Content-based structuring of still and animated images collections

Valérie Gouet-Brunet

Université Pierre et Marie Curie - Paris 6

Document de Synthèse

présenté pour obtenir

l'Habilitation à Diriger des Recherches

Mention Informatique

par

Valérie GOUET-BRUNET

Titre

Structuration par analyse du contenu des collections d'images fixes et animées

Content-based structuring of still and animated images collections

Soutenue le 2 décembre 2008

Président	Henri MAÎTRE	TELECOM ParisTech, Paris
Rapporteurs	Eric PAUWELS	Centre for Mathematics and Computer Science, Amsterdam
	Sylvie PHILIPP-FOLIGUET	Ecole Nationale Supérieure de l'Electronique et de ses Applications, Cergy-Pontoise
	Shin'ichi SATOH	National Institute of Informatics, Tokyo
Examinateurs	Patrick GALLINARI	Université Pierre et Marie Curie, Paris
	Geneviève JOMIER	Université Paris-Dauphine, Paris
	Nozha BOUJEMAA	Institut National de la Recherche en Informatique et en Automatique, Rocquencourt
	Michel SCHOLL	Conservatoire National des Arts et Métiers, Paris

Remerciements

Je tiens à exprimer ma reconnaissance aux personnes qui m'ont accompagnée tout au long des mes recherches et permis d'arriver jusqu'ici.

Monsieur Eric Pauwels, Professeur au Centre for Mathematics and Computer Science à Amsterdam, Madame Sylvie Philipp-Foliguet, Professeur à l'École Nationale Supérieure de l'Electronique et de ses Applications, et Monsieur Shin'ichi Satoh, Professeur au National Institute of Informatics de Tokyo, qui m'ont fait l'honneur d'être les rapporteurs de cette habilitation. Je les remercie pour le temps qu'ils ont consacré à juger ce travail.

Monsieur Patrick Gallinari, Professeur à l'université Pierre et Marie Curie, Madame Geneviève Jomier, Professeur à l'Université Paris-Dauphine, et Monsieur Henri Maître, Professeur à l'École TELECOM ParisTech, pour l'intérêt qu'il ont porté à ce travail en acceptant de faire partie du jury. Je remercie aussi Monsieur Philippe Chrétienne, Professeur à l'université Pierre et Marie Curie, pour avoir veillé au bon déroulement de la préparation de cette habilitation.

Michel, pour m'avoir initiée aux affres de la malédiction de la dimension mais surtout pour sa confiance en me recrutant dans son équipe au CNAM. Je le remercie aussi pour les sujets de recherche passionnants qu'il m'a proposés tout en me laissant la liberté de développer mes propres axes, pour son ouverture d'esprit, son dynamisme, sa disponibilité et sa gentillesse. Merci aussi à toute l'équipe Vertigo, les anciens comme les nouveaux : David, Bernd, Dan, Michel C. et Nicolas. Sans oublier tous les doctorants de l'équipe, en particulier les miens : Bruno, Nouha et maintenant Rin et Radhwane, avec qui j'ai plaisir à faire de la recherche et sans qui ces travaux n'auraient pas pu aboutir. Je n'oublie pas non plus mes nombreux collègues du laboratoire CEDRIC et de la Spécialité Informatique du CNAM qui m'ont aidée à accomplir le métier d'enseignant-chercheur.

Le travail présenté ici doit aussi beaucoup à l'équipe Imédia de l'INRIA Rocquencourt. Je remercie Nozha, pour m'y avoir accueillie il y a déjà huit ans, à l'époque en tant que postdoc et par la suite comme collaborateur extérieur. Merci de m'avoir donné l'opportunité de participer à des projets passionnants et d'avoir pu développer des axes de recherche porteurs. Ma sympathie va aussi aux membres passés et actuels de l'équipe : Anne, Minel, Julien F., Hichem S., Jean-Paul, Laurence, Jean-Philippe, Alexis et tous les autres.

