,1,4,6,3,2,7] (which is actually the same as the ideal ranking u provided by the assessors), then the precision values for both systems are 1 – see the keyword search result in Figure 2.29, since both systems retrieve *all relevant images, and only relevant images*. The example shows that in this specific experiment, the recall-precision measure is not able to point out that the ranking s' is better than the ranking s, and that, therefore, the second system performs better than the first one.

Note that the size of the collection and the number of relevant images per query are purposely kept small in our experiments so that images can be precisely annotated manually, and so that assessors can accurately rank images according to their relevance to queries. The system implements both each criterion individually and their combination. The divergence value (DV) averaged over the 20 queries is 0.24 when considering the surface only, 0.31 for the position, and 0.63 for the homogeneity. The latter DV is high because relevant images are sorted according to their homogeneity values, which do not take into account the query term. When combined together, the 3 criteria perform better that separately, as shown in Figure 2.30: DVs range from 0 (for *"Pool"*, *"River"* and *"Temple"*) to 0.26 (for *"Fortress"*), with an average of 0.13. In Figure 2.30, the number of relevant images for each query is shown

between parentheses. The results we obtained depend neither on the number of relevant images for the queries, nor on the type of objects considered (natural/non-natural, generic/specific or living/non-living). The high DVs (greater than 0.20) correspond to the queries *"Geyser"*, *"Motorbike"*, *"Sculpture"*, and *"Fortress"*. For these query results, some of the relevant images are very similar with respect to their visual configuration, despite aesthetic-level variations. Hence, the system is unable to discriminate between these images since the importance values, based on the visual configuration of images, are close to one another. However, the low average DV of 0.13 confirms that our system ranking is very close to the users' perception of aboutness. This experiment provides a validation to both our criteria modeling and their combination.

2.3.4 Weighting scheme for star graphs

As a last part related to weighting schemes for images, this section describes a weighting scheme designed for star graphs (star graphs are defined Section 2.2.3). The specificity is that they involve several image objects in a single descriptor – not just a single object. Besides, the document expansion that comes with the model has to be taken into account in the weighting scheme. The proposed method is once again inspired from the *tf-idf*, with two sources: one local value estimated from the image and one global value estimated for the collection. This section builds on the previous section, where we discussed a local importance model for a single object, equivalent to a *tf* measure for image objects.

Local importance of a star graph in an image

Star graphs associated to an image are composed with concepts attached to image regions. Hence, the importance of a star graph reflects the importance of the corresponding image parts. A function combining the *importance values* of the concepts must satisfy the following requirements:

- the importance value of a star graph should increase with the importance values of its concepts;
- the arity (number of concepts involved) of the relation of a specific star graph should not influence the importance value.

These requirements are satisfied by using an **average function**. We can define the local importance of a star graph s (composed of several concepts c_k) in an image j as:

$$local_{s,j} = \frac{1}{arity_s} \sum_{c_k \in s} importance_j(c_k)$$

where $importance_j(c_k)$ is the importance value of the image object associated to c_k (as defined in Section 2.3.3). Star graphs are considered as index terms, and there are cases where an image contains several instances of a given star graph. In such cases, the local value for the corresponding index term is obtained with a normalized sum of the values, which is consistent with the usual way of estimating the *tf* value of text index terms.

Global importance of a star graph in a collection

In text retrieval with the vector space model, the global importance (such as the *idf* value) of an index term aims at emphasising the impact of the most discriminating index terms. This factor for a star graph s is calculated according to the classical formula:

$$global_s = \log\left(\frac{N}{n_s}\right)$$

where N is the number of images in the dataset, and n_s is the number of images that are indexed by the star graph s.

An important aspect of this factor is that it is applied *after* the document expansion [1]. Therefore, the semantics is slightly different from the usual *idf*, since the extended index contains all generalizations of star graphs defined in the original index. The standard *idf* of a term i is calculated by counting the number of documents in which i occurs, while the proposed $global_s$ value is calculated for the star graph s by counting not only the documents in which s occurs, but also the documents in which any generic star graph of s occurs. For instance, in order to estimate the global importance value of the term *left(Boat,People)*, it is necessary to take into account the documents indexed by *left(Boat,People)*, but also the documents indexed by *left(Sailboat,People)*, *left(Boat,Jean)*, and *left(Boat,Matthieu)* – assuming that the concept type lattice contains the following relations: *Sailboat* $\leq_c$ *Boat*, *Jean* $\leq_c$ *People*, and *Matthieu* $\leq_c$ *People*. Indeed, all of these documents are **implicitly** indexed by *left(Boat,People)*, which is made explicit after the document expansion.

Weight of a star graph

The above definitions of the local and global values allow us to define a weighting scheme for star graphs, that is a combination of both values. We define the weight of a star graph s in a document j based on the two following sources: the $local_{s,j}$ value, that is calculated for s in the document j, and the $global_s$ value, that is calculated for s independently of the document in which it appears:

$$w_{s,j} = local_{s,j} \times global_s \tag{2.14}$$

The weight of a star graph combines its **local importance** in the document and its **global importance** in the collection.

Evaluation of the weighting scheme for star graphs

We tested the proposed weighting scheme with the same two collections as in Section 2.2.3 (Coll-1 and Coll-2), that we used to evaluate the benefits of integrating relations into the VSM. The results given in Figure 2.31 show the MAP values for several weighting settings for both collections, in the original space and the extended space.

1. The document expansion is described in Section 2.2.3.

First, we can notice the large difference in the results between the original and extended spaces, that consolidates the findings about the positive impact of the document expansion described in Section 2.2.3. Second, we observe no significant difference accros the weighting settings. In order to explain this result, it should be noticed that the sparseness of the vector spaces – in comparison to the size of the collections – makes the very occurrence or absence of a star graph in a document index decisive for the retrieval, regardless to its weight. Indeed, in Coll-1 (resp. Coll-2), the star graph vector space has 53,000 (resp. 7,000) dimensions, while the collection contains about 800 (resp. 2,000) images, which is about two orders (resp. one order) of magnitude. For the same reason as mentioned in the previous Section 2.3.3, this makes it difficult to evaluate the weighting method directly. Once again, the small differences that can be noticed in the MAP values for a given system are the results of slight changes in document rankings, since all weighting settings will obviously produce the same set of retrieved images – certainly ranked differently. Therefore, since the standard recall-precision measure evaluates system *rankings* based on a *binary* ground truth, ranking differences inside a set of mostly relevant documents have very little effect on the measure. However, we argue that the rankings produced with the *local* weighting model are closer to users' perception of relevance, since this model is based the geometry-based importance measure, which was proved to be a good estimator of this perception. Therefore, the top-ranked documents will contain star graphs where objects are big and in central position, inside low-clutter images. More details about this model and the experiments can be found in our **IPM'11 journal paper** [MCM11].

2.4 Conclusion

This chapter summarizes our works in image representations for objects, in which we develop three aspects: visual vocabularies, relations, and weights. We discussed in the first part (Section 2.1) several contributions in the bag-of-visual-words domain. We introduced the *Edge Context* descriptor, that refines SURF descriptor and makes it more discriminative. We also described two alternatives to the standard KMeans-based quantization step: one uses an iterative selection of candidate words among a random set of features, and the other consists of a split representation with a reference vocabulary and a reduced-size vocabulary that is projected onto the reference. Both approaches attempt to improve the standard way of building and using a visual vocabulary, in different contexts. Finally, the section ends with a discussion about the basis behind applying text processing techniques to visual words, and highlights the importance of both the quantization step and the words distribution. The second part of this chapter (Section 2.2) discusses the general notion of *relation*, taken at several levels – relations between low-level descriptors for building visual phrases to enrich and refine the representation, relations across modalities for annotating images, and relations between image objects for increasing the expressivity of description. The last part (Section 2.3) focuses on the question of weighting schemes for image parts. This part describes weighting strategies

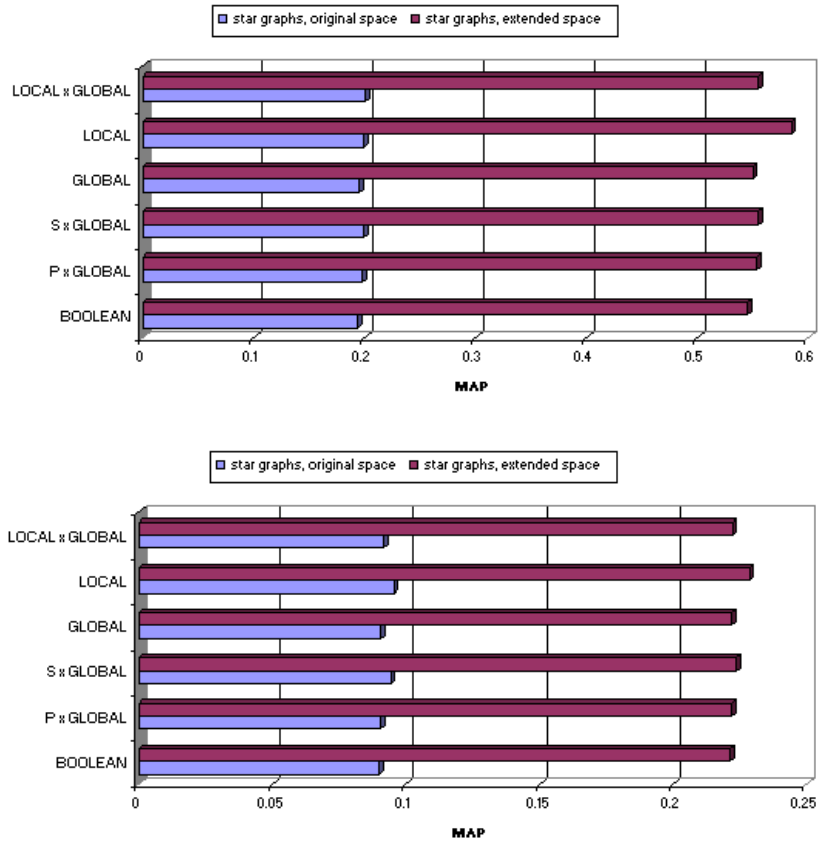


Figure 2.31 – Retrieval results with several weighting schemas for the two collections, in the original space and the extended space. Above: Coll-1, below: Coll-2. S refers to the Size criterion, and P refers to the Position criterion.

based on varied sources of information (spacial modeling, users' gaze, and geometry), applied at several levels of granularity (visual words, regions, objects, and star graphs).

These contributions mainly originate from the work done during my PhD at Grenoble 1 University, my postdoc at NII, and Ismail's PhD at Lille 1 University. This work also involve several colleagues: Ioan Marius Bilasco (Assistant Professor in the team), Taner Danişman (former postdoc of the team, now Assistant Professor at Akdeniz University in Turkey), José Mennesson (former postdoc of the team, now Lecturer at Télécom Lille 1), Pierre Tirilly (Assistant Professor in the team), and Thierry Urruty (former PhD student of the team, now Assistant Professor at University of Poitiers). The selection of referenced papers related to this chapter include 4 journal papers and 17 conference papers between 2002 and 2016.

The topics discussed in this chapter relate to image representations for general objects. The specificity of persons and faces compared to general objects requires dedicated features and techniques. While in object classification, faces would be a single class, the objective of person

recognition is to distinguish different classes among a set of faces. Moreover, since human faces are among the most useful semantic content of interest for users [Süs15], it is important that models and systems be able to finely represent persons and faces. Next chapter describes our work related to person representation and recognition.

Chapter 3

Person recognition: exploring depth and time

Person recognition is a general topic in computer vision that includes both *verification* (or *authentication*) and *identification*. Verification is concerned with validating a claimed identity based on the image of a person, and either accepting or rejecting the identity claim (one-to-one matching). The goal of identification is to identify a person based on the image of a face or body, that needs to be compared with all the registered persons (one-to-many matching). The general objective of person recognition is to automatically decide about a person identity by analyzing their appearance, namely facial and body features. The challenges in automatic face recognition arise from to two major problems. The first problem is the **inter-class similarity**, that is due to a structural similarity between faces. Indeed, two different faces are very close in structure since they are composed of the same parts (eyes, mouth, nose, etc.) whose location and shape vary slightly. The second problem is the **intra-class variability**, that is due to appearance changes of a given face obtained in different acquisition conditions (see Figure 3.1). Such changes are due to several factors such as changes in face expression, head pose, occlusion, or lighting conditions. As a consequence of these two problems, faces of different people acquired under identical conditions are often more similar one to another than faces of the same person taken under different conditions. As shown in Figure 3.2, a variation of the lighting or pose can drastically impact the appearance of a face: images in the top row belong to one person, and images in the bottom row belong to another person. However, the images in each column (a), (b) and (c) appear to be closer to each other.

In order to evaluate and compare the results of different face recognition approaches, several benchmarks were created [PWHR98, HRBLM07, LM14]. Among those, Labeled Faces in the Wild (LFW) [HRBLM07] is usually considered to be the current reference [ZLY+15, OMG+14, SKP15, SWT14, TYRW14]. It is a dataset of face photographs designed for studying the problem of unconstrained face recognition. It contains 13,000 photos of 1,680 individuals. The

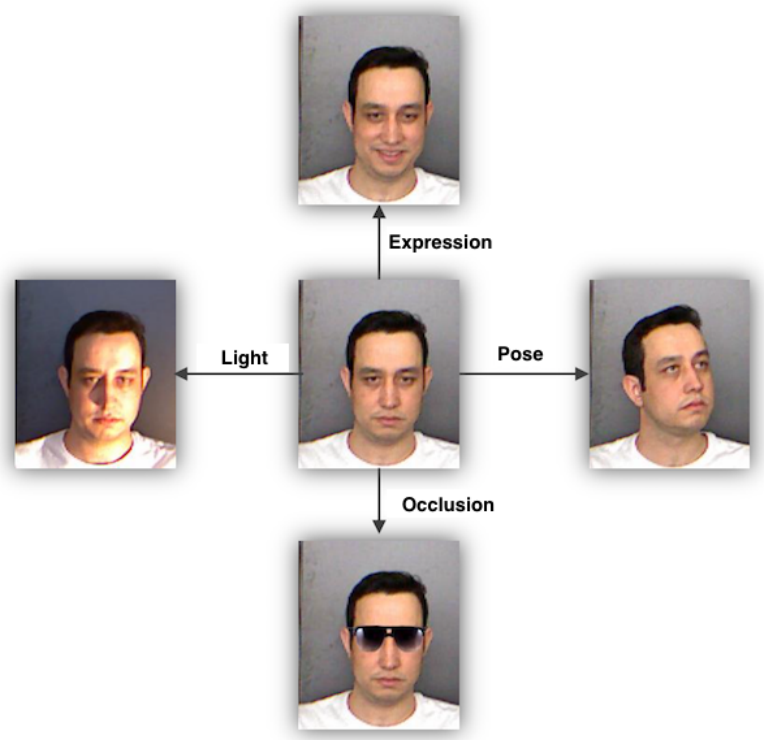


Figure 3.1 – Example of appearance variation of a face.

performances of major approaches are reported on the dataset website[1]. The best results are obtained with the *Unrestricted with labeled outside data* protocol, which means that (1) the identities are available for all dataset images, allowing one to potentially form a large number of matching/mismatching image pairs, and (2) using almost any kind additional data from outside LFW (such as external labeled images, patches, annotations, etc.) is allowed for system training. At the time of writing, the best systems achieve an accuracy of 0.9977 ± 0.0006 (Liu *et al.* [LDB+15] from Baidu) and 0.9977 ± 0.0009 (AuthenMetric), which is above the human performance of 0.9920 [KBBN09]. Such systems require huge training sets and have a high computational cost. Both systems use tremendous amounts of external data: 1.2M faces from 18K individuals for Baidu, and 500K faces from 10K individuals for AuthenMetric. More importantly, while these systems achieve very high recognition rates, such results on LFW should be taken in a specific context: the faces in the wild are actually not that wild, since the LFW dataset was built using the Viola & Jones frontal face detector [HRBLM07]. In other words, the poses, expression and illumination variations are unconstrained *to the limits* of the face detector. We can argue that such good results on the LFW dataset might not be obtained in real-life conditions, especially when the lighting conditions are extreme and would not even allow a face to be detected, or on video frames, since individuals are likely to appear with

1. URL: http://vis-www.cs.umass.edu/lfw/.

considerable variations in pose, illumination, expressions, and with occlusion or even blurry.

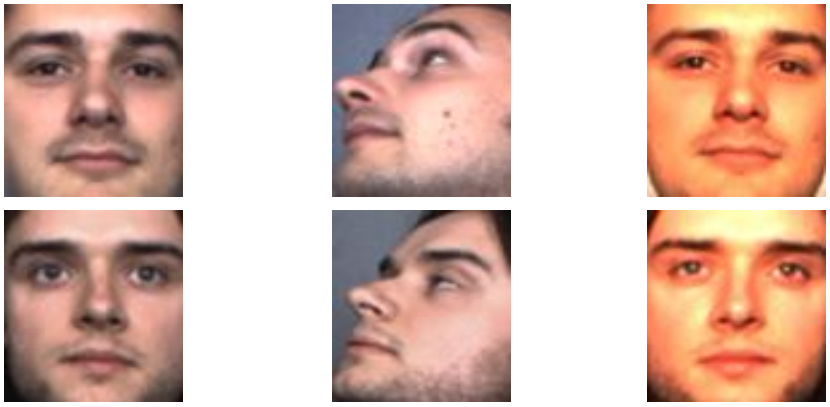


Figure 3.2 – Effects of changes in acquisition conditions: (a) frontal pose, (b) pose change, (c) light change. Each line shows a different face, however faces in each column look more similar one to another.

In this chapter, we present our contributions to improve person recognition by addressing these problems. In Section 3.1, we present our work for enhancing the precision of face recognition by using **depth information**, in the form of a 2D-3D bimodal approach for face recognition, with the motivation that the depth information (i.e. the 3D modality) is complementary to the 2D visual information.. Since the research domain of 2D face recognition has reached a certain maturity and numerous recognition methods give satisfactory results in controlled conditions [ZCPR03a, JA09, Kum13], the 2D modality is essentially processed with standard state-of-the-art methods, and our work focus on the 3D modality recognition, from the data acquisition to the fusion of 2D and 3D modalities. We introduce a 2D-3D bimodal approach that combines visual and depth features in order to provide higher recognition accuracy and robustness than classical monomodal approaches. This work was carried out during the PhD of Amel Aissaoui (September 2010 – June 2014), that I co-directed with Prof. Chaabane Djeraba. We made contributions in the following aspects:

— the **acquisition** of the face depth data: the choice was made to use a simple passive[1] stereo setting to reconstruct 3D faces from stereo image pairs. The originality lies in the use of an Active Shape Model (ASM) [MN08] on the faces in stereo images to estimate the depth of a limited number of fiducial points, with a high confidence. Such points are used to bound the depth of other neighbour points, yielding a high-quality reconstruction. This part of the work was published at **VISAPP'12** [**AAY+12**], **CORESA'12**

1. No infrared light is required.

[AAMD12a], and **ICIP'12** [AMD12] for the original idea and early results, and our **MTAP journal paper** [AMD13] describes the complete model and extended experimental results;

— the **representation** of face depth data by introducing a new descriptor: *Depth Local Binary Patterns* (DLBP), a specific descriptor for depth data, which was published at **ICIP'14** [AMD14];

— and finally, the **fusion** of 2D and 3D modalities for face identification, by designing a two-stage fusion method, which was published at **VISAPP'15** [AM15].

The experiments conducted with different public datasets allowed to validate our propositions. In particular, results shown the quality of the data obtained using the proposed reconstruction method, and also a gain in precision obtained by using the DLBP descriptor and the two-stage fusion. Beside these contributions, we designed and made publicly available a multipurpose face dataset dedicated to evaluate face analysis methods (recognition of identity, expression, and pose): FoxFaces [1]. This dataset was mainly designed by Amel Aissaoui and Afifa Dahmane [2] during their PhDs, under the direction of my colleague Ioan Marius Bilasco (Lille 1 University) and me, and it is described in details in our **VISAPP'16 paper** [ADMB16].

In the second part (Section 3.2), we focus on person recognition by exploiting the **variation of appearance over time**. This work is part of the PhD of Rémi Auguste (November 2010 – July 2014), under the direction of Prof. Chaabane Djeraba and me. The work was done in the context of an ANR-funded project, PERCOL [3], as a part of the french national REPERE challenge [GK13] targeting person spotting and naming from TV shows. A video corpus containing several hours of TV shows broadcasted from BFMTV and LCP TV French channels was provided to the participants of the challenge. The objective was to build a system able to answer the following questions: *Who is speaking? Who is present in the video? What names are cited? What names are displayed?* The challenge is to combine the various information coming from speech and images. Three consortia participated in this challenge; we participated as a partner of the PERCOL consortium, in charge of the visual modality – among other modalities such as audio, Optical Character Recognition, etc. In this context, we focused on *person re-identification* in the video, which is a fundamental task consisting in finding the occurrences of a given individual across shots of a single video or across various videos. The proposed contribution for recognizing persons from videos consists of two parts:

— **Person re-identification**: The first part extracts *persontracks* (short video sequences featuring a single person with no background) from the videos. The re-identification con-

1. The dataset is freely available on request at URL: http://www.lifl.fr/FOX/index.php?page=datasets

2. Afifa Dahmane is a former PhD student from our team, co-advised by Ioan Marius Bilasco. Her main research interest during her PhD was determining a precise estimation of the head pose from an image or video. She is now an Assistant Professor at University of Sciences and Technology Houari Boumediene, Algiers, Algeria.

3. PERson reCOgnition in audiovisuaL content, partners: France Telecom (Orange Labs Lannion), LIF (Aix-Marseille University), LIA (University of Avignon and the Vaucluse), LIFL (Lille 1 University). ANR-10-CORD-0102, Nov. 2010 – Jun. 2014.

sists in gathering persontracks according to their (estimated) identity. For this purpose, we defined a novel descriptor dedicated to persontracks: *Space-Time Histogram* (STH) and its associated similarity measure, that were introduced at **EGC'12** [AAMD12c] and **CORESA'12** [AAMD12b] (in french), and complexity issues and evaluation are discussed in our **ICMR'15 paper** [AMT15]. This descriptor is used in our approach for persontracks clustering.

— **Cluster labelling strategies for person recognition**: The objective of the second part is to put a name on persontrack clusters. It consists of various strategies to (1) assign an identity to a persontrack using its frames and (2) to propagate this identity to other members of its cluster, as described in our **PR journal paper** [AMT16] (submitted, under review at the time of writing).

These two parts form our proposed approach to automatically recognize persons in TV shows, that was defined and used during PERCOL project. Besides, we designed a re-identification benchmark, FoxPersonTracks [1], that is publicly available. It is composed of a manually-filtered subset of REPERE dataset, containing 4,604 persontracks (about 170 minutes of video) showing 266 individuals. The dataset comes with evaluation metrics and tools to easily compare systems' results. This benchmark is described in our **CBMI'15 paper** [ATM15]. Our proposals were evaluated using this benchmark and Buffy dataset [ESZ06, SEZ09]. The results of our experiments show that our approach significantly improves the precision and robustness of the recognition process thanks to the use of re-identification.

3.1 Face recognition by combining visual and depth data

Visual and depth features play complementary roles in the description and recognition of faces as they typically represent different faces' characteristics. The 2D image provides informations about textured regions with little geometric structure (e.g. hairy parts, eyes, eyebrows), and the 3D data provides information regarding less textured regions (e.g. nose, chin, cheeks). Hence, using the 3D shape information in addition to the intensity images allows a better face representation and therefore better precision and robustness in recognition.

3.1.1 Depth map generation with stereo cameras

The first step towards depth-based face recognition is data acquisition. The depth information can be obtained using different techniques. On one hand, 3D scanners are currently the most accurate and most widely used in 3D face recognition. Although such devices provide high-quality 3D data (point cloud and mesh), they can only be used in a limited number

1. FoxPersonTracks dataset, hosted and distributed by ELRA (http://catalog.elra.info), ISLRN: 168-132-570-218-1, ELRA ID: ELRA-S0374. Freely available for academic researchers.

of applications, mainly because of their cost, and also because of the long acquisition time, where the full cooperation of the subject is therefore mandatory. Note however that 3D data is richer than depth data, since it contains a full 3D representation of the face, and not just a distance-to-sensor information for 2D points. On the other hand, there exists other depth acquisition systems, either based on *structured light* like Microsoft's Kinect: fast but with a limited quality [1] for face analysis, *time-of-flight* like MESA IMAGING's SR4000: fast but the device is more expensive, or *stereo reconstruction* from image pairs. The data acquired with such alternative systems generally have lower quality and resolution than when obtained with 3D scanners, and yet it enables a fast depth acquisition with a reasonable cost and few constraints, and therefore they bring a promising alternative to 3D scanners. As a quality/cost tradeoff, we selected a passive stereo system, consisting of a simple pair of calibrated cameras.

Disparity maps

Stereo-based depth estimation requires to calculate a *disparity map* of the scene captured from two different (yet close) points of view, with a calibrated system. The disparity is the difference found in stereo images of the projections of a given 3D real-world point. In order to calculate the disparity map, a *stereo matching step* searches both images for pixels representing the projection of given 3D points. The disparity d of a point is calculated using the Euclidian distance as follows:

$$d = \sqrt{(x_i - x_i^{'})^2 + (y_i - y_i^{'})^2} \tag{3.1}$$

Since the location of corresponding points in the left and right images is unknown, and the search space for matching a point is relatively large, a rectification process that consists in projecting the stereo pair onto a common image plane (to align the y coordinates) helps reducing the search space to 1 dimension along epipolar lines. Therefore, when using a calibrated system and rectified stereo pairs, the y coordinates of each corresponding points are identical, and the disparity is calculated from a simple difference between the x coordinates:

$$d = \sqrt{(x_i - x_i^{'})^2} = |x_i - x_i^{'}| \tag{3.2}$$

Note that the disparity values in such settings are therefore integer values representing the pixel shifts on the x-axis. Once the disparity map is calculated, the depth z of a point $p(x, y, z)$ with a disparity value d is estimated as:

$$z = \frac{fb}{d} \tag{3.3}$$

where f and b are the camera focal and baseline (i.e. the distance between cameras), respectively – see Figure 3.3. In order to calculate the disparity map, many algorithms were proposed

1. We refer to the first version of the sensor, Kinect for Xbox 360, the only one available during Amel's work. Since September 2014, the second version, Kinect for Xbox One (or Kinect 2), provides much better depth information by using a time-of-flight sensor.