Tous les membres de la grande famille Wisdom, et en particulier Marta Rukoz et Maude Manouvrier, respectivement Professeur et Maître de Conférences au LAMSADE de l'Université Paris-Dauphine, avec qui de nouvelles aventures commencent, par le biais de l'encadrement de deux nouvelles thèses et de projets prometteurs.

Je tiens aussi à remercier Vincent Oria, Professeur au New Jersey Institute of Technology et Claudia Bauzer Medeiros, Professeur au Laboratoire des Systèmes d'Information à São Paulo, pour m'avoir invitée et accueillie dans leurs laboratoires cette année. Outre les activités enrichissantes que j'ai pu mener là-bas, ces deux séjours m'ont grandement aidée à prendre le recul nécessaire à la rédaction de cette habilitation.

Et enfin bien sûr, Eric, pour partager ma vie et illuminer mon quotidien. Comme j'aimerais être capable de comprendre tes patients aussi bien que tu comprends mes descripteurs d'images !

Résumé

Ce document présente une synthèse de mon activité de recherche depuis 2001, date qui correspond à la fin de ma thèse. Mon domaine de recherche est l'indexation par analyse du contenu visuel des grandes collections d'images fixes et animées. J'ai exploré cette problématique sous l'angle de l'analyse d'images en vue de proposer de nouveaux descripteurs des contenus visuels, mais aussi sous l'angle des bases de données par l'étude de nouvelles méthodes d'accès multidimensionnelles dédiées aux bases d'images. La plus grande partie de mon travail repose sur la notion de description locale par extraction de points d'intérêt. Populaire par sa robustesse aux transformations de l'image, cette catégorie d'approches souffre d'inconvénients que je me suis attachée à étudier et à minimiser pour plusieurs applications manipulant des contenus image et vidéo. Son premier défaut réside dans la relative pauvreté de la description mise en jeu, puisque extraite localement et donc représentant assez mal les entités décrites. Nous avons traité ce problème selon trois directions : (1) la proposition de mesures de similarité fines entre deux images décrites par des descripteurs locaux (2) la combinaison de descripteurs visuels hétérogènes, incluant plusieurs types de descripteurs locaux et globaux, avec comme objectif de mettre en avant les avantages de chaque type pour la reconnaissance d'objets et (3) la caractérisation du comportement spatio-temporel des descripteurs locaux pour améliorer la détection de copies dans les vidéos. Le second défaut majeur des approches locales porte sur l'énorme volume de caractéristiques multidimensionnelles généré, rendant la recherche dans les grandes collections d'images difficilement réalisable sans l'aide de méthodes d'accès dédiées. Après avoir revisité les phénomènes de la malédiction de la dimension pour les structures d'index classiques en bases de données appliquées aux bases de descripteurs d'images, nous avons proposé (1) des stratégies permettant d'optimiser la recherche à partir de requêtes multiples, i.e. les requêtes composées de plusieurs vecteurs (comme c'est le cas avec les descripteurs locaux) et (2) un modèle hiérarchique permettant d'accélérer la recherche exacte, approximative et progressive des plus proches voisins dans les grands volumes de données multidimensionnelles. Cette synthèse se termine par une présentation de mon travail actuel, dans la continuité des activités sus-citées ainsi que vers de nouvelles directions de recherche telles que l'intégration de l'information spatiale dans la représentation des contenus visuels.

Mots clés

Image, Vidéo, CBIR, Descripteurs locaux, Points d'intérêt, Reconnaissance d'objets, Détection de copies, Structures d'index multidimensionnelles, Malédiction de la dimension, Requêtes multiples, Passage à l'échelle.

Abstract

This document presents a synthesis of my research activity since 2001, which corresponds to the end of my PhD thesis. The research domain I investigate is content-based indexing of still and animated images. I have explored this area under the viewpoint of image analysis for the proposal of new descriptors of visual contents, as well as under the viewpoint of databases by studying new multidimensional access methods dedicated to visual contents collections. Most of my activity rests on the approaches of local description based on interest points extraction. Popular because of their robustness to image transformations, these approaches suffer from drawbacks that motivated my research for several kinds of applications manipulating image and video contents. Their first weakness concerns the relative poorness of the involved description, since locally extracted and then representing the whole object contents not sufficiently. We have addressed this problem according to three directions: (1) the proposal of fine similarity measures between two images described locally (2) the combination of heterogeneous visual descriptors, including several categories of local and global features, with the aim of exhibiting the richness of each category for object recognition applied to video surveillance and (3) the description of the spatio-temporal behavior of local descriptors for improving copy detection in video sequences. The second drawback of local approaches rests on the high volume of multidimensional features generated, making unachievable search in large collections of contents without dedicated access methods. After having revisited the curse of dimensionality phenomena for the state-of-the-art index structures on image databases, we proposed (1) strategies for improving retrieval when considering multiple queries, i.e. queries composed of several vectors (such as with local descriptors) and (2) a hierarchical model enabling to accelerate nearest neighbors search in high-dimensional feature spaces, under exact, approximate and progressive retrieval scenarios. This synthesis ends with the presentation of my actual work, in the continuity of the aforementioned activities as well as towards new research directions such as the description of the spatial layout of the visual contents.

Keywords

Image, Video, CBIR, Local descriptors, Interest points, Object recognition, Copy detection, Multidimensional index structures, Curse of dimensionality, Multiple queries, Scalability.

Contents

Introduction

This document gathers the main elements of my research activity since early 2001, which corresponds to the end of my PhD thesis. As introduction, in the following I relate the calendar of my research activities and the context of their execution.

I carried out my PhD thesis at the Research Center LGI2P of the Ecole des Mines d'Alès, under the supervision of Philippe Montesinos (LGI2P) and of Jean-Claude Bajard (LIRMM), during the period 1996-2000. I began studying the Computer Vision problem of image matching by content analysis, with applications to 3D reconstruction and synthesis of novel views. This work plunged me into the world of local descriptors based on the extraction of interest points. Our contributions were the exploitation of color information for improving point matching in stereo images, as well as the integration of geometry into an iterative image matching scheme designed for large sets of local descriptors.

At the end of 2000, my postdoctoral position in the research group Imedia at INRIA Rocquencourt (leaded by Nozha Boujemaa) allowed me applying and generalizing the results of my thesis on the problem of Content-Based Image Retrieval (CBIR). This work was challenging because matching one image against a database of images with interest points is much more difficult than matching stereo images. We had to adapt the local features and in particular to reorganize the geometrical constraints studied during the thesis into new similarity measures. Using local descriptors enabled querying a database by the way of the paradigm of retrieval by example with partial queries, i.e. the possibility to select a part of the image or an object as query. Such a paradigm was applied to several scenarios within the scope of the European Program "STOP" (2000-2001) and of the French project MediaWorks (RIAM, 2001-2004).

By managing descriptors in large volumes of images, and in particular local descriptors that provide a precise description of the content but that are reputed to be voluminous, we became aware that an efficient CBIR system has to consider the structuring of the high-dimensional feature spaces associated with descriptors in order to perform similarity search efficiently in terms of time retrieval. Fruitful discussions started with Michel Scholl, Professor of the Vertigo research group (and leader of) in the CEDRIC Laboratory in Computer Science at CNAM, that helped me investigating this topic deeper. Studying content-based image retrieval from the viewpoint of databases was innovative at this period. In particular, the main challenge was to analyze and exploit the specificities of

multidimensional spaces generated by visual descriptors for adapting the existing classical access methods or proposing more dedicated ones. The collaboration between Imedia and Vertigo was reinforced by the co-supervising of several master students and by the joint participation to the French project BIOTIM (ACI "Masses of Data", 2003-2006). In 2002, it also conducted to my hiring in Vertigo as assistant professor, for developing a research axis on "image databases". In 2004, Vertigo obtained a funding from the French government that allowed recruiting a PhD student, Nouha Bouteldja (2004-2008, MENRT funding, CNAM), who has been co-supervised by Michel Scholl (25%) and myself (75%). The theme of this thesis was the proposal of multidimensional index structures dedicated to local descriptors. At present, Nouha is writing her thesis report, she will defend it at the end of 2008.