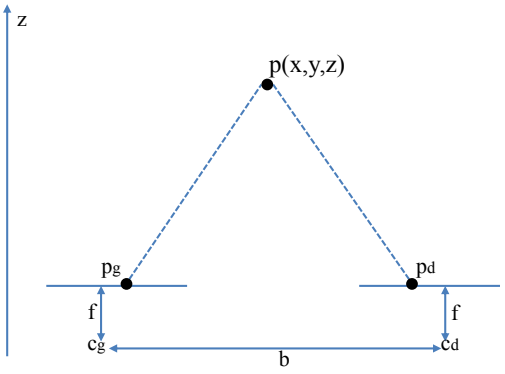


Figure 3.3 – Stereo cameras setting. p is a point in the 3D real world with corresponding projections p_g and p_d in the left and right images, respectively. f is the camera focal, and b is the *baseline*.

[SS02]. They can be classified into two categories: *global* and *local methods*. Global methods solve optimisation problems over energy minimisation constraints. Some popular global methods use graph cut [KZ02], belief propagation [SZS03] and dynamic programming [TV98]. Global methods provide a good accuracy since they process pixels in a holistic (therefore non independant) manner. However, they face problems when searching corresponding points for non-edge pixels because of the aperture problem, that is the lack of information in matching windows for homogenous regions, as illustrated in Figure 3.4. For example, there may exist

Figure 3.4 – Illustration of the aperture problem for homogeneous face regions: it is hard to find matching points.

many patches with a similar appearance, and therefore it is difficult to select the correct match. Another drawback of global methods is that they require a long processing time [KZ02].

Local methods (also called *block-matching* methods) are based on intensity correlation and they can be used in real-time applications because of their low complexity. Correlation-based stereo matching algorithms typically produce dense depth maps by calculating the matching cost for each pixel inside a small matching window, and by searching the lowest-cost match. Different similarity measures are used in correlation-based methods, including Sum of Absolute Differences (SAD):

$$SAD_{(I_L(x,y),I_R(x',y'))} = \sum_{u=0}^{m-1}\sum_{v=0}^{n-1} |I_L(x+u, y+v) - I_R(x^{'}+u+d, y^{'}+v)| \tag{3.4}$$

where I_L (resp. I_R) is the left (resp. right) image, and $m \times n$ is the size of the matching window w. These methods also suffer from the aperture problem. Note that in our case of rectified stereo pairs, the matching window is 1D since the search space is reduced to the epipolar line.

Proposed method

We describe below a method for stereo reconstruction that is specific to faces. Starting from a stereo pair showing a face, the process can be summarized as follows:

1. **Sparse points estimation**. An Active Shape Model is first fit to both the left and the right images. We obtain a set of N fiducial points (keypoints) with high location confidence, that serves as a reference to estimate disparity values as in Equation 3.2. In our approach, we use $N = 46$ fiducial points. Since the ASM estimation is robust, we can rely on such points to estimate sparse but reliable disparity values for a limited number of points.
2. **Dense map generation**. Because the 3D face surface is globally smooth and continuous, the N points can be used to set a disparity range for the remaining points. The key idea in this process is that the disparity of any face point should lie within the minimum and maximum disparity values of its neighbour fiducial points. This constraint helps achieving smooth estimates of dense disparity maps, that are calculated using the local method described above: SAD. From the disparity map, the depth map is generated using Equation 3.3.
3. **Noise removal**. The generated depth maps are likely to contain noise, such as holes or spikes, caused by missing values or wrong matches, mainly due to the aperture problem. In this post-processing step, a dedicated noise detector finds and corrects noisy values.

Disparity model for sparse points

The first step of our method consists in building a disparity model of the given face. This model gives a holistic representation of the disparity distribution of the face points, which will be used as a guidance in the dense disparity map generation. In order to build the disparity model, we start by fitting an ASM on both the left and the right images to identify a set

of corresponding points with high confidence (see Figure 3.5-left and centre). The ASM is a statistic shape model obtained after a learning process on annotated faces. Fitting the ASM to a face image consists in estimating the shape parameters of the model by minimizing a cost function defining how well a particular instance of the model fits the face.

The ASM guarantees a precise location of fiducial points in the stereo pair, and therefore a high disparity confidence for these points.

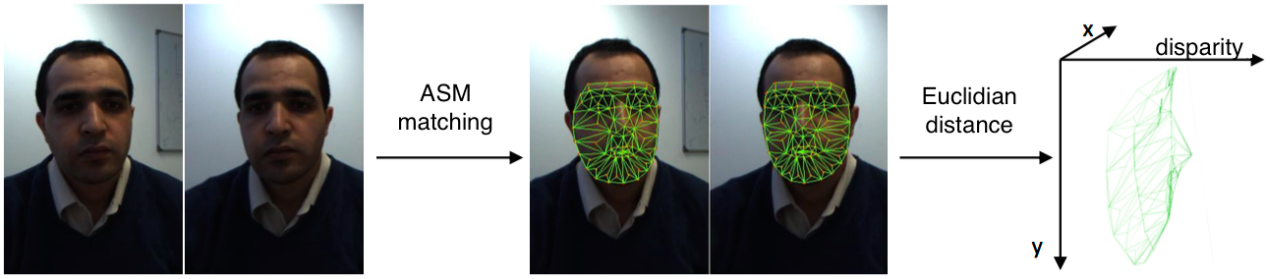


Figure 3.5 – Construction of the disparity model for a face.

After fitting the ASM to both images separately, we obtain the 2D coordinates of N face points in the right image $R = \{(x_i, y_i)|i \in [1, N]\}$ and in the left image $L = \{(x_i^{'}, y_i^{'})|i \in [1, N]\}$, that are used to obtain the set of 3D coordinates: $P = \{p_i(x, y, d)|i \in [1, N]\}$, that represents the disparity model for the face (Figure 3.5-right). This model is used in the next step as guidance for generating the dense depth map.

Dense depth map generation

The disparity values for all face points are calculated based on the disparity model. We begin by partitioning fiducial points into slices of homogeneous depth that we call level sets (or level planes), according to their disparity values (see Figure 3.6-left and center). The level sets are defined to be perpendicular to the normal vector centred on the point with the highest disparity (i.e. closest point to the cameras), as given by the model. This point corresponds to the nose tip when the head orientation is close to frontal, and defines the closest set. The farthest plane is defined by points with lowest disparity. Between these two sets, we define a number of intermediate sets, corresponding to frequent disparity values found in the model.

Figure 3.7 illustrates how the head pose variations impact the decomposition step of the disparity model. After the decomposition step, different areas can therefore be defined in the face image based on disparity ranges. We generate shape regions in one of the stereo images (left or right, chosen arbitrarily) that correspond to the points of the disparity model inside the slice (Figure 3.6-right). A disparity range is assigned to each shape according to the disparity values of the points belonging to the slice. The disparity range is used as a hard constraint when applying SAD to calculate the dense disparity map. This constraint is central since it

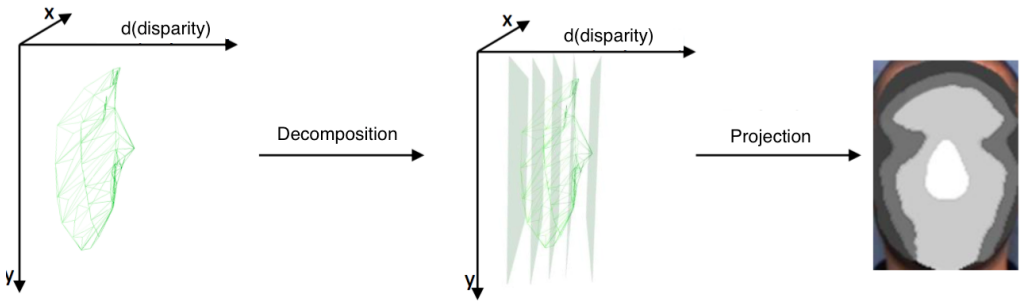


Figure 3.6 – Decomposition of the disparity model into slices.

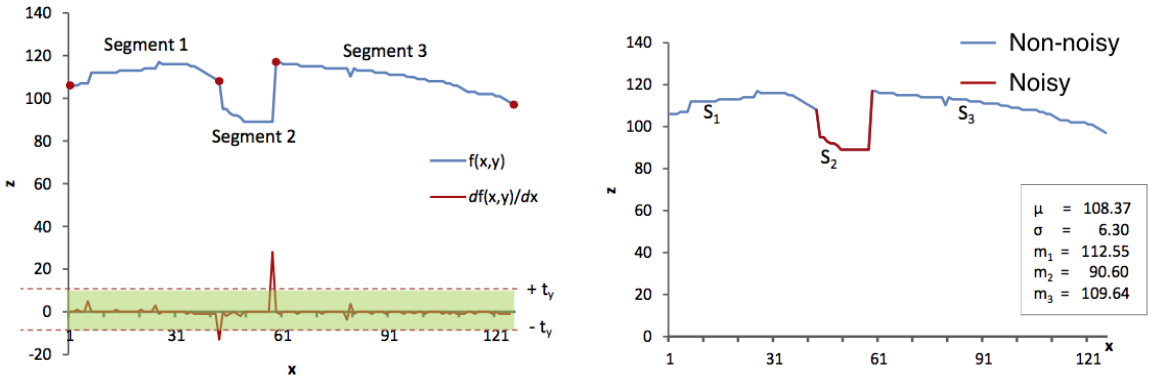
further reduces the search area to just a small segment (inside the shape), instead of the entire epipolar line. It also reduces the number of matching errors since the face disparity values are bound inside neighbour level planes. Therefore, it guarantees a consistent dense disparity map. Finally, the dense depth map is obtained by applying Equation 3.3 to the dense disparity map.

Figure 3.7 – Examples of model projection on a face with different poses.

Noise detection and removal

This post-processing step is intended to remove potential holes and spikes caused by uncertainties and wrong matching values. Since global methods affect the entire face data and

cause data loss, we follow a local methodology for noise detection and noise removal. The noise detection is easily performed for holes and small spikes but becomes difficult to perform when it comes to detect large spikes caused by wrong matches in homogenous surfaces. We introduce a simple method for addressing the noise detection problem, that consists in (1) segmenting depth rows based on the depth gradient – assuming a continuous face surface, and (2) classifying segments as noisy or non-noisy. Considering the depth curve as a smooth function, we use its first derivative to detect the main cut points that split it into a set of segments, as shown in Figure 3.8-left. Then, the segments are classified as noisy or non-noisy: given a depth row with a mean μ, a standard deviation σ and a set of segments $s_1, s_2, \ldots, s_n$ (obtained from the cut points) with associated means $m_1, m_2, \ldots, m_n$, we identify a given segment s_i with a mean value m_i as noisy if $m_i \notin [\mu - \sigma, \mu + \sigma]$. Figure 3.8-right illustrates this classification.



(a) (b)

Figure 3.8 – (a) Detection of cut points to split a depth row into segments. Red dots show cut points. (b) Noisy segments detection. The m_i values are the mean values of S_i segments.

Once the noisy regions are detected, they are removed and a hole-filling method based on cubic interpolation is applied. Figure 3.9 gives an example of the proposed depth map denoising process. This method is able to identify not only holes but also small and large spikes so that the filling step only impacts the detected noisy regions and does not affect the rest of the data (contrary to using e.g. a median filter convolution). Another advantage of this method is that it does not use any free parameter and is therefore fully automatic.

Evaluation of the generated depth map

In order to evaluate the quality of the depth map, we synthesized a stereo database of 105 faces from the Texas 3D Face Recognition Database [GCMB10]. This dataset contains 1,149 colour images from 118 persons, and corresponding depth maps that serve as a ground truth. Disparity maps are generated from the stereo pairs of faces from different persons using the proposed method, and also using two other methods for comparison: a standard

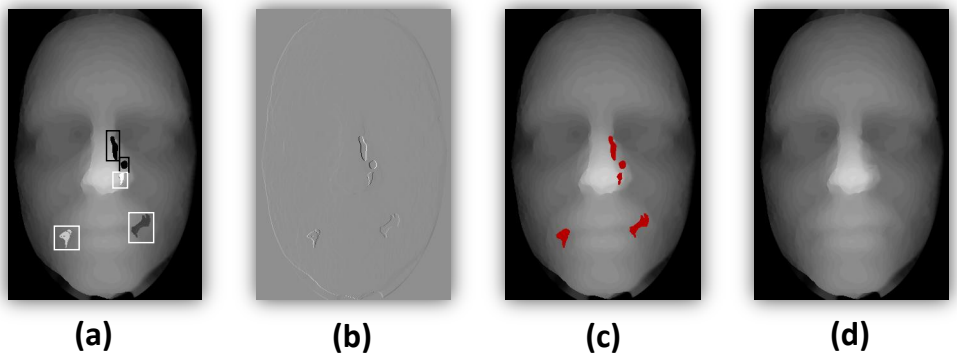


Figure 3.9 – Depth map denoising: (a) depth map with holes, (b) gradient, (c) noisy slice detection, (d) corrected depth map.

block-matching method [BT99] (that does not use the proposed disparity model) and the graph-cut method [KZ02] (that is global). Figure 3.10 gives an illustration of the results for the 3 methods. The results are compared using the Root Mean Squared error (RMS) [SS02] on disparity maps defined in Equation 3.5, and the Percentage of Bad Matching pixels (PBM) defined in Equation (3.6):

$$RMS = \left(\frac{1}{np} \times \sum_{x,y} (d_E(x,y) - d_T(x,y))^2 \right)^{\frac{1}{2}} \tag{3.5}$$

$$PBM = \frac{1}{np} \times \sum_{x,y} D(x,y) \quad , \quad D(x,y) = \begin{cases} 0 & if\ |d_E(x,y) - d_T(x,y)| \leq \delta_d \\ 1 & otherwise \end{cases} \tag{3.6}$$

where:

- np is the number of pixels in the depth map;
- $d_E(x,y)$ and $d_T(x,y)$ are the estimated disparity and the ground truth disparity, respectively, for the pixel (x,y);
- δ_d is a disparity error tolerance – in our experiments, we use $\delta_d = 1.0$ since it is the most commonly used value in the previously published studies [SS02].

Figure 3.11-left shows that the RMS error is reduced from 7.33 for the block-matching results to 4.95 for the proposed method that integrates the disparity model in the block-matching process. The PBM graph (Figure 3.11-centre) shows how the percentage of the bad matching pixels in our results is very small compared to that obtained using the block-matching method. Although the block-matching method is rapid, our method is faster and requires less time than both other methods. As shown Figure 3.11-right, our method requires about 50 times less processing time than the graph-cut method, for a 501×751-pixels image. This is because the disparity interval is restricted to a small segment from the epipolar line using the disparity model. More details about the evaluation of our depth maps can be found in our **ICIP'12**

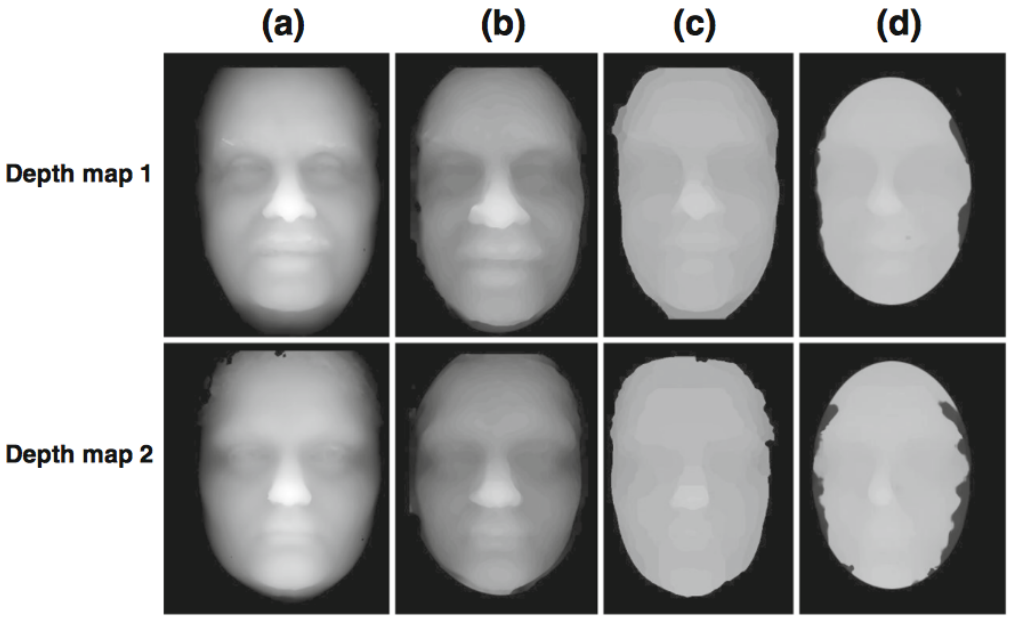


Figure 3.10 – Example of depth maps: (a) original (ground truth), (b) our method, (c) graph-cut and (d) block-matching.

paper [AMD12] and **MTAP journal paper** [AMD13]. The purpose of building depth maps is to use them to compare and identify faces, with the help of a new descriptor that we introduce in the next section.

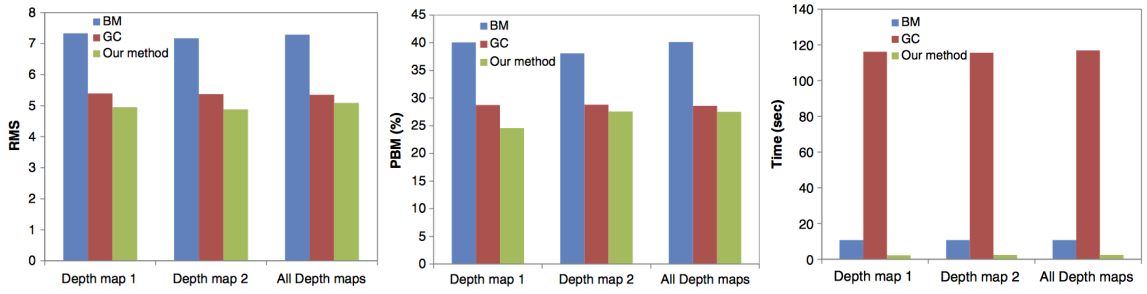


Figure 3.11 – Comparison of depth maps. Left: RMS values. Centre: PBM values. Right: Processing time.

3.1.2 DLBP: Depth Local Binary Patterns

The second contribution in this work is a descriptor for representing face depth data. While most 2D-3D bimodal approaches for face recognition use a single descriptor for each modality

[CBF03, XLTQ09, HAWC09, JCB11], we choose to use different descriptors for both modalities. For the 2D modality, among the most popular 2D face descriptors, the Local Binary Patterns (LBP) proposed by Ojala *et al.* [OVOP01, OPM02] are considered to be one of the simplest and most efficient local descriptors. Due to their computational efficiency and good discriminative capabilities, LBP were widely used in many face image analysis fields [HSA+11]. Regarding the 3D modality, we introduce a new descriptor : the **Depth Local Binary Patterns (DLBP)**, which are an extension of the LBP that we designed for depth images. This extension allows extracting more discriminative features from depth images.

Standard LBP

The LBP descriptor encodes pixel-wise information in a given image, and describes each pixel with the relative grey levels of its neighbour pixels. The $LBP_{R,V}$ code for a pixel $P(x, y)$ is calculated as follows:

$$LBP_{R,V}(P) = \sum_{i=0}^{V-1} s(p_i - p)2^i, \ with \ s(k) = \begin{cases} 1 & \text{if} \quad k \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

where :

— p is the grey-level value of the central pixel P;

— p_i is the grey-level value of the neighbour pixel P_i around the central pixel with a radius R. The position of the neighbour pixels are given by Equations 3.8 and 3.9 – with a bilinear interpolation in case the estimated position does not exactly match a pixel, allowing several radius values and neighbourhood sizes.

$$x_{P_i} = x + R\cos\left(\frac{i}{V} \times 2\pi\right) \tag{3.8}$$

$$y_{P_i} = y - R\sin\left(\frac{i}{V} \times 2\pi\right) \tag{3.9}$$

In practice, Equation 3.7 shows that the sign of differences in the neighbourhood are interpreted as a V-bit binary value, resulting in 2^V possible values for a pattern. Local histograms of LBP codes are often extracted from a grid of regions to form the final descriptor. Numerous approaches used LBP for depth image description [LZAL05, HAWC10, WRM10, XHL11, TYSH13]. When applying LBP to depth images, every one of the 2^V LBP codes represents a 3D pattern like a flat, convex, concave, or more complex shapes (see Figure 3.12). In [HAWC10, XHL11] for example, authors use LBP for both 2D and 3D face representation in a bimodal face recognition approach, yielding a high accuracy. A multi-scale extension *Multi-Scale LBP* (*MS-LBP*) was proposed by Di Huang *et al.* [HZA+10], in which several radius/neighbourhood parameter combinations were used. Authors shown that this multi-scale extension with 4 different radius values ($R \in \{3, 4, 5, 6\}$) gives better results than the original LBP descriptor where $R = 1$.

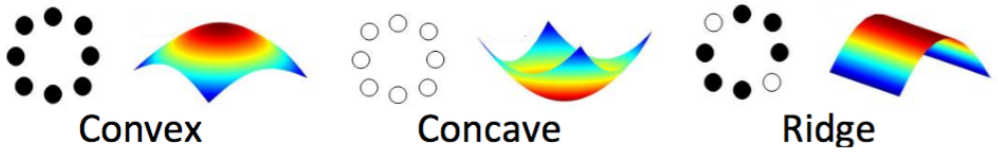


Figure 3.12 – Examples of shape patterns detected with $LBP_{1,8}$.

However, the direct application of LBP to depth images is likely to yield imprecise descriptions. For instance, Figure 3.13 gives some examples of similar shapes with different magnitudes that are encoded in the same way, which obviously decreases the discriminative power. The reason for this confusion originates from the very definition of LBP, that only considers the sign of the differences between the pixel and its neighbourhood.

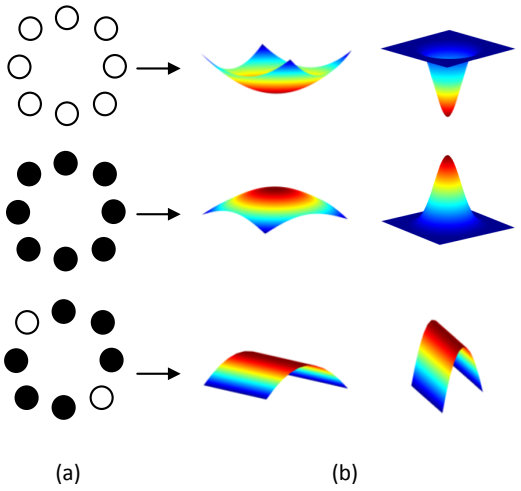


Figure 3.13 – Confusion between similar 3D shapes. (a) LBP codes. (b) Potential 3D shape matches.

Existing extensions of LBP for depth images

This lack of descriptive power is clearly problematic when one needs to derive a discriminative representation for a face recognition task. In order to address this problem, some LBP-based descriptors dedicated to depth face image were introduced. Yonggang Huang *et al.* [HWT06] proposed an extended LBP version, named 3DLBP. Beside the information provided by LBP, 3DLBP also considers the magnitude of the difference between the central pixel and its neighbourhood, called *Depth Difference* (DD). Authors found that the absolute value $|DD|$ seldom exceeds 7 in face depth images when the radius is small ($R = 2$), based on a statistical study. Therefore, the values are encoded on 3 bits – and values greater than 7 are

set (forced) to 7. Each neighbour is assigned a 4-bit code $i_1 i_2 i_3 i_4$, where i_1 encodes the sign like in the original LBP, and i_2, i_3, and i_4 encode $|DD|$. The concatenation of the i_1 bits from all neighbours form the first code c_1, which is actually the original LBP code. In addition, 3DLBP includes c_2, c_3, and c_4, made respectively of the concatenation of all i_2, i_3, and i_4 bits. Finally, a pixel is represented with the 4 codes c_1, c_2, c_3, c_4. Following the original way of building the final descriptor in LBP, local histograms are extracted from the 4 maps, and concatenated. Experimental results prove that considering the magnitude information somehow improves the discriminative power of the descriptor. However, the coding scheme of the 3DLBP suffers from 3 drawbacks:

1. The feature vector size is much larger. Each image is represented by 4 matrices of the same size. These matrices are segmented into regions and a histogram for each matrix region is calculated. This gives a feature vector of length $H \times 4 \times 256$, where H is the number of histograms extracted from each map.
2. The coding scheme is very sensitive to the depth variations. A small variation in the neighbourhood of a pixel leads to a big difference in the corresponding code. For example, for a given pixel with value 0 and neighbourhood values $(2, 1, 1, 1, 2, 5, 4, 3)$, the 3DLBP codes are $(251, 96, 145, 174)$. For a slight increase of the value of the last neighbour from 3 to 4, we obtain $(251, 224, 17, 46)$. This is due to the split of the binary sequences of each DD to 3 different layers.
3. 3DLBP is limited to a small radius. The DD values are encoded on 3 bits, in agreement with the maximal $|DD|$ value 7 found when $R = 2$. However, this maximal value does not hold anymore when $R > 2$, and therefore the method is not adapted to a multi-scale context. Although a close neighbourhood is suitable for LBP coding in 2D face recognition, we argue that it is not sufficient for depth faces representation because of the intrinsically low local constrast in face depth images. Unlike 2D face images (colour or grey level), face depth images are smooth and using large scales is essential for capturing depth contrast. One study proposed by Di Huang *et al.* [HAWC12] define a multi-scale generalization of 3DLBP. However, the $|DD|$ threshold of 7 is kept in this multi-scale extension, although this value was selected in 3DLBP based on $R = 2$. One can expect $|DD|$ values to be larger than 7 when $R > 2$, because the depth contrast is bigger at larger scales. Therefore, the descriptor "saturates" with values greater than 7, and such values are all set to the maximum value 7, and consequently this large constrast information lost. This possibly explains why results obtained in their work vary little with different radius values.