While working at CNAM, I continued my collaboration with Imedia as associate member: I took an active part in the European Network of Excellence MUSCLE (2004-2007) as deputy leader of the JPA2 work package on "Content-Based Description" (leader Nozha Boujemaa) and leader of the task on "Image and video processing". Within MUSCLE, I directed my research towards the description of video contents. Compared to images, such contents provide a lot of new relevant information, new challenges and applications that we explored with local descriptors. We started a collaboration with the INA (Institut National de l'Audiovisuel) which is a French organization which stores and preserves the national audiovisual heritage. Through the PhD thesis of Julien Law-To (2004-2007, CIFRE funding, Versailles Saint-Quentin University), co-supervised by Nozha Boujemaa (10%), Olivier Buisson, researcher at INA (45%) and myself (45%), we focussed on the challenging problem of video copy detection. These three years of research led to the proposal of the system ViCopT, whose main characteristic is to describe the temporal behavior of local descriptors along video sequences and then provides a very distinctive description of the video content. Julien defended his thesis on December 14, 2007 (distinction "mention très honorable").

In parallel at CNAM, I have investigated another orthogonal research direction for enriching the description of images with local descriptors: during the Master thesis (2003), CNAM engineer thesis (2004-2005) and then PhD thesis of Bruno Lameyre (2005-early 2009, Survision funding, CNAM, director Michel Crucianu 5%), we have been interested in the joint exploitation of local and global features. Such a combination provided a synergy that was first used for object tracking; Bruno obtained the award of the best CNAM engineer thesis in 2005 for this work. Then during his thesis, this work was generalized to object recognition and precise localization. It was applied to surveillance, within the scope of industrial contracts I have initiated and I am supervising with the French company Survision (2003-2004 and 2006-2009) which develops softwares dedicated to video surveillance.

Since 2007, I take part in the federation Wisdom (PPF, 2007-2010) which gathers three parisian research groups (from laboratories CNAM/CEDRIC, Paris-Dauphine/LAMSADE and Paris-6/LIP6) that share common research interests in databases. In particular, with Maude Manouvrier and Marta Rukoz of the LAMSADE, we are studying the description of the spatial relationships existing between objects of interest in images. This topic addresses problems inherent to image analysis as well as access methods, making the collaboration particularly relevant. At present, this collaboration has resulted in the proposal of the

French project DISCO (ANR MDCO, 2008-2010); in this project, I am the leader of the WP2 on "Search by content and centralized index structures". In parallel, I initiated and I am supervising a French project "Paris en images" funded by the "Mairie de Paris" (2007-2009) on content-based image indexing. At present, with Maude and Marta, we start to co-supervise two PhD students on the topic of the description of the spatial relationships in images: the first one, Nguyen Vu Hoang (2008-2011, MENRT funding, thesis director Marta Rukoz) will work on the use of spatial relationships between symbolic objects to define a context allowing the efficient retrieval of new objects. The second PhD student (Radhwane Kissi, funding from projects DISCO and "Paris en images" for three years) will study the definition of hierarchies to improve the description of spatial relationships in images, in terms of quality of the description as well as in retrieval time. According to my success to the "habilitation", I will be his thesis director.

In 2008, I obtained a sabbatical leave for six months from February to July 2008 ("CRCT", national campaign), that has allowed me to start two international collaborations. The first one is with Vincent Oria at the New Jersey Institute of Technology (New Jersey, USA) on the problem of the annotation of personal albums photos; I was invited one month in his labs (March 15 - April 15, 2008). The second collaboration is with Claudia Bauzer Medeiros and Ricardo Torres, from the LIS (Laboratory of Information Systems) of the UNICAMP University (São Paulo state, Brazil); I was invited three weeks in their labs in August 2008 to set up a collaboration on the theme of image content description and interrogation (in particular with spatial relationships) dedicated to remote sensing images (agriculture and biodiversity).