In [MdBJG14], authors proposed an LBP-based descriptor called Depth Local Quantized Pattern (DLQP) where a quantization step is introduced to capture the main depth difference values, and to increase its capacity to distinguish different depth patterns. This allows a finer and more flexible distinction of the different values found in depth images. We share the same objective of increasing the descriptor ability to distinguish different depth patterns.

DLBP

Inspired by the work described above, we proposed a new versatile descriptor called DLBP [AMD14]. Unlike 3DLBP and DLQP, DLBP is designed to consider large radius values in order to extract more discriminative features from smooth and low-contrast data, since it works on a multi-scale level:

$$\begin{aligned} DLBP_{R,V}(P) &= \begin{pmatrix} c^s_{R,V}(P) \\ c^m_{R,V}(P) \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=0}^{V-1} s(p_i - p)2^i, \ s(k) = \begin{cases} 1 & \text{if} \quad k \geq 0 \\ 0 & \text{otherwise} \end{cases} \\ \sum_{i=0}^{V-1} m(|p_i - p|)2^i, \ m(k) = \begin{cases} 1 & \text{if} \quad k \geq T^m_{R,V} \\ 0 & \text{otherwise} \end{cases} \end{pmatrix} \end{aligned} \tag{3.10}$$

where:

- p is the grey-level value of the central pixel P;
- p_i is the grey-level value of the neighbour pixel P_i around the central pixel, whose position relative to P is defined according to R and V (see Equations 3.8 and 3.9);
- $T^m_{R,V}$ is the magnitude threshold.

The magnitude threshold value $T^m_{R,V}$ is automatically estimated in order to take advantage of the depth information at several scales. For this purpose, the optimal value is statistically determined to maximize the discrimination power, while being robust to possible noise. The multi-scale and direction gradient $\textit{MSD-Gradient}_{R,V}$ in a depth map with parameters (R, V) for a pixel P is calculated as the average value of the depth difference:

$$\textit{MSD-Gradient}_{R,V}(P) = \frac{1}{V} \sum_{i=0}^{V-1} |p_i - p| \tag{3.11}$$

The magnitude threshold $T^m_{R,V}$ is obtained by taking the median of all $\textit{MSD-Gradient}_{(R,V)}$ non-null values from all depth maps. The proposed method for adaptive thresholding guarantees an efficient coding of the neighbours' magnitudes for different scales. In addition, it preserves only the relevant data by ignoring potential outliers in the depth map. It also offers more invariance to the depth maps resolution variations.

In the remainder, we follow the classical way of building the final descriptor that consists in concatenating local histograms extracted from a grid. Local histograms are extracted from H regions in both sign and magnitude matrices. Figure 3.14 illustrates how the final histogram is built, with $H = 2 \times 2$.

In summary, the proposed descriptor targets a discriminative feature extraction from depth images. It is based on a coherent coding of the sign and the magnitude of the pixel neighbourhood. The adaptive thresholding allows to extract discriminative variations from the depth maps at large scale. Like the classical LBP, the DLBP is simple and compact.

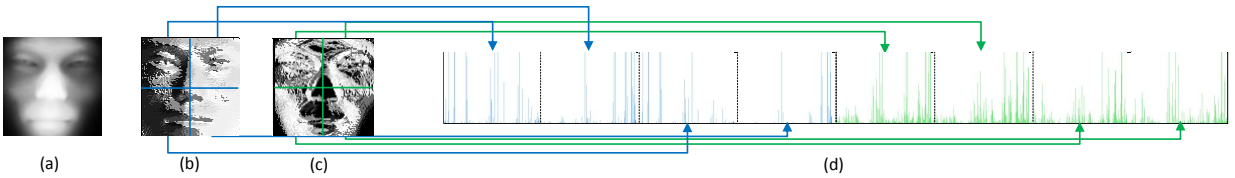


Figure 3.14 – Extraction of the DLBP descriptor from local histograms extracted from the sign and magnitude matrices with $H = 2 \times 2$: (a) depth map, (b) sign matrix, (c) magnitude matrix, (d) resulting histogram.

Evaluation of DLBP

We report below experimental results that demonstrate the efficiency and the behaviour of the proposed descriptor. We start by studying the impact of two parameters on the face recognition precision: the radius value R and the number H of local histograms – V is set to the constant value 8. Then we compare the results obtained with DBLP to the standard LBP and to 3DLBP. We used 4 (+2) datasets for this evaluation:

— **FRGC** [PFS+05]: the most used 3D faces dataset, composed of 4007 face colour images from 446 persons, with corresponding depth images.

— **Texas** [GCMB10]: this dataset, that we used in the previous Section 3.1.1 to evaluate the quality of generated depth maps, contains 1149 colour images from 118 persons, and corresponding depth maps. We also kept the depth maps generated with our proposed method, to form an extra dataset: **TexasStereo**.

— **Bosphorus** [SAD+08]: a rich dataset in terms of changes in face expressions and pose. It contains 4652 colour images from 105 persons, with corresponding depth maps. Images with large pose variations (i.e. rotations larger than 30°) are not considered in the experiments.

— **FoxFaces** [ADMB16]: a multi-purpose face dataset including colour and depth images from 64 persons, that we designed, as described later in Section 3.1.4. We used two sub-datasets from FoxFaces: **FoxStereo** and **FoxKinect** [1]. The depth maps in FoxStereo are generated with our stereo reconstruction method, and the depth maps in FoxKinect are generated from the Kinect sensor data.

As a result of the first experiment, in Figure 3.15, we display the the recognition rates obtained when varying R, using $1NN$ (i.e. k Nearest Neighbours with $k = 1$) and a 10-fold cross-validation. In this experiment, we used $H = 25$, meaning that 25 local histograms are extracted per map (5×5-grid) and concatenated to form the DLBP descriptor. These results show that, for most collections, when the radius gets large the precision increases, which validates our claim that considering large radius values helps building more discriminative descriptors. The increase is observed mainly for small radius values, indicating a kind of threshold whose

1. The third dataset FoxTOF was used in our experiments because the output is monomodal only.

value varies from 3 to 5, and considering larger neighbourhoods does not bring extra useful depth contrast information.

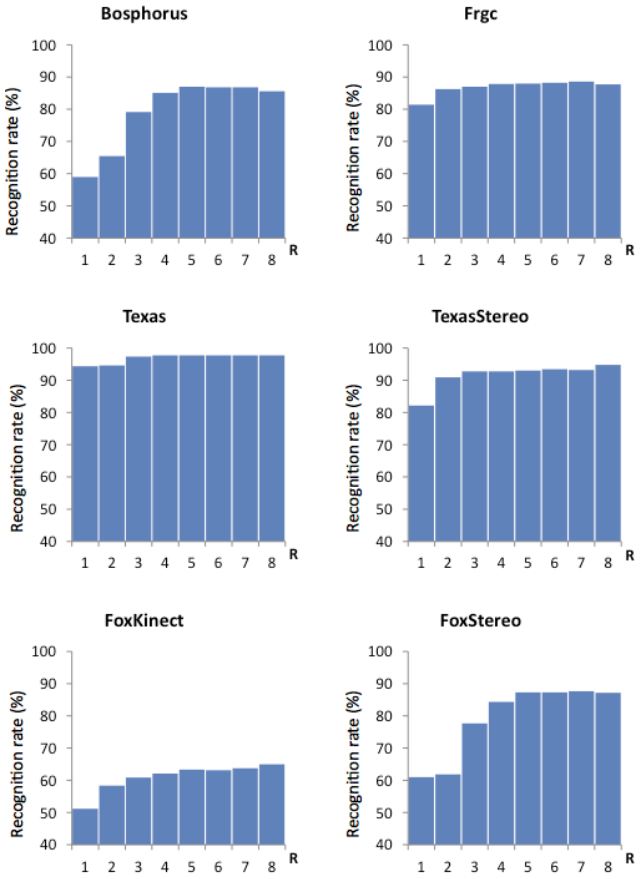


Figure 3.15 – Impact of varying the radius R on the precision of recognition ($H = 25$).

In a second experiment, we compare the recognition rates obtained when varying H, that is to say the number of local regions in the grid used to extract the histograms, with several values of R. Note that the results confirm the previous finding stated above about a possible threshold-like value for R. In Figure 3.16, we see that the number of local histograms has a certain impact on the precision: the larger H, the higher precision. However, the gain in precision tends to dim as H increases. Besides, using higher values of H brings a greater complexity (especially a longer matching time, using 1NN in a higher dimensional space) for a limited gain. Moreover, for all collections, the impact of H fades when R is large. This is explained by the low contrast in face depth images: when R is small, a large number of face points are likely to generate the same code, because in small neighbourhoods, the differences in shape are too small to be captured by the descriptor. Therefore, when using a unique global histogram ($H = 1$), we obtain a high number of similar descriptors distributed over only a few

bins, which does not help distinguishing faces. However, using a larger number of histograms ($H > 1$) when R is small helps to locally preserve spacial information about the location of the descriptors. On the contrary, when R is big, the depth contrast is more easily captured by the descriptors: the codes are less alike, and therefore they are distributed over a larger number of bins, allowing a more precise description, which means a more discriminative descriptor, even when $H = 1$. As a consequence, the difference in precision across the various settings of H is lowered when R is big.

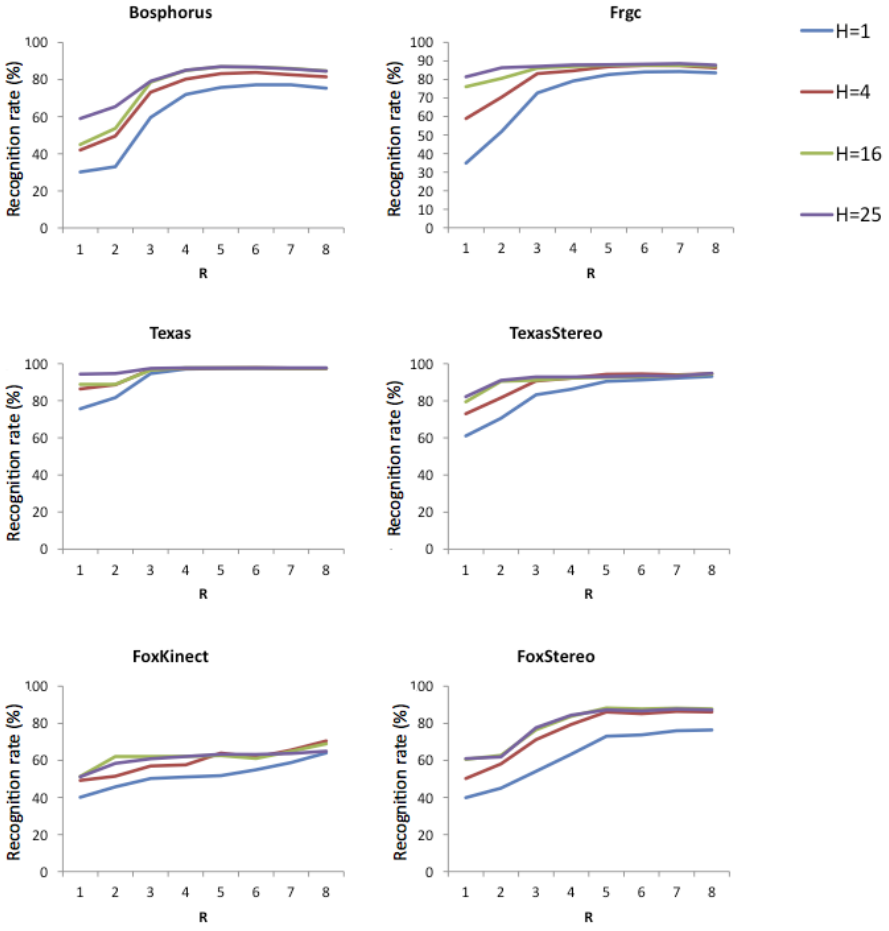


Figure 3.16 – Impact of varying the number H of local histograms to build DLBP vectors.

As a last experiment, Figure 3.17 shows a comparison of the precision of recognition for the standard LBP, the 3DLBP by Yonggang Huang *et al.* [HWT06] with the multi-scale extension proposed by Di Huang [HAWC12], and our proposed DLBP. Experiments were run using $H = 25$. We see that 3DLBP and DLBP generally both yield a higher precision than LBP, which confirms the positive contribution of the magnitude.

We observe a general tendency for 3DLBP and DLBP to give equivalent recognition rates on

Texas, TexasStereo, and FoxStereo datasets. However, for Bosphorus, FRGC, and FoxKinect, DLBP give higher accuracy than 3DLBP when $R > 3$. For LBP and DLBP, we note a slight increase of the precision when R increases. As for 3DLBP, a large scale does not seem to improve the precision because the coding scheme implemented by this descriptor does not exploit the depth contrast in large scales. Indeed, the multi-scale 3DLBP of Di Huang [HAWC12] is extracted in different scales, where the threshold of 7 is used, as in [HWT06]. However, this threshold was determined after a statistical study using $R = 2$, which does not hold for larger values. Therefore, large contrast values in large scales are not fully exploited by 3DLBP. In contrast, the proposed DLBP descriptor succeeds in exploiting the magnitude information with the use of the automatic multi-scale threshold selection. Moreover, the descriptor size for DLBP is half the size of 3DLBP. More details about DLBP and these experiments can be found in our **ICIP'14 publication** [**AMD14**]. Next section provides details about how our approach combines the 2D (visual) and 3D (depth) modalities for recognition.

3.1.3 Two-stage fusion of 2D and 3D modalities

As a last contribution in the proposed bimodal 2D-3D approach for face recognition, this section describes the two-stage fusion. Bimodal 2D-3D face recognition methods make use of both modalities to represent a face, with the objective of taking advantage of both 2D and 3D data in a complementary manner. Indeed, combining both types of data is likely to yield higher results than using a single type separately. When combining modalities, one always has to make the choice of *when* the fusion should take place. Indeed, several merging strategies exist, depending on whether the merging is applied before or after classification [SP02].

— **The early fusion** (fusion of descriptors, before classification) consists in merging descriptors extracted from each modality separately, and the new descriptor is used for classifier training. Another less popular way consists in merging raw data from the sensors to make new data, before extracting descriptors.

— **The late fusion** (fusion of decisions, after classification) comes after that separate classifiers are built for each modality; then the classifiers' outputs are combined.

The choice of the best fusion strategy is important in order to benefit from the complementarity of modalities, and therefore to enhance the results [PFS+05]. However, for most 2D-3D bimodal approaches, this choice is not obvious. Experimental studies were carried out in order to compare several strategies [BG05, LZAL05], and yet the results are not conclusive. While the decision (late) fusion yields better results than the descriptor (early) fusion in the work of Benabdelkader *et al.* [BG05], the opposite result was reported in the work of Li *et al.* [LZAL05]. In general, late fusion seems to be the most popular strategy in bimodal face recognition [BG05, HBGVdM05, GA06, WLCV07, SCLT07, ANRS07a, SHA+10, JCB11] – only a few methods implement early fusion [LZAL05, XLTQ09]. The late fusion can be performed at the score-level, with *class-wise similarity measures* by using merging rules such as a sum, a product, or a weighted sum, or at the decision-level, with *identities* by using a majority vote, or a weighted majority vote [GA06], where confidence values are assigned to classifiers according

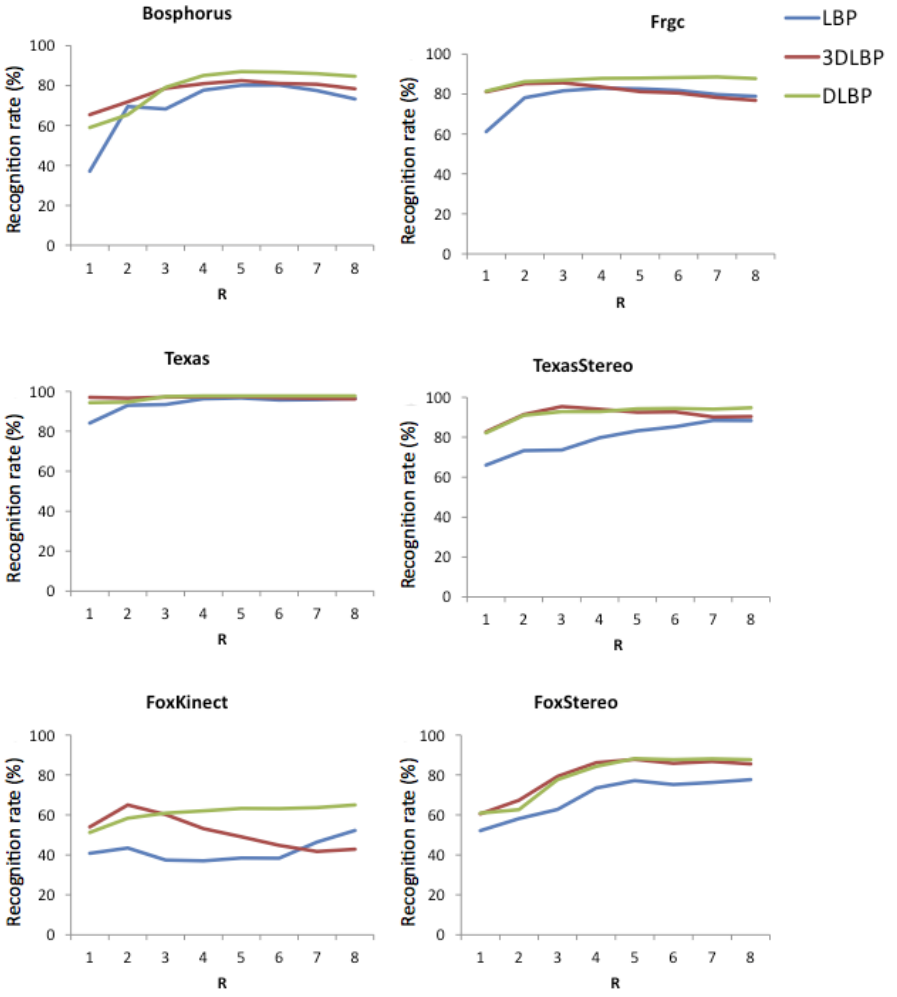


Figure 3.17 – Comparison of LBP, 3DLBP, and DLBP.

their precision for a given decision. In the latter setting, an important issue is that the systems are likely to give results similar to those obtained with a single modality, especially in situations where a given classifier always performs better than the others, and is therefore the unique contributor to the system results. Besides, most late fusion methods consider that modalities are independent and process them independently. However, the independence hypothesis for 2D and 3D face representations is arguable since the data is extracted from the same face. For example, as highlighted by Husken *et al.* [HBGVdM05], the position of fiducial points (eyes, nose tip, etc.) are identical. Late fusion does not allow a synergic processing since each modality is taken separately. On the contrary, it can be argued that early fusion is likely to perform better if the complementarity of 2D-3D face data is thoroughly exploited. In our work, we consider both early and late fusions by designing a two-stage fusion strategy that benefits from both strategies.

Description of the proposed two-stage fusion strategy

We describe below the proposed two-stage fusion strategy, as shown in Figure 3.18. The first

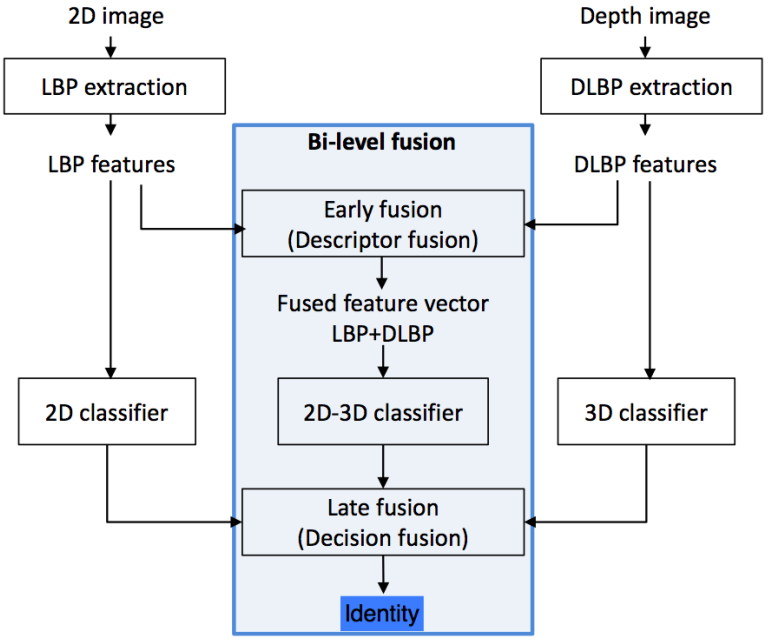


Figure 3.18 – Overview of the proposed two-stage fusion strategy for bimodal 2D-3D face recognition.

step consists in extracting the bimodal face representation made of LBP and DLBP feature vectors, as described previously. Three classifiers are trained:

- The first classifier is trained for the 2D modality, i.e. LBP vectors extracted from grey-level images.
- The second classifier is trained for the 3D modality, i.e. DLBP vectors extracted from depth images.
- The last classifier is trained for both modalities fused with an early fusion strategy, i.e. a new descriptor obtained by concatenating both LBP and DLBP feature vectors.

Classifiers are trained separately and used for test faces identification. In order to identify an unknown face, decisions returned from all classifiers are merged (late fusion) to output the identity, as described below. Given a face represented by a grey-scale image and a depth image, three descriptors D_1, D_2, and D_3 are extracted using LBP, DLBP and their fusion respectively. Three classifiers M_1, M_2, and M_3 trained separately provide three decisions $M_1(D_1)$, $M_2(D_2)$, and $M_3(D_3)$ respectively. We propose to use a decision fusion scheme based on the weighted majority algorithm [XKS92], which is among the most powerful, simplest, and easiest to implement. More formally, let $M_j(D_j) = i$ mean that classifier M_j assigns the descriptor D_j to class i, with $i \in \{1, 2, ..., n\}$, n is the number of classes, $j \in \{1, 2, ..., m\}$, and m is the number

of modalities. In our case, $m = 3$ with 2D, 3D and 2D-3D early fusion. An indicator function F, defined by the following Equation 3.12, is assigned to each classifier:

$$F_i^j(D_j) = \begin{cases} 1 & if \quad M_j(D_j) = i, \\ 0 & otherwise. \end{cases} \tag{3.12}$$

The combination F_i^E of the three classifiers for each class i is hence written as:

$$F_i^E = \sum_{j=1}^{m} \alpha_{ij} F_i^j(D_j) \tag{3.13}$$

The weight α_{ij} represents the reliability of classifier M_j for a given decision (i.e. class i). These weights are given by the classifiers' recognition rates obtained in the training step for each class. The final decision (identity) is given by $argmax_i(F_i^E)$.

Evaluation of the proposed two-stage fusion strategy

We report here experimental results to validate the combination of 2D and 3D data with the proposed two-stage fusion strategy. We used the same collection set as in the previous Section 3.1.2, that is to say Bosphorus, FRGC, Texas, TexasStereo, FoxKinect, and FoxStereo. Faces are located and normalized to 100×100 pixels from grey-level and depth images using the annotations provided by the datasets. We implemented the feature extraction process using different parameters for LBP and DLBP, and we used the settings that gave the best results. In this experiment, Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel are used for classification. Here again, the precision is evaluated with a 10-fold cross validation. The results of monomodal settings (2D, 3D) and early/late/two-stage fusion strategies for the six datasets are presented in Figure 3.19. These results show the following points:

— The **2D monomodal** setting yields a higher precision than the **3D monomodal** setting for Bosphorus, FoxKinect, FoxStereo, and TexasStereo datasets; for FRGC and Texas, however, it is the opposite. For Bosphorus, the 2D modality gives a high precision mainly because there is no light change in this dataset, the visual recognition benefits from favorable conditions. For FoxKinect, FoxStereo, and TexasStereo, the higher results of the 2D modality are explained by the lower quality of the 3D data. As for the higher results for the 3D modality with FRGC and Texas, they are explained by the very large lighting variations in these two datasets, as illustrated in Figure 3.20, resulting in a lower precision for the 2D modality.