This report is organized as follows: chapter 1 presents the scientific context of my research activities, the facts that motivated our choices in terms of problems investigated. Then it presents the outlines of my contributions to the domain of Content-Based Image Retrieval. This description refers to chapters 2, 3 and 4 which detail my past work. Finally, chapter 5 gives a panorama of the research activities that I am currently investigating and also presents the directions of research that motivate me for the future.

Chapter 1

Scientific context, motivations and main contributions to the domain

This chapter has for objective of presenting the scientific context of my research activities. In section 1.1, after a brief introduction to Content-Based Image Retrieval, I focus on local visual descriptors that have been the leitmotiv of my research, by revisiting the literature on this topic, describing the main advantages and inconvenients of such features and finally giving an idea of the current related trends of research. Then section 1.2 presents an overview of my contributions with the associated publications since early 2001. This description mainly refers to chapters 2, 3 and 4 where the main points of these contributions are detailed.

Contents

1.1 Structuring collections of visual contents

Content-Based Image Retrieval (CBIR) is a domain of research that expanded in the early 90's and that has grown tremendously after the year 2000 in terms of the people involved and published papers [Datta et al., 2008]. Its objective is to help in organizing multimedia documents archives by analyzing their visual content. This objective addresses the fundamental open problem of image understanding as well as the problems of management and interrogation of collections of visual contents. The domain gathers researchers of several fields of computer science, e.g. image analysis, computer vision, machine learning, information retrieval, human-computer interaction, database systems, Web and data mining, information theory, statistics, with also more recently the contribution of psychology and neurosciences [Gouet-Brunet, 2006; Wang et al., 2006; Datta et al., 2008].

The core task of CBIR is related to the extraction of features that encapsulate the content of the image, according to a given application. In the community of CBIR, these features are sometimes called *indexes*, *signatures* or *fingerprints*. Because term *index* has several significations according to the community, here we will refer as *signatures* and the term *index* will be employed in the context of *index structures* that refer to the databases terminology designing access methods to data. Since the beginning of CBIR, the researchers that focus on the design of image descriptors are driven by the very challenging objective of reducing the well-known semantic gap, i.e. of proposing descriptors which produce signatures embedding an information related to the image content that coincides at best with what the user sees of this image in a given situation, and defining associated similarity measures that agree the user's perception of similarity. There exist a large palette of content-based descriptors: images can be described either by low-level features that represent information about the signal, such as color, texture, shapes, regions, interest points, etc. or by higher level visual concepts such as trained and recognized objects or specific objects related to a domain (e.g. faces).

Color, texture and shape have been identified as the main low-level descriptors that can characterize an image. For example, the visual features included in the MPEG7 standard consist of histogram-based descriptors, spatial color descriptors and texture descriptors [Manjunath et al., 2002]. They are called *global descriptors* because they resume in one feature vector all the image content. Such kind of features has been used for a long time to characterize the visual aspect of images. They have the advantage of characterizing the main aspect of images and encapsulating some global semantics or ambiance such as "indoor" or "painting", while requiring a low amount of data to describe it (in general one high-dimensional vector per image). Screenshot (a) in figure 1.1 illustrates a scenario of retrieval by example with global descriptors. Despite these advantages, nowadays they seem to be less studied because many people are focussing on *local descriptors*, as stated in [Datta et al., 2008]. These ones present different advantages that set them aside for other kinds of topical applications. Because my research activity is funded on such descriptors, I detail them more precisely in the following sections.

1.1.1 Local features as signature of visual contents

When considering applications where parts of the image have to be precisely and robustly described, a popular solution is to capture the image content by extracting interest points

from the image and characterize them with a local description, i.e. by extracting informations on the signal locally around the point. Traditionally, interest points are extracted in sites of the image that are supposed to catch the attention, i.e. where the signal varies. Such approaches are said to be *salient features-based*, because the information extracted is condensed into limited but salient sites. Since the 70's, literature on image description with local features is abundant. They were initially employed in robotics for stereo image matching with application to robot localization or scene reconstruction. Since the PhD thesis of C. Schmid in 1996 on their adaptation to CBIR, their use is widespread in this domain. In particular, their robustness to occlusion, to background clutter and to several image transformations makes them useful for tasks such as sub-image retrieval or object recognition in images with heterogeneous visual contents, as illustrated with the query example (b) of figure 1.1.

(a) Image retrieval with global descriptors (color, texture, shape)

(b) Sub-image retrieval with local descriptors

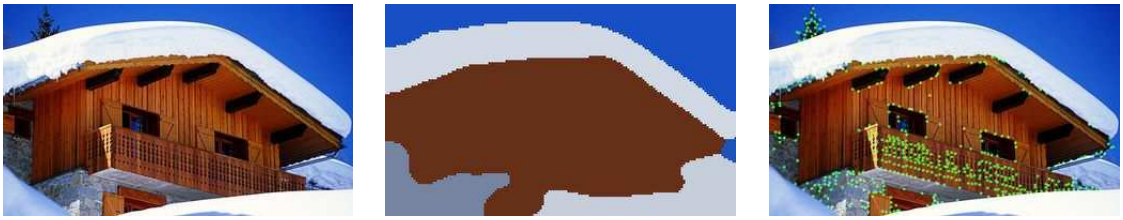
Figure 1.1: *Two scenarios of image retrieval in a database of generic images (sources: Images Du Sud) by using the CBIR system IKONA developed at Imedia. Each query is represented by a white rectangle on the upper left picture in scenarios (a) and (b). Query (a) is done on the whole image and involves global descriptors while query (b) corresponds to a sub-part of the image (a sunflower) manually surrounded by the user at query time, and described with local descriptors. The nine best responses are displayed from the right of the query, sorted by decreasing order of similarity. We observe that query (a) returns images showing an global ambiance of flowered gardens while query (b) enables to retrieve images containing the object of interest precisely.*

There exist many solutions for image description based on sets of sparse local features, see for example [Montesinos et al., 2000; Dorkó and Schmid, 2003; Wallraven et al., 2003; Fergus et al., 2003; Lowe, 2004; Agarwal et al., 2004; Ferrari et al., 2006; Opelt et al., 2006; Joly, 2007]. They mainly differ in the nature of the interest point, the local description used, the image transformations authorized, and the classifier considered to distinguish between

categories of training data when recognition is the goal. More recently, other approaches have exploited a bag-of-features representation of the local descriptors [Sivic and Zisserman, 2003; Willamowski et al., 2004; Sivic et al., 2005b]. This concept comes from text retrieval and consists in building a visual vocabulary (a codebook) from quantized local descriptors and in representing the image by a vector of fixed size involving the frequency of each visual word of this vocabulary. The concept is usually applied to object recognition, after having trained descriptors to be tolerant to inter-class variability and also intra-class variability when dealing with recognition of classes of objects. This representation has the advantage of locally describing the image content while only involving one feature vector per image, and of generating a very sparse feature space that can be accessed quickly with inverted files. One drawback is that it cannot easily integrate a spatial description of the points distribution in the image. The reader can see for example the report of the European PASCAL Challenge on Visual Object Classes recognition [Everingham et al., 2006], where most of the approaches of object recognition proposed are based on local descriptors; and the recent survey on local invariant feature detectors [Tuytelaars and Mikolajczyk, 2008].

1.1.2 Limitations of local descriptors

Since interest point are traditionally extracted on high-frequency regions of the image, local descriptors do not characterize well homogeneous areas. Here, other categories of local approaches such as a segmentation in regions should be better appropriated, as we studied it in publication [Boujemaa et al., 2004a]. The examples of figure 1.2 illustrate the complementarity of these local approaches by showing the different supports of description exhibited.