— The **descriptor fusion** (early fusion) slightly enhances the system performance for individual modalities for FRGC and Texas datasets, where the 3D modality gives higher results than the 2D modality. This does not hold for the other four datasets (Bosphorus, FoxKinect, FoxStereo, and TextasStereo), where no improvement is observed; on the contrary, the early fusion results brings down the 2D modality. In Bosphorus, for

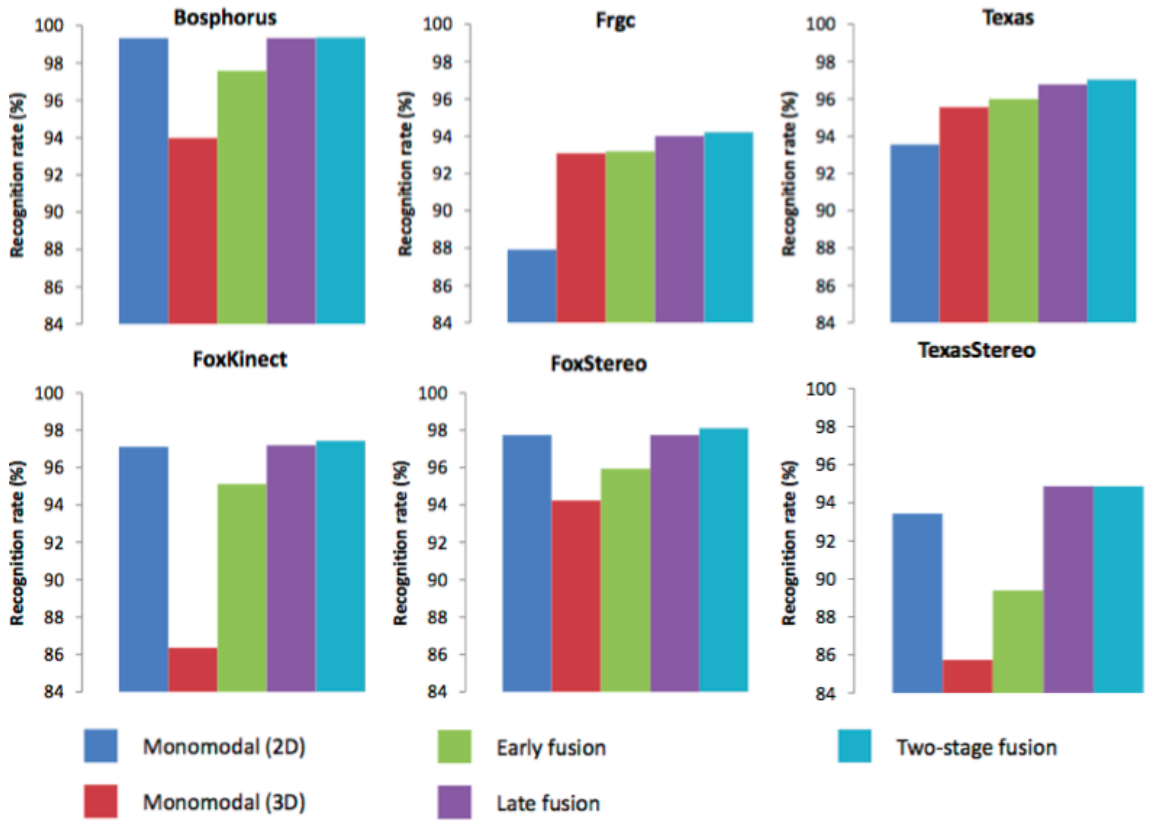


Figure 3.19 – Comparison between several face recognition methods: monomodal 2D, monomodal 3D, early fusion, late fusion, and the proposed two-stage fusion strategy, for the five collections.

instance, the descriptor fusion yields a kind of average value of both modalities taken separately. This is explained by the lower quality of 3D data (especially for FoxKinect, FoxStereo, and TexasStereo), compared to FRGC and Texas, which brought down the results. Indeed, one issue with the early fusion is that merging the descriptors is likely to accumulate possible noise from both modality vectors. As a consequence, if the data from one modality is noisy or low in quality, it impacts the whole final descriptor, decreasing the global system performance. Besides, we also noticed that the high results of early fusion for FRGC and Texas comes from the the *complementarity* of the modalities in such datasets, where the lighting variations are large (Figure 3.20). In such a case, the 3D modality, which is robust to light change, complements the 2D modality.

— The **decision fusion** (late fusion) does either *enhance* or *preserve* monomodal system performances for all collections. Indeed, as stated before, each descriptor is processed separately, and if a descriptor is better than the other for a given class, the corresponding

decision will not be influenced by the other descriptor. This result indicates that in this case of bimodal face recognition, decision fusion is a better option than descriptor fusion, which is consistent with most results found in the literature.

— The proposed **two-stage fusion** turns out to yield the highest recognition rate. It results in a higher precision than those obtained with descriptor fusion or decision fusion for most datasets, and it guarantees a precision identical to decision fusion when no improvement is brought (i.e. for Bosphorus and TexasStereo).

Figure 3.20 – Examples showing some illumination variations in FRGC dataset (top row) and Texas dataset (bottom row).

The experiments carried out with several datasets indicate that the proposed bimodal approach based on the two-stage fusion strategy allows to enhance the face recognition precision. Even when the descriptor fusion does not perform well, no loss in precision is observed. This is due to the important fact that the decisions used in the two-stage fusion (resulting from the 2D modality, the 3D modality, and the descriptor fusion) are considered independently. If one modality performs better than the other for a given class, the corresponding classifier decision will not be influenced by the other. The increased results for the two-stage fusion strategy, however, have to be weighted against the need to train 3 classifiers instead of 1. Indeed, when using this strategy, the off-line training phase requires more time – when using SVM like in our experiments, or even the testing phase – when using a lazy classifier such as kNN.

3.1.4 FoxFaces multi-purpose dataset

In this last part of Section 3.1, we describe the **FoxFaces** dataset, that we collected during the PhDs of Amel Aissaoui and Afifa Dahmane, with the contribution of Ioan Marius Bilasco. The creation of this dataset was motivated by a lack encountered in the existing 3D/4D datasets (4D refers to 3D + time). The proposed dataset features face data (color/stereo/range images, and videos) from 64 subjects, captured with different changes in pose, expression

and illumination. This dataset is unique in two aspects: the acquisition is performed using 3 little-constrained devices allowing to capture 2D, stereo and depth face data. In addition, it contains both still images and videos allowing static and dynamic face analysis. All faces are labeled with gender, facial expression, approximate pose orientation and the coordinates of some manually annotated fiducial points. Hence, our dataset is a useful resource dedicated to researchers in face recognition and analysis. The static and dynamic data can be used for the evaluation of 2D, 3D and bimodal algorithms for face recognition under various conditions, and also facial expression recognition and pose estimation algorithms. We set up an acquisition system composed of 3 sensors:

— An **infrared sensor**: we used Microsoft Kinect, that contains a color camera, an infrared light, and an infrared CMOS sensor (QVGA 320x240, 16 bits). The depth is inferred with structured light, by analyzing a known infrared speckle pattern with triangulation.

— A **time-of-flight sensor**: we used Mesa Imaging SR4000 sensor, that illuminates the scene with a modulated IR light. By measuring the phase change of the reflected signal, the distance can be determined precisely for every pixel in the sensor, creating a 3D depth map of the subject or scene.

— **Stereo cameras**: we used Point Grey Bumblebee XB3, a multi-baseline sensor equipped with three 1.3-megapixel cameras. The large baseline offers a high precision in higher distances from camera, and the small baseline improves close range matching and minimum-range limitations.

Figure 3.21 – Sample images from the 3 sensors. Top row: infrared sensor images from Kinect: (a) color (b) depth. Middle row: time-of-flight sensor images from SR4000: (a) infrared (b) depth (c) confidence matrix (bright pixels mean high confidence). Bottom row: image triple from the stereo camera Bumblebee XB3.

FoxFace was acquired with this 3-sensor system, therefore it contains 3 sub-datasets: FoxKinect (bimodal 2D-3D data), FoxTOF (monomodal depth data), and FoxStereo (bimodal stereo images + estimated depth maps).

Methodology

The data acquisition was carried out indoor, in our office rooms at CRIStAL, with 64 subjects (46 males[1], 18 females) aged 22-59. Subjects are located 1 meter away from the cameras[2]. Three parameters (lighting conditions, face expression, and head pose) are varied throughout the data acquisition. For each subject, 40 images are recorded, corresponding to:

— 3 lighting conditions: ambient, frontal, side;
— 7 face expressions: neutral, joy, sadness, hanger, disgust, fear, surprise;
— 30 head poses resulting from a combination of 9 positions in *yaw* (from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$, using $\frac{\pi}{8}$ steps), with 3 *pitch* directions (downwards, frontal, upwards), plus 2 *roll* positions (left and right).

In total, the collection contains 2560 images. Figure 3.22 shows an example of all possible variations for a subject. In addition to static data (images and depth maps), video sequences of images and depth maps containing all variations are also recorded for each subject.

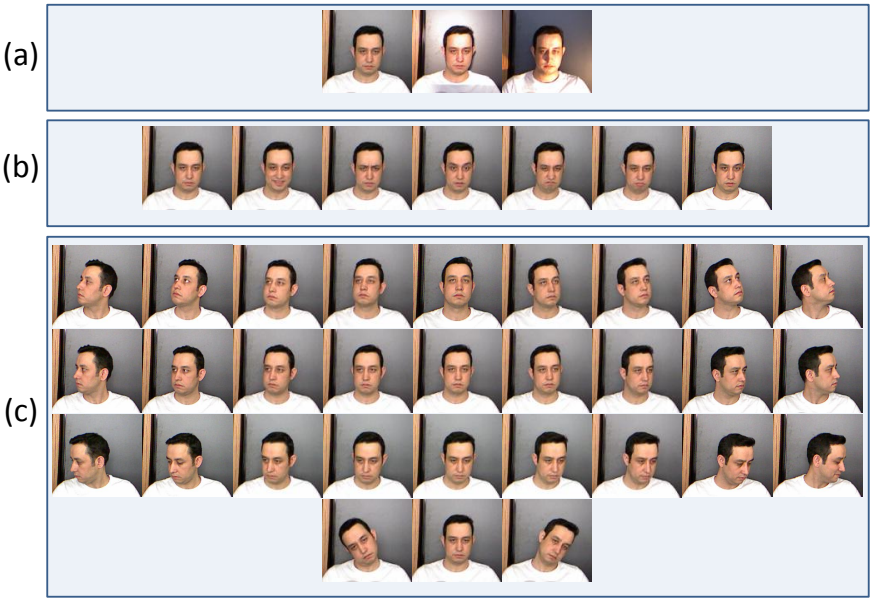


Figure 3.22 – Example of all possible variations for a subject: (a) 3 lighting conditions (b) 7 face expressions (c) 30 head poses.

1. Note that among the 46 male persons, two are twin brothers.
2. It is the minimal distance for the Kinect.

Annotation

A manual annotation was performed for 4 main face interest points (eyes, nose tip, mouth center) in order to enable a precise face localisation. Figure 3.23 shows annotated points on a face from the dataset.

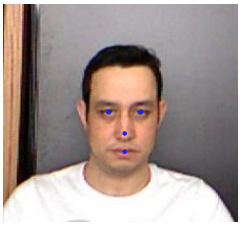


Figure 3.23 – Annotated interest points on a face.

While this dataset is relatively small compared to large scale face datasets such as LFW [HRBLM07, LM14], it still offers a rich evaluation resource for researchers in face recognition using different modalities. In addition to the face recognition task, the available data allows the evaluation of algorithms in a large range of research fields:

— **Face detection**: the annotation information can be used for evaluating face and fiducial points (eyes, nose and mouth) detection methods in both 2D and 3D modalities.
— **3D face reconstruction**: the stereo pairs can be used in order to estimate depth maps of the face using stereo-based reconstruction algorithms. 3D face model reconstruction algorithms based on combining depth maps captured from different point of views can also be applied on this dataset.
— **Head pose estimation**: the dataset is rich in terms of changes in pose. Faces are captured under 30 different poses, therefore data can be used for 2D, 3D and bimodal head pose estimation.
— **Facial expression recognition**: Several expressions for each identity are acquired in the collection. Facial expression recognition methods (2D, 3D or bimodal), can therefore be evaluated using this collection.
— **2D, 3D and 2D-3D bimodal face recognition**: images and depth maps can be used to evaluate 2D (resp. 3D) face recognition methods. They can also be combined for bimodal recognition. Moreover, the recognition performance can be evaluated across changes in pose, expression and illumination.
— **Dynamic face analysis**: When using motion for identification and expression recognition (e.g. using action units), the images and depth sequences provided in the dataset can be used for model training and evaluation in these research areas.

3.2 Dynamic person recognition in TV shows

In this second and last section of Chapter 3, we describe the work from the PhD of Rémi Auguste, who explored the time dimension in person recognition. We adress here the specific case of identifying persons from videos; indeed, since persons are generally central to video contents (e.g. TV shows, YouTube videos, movies, etc.), the ability to identify them is an essential step to analyze, index and organize videos.

A naive way to identify persons in videos is to apply a person recognition algorithm to every single frame of the video. This approach has two drawbacks. First, effective person recognition algorithms are computationally expensive (both in training and prediction). Moreover, these algorithms are usually based on face recognition, since face is one of the most distinctive features of human identity. Face recognition is known to be difficult [ZCPR03b, ANRS07b, ZKM07] when:

— visual conditions (expression, pose, occlusion, lighting conditions, etc.) vary;
— the number of identities to recognize is high.

These two conditions are typically met in large real-world video databases. To overcome these issues, large amounts of training data are required. Therefore, frame-based face recognition is expensive both in computational power and annotated training data, so it cannot scale to very large video databases.

The core idea in the work of Rémi Auguste during the ANR-PERCOL project is to bypass these issues by using the structure and redundancy of video data. In a video, a given person usually appears several times, both in consecutive frames and separated frames spread over the whole video. If these frames can be grouped, then the recognition problem is reduced to classifying groups of frames, which allows to:

— decrease the computational cost by predicting identities only for a few frames within each group;
— increase precision by reducing the data variability, when identities are predicted after selecting only visually uniform frames, in terms of illumination, pose, etc.;
— increase robustness by propagating identities to other frames within the groups, thereby allowing predictions in difficult cases (e.g. non-frontal poses, blurry frames, severe illumination changes, or occlusions). Moreover, a given identity can be assigned based on several frames instead of a single one, which further improves robustness.

In the proposed approach, consecutive frames are grouped using a tracking algorithm; non-consecutive frames are grouped using a re-identification algorithm. These kinds of algorithms can be unsupervised and have shown to be more robust to visual variations than face recognition [BGS14]. For re-identification, we designed and used a novel descriptor, the *Space-Time Histograms* (STH) [AMT15], to represent individuals found in the videos, and to match their occurrences based on the visual appearance. This descriptor, that is the first contribution in this work, allows to build clusters of individuals, as an intermediate step towards recognition.

Once the frames are grouped, persons are recognized using a subset of frames only, then their identities are propagated to the other frames. Experiments show that given a standard

face recognition algorithm, our strategy, that is the second contribution in this work, allows to identify persons in more frames, with a higher precision, while requiring less training data and computational power. The proposed recognition architecture and propagation strategies are depicted in Figure 3.24, and are independent of the subsystems used for tracking, re-identification and recognition.

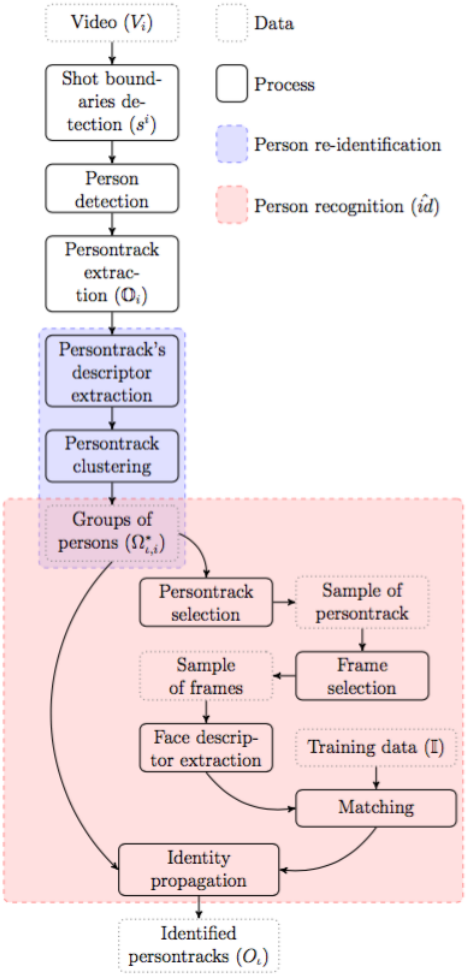


Figure 3.24 – Global illustration of the proposed system for person recognition in video streams.

3.2.1 Problem formulation

Naive approaches to person recognition in videos would simply apply a recognition algorithm to each frame or frame region [1], considered as independent from one another. Formally, the goal of such approaches is to propose a function $\hat{id_f}$ that approximates the ground truth function id_f associating an identity $\iota \in \mathbb{I}$ to every frame or frame region $f \in \mathbb{F}$ featuring a person:

$$\begin{array}{rl} id_f : & \mathbb{F} \to \mathbb{I} \\ & f \to \iota \end{array} \tag{3.14}$$

where $\mathbb{I} = \{\iota_0, \iota_1, \ldots, \iota_{|\mathbb{I}|-1}, \varnothing\}$ is the set of possible identities in the dataset and $\mathbb{F}$ is the set of frames or frame regions featuring a person. The set of identities $\mathbb{I}$ includes the special unknown identity $\varnothing$, which may occur in practice when ground truth labels are incomplete or when recognition cannot be performed. Assuming that frames are independent prevents from leveraging the fact that a given person usually appears multiple times in a video, both in successive and non-successive frames.

By contrast, our approach first defines a video as an ordered sequence of shots, where a shot is a sequence of consecutive frames from the video recorded by a single camera [ZCPR03b]. Therefore, we can formally define a video V_i from of a corpus of videos $\mathbb{V}$ as:

$$V_i = (s_0^i, s_1^i, \ldots, s_{|V_i|-1}^i) \tag{3.15}$$

where s_j^i is the j-th shot of video V_i, and $|V_i|$ is the number of shots in V_i. A shot s_j^i is defined as:

$$s_j^i = (f_t, f_{t+1}, \ldots, f_{t+|s_j^i|-1}) \tag{3.16}$$

where f_t is the first frame of the shot and $|s_j^i|$ is the length (number of frames) of shot s_j^i.

Shots allow us to define spatio-temporal video regions featuring a single person, from which we will be able to leverage temporal information or visual redundancy to perform recognition. Such regions, called *persontracks* [DDLS08], are defined as sequences of frames cropped spatially and temporally around a specific person (see Figure 3.25). Persontracks can typically be extracted by detecting faces or bodies and tracking them over the shot. Therefore, to every video V_i corresponds a set of persontracks $\mathbb{O}_i = \{o_0, o_1, \ldots, o_{|\mathbb{O}_i|-1}\}$. The person recognition problem boils down to finding a function $\hat{id_p}$ approximating the ground truth function id_p that associates every persontrack o to the identity ι of the person it features :

$$\begin{array}{rl} id_p : & \bigcup_i \mathbb{O}_i \to \mathbb{I} \\ & o \to \iota \end{array} \tag{3.17}$$

This process can be divided into three simpler sub-problems:

1. For the sake of simplicity, we will assume in the following that any frame or frame region can contain at most one single person; This assumption is valid without loss of generality, assuming that there exists a way to extract such regions from the raw frames.

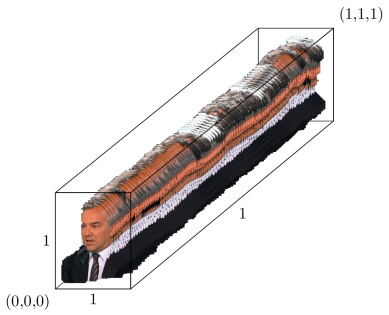


Figure 3.25 – Actual example of a persontrack.

1. Segmenting videos into shots.
2. Extracting persontracks from shots.
3. Associating identities to persontracks.

This formulation of person recognition allows us to increase the efficiency and effectiveness of standard face recognition algorithms by leveraging:

— intra-persontrack information, by identifying only a few well-chosen frames within persontracks;

— inter-persontrack information, by grouping persontracks of a single person in an unsupervised way, and then identifying only a few well-chosen persontracks.

To address this formulation of person recognition, we propose the architecture depicted in Figure 3.24. In this architecture, the first two subproblems (namely, shot segmentation and persontrack extraction) are simply handled using any standard algorithms for shot boundary detection [CNP06], then for face/body detection [YKA02, DWSP12] and for tracking [YS06]. The last subproblem, persontrack identification, is performed in two major stages:

1. Grouping persontracks (in blue in Figure 3.24).
2. Identifying groups of persontracks (in red in Figure 3.24).

Grouping persontracks is performed by the proposed re-identification algorithm [AMT15]. It provides groups of persontracks $\Omega^*_{\iota,i}$ formally defined as follows:

$$\Omega^*_{\iota,i} = \{o \in \mathbb{O}_i | id_p(o) = \iota\} \tag{3.18}$$

Given these groups of persontracks, persontracks can be identified in the second stage following a two-step process in which (1) the persontracks are identified based on their frames, and (2) an identity is assigned to each group based on the identities of the persontracks it contains. In other words, we address in each step the more general problem of assigning an identity (label) to a group (or cluster) according to the identity of its members. In our case, in which determining the identity is costly, the idea is to select a small subset of representative members (called *seed members*), so that identifying them would be sufficient to reliably determine the identity of the group. Therefore, we implement in both cases the following strategy:

1. Select a subset of representative seed members.
2. Generate identity hypotheses based on the seed members.
3. Select the most likely label among the hypotheses.
4. Propagate the selected label to the rest of the group.

This strategy is implemented at both the persontrack level and the persontrack cluster level: an identity is assigned to a persontrack (resp. a cluster) by identifying a subset of seed frames (resp. seed persontracks). As shown in the experiments, this approach increases both the robustness and the precision of identification. Grouping persontracks requires to be able to decide the similarity of two given persontracks. We define and use a dedicated descriptor, the space-time histogram, with a similarity measure for this task.

3.2.2 Space-time histograms

Space-time histograms are an extension of the spatiograms proposed in [BR05], which are themselves an extension of the classic colour histograms. In order to benefit from colour, geometry and motion information from videos, we extended the spatiogram to the temporal dimension. A space-time histogram is typically extracted from the 3D volume of a persontrack (as illustrated in Figure 3.25).

Definition

The data structure of a space-time histogram sth_o built on a persontrack o is defined as:

$$sth_o(b) =< n_b, \mu_b, \Sigma_b >, \quad b = 1, \ldots, B \tag{3.19}$$

where n_b is the number of pixels in bin b and B is the total number of bins. The average position in space and time, μ_b, is defined as:

$$\mu_b = (\bar{x}_b, \bar{y}_b, \bar{t}_b) \tag{3.20}$$

where $\bar{x}_b$, $\bar{y}_b$ and $\bar{t}_b$ are the average normalised positions of the pixels in space and time. Σ_b is the covariance matrix the space-time positions:

$$\Sigma_b = \begin{pmatrix} cov(x_b, x_b) & cov(x_b, y_b) & cov(x_b, t_b) \\ cov(y_b, x_b) & cov(y_b, y_b) & cov(y_b, t_b) \\ cov(t_b, x_b) & cov(t_b, y_b) & cov(t_b, t_b) \end{pmatrix} \tag{3.21}$$

Note that this covariance matrix is symmetric since $cov(a, b) = cov(b, a)$. We see spatiograms as a generalization of histograms; they actually *contain* histograms, since they hold the same pixels counts histograms. In addition, spatiograms hold extra information about the pixels spacial distribution. For the same reason, we see space-time histograms (as defined in Equation 3.19) as a generalization of spatiograms; they also actually *contain* spatiograms, since they hold the same pixels counts and spatial distribution data as spatiograms. In addition, they hold extra information about the pixels distribution over time. If we follow the terminology of [BR05], space-time histograms are seen as *third order spatio-tempo-gram*.

Similarity measure

In order to compare space-time histograms, we designed a similarity measure inspired from the measure used for spatiograms [TCAKD09], that is extended with a temporal dimension. The similarity measure is made of two components.

— The Mahalanobis distance is used to estimate whether space-time histograms come from the same statistical distribution. Since the pixels distributions along x, y, and t are likely to have different variances, in addition to comparing the average values, it is necessary to include the covariance matrix into the estimation. This measure is used to estimate a similarity: for this reason, we use its complement to 1. Let ψ_b be the similarity measure based on the Mahalanobis distance, measuring the similarity of the bins of index b coming from two space-time histograms sth_o and $sth_{o'}$:

$$\psi_b = 1 - \sqrt{(\mu_b - \mu'_b)^t \hat{\Sigma}_b^{-1} (\mu_b - \mu'_b)} \tag{3.22}$$

In Equation 3.22, the covariance matrix $\hat{\Sigma}_b^{-1}$ is estimated using the following formula:

$$\hat{\Sigma}_b^{-1} = (\Sigma_b^{-1} + (\Sigma'_b)^{-1}) \tag{3.23}$$

— The χ^2 distance is used to measure the difference in pixels counts between two bins, with the interesting property that the value is proportional to their sizes. Therefore, large bins with small differences will have little impact on the value. The χ_b^2-based similarity measure (once again, the complement to 1 is used) between two bins of index b is defined as:

$$\chi_b^2(n_b, n'_b) = 1 - \frac{(n_b - n'_b)^2}{n_b + n'_b} \tag{3.24}$$

The similarity measure between two space-time histograms sth_o and $sth_{o'}$ of identical size is defined as:

$$s(sth_o, sth_{o'}) = \sum_{b=1}^{B} \psi_b \times \chi_b^2(n_b, n'_b) \tag{3.25}$$

The combination of Mahalanobis and χ^2 allows to take into account the various aspects of the space-time histograms, that is to say: differences in pixels counts, average values, covariance matrices. We combine them in a product for a given bin, and sum the overall results to obtain the similarity measure. Note that both measures will give values in the interval $[0, 1]$ when space-time histograms are normalized. Their product is therefore also in the same interval. Naturally, the similarity mesure between a space-time histogram and itself is 1. More details regarding space-time histograms and the similarity measure (namely about construction cost, storage space, and matching complexity) can be found in our **ICMR'15 publication** [**AMT15**].

3.2.3 Persontrack clustering for re-identification

The space-time histograms and similarity measure defined in the previous section are used for person re-identification, by comparing space-time histograms extracted from persontracks. Our hypothesis is that the similarity $s(sth_o, sth_{o'})$ is high (close to 1) when $id_p(o) = id_p(o')$, that is to say, when two persontracks contain the same person. On the contrary, this similarity is low (close to 0) when $id_p(o) \neq id_p(o')$. Note that this hypothesis requires that the compared persontracks are taken from a unique video V_v, so that the global appearance of the persons shows minor variations.

The first step in our approach is to build a space-time histogram for each persontrack. Based on all space-time histograms, the similarity measure (Equation 3.25) is used to generate a similarity matrix M from the video V_v, of size $|\mathbb{O}_v| \times |\mathbb{O}_v|$, where $|\mathbb{O}_v|$ is the number of persontracks in V_v:

$$M = \begin{bmatrix} S_{11} & S_{12} & \dots \\ S_{21} & \ddots & \vdots \\ \vdots & \dots & \end{bmatrix} \tag{3.26}$$

where $S_{ij} = s(\text{sth}_{o_i}, \text{sth}_{o_j})$. Because the similarity measure is symmetric, the matrix is also symmetric. The similarity matrix's main diagonal is therefore filled with the value 1. Based on this matrix, the persontracks are then grouped using a hierarchical agglomerative clustering (HAC) algorithm [JD88].

Evaluation metric for the re-identification

The similarity matrix M, that is generated for clustering, is used to evaluate the precision of re-identification. Let us consider the matrix' rows:

$$M = \left[M_1, \dots, M_{|M|}\right]^T \tag{3.27}$$

where each row M_i holds the similarity measures between a space-time histogram sth_{o_i} and all other space-time histograms (including o_i). By sorting each row in descending order of similarity, we define a new matrix R containing, in each row, all persontracks (actual persontracks, not their similarity values) sorted in a decreasing order of similarity:

$$R = \{r_{ij} = o_k | \, \text{rank}(o_k, M_i) = j\} \tag{3.28}$$

where $\text{rank}(o_k, M_i)$ is the rank of $s(\text{sth}_{o_i}, \text{sth}_{o_k})$ in M_i (rank 1 for the highest similarity, rank 2 for the second higher similarity, etc.) Of course, the first value is a self-match: $r_{i1} = o_i$ because $s(\text{sth}_{o_i}, \text{sth}_{o_i}) = 1$ is the highest possible match value.

One problem with the classic average precision measure is that the number of occurrences per identity may vary. Therefore, the identities with a small number of persontracks would contribute too much to the average. For this reason, we use the *precision at n* ($P@N$) [RBJ89],

that is estimated by considering the n first answers, where n is the expected number of answers, as given by the ground truth. In our case, given a persontrack o_i, n_i is the number of persontracks of a video V_v bearing the same identity as o_i:

$$n_i = |id_p^{-1}(id(o_i))|, \forall o_i \in \mathbb{O}_v \tag{3.29}$$

where id_p^{-1} is the inverse function of id_p that gives retrieves the set of persontracks corresponding to a given identity. This precision at n_i for o_i is given by P_i:

$$P_i = \frac{|\{r_{ij} \in R | j \leq n_i\} \cap id_p^{-1}(id_p(o_i))|}{n_i} \tag{3.30}$$

We then calculate the weighted average $\bar{P}$ of the precisions values:

$$\bar{P} = \frac{\sum_{i=1}^{|M|} P_i n_i}{\sum_{i=1}^{|M|} n_i} \tag{3.31}$$

This average is weighted by the number of persontracks by identity. It avoids the introduction of a bias in the final metric.

Comparison with spatiograms and histograms

We report the results of a comparison between space-time histograms, spatiograms, and color histograms. We use a dataset that we designed specifically for this purpose by manually filtering the original REPERE dataset (FoxPersonTracks is described in details the next Section 3.2.7). Space-time histograms, spatiograms and colour histograms are built for all persontracks in the dataset. We start by plotting the evolution of the precision of the 3 descriptors when varying the number of bins, in order to compare their behaviour. We observe in Figure 3.26 that the precisions for all descriptors evolve in parallel. The precision for each approach increases rapidly for a low number of bins (between 10 and 1,000), it reaches a maximum and starts decreasing at around 5,000 bins. The proposed space-time histograms (sth) obtain a better precision than color histograms (h) or spatiograms (sp). The very low *p-value* (under 0.005) of a student-test indicates that this improvement is *very significant*, and garanties the reproducibility of the experiments. The high precision confirms our hypothesis that space-time information is important to re-identify the persons in persontracks. Besides, it is interesting to observe that the precision of spatiograms is almost identical to that of color histograms. This indicates that space information along with color is not better at distinguishing between the persons featured in the persontracks than color alone.

In the next experiment, we compare the *memory cost* for each approach. Color histograms have a memory cost of 1 per bin (pixels count). Spatiograms have a memory cost of 6 and space-time histograms have a memory cost of 9 [AMT15]. Figure 3.27 shows the precision of the different approaches with their memory cost (relatively to color histograms). We observe

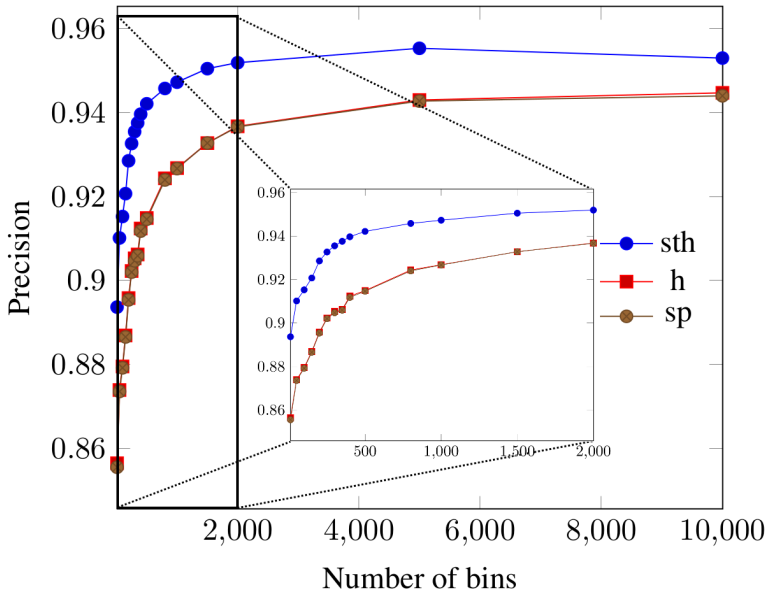


Figure 3.26 – Evolution of the precision as the number of bins in the descriptors increases, from 10 to 10,000 bins. Plots for color histograms and spatiograms are superimposed.

that for a memory cost lower than 4,500, color histograms yield the best precision. For a memory cost of 4,500, space-time histograms and the color histograms give a similar precision. This means that a 500-bin space-time histogram is equivalent in precision to a 4,500-bin color histogram. With a memory cost higher than 4,500, space-time histograms give the highest precision, that keeps increasing steadily, whereas the precision of color histograms decreases slowly. Spatiograms give a much lower precision, relative to their memory cost, than the other approaches. It is only with a memory cost of over 30,000 that the precision of the spatiograms reaches and overtakes that of the color histograms. Therefore a 5,000 bins spatiogram is equivalent in precision to a 30,000 bins color histogram. For memory cost values over 2,000, the precisions of spatiograms and space-time histograms seem to evolve parallel one to another, and the precision of spatiograms remains below that of space-time histograms for all memory cost settings. In summary, the space-time histograms yield a higher precision, for an equivalent memory cost, than color histograms and spatiograms. This result shows again the contribution of temporal and spatial information to re-identify the persons featured in the persontracks.

Given the internal representation of persontracks gathered in clusters according to their (estimated) identity, the next steps towards person recognition consist in:

1. identifying persontracks based on their frames (Section 3.2.4)
2. assigning identities to clusters based on the identity of their members (Section 3.2.5)

In other words, we address in each step the more general problem of assigning an identity/label

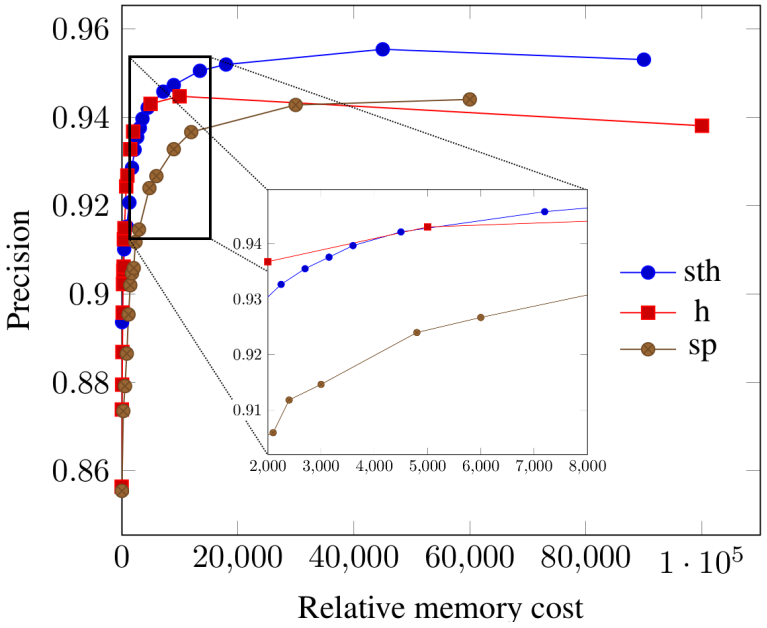


Figure 3.27 – Evolution of the precision as the memory cost increases for each approach.

to a group according to the identity of its members. In our case where determining the identity is costly, the idea is to select a small subset of representative members, so that identifying them would be sufficient to reliably determine the identity of the group. Therefore, we implement in both cases the following strategy:

1. Select a subset of representative *seed members*.
2. Generate identity hypotheses based on the seed members.
3. Select the most likely label among the hypotheses.
4. Propagate the selected label to the rest of the group.

This strategy is implemented at both persontrack and cluster levels: an identity is assigned to a persontrack (resp. a cluster) by identifying a subset of seed frames (resp. seed persontracks). In the next two sections, we discuss several ways to select seeds and propagate identities at each level.

3.2.4 Frame-based persontrack identification

At this level, the objective is to assign identities to persontracks by selecting a subset of representative seed frames, identifying each of them to generate identity hypotheses, then selecting and propagating the most likely identity.

Seed frames selection and identification

Several strategies are defined to select seed frames:
— sample the first n frames of the persontrack;
— sample the last n frames of the persontrack;
— sample the n central frames of the persontrack;
— sample any n frames randomly.
Once the seed frames are selected, they can be identified using any standard face recognition algorithm. This generates a set of k ($k \leq n$) identity hypotheses for the persontrack.

Identity selection and propagation

A voting process is used to select the most likely identity from the hypotheses. It defines the identity $\hat{id}_p(o)$ of a persontrack o, noted ι_o, as the identity found most frequently among its seed frames by the recognition function $\hat{id}_f$. Formally:

$$\iota_o = \arg\max_{\iota \in \mathbb{I}} |\{f_k \in o | \hat{id}_f(f_k) = \iota\}| \tag{3.32}$$

A confidence score $\text{conf}(o, \iota)$ can also be associated to the predicted identity ι_o, computed as the relative frequency of this identity in the persontrack:

$$\text{conf}(o, \iota) = \frac{|\{f_k \in o | \hat{id}_f(f_k) = \iota\}|}{|\{f_k \in o | \hat{id}_f(f_k) \neq \varnothing\}|} \tag{3.33}$$

In this case, the prediction of the unknown identity $\hat{id}_f(f_k) = \varnothing$ happens when the recognition algorithm has rejection capabilities or fails on some frames for some reason (e.g. when it is unable to compute required keypoints). This score provides a measure of the confidence associated to the proposed identity. It can be used for instance to reject predicted identities that have a score below a threshold θ, which can be considered as insufficiently reliable. In particular, using $\theta = 0.5$ implements majority rule. Once the identity is selected, it is assigned to all the frames of the persontrack, thus labeling the persontrack itself.

Evaluation of the persontrack identification: dataset and evaluation measures

We used the annotated data corpus provided for the REPERE challenge [BGK14], which consists of several hours of annotated news videos. The videos were initially broadcasted by two French TV channels: BFMTV and LCP. The dataset features a variety of shows: news, outdoor reportages, public debates and interviews. The videos offer a wide range of lengths, shooting styles, and locations (indoor, outdoor, studio settings, etc.) A part of this dataset focusing on persontracks was released as a standalone benchmark for person re-identification and recognition under the name *FoxPersonTracks* [ATM15], as described in the next Section 3.2.7. This dataset provides a filtered and clean subset of REPERE persontracks along with their

identities. It is composed of 4,604 persontracks (about 170 minutes of video) featuring 266 identities. Each person appears on average in 17 persontracks.

The training data for the person recognition module is based on another subset of the REPERE corpus containing only videos that were not used in the *FoxPersonTracks* dataset, to ensure a valid evaluation. Some persons featured in the *FoxPersonTrack* dataset do not exist in this training set; the corresponding persontracks were removed from the dataset, leaving 2,316 persontracks in the test dataset.

The systems are evaluated based on the number (or ratio) of correct, incorrect or unknown predictions, defined as follows:

— correct: the predicted identity matches the identity provided in the ground truth;
— incorrect: the predicted identity does not match the identity provided in the ground truth;
— unknown: the system could not predict any identity.

From these measures, two standard evaluation measures, precision and recall, are computed, as follows:

— precision (P):

$$P = \frac{\#correct}{\#correct + \#incorrect};$$

— recall (R):

$$R = \frac{\#correct}{\#correct + \#incorrect + \#unknown}.$$

Depending on the context, these evaluation measures are computed either on a frame-by-frame or persontrack-by-persontrack basis: in the remainder of this paper, precision (resp. recall) is noted P_f (resp. R_f) when it is computed on a frame-by-frame basis, and P_p (resp. R_p) when it is computed on a persontrack-by-persontrack basis.

Since the similarity measure provided for STH does not ensure the triangle inequality, computing seed persontracks as cluster centers requires in this case to select persontracks that are the closest to each other within the cluster (as mentioned in Section 3.2.5); this can be done at a limited additional cost since HAC already produces a complete similarity matrix for all persontracks. Table 3.1 provides the baseline results of this re-identification method on the dataset considered.

Purity is 92.1%, i.e. 7.9% of the persontracks fell into the wrong cluster. As a consequence, identification strategies based on intra-cluster propagation, like ours, can achieve at best an accuracy of 92.1% when building clusters using this re-identification method; this accuracy will therefore be the gold standard of our evaluation.

Frame-based person recognition is based on the face detector proposed in [DBID10]. It is based on the following steps:

1. Face normalization (rotation, crop, color conversion).
2. Linearization into a vector of pixel values.

Clustering	Purity	Fragmentation	Clusters number
	92.1%	16.95	562

Table 3.1 – Baseline performance of the re-identification method [AMT15].

3. Training/prediction using an SVM (libSVM).

This algorithm can reject a face (no identity produced) when the face normalization step fails [1].

In this first experiment, only the method for frame-based persontrack identification is evaluated. This is done in two rounds: first by considering a majority vote over all the frames, then by considering a majority vote over subsets of seed frames selected following the strategies described in Section 3.2.4.

Persontrack identification using all frames

Table 3.2 presents the results obtained when all the frames of a persontrack vote for its identity. The proposed approach yields better results in both precision P_f and recall R_f. The large increase in recall can be explained by the fact that many frames that could not be labeled initially (because of their difficult conditions of pose, lighting, etc.) can now be annotated thanks to re-identification and propagation. The increase in precision can be explained in two ways: either the previously unlabelled frames are "easier" to recognize, leading to less errors on average, which is unlikely because the recognition software failed on those frames, or the propagation process allows to compensate for errors made by the recognition software, by eliminating identities that acted as outliers within the persontracks. Comparing precisions of both approaches only on the frames that were initially labeled by the raw recognition approach shows an increase in precision (85.27%), i.e. propagation does compensate for initial annotation errors. P_p and R_p show the performance of the propagation system in terms of labeled persontracks. These measures do not fit raw recognition as it does not use the notion of persontrack. Their values are consistent with P_f and R_f. The precision score is due to initial annotation errors that are propagated within the persontracks. Recall values are both due to annotation errors and to the fact that a number (484, i.e. over 20% of the dataset) of persontracks could not be assigned any identity, because none of their frames could be labeled by the recognition system.

Persontrack identification using a subset of seed frames

Now, subsets of frames are selected using the frame-selection strategies presented in Section 3.2.4: (1) first n frames (2) last n frames, (3) n central frames, and (4) n random frames. Figure 3.28 shows the precision R_p obtained with respect to the ratio of frames used to identify

1. Such failures occur when the algorithm is unable to detect the eyes, which are required for rotation-wise and scale-wise normalizations.

Strategy	P_f	R_f	P_p	R_p
Raw recognition	83.27%	37.89%	N/A	N/A
Majority vote	85.28%	66.06%	84.22%	66.62%

Table 3.2 – Precision and recall of the standard recognition framework and the proposed approach based on intra-persontrack propagation. All the frames of the dataset are used.

the persontracks; as a recall, Table 3.2 presents the scores obtained when using all the frames in the propagation process.

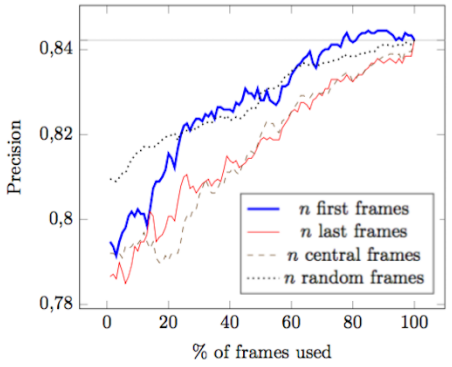


Figure 3.28 – Precision P_p w.r.t. the number of frames used and the strategy considered to identify the persontracks. The score for random sampling is averaged over 100 runs.

Results show that selecting frames at random yields the best results (0.81 to 0.82 precision) when using 1% to 25% of the frames. In the 25%-70% range, random selection is competitive with n-first selection. The latter provides the best performance (over 0.84 precision) in the 70%-100% range. The two other strategies both offer a lower precision in any case.

This difference can be explained by the fact that head poses tend to be non-frontal at the end of many persontracks, whereas recognition algorithms are usually more effective on frontal faces. More specifically, in the TV shows considered here, anchormen occur more frequently than other individuals, and have a specific behaviour, as illustrated in Figure 3.29: during the introduction, they usually face the camera first to catch the attention of the audience, then turn their face towards the next speaker or another camera. Such persontracks represent a large ratio of the whole dataset, so it has a significant effect on the results.

An alternative method to improve the seed frame selection is to use the frames' content to select the most representative ones, such as the most frontal pose (based on geometrical criteria), or select *feyfaces* from the set as suggested by El Khoury *et al.* [KSJ10].

Figure 3.29 – Sample frames taken from a debate show illustrating a specific journalist behaviour: the journalist introduces the subject (frontal pose), then asks a question to one of the guests (non-frontal pose).

3.2.5 Cluster labeling

At this level, the objective is to assign identities to groups of persontracks by selecting a subset of representative seed persontracks, identifying each of them to generate identity hypotheses, then selecting and propagating the most likely identity to all the persontracks belonging to the same group. This is the final step of the proposed person recognition approach, since at the end of this step, an identity is assigned to each cluster, and therefore all persontracks and all their frames are identified, thereby fully defining the functions $\hat{id}_p$ and $\hat{id}_f$.

Selection and identification of seed persontracks

We define three strategies to assign an identity to a cluster of persontracks, based on the identity of its members.

1. **Center of the cluster:** One approach is to select the persontracks of a group that are located near its center. It retains the "average" persontracks, which can be considered as the most representative of the group, putting aside outliers that may be more difficult to identify due to their non-standard properties (lighting conditions, head or body pose, etc.) The center of the group can be simply computed as the centroid of the cluster, when clusters are built based on a similarity measure that ensures triangle inequality. Alternatively, seeds can be computed as the persontracks that are the closest on average to every other member of the group. The computational cost of computing the center (which can be expensive, especially in the second case) can be reduced by making use of the intra-cluster distances that were already computed during re-identification.
2. **Confidence value:** The identities with the highest confidence values $\text{conf}(o, \iota)$ (see Eq.(3.33)), ideally above the threshold θ (typically, $\theta = 0.5$), are selected. The advantage of this approach is that only the most reliable persontracks are used to decide the identity of the cluster. The drawback is that it is preferable to assign an identity to all persontracks to be able to keep only the actual best ones.
3. **Random selection:** Another simple strategy is to randomly select the seed persontracks. This approach is computationally inexpensive, however there is no guarantee

that selected persontracks are not outliers. Such a strategy would typically yield good results when the clusters are compact enough.

Seed persontracks can then be identified using the procedure described in the previous Section 3.2.4 to generate a set of identity hypotheses to be assigned to the persontrack group.

Identity selection and propagation

Here again, a voting process is used to select the most likely identity among the identity hypotheses. The identity ι_Ω of a group Ω is given by:

$$\iota_\Omega = \arg\max_{\iota \in \mathbb{I}} |\{o \in \Omega | \hat{id}_p(o) = \iota\}| \tag{3.34}$$

The identity $\hat{id}_p(o)$ may be unknown ($\varnothing$), if o could not be identified. A confidence score $\text{conf}(\Omega, \iota)$ can also be assigned to the predicted identity:

$$\text{conf}(\Omega, \iota) = \frac{|\{o \in \Omega | \hat{id}_p(o) = \iota\}|}{|\{o \in \Omega | \hat{id}_p(o) \neq \varnothing\}|} \tag{3.35}$$

Once the identity is selected, it is assigned to all the persontracks within the group, thus labeling the group itself.

Evaluation of the cluster labeling: using all the persontracks

In this second experiment, we evaluate the second step of our approach: person identification based on identity propagation over groups of persontracks. To do so, the individual persontracks are identified using a voting process based on all their frames (which yields the previous precision P_p of 84.22% given in Table 3.2). Here again, the proposed approach is evaluated in two rounds: one round in which all persontracks within each group are considered as seed persontracks, and another round in which only a subset of seed persontracks are selected using the strategies presented in Section 3.2.5.

The results of the first evaluation round are presented in Table 3.3. They show that propagation improves both precision and recall. The increase in recall is high, mostly because propagation allows to annotate many persontracks that were initially labeled as unknown. This is the case for persontracks in which no frame could be identified, but that could be assigned the dominent identity of their cluster. After propagation, only 5 persontracks are still unknown: they fell into clusters containing only unknown persontracks. Here again, the increase in precision can be explained in two ways: either initially unknown persontracks are easier to recognize or propagation compensates for initial labeling errors. Among the 1,832 persontracks that can be annotated without propagation, 1,543 (84.22%) are correctly annotated before propagation and 1718 (93.78%) after propagation; it confirms that the increase in precision is mostly due to the ability of the propagation technique to fix initially incorrect labels.

	Correct	Incorrect	Unknown	P_p	R_p
Before	1,543	289	484	84.22%	66.62%
After	2,001	310	5	86.59%	86.40%

Table 3.3 – Results before and after propagating the identity of the persontracks. Propagation is based on majority vote within the clusters.

Using a subset of seed persontracks

Figure 3.30 presents the precision obtained with respect to the ratio of seed persontracks used to propagate identities. The persontracks are selected using the strategies presented in Section 3.2.5. Here again, the precision of the *random strategy* is averaged over 100 runs.

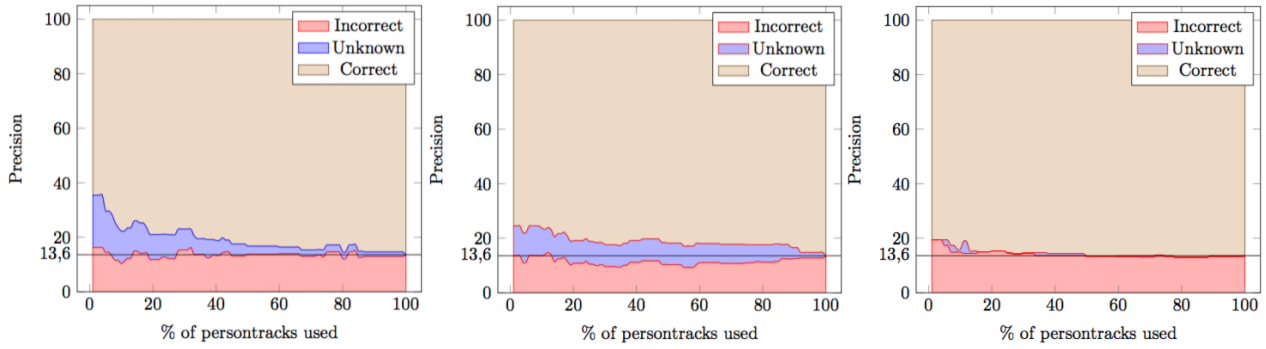


Figure 3.30 – Precision of recognition P_p w.r.t. the ratio of seed persontracks used to identify the persontracks, using (left) random selection (center) similarity-based selection and, (right) confidence score-based selection.

Random selection of seed persontracks (Figure 3.30-(a)) results in a steady ratio of incorrect identifications, about 13%, and a ratio of unknowns (i.e., clusters containing no identified seed) that diminishes as the number of seed persontracks increases, from 19.17% to 0.22%.

Similarity-based selection makes the number of unknowns smaller, and decrease slowly, from 10.88% to 0.22%. It allows the number of incorrectly recognized persontracks to decrease as the ratio of seed persontracks increases (from 13.6% to 9.28%, when using 55% of seed persontracks), then increase again to reach its final value. When using just 1% of seed persontracks, the ratio of correct identifications (75.52%) is higher than previously with the random selection (64.59%).

Finally, the approach based on confidence scores offers the most steady results: the ratios of correct, incorrect and unknown persontracks remain constant after about 15% of seed persontracks are used; before this, the number of incorrectly recognized persontracks decreases slightly, from 19.34% to about 14%.