(a) Original image (b) Image segmentation (c) Interest point extraction (in green)

Figure 1.2: *Two examples of local image descriptors.*

But the most significant inconvenients of local descriptors come from their advantages: their genericity and robustness to several image transformations cannot make them highly informative. In addition, their local nature resumes their visual relevance to a low-level visual cue, allowing description of local parts of objects of interest without being able to integrate its global appearance. By integrating a description of the relationships existing between interest points, they can describe a certain degree of structural information, but when used for object recognition purposes, the segmentation of the recognized object obtained clearly remains rough.

Another important drawback of these features is that they produce a significant larger volume of data than global features. The visual content of an image is not described by

a single vector but rather by a set of vectors describing different regions of the image; for example, the popular SIFT approach [Lowe, 2004] produces several hundreds of 128-dimensional vectors on images of natural content. Here, the use of efficient access methods and multidimensional indexing structures is essential to perform similarity search in large sets of contents or to deal with time dependent applications. In contrast, global features can often be processed with a simple sequential scan of the data set (up to 1 million of images with current material). But generally speaking, any kind of descriptor will suffer from the curse of dimensionality phenomena beyond a certain volume of multidimensional features. Consequently, since few years, *scalability* has become an important issue in CBIR.

1.1.3 Recent improvements on local descriptors

Because of the topical applications associated with local descriptors and their numerous limitations, research on this topic is still very active. In this section, I address two directions of research that motivate a lot of researchers and provided many recent publications.

Contribution of video

A majority of the state-of-the-art approaches of local description involves sparse and still images. Since few years, the facility to produce and disseminate large volumes of videos and the new associated applications in several domains of the industrial world (audiovisual, Internet, surveillance, etc) create new scientific challenges and make these contents of great interest in the research community. Recent approaches [Sivic and Zisserman, 2003; Sivic and Zisserman, 2004; Sivic et al., 2004; Grabner and Bischof, 2005; Opelt et al., 2005] proposed to exploit the richness of video sequences for improving description with local features (mainly motivated by object recognition purposes), based on the two following observations: humans can recognize animated objects better than objects from still images; technically, videos provide multiple samples of object's appearance in a form that can be easily linked using straightforward tracking. The motivation for tracking features is to provide more robust visual features by modeling their variability along trajectories, while generating a compact feature space by grouping redundant features. Moreover, incorporating temporal information in the image description can compensate the poor quality of video images, in particular of video surveillance images.

Combination with other features

With the recent proposal of new local detectors and descriptors involving different topological natures of support exhibiting either discontinuity or homogeneity (including textured patches, homogeneous regions, local shapes and symmetry points), some works [Sivic and Zisserman, 2003; Jurie and Schmid, 2004; Opelt et al., 2005; Opelt et al., 2006] proposed to improve image description, mainly motivated for object recognition in still images or videos, by exploiting a combination of such complementary local descriptors. Also with the aim to enrich local descriptors, other approaches proposed recently consist of associating a more global *context* to the local description. In [Mortensen et al., 2005], the SIFT vector describing each feature point is reinforced with a vector containing curvilinear shape

information from a larger neighborhood. In [Amores et al., 2005], a context based on correlograms is added to the description of the interest point, which integrates mutual spatial relations between the points of the object; the approach avoids the use of graphs that usually characterize spatial relations and have a high cost. The problem of object precise localization after recognition with local features motivates several researchers [Leibe et al., 2004; Sivic et al., 2005b; Cao and Fei-Fei, 2007; Larlus and Jurie, 2008]. For example, in [Sivic et al., 2005b], couples of points which co-occur in a local neighborhood are proposed in a bag-of-features representation for better performing object localization, in addition to recognition.