In conclusion, using confidence scores yields the best results, especially as it limits significantly the number of unknowns; however, it should be noted that it requires to have reliable

confidence scores, which may not be the case when the number of seed frames used to identify seed persontracks is low. Using similarity-based selection can also be interesting as it can reduce the number of incorrects, to some extent: it shows better rejection capabilities that may be useful when precision is preferred over recall. It should be noted that both these strategies perform better than a simple random selection.

3.2.6 What can be done when using only 0.1% of the frames?

In this last experiment, the two stages of the proposed architecture are combined: intra-persontrack propagation and inter-persontrack propagation. To demonstrate the effectiveness of the proposed method, only 1% of frames and 1% of persontracks are used. It should be noted that the inter-persontrack propagation method based on confidence scores was not used here, to avoid the issue of unreliable confidence scores, as mentioned in the previous section.

Table 3.4 presents the performance of the system for every combination of an intra-persontrack and inter-persontrack propagation technique. The results are consistent with the observations from the previous experiments: random sampling works best for frame selection (intra-persontrack propagation), and similarity-based selection works best for persontrack selection (inter-persontrack propagation): all correct scores are over 70% vs. 65.98% at best, and unknown scores are 10.88%, vs 19.57% at best. The main result here is that the proposed method is able to recognize persontracks with a 84.26% precision and a 75.09% recall while performing actual face recognition on only a very small subset of the initial frames to be recognized (about 0.1%). Both scores are better than what the initial recognition system was able to offer; in particular, recall is significantly improved, showing that the proposed approach is effective to label frames that offered too bad conditions (pose, lighting, etc.) to be recognized by a conventional system. Finally, it should also be noted that random frame selection reaches results that are very close to those obtained using all the frames (75.09% vs. 75.52%): since using 1% of seed frames or using all the frames of the persontracks can yield similar results, a simple way to increase the system precision is to label more persontracks rather than individual frames.

3.2.7 FoxPersonTracks benchmark for re-identification

We describe in this section a dataset dedicated to the training and evaluation of methods for person re-identification in TV broadcast shows [ATM15]: FoxPersonTracks. The original data used in this work was distributed as part of the REPERE challenge. During the challenge, 299 videos were released, covering 9 TV shows originally broadcasted by the French TV channels LCP and BFMTV. This data came together with manually obtained ground truth for speech recognition and speaker recognition over the whole data, and OCR and face detection over a limited number of keyframes. The dataset described here was produced based on 134 videos [1]

1. This corresponds to all videos available during Phase 0 and Phase 1 of the challenge.

Persontracks selection	Frame Selection	Correct	Incorrect	Unknown
Random	n first	63.29%	16.51%	20.21%
	n last	62.35%	18.08%	19.58%
	n central	63.43%	16.72%	19.85%
	n random	65.98%	14.44%	19.57%
Similarity	n first	73.70%	15.41%	10.88%
	n last	71.76%	17.36%	10.88%
	n central	70.08%	19.04%	10.88%
	n random	75.09%	14.03%	10.88%

Table 3.4 – Comparison of the propagation results of each proposed strategy combinaison using only 1% of the frames of 1% the persontracks of each group.

from the REPERE dataset and their associated ground truth.

The main purpose of building this dataset is to evaluate re-identification algorithms, such as the approach presented in previous sections. A baseline for face and upper body detection and segmentation provides a common ground to compare the absolute performance of various algorithms. The proposed dataset is composed of:

— 4,604 persontracks from various TV shows, each featuring one of 266 identities;
— ground-truth data providing the full name of the person for each shot;
— ground-truth data providing the face position for a subset of 2081 keyframes extracted from the shots;
— baseline face detections for all frames of the shots;
— baseline detections of faces;
— baseline background subtraction data based on the detected face;
— evaluation software to compute the metrics.

All the data referred to as *ground truth* was obtained manually, and the data referred to as *baseline* was obtained automatically.

Persontracks extraction

The process used to generate the persontracks from the original videos is illustrated in Figure 3.31. First, the REPERE ground truth is used to extract sequences of consecutive frames annotated with names. The ground truth directly provides the first and last frames of such sequences together with the name of the person appearing in it. Two or more of these sequences may overlap, meaning that more than one person appear in the sequences. Also, some persons appearing in the sequence may not be annotated (e.g. distant persons in an audience) according to annotation guidelines. The remaining steps of this process aim at removing these sequences to provide filtered set of persontracks containing only a single person.

Persontracks are extracted from videos by using Viola and Jones face detector [VJ02],

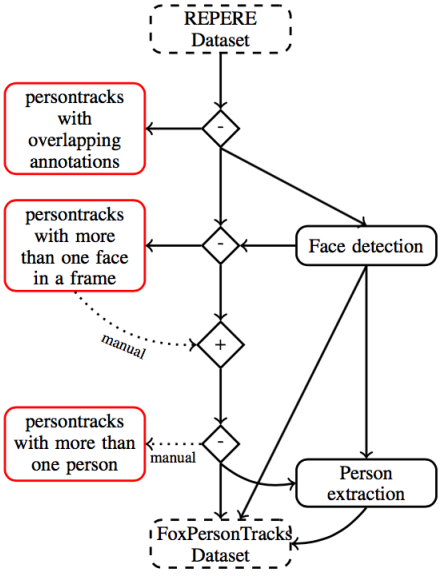


Figure 3.31 – Steps for building FoxPersonTracks dataset from the REPERE dataset.

whose results are optimised to maximize the proportion of skin colour in the detection. This helps eliminating most of the background from the detection when the person is not perfectly facing the camera. The estimated face position and size are used to initialize a set of masks provided to Grabcut algorithm [RKB04] in order to separate the person from the background (see Figure 3.32-left). The masks are initialised from ellipses calculated from the detected face (see Figure 3.32-center). The extracted person's image is then resized to a fixed size. This size was empirically deduced from our experiments, we noticed that using 50% of the original frame size was good to fit most of the occurrences without resizing them. Figure 3.32-right shows the result of the Grabcut algorithm applied to the original image using the mask. A normalization step is applied so that the persons' position is centred and the number of pixels in each frame is normalized. Normalising the illumination was not required in our approach since the illumination is stable enough throughout a TV show. Persontracks are built by stacking consecutive detections over time, so as to make a 3D volume (2D+time), such as shown in Figure 3.25. The depth of the volume is the temporal dimension, which is also normalized. After the persontrack extraction, we obtain short videos, each centred on one person, with a clean dark background (the pixel value is 0).

Post-processing and manual filtering

Overlapping sequences are then removed from the dataset, based on the ground-truth information. After this second step, the only sequences that may contain more than one person are the ones that were not fully annotated (in other words: sequences showing more than

Figure 3.32 – Example of person extraction process on a single frame and a single person. Left: Original frame with the detected face. Center: Mask calculated from the detected face. Right: Result of the Grabcut algorithm applied on the original image using the mask.

one person). To deal with such sequences, we manually inspect all frames in which the face detector spotted more than one face. If the frame actually contains more than one face, the whole corresponding sequence is discarded. After this third step, the only remaining sequences that may contain more than one person are those where the ground truth was incomplete and the face detector failed. All these sequences are once again manually inspected and every sequence containing more than one person is discarded. This last filtering step ensures the accuracy and quality of the dataset. At the end of the filtering, we obtain 4,604 persontracks of 266 different persons.

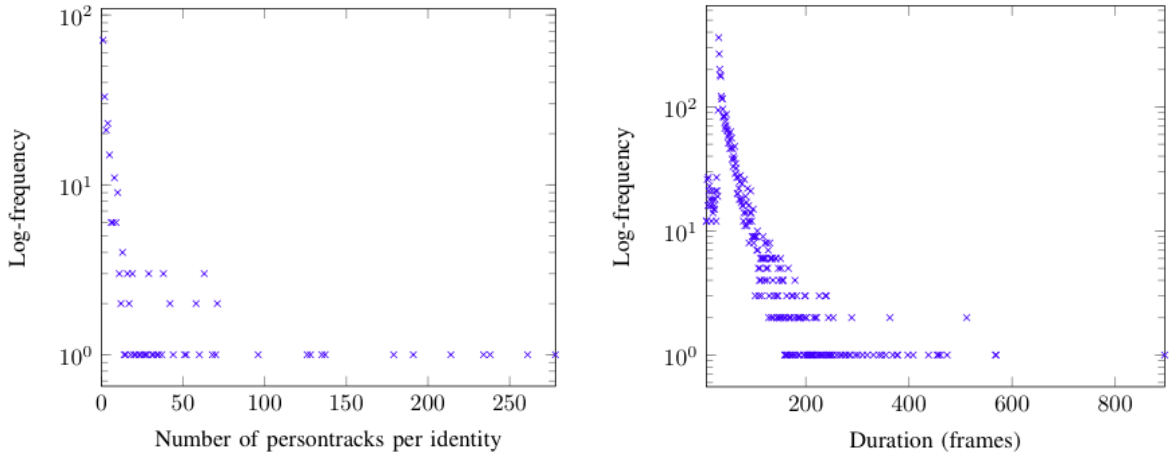
Dataset statistics

In average, there are 5 individuals per video (min 1, max 12), each individual appears in 1.2 TV shows among 9 (min 1, max 4), and each identity appears in 1.7 videos (min 1, max 15). Figure 3.33-(a) shows the occurrence distribution in a log-scale. Since the original videos are broadcast TV shows, the journalists (anchorpersons) appear more frequently than any other person. Some journalists appear in more than 50 shows, while many other persons appear only once. Figure 3.33-(b) shows the persontrack length distribution in the dataset. It shows that most persontracks are short in length, with an average of 55 frames. Note that as the peak shows, one third of the persontracks has a length between 31 and 38 frames.

As compared to existing benchmarks, our dataset:

— covers an original yet significant use case of re-identification: TV broadcast shows;
— provides a large number of tracks (over 4,600) and a large number of individuals to be re-identified (266);
— provides full video shots rather than individual images, allowing to evaluate motion-based methods in addition to visual-based ones;
— is based on real-world data: excerpts from actual French TV broadcast shows.

Other related tasks that can be evaluated using this dataset includes: person identification from video, face detection from video, relative impact of external components on the final re-identification performance (face detection, upper body detection, face/upper body segmen-



(a) (b)

Figure 3.33 – (a) Number of occurrences per identity distribution in the dataset ranging from 1 to 278 (x-axis) with an average of 17. (b) Duration in frames distribution of our datasets ranging from 7 frames to 896 frames (x-axis) with an average of 55 frames. Note that the y-axis is logarithmic.

tation).

3.3 Conclusion

This chapter gives a description of our contributions in the domain of person recognition. We explored two directions during the PhD of Amel Aissaoui and Rémi Auguste: the use of depth and the use of time.

The first part describes a 2D-3D bimodal recognition approach with the following main contributions: (1) a stereo reconstruction method that uses an ASM to obtain a set of high-confidence depth points, that is used to bound the depth estimates for all other points, (2) an original descriptor (DLBP) designed to represent and compare depth faces, in a multi-scale manner in order to capture the low depth contrast in such data, and (3) a two-stage fusion strategy, that combines advantages of early and late fusions.

The second part describes a dynamic person recognition approach in TV shows, that is based on re-identification. The approach is grounded on the use of a novel descriptor (Space-Time Histograms) with a dedicated similarity measure in order to cluster similar persontracks. This descriptor helps capturing the space-time distribution of the pixels in persontracks, offering a higher clustering precision. Using a state-of-the-art frame-based face recognition algorithm, several strategies are defined to identify persontracks based on their frames, and to label clusters based on the identity of their members. Because the complexity of the proposed approach is reasonably low, it can be used to annotate very large collections, e.g. french INA

archives, in a reasonable time.

These contributions originate from the work of Amel and Rémi, and several colleagues were also great contributors, such as Ioan Marius Bilasco, Afifa Dahmane, Taner Danişman, Pierre Tirilly, and Tarek Yahiaoui (former postdoc of the team, now full-time R&D engineer for ANAXA-VIDA). The selection of referenced papers related to this chapter include 1 journal paper and 12 conference papers between 2012 and 2016. This chapter ends the presentation of our work.

Chapter 4

Conclusion

This document presents a summary of my research activities from 2002 to 2016. I summarize below the contributions presented throughout this document, that are directed towards defining and validating advanced image features for image representation.

4.1 Summary of contributions

Chapter 2 is dedicated to image representations. In the first part, Section 2.1, we introduced the edge context descriptor (as a part of Ismail's PhD), that brings a rich context to SURF descriptors, and yields over 30% relative gain over SURF on Caltech-101 dataset. We also described two alternative approaches to KMeans-based vocabularies. The first alternative is an information-gain-based selection of visual words among a large set of randomly selected features. This representation brought an 8% increase of the retrieval score on UKB dataset (in addition to being computationally much simpler) compared to the KMeans baseline. The second alternative is a split representation for mobile image search, where a query wise bag of words is used to build compact representation of the query to be matched against a reference vocabulary. By using as little as a third of the query descriptors, we achieved similar and even sometimes higher results than the baseline on Paris and Oxford datasets, thereby reducing the amount of data that should be transmitted to the remote server for matching. The final discussion in this section investigates a possible relation between word distributions and system' effectiveness. The study presented here indicates that higher retrieval results can be obtained when the visual words follow a distribution similar to text words, possibly because applying text techniques to visual words only makes sense in such situations. This initial work calls into question the very paradigm of visual words.

In the second part, Section 2.2, we discussed the general notion of relation. At a low level, the use of visual phrases (made of frequently occurring patterns of words) for image representation brings a refined representation, inspired from text phrases. A combination of words and phrases proved a higher accuracy for most classes of Caltech-101 dataset. At a transversal

level, a cross-modal annotation model can benefit from exploiting the relations between visual and textual modalities. At a high level, we described our work for integrating relations between image objects in the vector space model, by using conceptual graphs. The objective is to combine the efficiency of the vector space model with the expressive power of the rich knowledge representation formalism of conceptual graphs. Experiments on two image collections showed an increase in precision (respectively +27% and +9% of relative gain) when including the relations, compared to using the bare object labels. Moreover, the proposed approach enables (1) a much faster matching – by 3 orders of magnitude, than using the projection operator for conceptual graph matching (isomorphism with exponential complexity), and (2) a flexible matching allowing to retrieve and rank partially relevant documents, wich is not possible with graph matching.

In the third part, Section 2.3, four models for weighting image parts are discussed. The four models follow an increasing level of granularity: visual words, image regions, image objects, and star graphs. We start with a spatial weighting scheme for the visual words, that integrates the spatial distribution of words in the image. The use of this weighting scheme with a vocabulary based on the edge context descriptor yields retrieval results with almost 60% relative gain over a standard vocabulary with binary weights on Caltech-101 dataset. The second model is a gaze-based weighting scheme for image regions, that is based on users' perception. This model was originally designed to assess the quality of advertising videos. It gives more importance to most seen regions, and an image search system could benefit from using such weights. The two last models consider the geometrical properties of image objects to define a weight for objects and star graphs. After a thorough validation of hypotheses regarding geometrical criteria, the weights are defined according to the size and position of objects, and also the image homogeneity. This model helps to accurately rank the retrieved images according to users' perception of relevance.

Chapter 3 is dedicated to person recognition. The objective is to address the limitations of *static 2D* approaches (i.e. those based on a face picture), and we described two directions for enhancing the accuracy and robustness of such approaches. In the first part, Section 3.1, we explored the use of depth for face recognition during Amel's PhD by introducing a 2D-3D bimodal face recognition approach that combines visual and depth features. The approach includes a stereoscopic reconstruction method that successfully benefits from the use of an active shape model to guide the depth map estimation, reducing by over 32% the root mean squared error of reconstruction to compared to block-matching methods, and in the same time reducing the processing time by an order of magnitude compared to graph-cut methods. The second contribution is a new descriptor dedicated to face depth maps: DLBP, whose originality is to operate in large neighbourhoods to better capture the low depth contrast of face. This descriptor achieves higher recognition rates than the original LBP and than 3DLBP in most settings. Finally, a two-stage fusion strategy combines the 2D and 3D modalities, and allows to benefit from early and late fusion schemes. In our experiments, this strategy always produced the highest recognition rates with the 6 datasets used for validation. Besides, during the PhD of

Amel and Afifa, we created a multi-purpose face dataset (FoxFaces) by collecting face images and videos from 64 subjects, with 3 different types of visual and depth sensor. This dataset is publicly and freely available for academic purposes.

In the second part, Section 3.2, we explored the use of time for person recognition during Rémi's PhD, in the context of an ANR project (PERCOL) that is part of the REPERE challenge targeting person spotting and naming from TV shows. The key idea in this approach is to take advantage of the variation of people's appearance over time. The proposed approach is based on re-identification, which is the fundamental task consisting in finding the occurrences of a given individual across shots of a single video or across various videos. In a first contribution, a new descriptor and its similarity measure are defined to represent and match persontracks (occurrences of individuals): the space-time histogram. It consists of an extension of color histograms and spatiograms to take into account the space-time distribution of pixels in the 3D volume of a video sequence. Experiments on REPERE dataset show that they allow a higher precision in persontrack similarity matching. Once the persontracks are organized into clusters, in the second contribution, we define several strategies to label persontracks based on their frames, and also to label clusters based on their members. A state-of-the-art face recognition algorithm is used to identify individuals in carefully selected seed frames from a persontrack, and the dominent identity is propagated to the entire persontrack with a voting method. In a similar way, the dominent identity found in carefully selected persontracks among a cluster are propagated to the entire cluster. We show in the experiments that one can successfully annotate over 84% of the persontracks by performing actual face recognition on a very small subset of the frame set (about 0.1%). Besides, we designed and made publicly available a re-identification benchmark, FoxPersonTracks, that is free for academic purposes. It is composed of a manually-filtered subset of REPERE dataset, containing 4,604 persontracks (about 170 minutes of video) showing 266 individuals. The dataset comes with evaluation metrics and tools to easily compare systems' results. Finally, we highlight that the first part of this work (the space-time histogram and its similarity measure) was implemented in PERCOL system during REPERE challenge. This system was built by the project consortium to perform a multimodal person clustering. The integration in the system of our persontracks similarity matrices resulted in a sharp decrease of the Estimated Global Error Rate (EGER) [KGQ+12], from 41.1 to 32.8, that is to say over 20% of relative error reduction.

When coming back to the issues of inter-class similarity and intra-class variability that we mentioned in Chapter 1, the objective of both DLBP and space-time histograms is to bring more discriminance to the systems. Indeed, by helping to distinguish persons who look alike in 2D static images, they contribute to reduce the inter-class similarity. Besides, within their respective context of use, they help smoothing the intra-class variability by bringing more invariance to the systems.

4.2 Impact

Gathering all results and writing up this habilitation thesis has been a good opportunity for a self assessment. I believe that it is always a wise move to take a step back and watch the path one has come, and put it into perspective with the current state of the domain, and the current directions that one follows. When publishing research, it is interesting to have a kind of feedback regarding one's impact. Table 4.1 shows citations' distribution at the time of writing, for each topic presented in this document. This table allows to compare the respective impact of the presented topics. The impact of my research in image representations (Chapter 2) is clearly higher than the one of person recognition (Chapter 3), not only because the research is older, but also because the number of related publication is higher. The most cited journal paper is the IPM journal paper [MCM11] that describes the relational vector space model using an advanced weighting scheme for image retrieval, that follows on from my PhD works. If we consider only publications after my PhD work, the most cited journal paper is the MTAP journal paper [ESMUD12] that describes the edge context descriptor, the spatial weighting scheme for visual words, and the combination of visual words into visual phrases. This reflects the fact that relations between descriptors and weighting schemes for images are central among my contributions. Note that regarding the dynamic person recognition topic, I did not include the citations (24) resulting from 4 consortium publications describing PERCOL system, in which our approach is implemented.

Topic	# publications	# citations	# citations per year
Visual vocabularies	6	39	10.33
Relations	13	83	12.67
Weights	16	111	16.08
2D-3D face recognition	7	8	4.25
Dynamic person recognition	6	9	3.75

Table 4.1 – Number of publications per topic (related publications), number of citations and average number of citations per year as of October 2016. Source: Google Scholar.

Finally, since a large part of the contributions presented in this document result from the PhDs of Ismail El Sayad, Amel Aissaoui, and Rémi Auguste, I am happy to share that, at the time of writing, Ismail El Sayad is an Assistant Professor at the Lebanese International University in Lebanon, Amel Aissaoui is an Assistant Professor at USTHB in Algeria, and Rémi Auguste founded a web service company providing an online collaborative video editor: Weaverize [1], in which he serves as a CEO.

1. URL: http://www.weaverize.com/en/.

4.3 Future work

Twelve years after defending my PhD in Grenoble, the backbone of my research interests is still made of image representation features, with several variations. Regarding the first part of my work, a large inspiration comes from the link between text and vision, including for visual vocabularies, relations between descriptors, and weighting models. It has been a core motivation in my research from the beginning. It is also with no doubt the most impactful part. One general direction of my research is to continue studying this link. This direction is consistent with the orientations taken by the new European Network on Integrating Vision and Language (iV&L Net [1]). Regarding person recognition, a natural yet challenging long term goal would be to explore the use of depth+time. Also, it would be interesting to explore how the proposed descriptors (DLBP, HST) would perform in other tasks than what they were originally made for (e.g. face expression recognition for DLBP, action or gesture recognition for HST, etc.). Besides, because of the specificity of persons and faces among general objects, dedicated representation models are developed for them. At a meta-level, a general methodology consists in investigating how to apply to general images the lessons learnt from person recognition, and vice versa; the idea is that each domain/topic can learn from the other. For instance, the use of the active shape model in the stereoscopic face reconstruction step (Section 3.1.1) allowed to generate higher quality depth maps while speeding up the reconstruction process. This was possible because a specific model for faces could provide constraints to guide the process. Since the ASM is a general statistical model that can be designed for e.g. medical images or mechanical assemblies, it would be relevant to apply the same methodology to such images, taking advantage of their own specificity.

In October 2015, two students started a PhD under my co-direction. Jalila Filali started a PhD at RIADI lab (ENSI, Tunisia) under the direction of Dr. Hajer Baazaoui and myself, and Cagan Arlsan started a PhD at CRIStAL under the direction of Prof. Laurent Grisoni (MINT research team, focus on gestural interaction) and myself.

With Jalila, we explore how textual ontologies and visual features can be combined to create a rich annotation model for image representation. One general direction is to study how the link between textual features (coming from an ontology) and visual features can be learnt, so that the annotation model is able to generate a structured image description. Regarding the visual features, we will explore strategies and models to learn features from the pixels (such as the HMAX model by Riesenhuber and Poggio [RP99]), rather than using a variation of the bag of words. Machine learning, that was long used in tasks such as classification, is now increasingly used to learn features, as a general trend. The idea is to let the system decide what feature is worth learning directly from the pixels. This way, we avoid using hand-crafted features that always require effort and expertise to tune, and show little ability to generalize well.

1. iV&L Net is implemented as the ICT COST Action IC1307, funded by the European Science Foundation, since 2014. URL http://ivl-net.eu.

In Cagan's work, we aim to design a captation environment to enable multimodal interaction by combining touch and in-air gestures in a seamless way. As for the computer vision aspects, we wish to explore features dedicated for this context, in an unconstrained environment. The idea is to design a flexible captation environment to analyze gesture with no specific device such as Kinect, Leap Motion (short range sensor for hand and finger motion tracking), or OptiTrack (high-capture-rate camera for motion capture). A scenario would be to use a set of standard RGB cameras (e.g. a laptop webcam and smartphone camera capturing a given scene, placed at free locations) to extract useful gesture information to allow such a multimodal interaction with the system. Several issues such as the system calibration, the definition of dedicated features, the fusion model, and the share of processing between the device and the central server (similarly to the discussion in the split representation of mobile image search in Section 2.1.2) will need to be addressed.

Finally, one long term direction of our research team (FOX) is to explore the use of spiking neural networks (SNN) in computer vision. Such type of neuro-inspired network differs in several aspects from multilayer perceptrons based on convolutional neural networks (CNN) used in deep learning. Contrary to CNN, spiking neurons do not fire at each propagation cycle, but rather fire only when their activation level (or membrane potential, an intrinsic quality of the neuron related to its membrane electrical charge) reaches a specific threshold value. When a neuron fires, it generates a non binary signal which travels to other neurons, which in turn increases their potentials. The activation level, modeled with differential equations, either increases with incoming spikes, or decays over time. Instead of relying on stochastic gradient descent and backpropagation for training such networks, the Spike-Timing-Dependent Plasticity (STDP) is used to adjust the weights (strength of connections) between neurons, based on the relative timing of a particular neuron's output and input spikes. Therefore, the network is asynchronous and the time matters in SNN, unlike in CNN. For example, when training an SNN, the order, duration, and frequency of sample feeds play an important role.

In order to use SNN in computer vision, a number of issues and questions need to be addressed, such as the design of SNN architectures, the understanding and control of the learning process, coding the input (grey-level pixel values to spikes) and decoding the output (interpret the output layer' spikes), to cite but a few of them. Since 2013, we are exploring this direction, in the context of an IRCICA-supported project dedicated to neuro-inspired information processing approaches. Other participants to the project develop SNN simulation software, and also investigate neuromorphic hardware implementations [1]. This project fosters a *change of paradigm* in information processing, that is motivated by several factors, including the analysis that current techniques are too energy-consuming to scale and meet the needs of the future. Indeed, we seem to have entered a new era where data has become *big* and therefore learning has become *deep* [2]. The research community witnesses a race towards the deepest

1. Spiking-network-based architectures have shown great potential as a solution for realizing ultra-low power consumption using spike-based neuromorphic hardware.

2. I wrote "therefore" because deep learning methods require tremendous amounts of data, and they can be

system that learns from the biggest data. Hawkins says Google's attitude is: "lots of data makes up for everything"[1]. Such deep learning systems now achieve the highest performance in a wide range of benchmarks, such as the object recognition [WYS+15] and face recognition [LDB+15] systems from Baidu that both achieve the highest performance at the time of writing on the reference datasets ImageNet (objects) and LFW (faces).

However, we believe that the energy and environment impact of big data (from production to media consumption, via transmission and processing) will be a major challenge in the next decades, despite the little information available regarding this issue. To give a few examples, the total amount of numerical data produced and stored in the world is expanding fast. This amount was estimated to be 4.4 ZB[2] in 2013, and is expected to reach 40 ZB by 2020 (which all the same is a good news for deep learning methods). In 2012, *cloud computing* already represented the fifth "country" for electrical consumption, after USA, China, Russia, and Japan [GI12]. A year later, in 2013, the world's ICT ecosystem was estimated to use about 1,500 TWh of electricity annually, equal to all the electric generation of Japan and Germany combined [Mil13]. This is about 10% of world electricity generation. DARPA Synapse project uses 10^{13} synapses (which is equivalent to a cat brain), with 2MW of electrical consumption, while a cat brains consumes about 2W. When considering the major milestones in the domain in the last two decades (local descriptors in the 1995', bag of words in the 2000', big data and deep learning in the 2010'), we believe that such a change of paradigm is likely to be the next major change in the 2020'.

successful only if such amounts of data are available.

1. MIT Technology Review, URL: https://www.technologyreview.com/s/513696/deep-learning/.

2. 1 zettabyte is 10^{21} bytes, equivalent of a 300,000 km high stack of DVD, which is the Earth-Moon distance.

Chapter 5

Curriculum Vitæ

5.1 General details

First name:	MARTINET
Last name:	JEAN
Status	Associate professor (CNU 27)
Institution	CRIStAL – IUT "A" – University of Lille
Birth date and place	11/11/1978 in Mopti (MALI)
Post address	CAMPUS Haute-Borne CNRS IRCICA-IRI-RMN Parc Scientifique de la Haute Borne 50 Avenue Halley – BP 70478 – 59658 Villeneuve d'Ascq Cedex
Email address	jean.martinet@univ-lille1.fr
Web page and phone	www.cristal-univ-lille.fr/∼martinej/, +33-(0)-65969-1191

5.2 Career

Since dec. 2007 :	**Associate professor CNU 27**, Image and video indexing. CRIStAL (UMR Lille 1/CNRS 9189) – IUT "A" Lille 1 University **Team leader of FOX research team**: Since March 2014. Holder of a *Bonus for Scientific Excellence* (PES B): 2010–2014 Holder of a *Bonus for Ph.D. and Research Supervising* (PEDR B): 2014–2018 6-month research leave (CRCT): Feb. 2015–July 2015
Dec. 2005 - nov. 2007 :	**Postdoctoral researcher (24 months)** National Institute of Informatics, Tokyo, Japan.
Sept. 2004 - aug 2005 :	**Teaching and Research Assistant (ATER)**, CLIPS-IMAG laboratory (UMR CNRS/UJF/INPG 5524) Pierre-Mendès-France University, Grenoble III.
Oct. 2001 - dec. 2004 :	**PhD in Computer Science**, **A relational vector space model adapted to images** CLIPS-IMAG laboratory, Joseph Fourier University Grenoble I (Government grant) Advisors: Pr. Yves Chiaramella, Dr. Philippe Mulhem, Reviewers: Pr. J.-M. Pinon, Pr. M. Boughanem, President of jury: Pr. C. Garbay.

5.3 Research directions

My research activities in FOX [1] research group in CRIStAL [2] are related to multimedia indexing/retrieval and computer vision. The main objectif is the efficient representation of visual contents (images, videos), in order to allow a content-based search. My works are more specifically directed towards two axis:

- image representations for indexing and retrieval: bag of words, relations between descriptors, weighting schemes for image parts,
- person recognition by exporing the use of depth and time.

The results of these works are generally implemented in applicative domains such as content-based access to image and video documents, image and video understanding.

5.4 Teaching activities since 2013

I give the details of my teaching activities since 2013, amounting to an average volume of 230 hours/year.

- **Object-oriented programming and design (IUT M2103, M2104, and M3105)**: Advanced course of object-oriented design and programming, design patterns, Java language. DUT level, 1st year (72 hours/semester, with an average of 25 students/semester) and 2nd year (32 hours/semester, with an average of 90 students/semester). Period: 2013–now.
- **Human-Computer Interfaces (IUT M2105)**: Introduction to HCI, event-based programming, MVC, Swing library, JavaFX. DUT level, 1st year (32 hours/semester, with 115 students in 2013/2014, 20 continuous education students in 2015/2016, 17 continuous education students in 2016-2017). Period: 2013–now.
- **Mobile programming (IUT M4104C)**: Design and development of mobile applications. DUT level, 2nd year (18 hours/semester with 14 continuous education students in 2013-2014, 24 students in 2015-2016, 16 continuous education students in 2016-2017, and 101 students in 2016-2017). Period: 2013–now.
- **Responsible for student projects**: Coordination of students project (free small-scale software development by groups of 2-3). DUT level, 1st year (16 hours/semester with an average of 50 students/semester. Period: 2010–2014.
- **Pattern recognition**: Extraction, representation and coding of attributes, statistical methods – decision theory, multidimensional data classification, syntactic methods – strings, trees, languages. Master Image, Vision, Interaction (IVI) of Computer Science/Automatic and Electrical Systems, 1st year: (16 hours/semester, with an average of 20 students/semester). Period: 2009-now.

The period starts in 2013 from the application of the new *National Educational Program* (Programme Pédagogique National). I participated in the elaboration of this program during several national meetings, namely regarding object-oriented programming and design (meetings in Dijon, 2011 and Orléans, 2012), and student projects (meeting in Bordeaux, 2012).

5.5 PhD students supervision

Current

- **Mr Cagan Arslan (Lille 1)** (50%, with Pr. Laurent Grisoni):
  - "Data fusion for man-machine interaction"

1. Fouille et indexation de dOcuments compleXes et multimedia
2. Centre de Recherche en Informatique, Signal et Automatique de Lille

— Government grant (Oct. 2015 – now).

— **Miss Jalila Filali (ENSI Tunis)** (50%, with Dr. Hajer Baazaoui):
  — "Image retrieval by learning ontologies and visual features"
  — Government grant (Tunisia). (Oct. 2015 – now).

Past

— **Mr Rémi Auguste (Lille 1)** (75%, with Pr. Chaabane Djeraba):
  — "Dynamic person recognition in audio-visual TV shows"
  — ANR-PERCOL project funding (ANR-DGA REPERE challenge). (Nov. 2010 – July 2014).
  — $\Longrightarrow$ Rémi has created and is CEO of a collaborative video editing platform (www.weaverize.com).

— **Miss Amel Aissaoui (Lille 1)** (75%, with Pr. Chaabane Djeraba):
  — "Bimodal face recognition by merging visual and depth features"
  — Government funding (Algeria). (Sept. 2010 – June 2014).
  — $\Longrightarrow$ Amel is now Assistant Professor at University of Sciences and Technology HOUARI BOUMEDIENE, Bab Ezzouar, Algeria.

— **Mr Ismail El Sayad (Lille 1)** (75%, with Pr. Chaabane Djeraba):
  — "A higher-level visual representation for semantic learning in image databases"
  — Government funding (Algeria). (July 2008 – Dec. 2011).
  — $\Longrightarrow$ Ismail is now Assistant Professor at Lebanese International University (LIU), Beirut, Lebanon.

5.6 Research projects, relation with industry

— **PERCOL (nov 2010 – jun 2014, ANR-10-CORD-0102)** – PERson reCOgnition in audiovisuaL content
  — Context of the national REPERE challenge targeting person spotting and naming from TV shows.
  — I was leader of the WP "Video analysis for person recognition".
  — Partners: France Telecom (Orange Labs Lannion), LIF (Aix-Marseille University) – project leader, LIA (University of Avignon and the Vaucluse), LIFL (Lille 1 University).
  — Total project grant: 397 800 euros.
  — Lille 1 University grant: 96 620 euros (funded 2 years of Rémi Auguste PhD).
  — URL: http://www.agence-nationale-recherche.fr/?Project=ANR-10-CORD-0102
— **TWIRL (jun 2012 - oct 2014, ITEA 2 Call 5 10029)** – Twinning virtual World (on-line) Information with Real world (off-Line) data sources
  — Ioan Marius Bilasco and I were task leader for two tasks: "Mobile video processing for location and people recognition", and "Dissemination activities and project website".
  — Partners from 3 countries: Cassidian Cybersecurity (France) – project leader, Pertimm (France), Ipernity (France), Mondeca (France), Lille 1 University (France), Telecom SudParis (France), TelecomParisTech (France), Smartsoft (Turkey), Tilda (Turkey), Tmob (Turkey), Siveco (Romania), Altfactor (Romania).
  — Total project grant: 5,762 Keuros.
  — Lille 1 University grant: 245 Keuro (funded a total of 45 person.month for 4 engineers/postdocs).
  — ULR: https://itea3.org/project/twirl.html and http://twirl.lifl.fr/wiki/index.php/Main_Page

5.7 Other activities and contributions

— **Workshop organization**

1. International Workshop on Multimedia Analysis of User Behaviour and Interactions (MAUBI), satellite event of the IEEE International Symposium on Multimedia (ISM'08). One-day event, 4 accepted papers. 16 participants.
2. International Workshop on Multimodal Interactions Analysis of Users a Controlled Environment (MIAUCE), satellite event of the International Conference on Multimodal Interfaces (ICMI'08). 6 accepted papers. 25 participants.

— **International journal review** ACM Transactions on Information Systems 2008 (TOIS 2008), IEEE Transactions on Multimedia 2011, International Journal of Computer Applications 2011 (IJCA 2011), Annals of telecommunications 2012, Multimedia Tools and Applications (14 papers between 2009 and 2016).

— **International conference review** for over 30 conferences between 2006 and 2016, including ACM MM'06, IEEE ICME'07, ECIR'08, ICPR'2012, CORIA'13, CORIA'14, VISAPP'14, ICPR'14, MMM'15, ACM MM'16, ICPR'16.

— **Organizing committee** of CORIA'05 (Grenoble) and CORESA'12 (Lille): participation to local organization and logistics.

— **Co-organizer** of SIMIE'11 ("Special Session on Simulation and Interaction in Intelligent Environments" – satellite event of the International Conference on Pervasive and Embedded Computing and Communication Systems (PECS'11), Vilamoura, Portugal), March 5-7, 2011.

— **Guest editor** of the ITE Transactions on Media Technology and Applications, special issue on "Multimedia Content Analysis", 2013.

— **Member of Selection Committee (COS Lille 1)**, 2012.

— **Technical expertise for an agreement committee** for a start-up creation at Chambre de Commerce et d'Industrie Grand Lille, March 2012.

— **Member of the Management Committee** of ICT COST Action IC1307 – European Network on Integrating Vision and Language (iV&L Net), since January 2014.

— **Reviewer and PhD examination board member** for the PhD of Francis Deboeverie, Ghent University - Vision Systems/IPI/TELIN/iMinds, 2014.

5.8 Selected publications between 2008 and 2016

International journals

1. **Improving Retrieval Framework using Information Gain Models**. Huu Ton Le, Syntyche Gbehounou, Francois Lecellier, Thierry Urruty, Jean Martinet, Christine Fernandez-Maloigne. Accepted in Journal of Signal, Image and Video Processing. 2016. (Impact Factor 2014=1.430).
2. **Boosting gender recognition performance with a fuzzy inference system**. Taner Danisman, Ioan Marius Bilasco, Jean Martinet. Expert Syst. Appl. 42(5): 2772-2784. 2015. (Impact Factor 2015=2.240).
3. **Rapid and accurate face depth estimation in passive stereo systems**. Amel Aissaoui, Jean Martinet, and Chabane Djeraba. In: Multimedia Tools and Applications (Jan. 2013). (Impact Factor 2013=1.058). URL: http://hal.inria.fr/hal-00834474.
4. **Intelligent Pixels of Interest Selection with Application to Facial Expression Recognition using Multilayer Perceptron**. Taner Danisman, Marius Bilasco, Jean Martinet, and Chaabane Djer-

aba. Signal Processing, pp. 1547-1556, Jan. 2013. (Impact Factor 2013=2.238) URL: http://hal.inria.fr/hal-00804171.

5. **Toward a higher-level visual representation for content-based image retrieval**. Ismail El Sayad, Jean Martinet, Thierry Urruty, and Chabane Djeraba, Multimedia Tools and Applications, pp. 455-482, Jan. 2012. (Impact Factor 2012=1.014).
6. **A relational vector space model using an advanced weighting scheme for image retrieval**. Jean Martinet, Yves Chiaramella, and Philippe Mulhem. Information Processing and Management, pp. 391-414, May 2011. (Impact Factor 2011=1.119). URL: http://hal.inria.fr/hal-00730563.
7. **Media objects for user-centered similarity matching**. Jean Martinet, Yves Chiaramella, Philippe Mulhem, Shin'ichi Satoh, Multimedia Tools and Applications Journal, Special Issue on "Semantic Multimedia", 2008.

International and national conferences/workshops

2016

1. **Introducing FoxFaces: a 3-in-1 head dataset**. Amel Aissaoui, Afifa Dahmane, Jean Martinet, Marius Bilasco. 11th International Conference on Computer Vision Theory and Applications (VISAPP 2016), volume 4, pp. 533-537, February 2016, Rome, Italy.
2. **Towards visual vocabulary and ontology-based image retrieval system**. Jalila Filali, Hajer Baazaoui Zghal, Jean Martinet. 8th International Conference on Agents and Artificial Intelligence (ICAART 2016), volume 2, pp. 560-565, February 2016, Rome, Italy.

2015

1. **Space-time Histograms And Their Application To Person Re-identification In TV Shows**. Rémi Auguste, Jean Martinet, Pierre Tirilly. 5th ACM International Conference on Multimedia Retrieval (ICMR 2015), pp. -, June 2015, Shanghai, China.
2. **Bimodal 2D-3D face recognition using a two-stage fusion strategy**. Amel Aissaoui, Jean Martinet. 5th International Conference on Image Processing Theory, Tools and Applications (IPTA 2015), pp. 279-284, November 2015, Orléans, France.
3. **Introducing FoxPersonTracks: a Benchmark for Person Re-Identification from TV Broadcast Shows**. Rémi Auguste, Pierre Tirilly, Jean Martinet. 13th International Workshop on Content-Based Multimedia Indexing (CBMI 2015), pp. -, June 2015, Prague, Czech Republic.
4. **Pruning near-duplicate images for mobile landmark identification: a graph theoretical approach**. Taner Dansiman, Jean Martinet, Marius Bilasco. 13th International Workshop on Content-Based Multimedia Indexing (CBMI 2015), pp. -, June 2015, Prague, Czech Republic.
5. **Identification de personnes dans des flux multimédia**. Frederic Bechet, Meriem Bendris, Delphine Charlet, Geraldine Damnati, Benoit Favre, Mickael Rouvier, Rémi Auguste, Benjamin Bigot, Richard Dufour, Corinne Fredouille, Georges Linarès, Jean Martinet, Gregory Senay, Pierre Tirilly. COnférence en Recherche d'Information et Applications (CORIA 2015), pp.-, March 2015, Paris, France.
6. **Bi-modal face recognition – How combining 2D and 3D clues can increase the precision**. Amel Aissaoui, Jean Martinet. 10e International Conference on Computer Vision Theory and Applications (VISAPP 2015), pp. -, March 2015, Berlin, Germany.

2014

1. **DLBP: a novel descriptor for depth image based face recognition**. Amel Aissaoui, Jean Martinet, Chaabane Djeraba. IEEE International Conference on Image Processing (ICIP 2014), October 2014, Paris, France.
2. **Elementary Block Extraction for Mobile Image Search**. Jose Mennesson, Pierre Tirilly, Jean Martinet. IEEE International Conference on Image Processing (ICIP 2014), October 2014, Paris, France.
3. **Iterative Random Visual Word Selection**. Thierry Urruty, Syntyche Gbehounou, Huu Ton Le, Jean Martinet, Christine Fernandez. 4th International Conference on Multimedia Retrieval (ICMR 2014), pp. 249, April 2014, Glasgow, Scotland.
4. **Multimodal understanding for person recognition in video broadcasts** Frederic Bechet, Meriem Bendris, Delphine Charlet, Geraldine Damnati, Benoit Favre, Mickael Rouvier, Rémi Auguste, Benjamin Bigot, Richard Dufour, Corinne Fredouille, Georges Linarès, Jean Martinet Gregory Senay, Pierre Tirilly. 15th Annual Conference of International Speech Communication Association (Interspeech 2014), Singapore, September 2014.
5. **From text vocabularies to visual vocabularies: what basis?** Jean Martinet. 9e International Conference on Computer Vision Theory and Applications (VISAPP 2014), pp. 668-675, January 2014, Lisbon, Portugal.

2013

1. **Human-centered region selection and weighting for image retrieval**. Jean Martinet. 8e International Conference on Computer Vision Theory and Applications (VISAPP 2013), pp. 729-734, Barcelona, Spain. Feb. 2013. URL: http://hal.inria.fr/hal-00812320.
2. **Unsupervised Face Identification in TV Content using Audio- Visual Sources**. Meriem Bendris, Benoit Favre, Delphine Charlet, Géraldine Damnati, Rémi Auguste, Jean Martinet, and Gregory Senay. 11th International Workshop on Content-Based Multimedia Indexing (CBMI 2013). June 2013.
3. **PERCOLI: a person identification system for the 2013 REPERE challenge**. Benoit Favre, Geraldine Damnati, Frederic Bechet, Meriem Bendris, Delphine Charlet, Rémi Auguste, Stéphane Ayache, Benjamin Bigot, Alexandre Delteil, Richard Dufour, Corinne Fredouille, Georges Linarès, Jean Martinet, Gregory Senay, Pierre Tirilly. First Workshop on Speech, Language and Audio in Multimedia, InterSpeech satellite event, August 2013, Marseille, France. URL: http://hal.inria.fr/hal-00812334.
4. **Human-centered region selection and weighting for image retrieval**. Jean Martinet. 8e International Conference on Computer Vision Theory and Applications (VISAPP 2013), pp. 729-734, February 2013, Barcelona, Spain.

2012

1. **Ré-identification de personnes dans les journaux télévisés basée sur les Histogrammes spatio-temporels**. Auguste, Rémi and Aissaoui, Amel and Martinet, Jean and Djeraba, Chabane. 12e Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC 2012), pp. 547-548, Janvier 2012, Bordeaux, France.
2. **Percol0 - un système multimodal de détection de personnes dans des documents vidéo**. Frédéric Bechet, Rémi Auguste, Stéphane Ayache, Delphine Charlet, Géraldine Damnati, Benoit Favre, Corinne Fredouille, Christophe Levy, Georges Linarès, Jean Martinet. JEP-TALN-RECITAL 2012, pp. 553-560. 4-8 juin 2012, Grenoble, France. 2012. URL: http://hal.inria.fr/hal-00812159.
3. **Construction de masques faciaux pour améliorer la reconnaissance d'expressions**. Taner Danisman, Ioan Marius Bilasco, Jean Martinet, and Chabane Djeraba. In: COmpression et REpresentation des Signaux Audiovisuels (CORESA). May 2012.

4. **Les histogrammes spatio-temporels pour la ré-identification de personnes dans les journaux télévisés**. Rémi Auguste, Amel Aissaoui, Jean Martinet, and Chabane Djeraba. In: Compression et Représentation des Signaux Audiovisuels (CORESA). May 2012.
5. **Fast Stereo Matching Method based on Optimized Correlation Algorithm for Face Depth Estimation**. Aissaoui, Amel and Auguste, Rémi and Yahiaoui, Tarek and Martinet, Jean and Djeraba, Chabane. 7e International Conference on Computer Vision Theory and Applications (VISAPP 2012), pp. 377-380, February 2012, Rome, Italy.
6. **3D face reconstruction in a binocular passive stereoscopic system using face properties**. Aissaoui, Amel and Martinet, Jean and Djeraba, Chabane. IEEE International Conference on Image Processing (ICIP 2012), pp. 1789-1792. Oct. 3, 2012, Orlando, Florida, U.S.A. URL: http://hal.inria.fr/hal-00812273.
7. **Reconstruction 3D de visages dans un système de stéréovision basée sur les propriétés du visage**. Amel Aissaoui, Rémi Auguste, Jean Martinet, and Chabane Djeraba. In: Compression et Représentation des Signaux Audiovisuels (CORESA). May 2012.

2011

1. **A Semantically Significant Visual Representation For Social Image Retrieval**. Ismail El Sayad, Jean Martinet, Thierry Urruty, Samir Amir, and Chabane Djeraba. In IEEE International Conference on Multimedia and Expo (IEEE ICME'11), pages 1-6, 2011. URL: http://hal. inria.fr/hal-00812291.
2. **A semantic higher-level visual representation for object recognition**. Ismail El Sayad, Jean Martinet, Thierry Urruty, Samir Amir, and Chabane Djeraba. Proceedings of the 17th international conference on Advances in multimedia modeling (I) (MMM'11), Springer LNCS 6523, Pages 251-261, 2011. URL: http://hal.inria.fr/hal-00812293.

2010

1. **Visual Sentence-Phrase-Based Document Representation for Effective and Efficient Content-Based Image Retrieval**. Ismail Elsayad, Jean Martinet, Thierry Urruty, Chabane Djeraba, 10ième Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC'2010), pp. 157-162. 26-29 January 2010, Hammamet, Tunisia. 2010. URL: http://hal.inria.fr/hal-00730579.
2. **A New Spatial Weighting Scheme for Bag-of-Visual-Words, Ismail Elsayad**, Ismail Elsayad, Jean Martinet, Thierry Urruty, Chabane Djeraba, pp. 1-6, CBMI'2010. 2010. URL: http://hal.inria.fr/hal-00730581.
3. **Semantics for intelligent delivery of multimedia content**. Marius Bilasco, Samir Amir, Patrick Blandin, Chabane Djeraba, Juhani Laitakari, Jean Martinet, Eduardo Martinez Gracia, Daniel Pakkala, Mika Rautiainen, Mika Ylianttila, and Jiehan Zhou. In: Proceedings of the International Symposium On Applied Computing. Mar. 2010, URL: http://hal.inria. fr/hal-00730593.
4. **Effective object-based image retrieval using higher-level visual representation**. Ismail El Sayad, Jean Martinet, Thierry Urruty, Samir Amir, and Chabane Djeraba. In IEEE International Conference on Machine and Web Intelligence (IEEE ICMWI), pages 218-224, 2010.
5. **Using association rules and spatial weighting for an effective content based- image retrieval**. Ismail El Sayad, Jean Martinet, Thierry Urruty, Taner Danisman, Md. Haidar Sharif, and Chabane Djeraba. In: VISAPP 2010. Jan. 2010, URL: http://hal.inria.fr/hal-00730580.
6. **Toward a higher-level visual representation for content-based image retrieval**. Ismail El Sayad, Jean Martinet, Thierry Urruty, Samir Amir, and Chabane Djeraba. In ACM International Conference on Advances in Mobile Computing and Multimedia (ACM MOMM), pages 221-228, 2010.

References

[AAMD12a] Amel Aissaoui, Rémi Auguste, Jean Martinet, and Chaabane Djeraba. Reconstruction 3D de visages dans un système de stéréovision basée sur les propriétés du visage. In *15e colloque COmpression et REprésentation des Signaux Audiovisuels (CORESA 2012)*, pages –, Berlin, Germany, May 2012. 58

[AAMD12b] Rémi Auguste, Amel Aissaioui, Jean Martinet, and Chabane Djeraba. Les histogrammes spatio-temporels pour la ré-identification de personnes dans les journaux télévisés. In *15e colloque COmpression et REprésentation des Signaux Audiovisuels (CORESA 2012)*, pages –, Lille, France, 2012. 59

[AAMD12c] Rémi Auguste, Amel Aissaioui, Jean Martinet, and Chabane Djeraba. Réidentification de personnes dans les journaux télévisés basée sur les histogrammes spatio-temporels. In *Extraction et gestion des connaissances (EGC'2012)*, pages 547–548, 2012. 59

[AAY+12] Amel Aissaoui, Rémi Auguste, Tarek Yahiaoui, Jean Martinet, and Chaabane Djeraba. Fast stereo matching method based on optimized correlation algorithm for face depth estimation. In *7th International Conference on Computer Vision Theory and Applications (VISAPP 2012), (POSTER)*, pages 377–380, Rome, Italy, February 2012. 57

[ADMB16] Amel Aissaoui, Afifa Dahmane, Jean Martinet, and Marius Bilasco. Introducing FoxFaces: a 3-in-1 head dataset. In *11th International Conference on Computer Vision Theory and Applications (VISAPP 2016)*, pages 533–537, Rome, Italy, November 2016. 58, 72

[AIS93] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, New York, NY, USA, 1993. 25

[AM15] Amel Aissaoui and Jean Martinet. Bi-modal face recognition - How combining 2D and 3D clues can increase the precision. In *10e International Conference*

on Computer Vision Theory and Applications (VISAPP 2015), pages –, Berlin, Germany, March 2015. 58

[AMD12] Amel Aissaoui, Jean Martinet, and Chaabane Djeraba. 3D face reconstruction in a binocular passive stereoscopic system using face properties. In *Proceedings of the IEEE International Conference on Image Processing (ICIP 2012), Orlando, Florida, USA*, pages 1789 – 1792, October 2012. 58, 67

[AMD13] Amel Aissaoui, Jean Martinet, and Chaabane Djeraba. Rapid and accurate face depth estimation in passive stereo systems. *Multimedia Tools and Applications*, 72(3):2413–2438, 2013. 58, 67

[AMD14] Amel Aissaoui, Jean Martinet, and Chaabane Djeraba. DLBP: A novel descriptor for depth image based face recognition. In *Proceedings of the 21th IEEE International Conference on Image Processing (ICIP 2014), Paris, France*, pages 298 – 302, October 2014. 58, 71, 75

[AMT15] Rémi Auguste, Jean Martinet, and Pierre Tirilly. Space-time Histograms And Their Application To Person Re-identification In TV Shows. In *5th ACM International Conference on Multimedia Retrieval (ICMR 2015)*, pages 91–97, Shanghai, China, June 2015. viii, 59, 84, 87, 89, 91, 96

[AMT16] Rémi Auguste, Jean Martinet, and Pierre Tirilly. Re-identification-based person recognition in tv shows. *Submitted to Pattern Recognition*, –(–):–, 4 2016. 59

[ANRS07a] Andrea F. Abate, Michele Nappi, Daniel Riccio, and Gabriele Sabatino. 2D and 3D face recognition: A survey. *Pattern Recognition Letters*, 28(14):1885 – 1906, 2007. 75

[ANRS07b] Andrea F. Abate, Michele Nappi, Daniel Riccio, and Gabriele Sabatino. 2d and 3d face recognition: a survey. *Pattern Recognition Letters*, 28:1885–1906, 2007. 84

[AS94] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *Int. Conf. Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Citeseer, 1994. 26

[ATM15] Rémi Auguste, Pierre Tirilly, and Jean Martinet. Introducing FoxPersonTracks: a Benchmark for Person Re-Identification from TV Broadcast Shows. In *13th International Workshop on Content-Based Multimedia Indexing (CBMI 2015)*, pages 1–4, Prague, Czech Republic, June 2015. 59, 94, 101

[AWH15] Chedi Bechikh Ali, Rui Wang, and Hatem Haddad. A two-level keyphrase extraction approach. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II*, pages 390–401, 2015. 25

[BETVG08] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *CVIU*, 110(3):346–359, 2008. 7, 9, 10

[BG05] Chiraz BenAbdelkader and Paul A. Griffin. Comparing and combining depth and texture cues for face recognition. *Image Vision Computing.*, 23(3):339–352, 2005. 75

[BGK14] G. Bernard, O. Galibert, and J Khan. The second official REPERE evaluation. In *Workshop on Speech, Language and Audio for Multimedia (SLAM)*, Penang, Malaysia, 2014. 94

[BGS14] Apurva Bedagkar-Gala and Shishir K. Shah. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, 32(4):270–286, 2014. 84

[BKB00] A. Bouch, A. Kuchinsky, and N. Bhatti. Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service. In *Conference on Human Factors in Computing Systems - CHI'00*, pages 297–304, 2000. 31

[BMP02] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002. 10

[BR05] Stan T. Birchfield and Sriram Rangarajan. Spatiograms versus histograms for region-based tracking. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1158 – 1163. IEEE, june 2005. 88

[BT99] Stan Birchfield and Carlo Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, December 1999. 66

[BTVG06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European Conference on Computer Vision (ECCV)*, pages 404–417. Springer, 2006. 7

[BYRN99] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999. 38

[CBF03] KBKI Chang, Kevin Bowyer, and Patrick Flynn. Face recognition using 2D and 3D facial data. In *ACM Workshop on Multimodal User Authentication*, pages 25–32. Citeseer, 2003. 68

[CHS09] Xin Chen, Xiaohua Hu, and Xiajiong Shen. Spatial weighting for bag-of-visual-words and its application in content-based image retrieval. In *PAKDD '09: Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 867–874, Berlin, Heidelberg, 2009. Springer-Verlag. 39

[CM92] M. Chein and M.-L. Mugnier. Conceptual graphs: Fundamental notions. *Revue d'intelligence artificielle*, 6(4):365–406, 1992. 31, 34

[CNP06] C. Cotsaces, N Nikolaidis, and I Pitas. Video shot detection and condensed representation: a review. *IEEE Signal Processing Magazine*, 23(2):28–37, March 2006. 87

[CTC+09] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. Chog: Compressed histogram of gradients a low bit-rate feature descriptor. In *CVPR*, pages 2504–2511, 2009. 17

[DBID10] Taner Danisman, Ioan Marius Bilasco, Nacim Ihaddadene, and Chabane Djeraba. Automatic facial feature detection for facial expression recognition. In *Proceedings of the Fifth International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 407–412, 2010. 95

[DDLS08] Xiaomeng Wu Duy-Dinh Le and Shin'ichi Satoh. *Encyclopedia of multimedia*, chapter Face Detection, tracking, and recognition for broadcast video, pages 228–238. Springer-Verlag New York Inc, 2008. 86

[DWSP12] Piotr Dollàr, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, April 2012. 87

[EMUD12] Ismail Elsayad, Jean Martinet, Thierry Urruty, and Chabane Djeraba. Toward a higher-level visual representation for content-based image retrieval. *Multimedia Tools and Applications*, 60(2):455–482, 2012. 24

[ESMU+10] Ismail El Sayad, Jean Martinet, Thierry Urruty, Taner Danisman, Md. Haidar Sharif, and Chabane Djeraba. Using association rules and spatial weighting for an effective content based-image retrieval. In *VISAPP 2010*, pages –, Angers, France, 2010. 24

[ESMU+11] Ismail El Sayad, Jean Martinet, Thierry Urruty, Yassine Benabbas, and Chabane Djeraba. A semantically significant visual representation for social image retrieval. In *International Conference on Multimedia and Expo (ICME) 2011*, pages 1–6, Barcelona, Spain, 2011. 24

[ESMUD10a] Ismail El Sayad, Jean Martinet, Thierry Urruty, and Chabane Djeraba. A new spatial weighting scheme for bag-of-visual-words. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2010. 7, 12, 39, 40

[ESMUD10b] Ismail El Sayad, Jean Martinet, Thierry Urruty, and Chabane Djeraba. Visual sentence-phrase-based document representation for effective and efficient content-based image retrieval. In *EGC*, pages 157–162, Hammamet, Tunisia, 2010. 24

[ESMUD11] Ismail El Sayad, Jean Martinet, Thierry Urruty, and Chabane Djeraba. A semantic Higher-Level Visual Representation for Object Recognition. In *Advances in Multimedia Modeling - 17th International Multimedia Modeling Conference (MMM) 2011*, volume 1, pages 251–261, Taipei, Taiwan, Province Of China, 2011. 24, 28

[ESMUD12] Ismail El Sayad, Jean Martinet, Thierry Urruty, and Chabane Djeraba. Toward a higher-level visual representation for content-based image retrieval. *Multimedia Tools and Applications*, pages 1–28, 2012. DOI:10.1007/s11042-010-0596-x. 7, 12, 110

[ESZ06] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference*, 2006. 59

[EVGW+] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/-VOC/voc2012/workshop/index.html. 14, 22

[FFFP07] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007. 11, 22

[GA06] B. Gokberk and L. Akarun. Comparative analysis of decision-level fusion algorithms for 3D face recognition. In *Pattern Recognition. 18th International Conference on*, volume 3, pages 1018–1021, 2006. 75

[GCC+11] B. Girod, V. Chandrasekhar, D. M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. A. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham. Mobile visual search. *IEEE Signal Process. Mag.*, 28(4):61–76, 2011. 19

[GCGR11] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y.A. Reznik. Mobile visual search: Architectures, technologies, and the emerging mpeg standard. *MultiMedia, IEEE*, 18(3):86–94, 2011. 16

[GCMB10] S. Gupta, K.R. Castleman, M.K. Markey, and A.C. Bovik. Texas 3D face recognition database. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 97–100. IEEE, 2010. 65, 72

[GI12] Greenpeace-International. How Clean is Your Cloud? *Campaign reports, URL=http://www.greenpeace.org/international/en/publications/Campaign-reports/Climate-Reports/How-Clean-is-Your-Cloud/*, April 2012. 113

[GK13] Olivier Galibert and Juliette Kahn. The first official repere evaluation. *First Workshop on Speech, Language and Audio for Multimedia (SLAM 2013)*, 2013. 58

[Had03] Hatem Haddad. French noun phrase indexing and mining for an information retrieval system. In *String Processing and Information Retrieval, 10th International Symposium, SPIRE 2003, Manaus, Brazil, October 8-10, 2003, Proceedings*, pages 277–286, 2003. 25

[HAWC09] Di Huang, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Asymmetric 3D/2D face recognition based on lbp facial representation and canonical correlation analysis. In *Proceedings of the 16th IEEE international conference on Image processing*, ICIP'09, pages 3289–3292, Piscataway, NJ, USA, 2009. IEEE Press. 68

[HAWC10] Di Huang, M. Ardabilian, Yunhong Wang, and Liming Chen. Automatic asymmetric 3D-2D face recognition. In *Pattern Recognition. 20th International Conference on*, pages 1225–1228, 2010. 68

[HAWC12] Di Huang, M. Ardabilian, Yunhong Wang, and Liming Chen. 3-d face recognition using elbp-based facial description and local feature hybrid matching. *Information Forensics and Security, IEEE Transactions on*, 7(5):1551–1565, 2012. 70, 74, 75

[HBGVdM05] Michael Husken, Michael Brauckmann, Stefan Gehlen, and Christoph Von der Malsburg. Strategies and benefits of fusion of 2D and 3D face recognition. In *Computer Vision and Pattern Recognition-Workshops. IEEE Computer Society Conference on*, pages 174–174. IEEE, 2005. 75, 76

[HL08] Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM. 14

[HRBLM07] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained

environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 55, 56, 83

[HSA+11] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 41(6):765–781, 2011. 68

[HWT06] Yonggang Huang, Yunhong Wang, and Tieniu Tan. Combining statistics of geometrical and correlative features for 3D face recognition. In *Proceedings of the British Machine Vision Conference*, pages 879–888, 2006. 69, 74, 75

[HZA+10] Di Huang, Guangpeng Zhang, M. Ardabilian, Yunhong Wang, and Liming Chen. 3D face recognition using distinctiveness enhanced facial representations and local feature hybrid matching. In *Biometrics: Theory Applications and Systems. Fourth IEEE International Conference on*, pages 1–7, 2010. 68

[IK99] L. Itti and C. Koch. Learning to detect salient objects in natural scenes using visual attention. In *In Image Understanding Workshop*, 1999. 42

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 13, 42

[JA09] Rabia Jafri and Hamid Arabnia. A survey of face recognition techniques. *Journal of Information Processing Systems*, 5(2):41–68, 2009. 57

[JCB11] S. Jahanbin, Hyohoon Choi, and A.C. Bovik. Passive multimodal 2-d+3-d face recognition using gabor features and landmark distances. *Information Forensics and Security, IEEE Transactions on*, 6(4):1287–1304, Dec 2011. 68, 75

[JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. 90

[JDS08] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *ECCV*, pages 304–317, Berlin, Heidelberg, 2008. Springer-Verlag. 16, 19

[JDS09] H. Jegou, M. Douze, and C. Schmid. On the burstiness of visual elements. In *CVPR*, pages 1169–1176, 2009. 17

[JDSP10] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010. 16

[KBBN09] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009. 56

[KGQ+12] Juliette Kahn, Olivier Galibert, Ludovic Quintard, Matthieu Carré, Aude Giraudel, and Philippe Joly. A presentation of the repere challenge. In *10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2012. 109

[KSJ10] Elie El Khoury, Christine Senac, and Philippe Joly. Face-and-clothing based people clustering in video content. In *Proceedings of the international conference on Multimedia Information Retrieval*, pages 259–303, 2010. 97

[Kum13] Vijaya Kumari. Face recognition techniques: A survey. *World Journal of Computer Application and Technology*, 1(2):41–50, 2013. 57

[KZ02] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *Computer Vision, European Conference on*, pages 82–96, 2002. 61, 66

[LAS97] Brian Lent, Rakesh Agrawal, and Ramakrishnan Srikant. Discovering trends in text databases. In *Proc ACM SIGMOD International Conference on Knowledge Discovery and Data Mining (KDD'93), ACM*, pages 227–230, 1997. 25

[LDB+15] J. Liu, Y. Deng, T. Bai, Z. Wei, and C. Huang. Targeting Ultimate Accuracy: Face Recognition via Deep Embedding. *ArXiv e-prints*, June 2015. 56, 113

[Lim01] J.-H. Lim. Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications (Special Issue on Image Indexation)*, 4(2/3):125–139, 2001. 34

[LJSP07] Franco Rojas López, Héctor Jiménez-Salazar, and David Pinto. A competitive term selection method for information retrieval. In *CICLing*, pages 468–475, 2007. 16

[LM14] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014. 55, 83

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 7

[LRH09] R. Lienhart, S. Romberg, and E. Hörster. Multilayer plsa for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and*

Video Retrieval, CIVR '09, pages 9:1–9:8, New York, NY, USA, 2009. ACM. 12, 27

[LUG+16] Huu Ton Le, Thierry Urruty, Syntyche Gbèhounou, François Lecellier, Jean Martinet, and Christine Fernandez-Maloigne. Improving retrieval framework using information gain models. *Signal, Image and Video Processing*, pages 1–8, 2016. 16

[Luh58] H. P Luhn. The automatic creation of literature abstracts. *IBM Journal of Research Development*, 2(2):159–165, 1958. 22

[LZAL05] StanZ. Li, ChunShui Zhao, Meng Ao, and Zhen Lei. Learning to fuse 3D+2D based face recognition at both feature and decision levels. In Wenyi Zhao, Shaogang Gong, and Xiaoou Tang, editors, *Analysis and Modelling of Faces and Gestures*, volume 3723 of *Lecture Notes in Computer Science*, pages 44–54. Springer Berlin Heidelberg, 2005. 68, 75

[Mar13] Jean Martinet. Human-centered region selection and weighting for image retrieval. In *8th International Conference on Computer Vision Theory and Applications*, page id 322, Barcelona, Spain, 2013. 39, 41

[Mar14] Jean Martinet. From text vocabularies to visual vocabularies : what basis? In *VISAPP 2014*, pages 668–675, Lisbon, Portugal, 2014. 8, 23

[MC96] M.-L. Mugnier and M. Chein. Représenter des connaissances et raisonner avec des graphes. *Revue d'intelligence artificielle*, 10(1):7–56, 1996. 31

[MCM02] J. Martinet, Y. Chiaramella, and P. Mulhem. Un modèle vectoriel étendu de recherche d'information adapté aux images. In *INFORSID'02*, pages 337–348, Nantes, 2002. 24, 39

[MCM05] J. Martinet, Y. Chiaramella, and P. Mulhem. A model for weighting image objects in home photographs. In *ACM-CIKM'2005*, pages 760–767, Bremen, Germany, 2005. 39

[MCM11] Jean Martinet, Yves Chiaramella, and Philippe Mulhem. A relational vector space model using an advanced weighting scheme for image retrieval. *Information Processing and Management*, 47(3):391–414, May 2011. 30, 43, 51, 110

[MCMO03] J. Martinet, Y. Chiaramella, P. Mulhem, and I. Ounis. Photograph indexing and retrieval using star-graphs. In *Proceedings of CBMI'03 - Third International Workshop on Content-Based Multimedia Indexing*, pages 335–341, Rennes, 2003. 24, 39

[MdBJG14] Tomas Mantecon, Carlos R. del Blanco, Fernando Jaureguizar, and Narciso N. Garcia. Depth-based face recognition using local quantized patterns adapted for range data. In *Proceedings of the 21th IEEE international conference on Image processing*, pages 293–297, october 2014. 70

[Mec95] M. Mechkour. Emir2: an extended model for image representation and retrieval. In *Database and Expert Systems Applications Conf., London*, pages 395–404, 1995. 31

[Mil13] Mark P. Mills. The cloud begins with coal. *Digital Power Group report, URL=http://www.tech-pundit.com/wp-content/uploads/2013/07/Cloud_Begins_With_Coal.pdf*, August 2013. 113

[ML02] P. Mulhem and J.H. Lim. Symbolic photograph content-based retrieval. In *In CIKM'2002 - ACM Conference on Information and Knowledge Management*, pages 94–101, Virginia, USA, 2002. 34, 35

[MLLD09] Jean Martinet, Adel Lablack, Stanislas Lew, and Chabane Djeraba. Gaze based quality assessment of visual media understanding. In *IEEE PSIVT-CVIM'09*, Tokyo, Japan, 2009. IEEE. 41

[MN08] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. *Computer Vision. European Conference on*, pages 504–513, 2008. 57

[MOCM03] J. Martinet, I. Ounis, Y. Chiaramella, and P. Mulhem. A weighting scheme for star-graphs. In *In Proceedings of ECIR'03 - 25th BCS-IRSG European Conference on Information Retrieval Research*, pages 546–554, Pisa, 2003. 24, 39

[MS07a] J. Martinet and S. Satoh. An information theoretic approach for automatic document annotation from inter-modal analysis. In *IJCAI-MIR'07*, Hyderabad, India, 2007. 24

[MS07b] J. Martinet and S. Satoh. Using visual-textual mutual information for inter-modal document indexing. In *ECIR'07*, Rome, Italy, 2007. 24, 30

[MS07c] Jean Martinet and Shin'ichi Satoh. A study of intra-modal association rules for visual modality representation. In *Fifth IEEE International Workshop on Content-Based Multimedia Indexing (CBMI'07), June 25-27 2007, Bordeaux, France*, pages 344–350, June 2007. 24, 27

[MSCM08] J. Martinet, S. Satoh, Y. Chiaramella, and P. Mulhem. Media objects for user-centered similarity matching. *Multimedia Tools and Applications – Special Issue on Semantic Multimedia*, 2008. 39, 42, 44

[MTM14] José Mennesson, Pierre Tirilly, and Jean Martinet. Elementary block extraction for mobile image search. In *IEEE International Conference on Image Processing (ICIP)*, pages 3958 – 3962, 2014. 8, 21

[NCS06] A. Nguyen, V. Chandran, and S. Sridharan. Gaze tracking for region of interest coding in jpeg 2000. *Signal Processing: Image Communication*, 21(5):359–377, June 2006. 42

[Nie94] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, San Francisco, 1994. 31

[NS06] David Nistér and Henrik Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2161–2168, 2006. 14

[OMG+14] Abdelmalik Ouamane, Bengherabi Messaoud, Abderrezak Guessoum, Abdenour Hadid, and Mohamed Cheriet. Multi scale multi descriptor local binary features and exponential discriminant analysis for robust face authentication. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 313–317. IEEE, 2014. 55

[OP98a] I. Ounis and M. Paşca. Finding the best parameters for image ranking: a user-oriented approach. In *Proceedings of The IEEE Knowledge and Data Engineering Exchange Conference (KDEX'98), Taipei, Taiwan*, pages 50–59, 1998. 47

[OP98b] I. Ounis and M. Paşca. Relief: Combing expressiveness and rapidity into a single system. In *SIGIR'98*, pages 266–274, 1998. 30, 31

[OPM02] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, 24(7):971–987, July 2002. 68

[OVOP01] Timo Ojala, Kimmo Valkealahti, Erkki Oja, and Matti Pietikäinen. Texture discrimination with multidimensional distributions of signed gray-level differences. *Pattern Recognition*, 34(3):727 – 739, 2001. 68

[Pat09] A.L. Patterson. Phrase-based indexing in an information retrieval system, May 19 2009. US Patent 7,536,408. 25

[PCI+07] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 19

[PCI+08] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1 –8, june 2008. 18, 19

[PD07] Florent Perronnin and Christopher R. Dance. Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*. IEEE Computer Society, 2007. 16

[PFS+05] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *Computer vision and pattern recognition. IEEE computer society conference on*, volume 1, pages 947–954. IEEE, 2005. 72, 75

[PWHR98] P. Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998. 55

[RBJ89] Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989. 90

[RKB04] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*, volume 23, pages 309–314, 2004. 103

[RKS+10] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision (ECCV)*, pages 30–43, 2010. 40

[RP99] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999. 111

[RWH00] Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. Experimentation as a way of life: Okapi at TREC. *Inf. Process. Manage.*, 36(1):95–108, 2000. 16

[SAD+08] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer, 2008. 72

[Sal71] G. Salton. *The SMART Retrieval System*. Prentice Hall, 1971. 4, 30, 31, 38

[SB88] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988. 16, 38

[SCLT07] Te-Hsiu Sun, Mingchih Chen, Shuchuan Lo, and Fang-Chih Tien. Face recognition using 2D and disparity eigenface. *Expert Systems with Applications*, 33(2):265 – 273, 2007. 75

[SEZ09] J. Sivic, M. Everingham, and A. Zisserman. "who are you?" – learning person specific classifiers from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1145âĂŞ–1152, 2009. 59

[Sha48] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948. 22

[SHA+10] Wael Ben Soltana, Di Huang, Mohsen Ardabilian, Liming Chen, and Chokri Ben Amar. Comparison of 2D/3D features and their adaptive score level fusion for 3D face recognition. *3D Data Processing, Visualization and Transmission (3DPVT)*, 2010. 75

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *CoRR*, abs/1503.03832, 2015. 55

[SM83] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983. 30, 38

[Sow84] J. F. Sowa. *Conceptual Structures*. Addison-Wesley, Reading, MA, 1984. 30, 31, 32, 34

[SP02] Conrad Sanderson and Kuldip Paliwal. Information fusion and person verification using speech & face information, 2002. 75

[SS02] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision.*, 47:7–42, April 2002. 61, 66

[Süs15] B. Jin ; G. Yildirim ; C. Lau ; A. Shaji ; M. Ortiz Segovia ; S. Süsstrunk. Modeling the importance of faces in natural images. In *Proc. SPIE 9394, Human Vision and Electronic Imaging XX*, pages 1–11, 2015. 53

[SW05] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005. 28

[SWS+00] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000. 1

[SWT14] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1891–1898. IEEE, 2014. 55

[SWY75] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975. 38

[SYM+07] M. Sano, N. Yagi, J. Martinet, N. Katayama, and S. Satoh. Image-based quizzes from news video archives. In *ICME'07*, Beijing, China, 2007. 24, 30

[SZ03] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Ninth IEEE International Conference on Computer Vision*, pages 1470–1477, vol.2, 2003. 40

[SZS03] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):787 – 800, july 2003. 61

[TCAKD09] Dung Truong Cong, Catherine Achard, Louahdi Khoudour, and Lounis Douadi. Video sequences association for people re-identification across multiple non-overlapping cameras. In Pasquale Foggia, Carlo Sansone, and Mario Vento, editors, *Image Analysis and Processing (ICIAP)*, volume 5716 of *Lecture Notes in Computer Science*, pages 179–189. Springer Berlin / Heidelberg, 2009. 89

[TV98] Emanuele Trucco and Alessandro Verri. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1998. 61

[TYRW14] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 55

[TYSH13] Hengliang Tang, Baocai Yin, Yanfeng Sun, and Yongli Hu. 3D face recognition using local binary patterns. *Signal Processing*, 93(8):2190 – 2198, 2013. 68

[UGL+14] Thierry Urruty, Syntyche Gbèhounou, Huu Ton Le, Jean Martinet, and Christine Fernandez-Maloigne. Iterative Random Visual Word Selection. In *International Conference on Multimedia Retrieval (ICMR 2014)*, page 249, Glasgow, Scotland, 2014. 8, 16

[vdSGS10] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010. 14

[VJ02] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision (IJCV)*, 57(2):137–154, 2002. 102

[vR79a] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2 edition, 1979. 13

[vR79b] C.J. van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworths, London, 1979. 38

[WLCV07] Jian-Gang Wang, EngThiam Lim, Xiang Chen, and Ronda Venkateswarlu. Real-time stereo face recognition by fusing appearance and depth fisherfaces. *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, 49(3):409–423, 2007. 75

[WRM10] Xueqiao Wang, Qiuqi Ruan, and Yue Ming. 3D face recognition using corresponding point direction measure and depth local features. In *Signal Processing. IEEE 10th International Conference on*, pages 86–89, 2010. 68

[WYS+15] Ren Wu, Shengen Yan, Yi Shan, Qingqing Dang, and Gang Sun. Deep image: Scaling up image recognition. *arXiv:1501.02876 [cs.CV]*, 2015. 113

[XHL11] Pengfei Xiong, Lei Huang, and Changping Liu. Real-time 3D face recognition with the integration of depth and intensity images. In Mohamed Kamel and Aurélio Campilho, editors, *Image Analysis and Recognition*, volume 6754 of *Lecture Notes in Computer Science*, pages 222–232. Springer Berlin Heidelberg, 2011. 68

[XKS92] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(3):418–435, May 1992. 77

[XLTQ09] Chenghua Xu, Stan Li, Tieniu Tan, and Long Quan. Automatic 3D face recognition from depth and intensity gabor features. *Pattern Recognition*, 42(9):1895 – 1905, 2009. 68, 75

[Yar67] A. L. Yarbus. *Eye Movements and Vision*. Plenum Press, New York, 1967. 41

[YKA02] Ming-Hsuan Yang, David J. Kriegman, and Narendra Ahuja. Detecting faces in images: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(1):34–58, 2002. 87

[YS06] Omar Yilmaz, Alper amd Javed and Mubarak Shah. Object tracking: a survey. *ACM Computing Surveys*, 38(4), 2006. 87

[YWY07] Junsong Yuan, Ying Wu, and Ming Yang. Discovery of collocation patterns: from visual words to visual phrases. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. 27

[ZCPR03a] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys*, 35(4):399–458, 2003. 57

[ZCPR03b] Wenyi Zhao, Rama Chellappa, P. Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys (CSUR)*, 35(4):399–458, 2003. 84, 86

[ZG08] Qing-Fang Zheng and Wen Gao. Constructing visual phrases for effective and efficient object-based image retrieval. *TOMCCAP*, 5(1), 2008. 27

[Zip32] G.K. Zipf. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press, 1932. 22

[ZKM07] Xuan Zou, Josef Kittler, and Kieron Messer. Illumination invariant face recognition: a survey. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2007. 84

[ZLY+15] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015. 55