1.2 Overview of the contributions

Since 2001, my research activity rests on the structuring of large collections of still and animated images, based on the analysis of their visual content with local descriptors. This thematic gathers two underlying fundamental problems: the first concerns the proposal of more efficient descriptors with the objective of reducing the unsolved problem of semantic gap. My contributions on this field are summarized in section 1.2.1 and described in chapters 2 and 3. Here, I have investigated several axes aiming at reducing the inconvenients of local descriptors related to their distinctiveness. Secondly, because these features (and, in a certain extent, global descriptors with very large sets of images) involve high-dimensional feature spaces, employing them imperatively requires the use of access methods with multi-dimensional index structures. My contributions on the proposal of appropriated structures are presented in section 1.2.2 and described in chapter 4. Most of these activities were done in collaboration with other researchers and students at CNAM and INRIA; they are indicated in the associated descriptions.

Since 2007, an important part of my current activity is dedicated to the description of the spatial relationships that exist between objects of interest in images. This work is ongoing and involves two new PhD theses, I briefly present it here and describe it more deeply in chapter 5 dedicated to my new activities and the perspectives of my research.

According to the context, the solutions proposed in my work have been applied to several applications and scenarios: content-based retrieval by example (for sub-image retrieval and image/video annotation), object recognition for surveillance purposes and video copy detection. All this work was published in several papers, listed in section 1.2.3.

1.2.1 Bringing distinctiveness to visual features *(Chapters 2 & 3)*

I have investigated two orthogonal axes with the aim of bringing distinctiveness to local descriptors: from one side, the enrichment of these descriptors in images by combining them with other categories of descriptors (chapter 2), and on the other side, their enrichment by considering their spatio-temporal behavior in video sequences (chapter 3). Because the relevance of a signature is intimately linked to its similarity measure, we also studied the way to compare local features and group of local features. Work on similarity is presented before the other contributions because it was the work I investigated first, at the beginning of my postdoc (chapter 2).

Similarity between images *(Chapter 2)*

During period 2001-2004, one part of my research work at INRIA and at CNAM concerned the study of similarity measures for improving matching of images described with local descriptors. This work is presented in details in chapter 2 dedicated to the content description of images. We tackled the problem at two levels of description: "interest point" and "groups of interest points" levels, presented in the two following paragraphs.

Local descriptors finely describe the signal locally around points, making them very discriminant and then potentially very sensitive to image transformations and noise. I began a study on the modeling of their variability for these transformations. This work was done at INRIA for still images where its was experimented against JPEG compression [Gouet and Boujemaa, 2002] and against synthetic transformations with Jean-Philippe Tarel, and also at CNAM with Bruno Lameyre for tracking of points in video sequences [Gouet and Lameyre, 2004].

In parallel, in the continuity of the work realized during my PhD thesis on robust stereo image matching with geometrical information [Montesinos et al., 2000], I also focussed on the study of solutions for integrating the spatial distribution of sets of interest points into the image description, adapted to the matching of large sets of images. This work was done at INRIA and applied to two scenarios requiring different solutions. The first scenario was 2D logo detection where transformations between images are known and then spatial registration is possible [Boujemaa et al., 2004b]; this scenario was investigated within the French project MediaWorks (2001-2004, RIAMM) with French TV channel TF1 that addressed the proposal of an hybrid text-image indexing and retrieval platform for video news. The second scenario dealt with non-rigid scenes; the corresponding work was published in [Gouet and Boujemaa, 2001] and was realized within the European program "STOP" with the French Judicial Police, where it served as Police investigation aids.

Synergies between heterogeneous visual features *(Chapter 2)*

As mentioned in section 1.1.3, the recent proposal of new local descriptors involving different topological natures of image support has given the possibility to study some combinations of them to improve image description, most of the time for object recognition purposes. A great part of my research concerned the idea of combining heterogeneous visual features to enrich image description, and is presented in details in chapter 2. I investigated this challenging direction through two activities: first by considering the combination of local descriptors of different topological natures with application to copy detection, at INRIA during the PhD thesis of Julien Law-To. We combined two extractors of interest points that differently exhibit visual saliency of the image structures: the first is the Harris detector, well-known for exhibiting sites characterized with high-frequency signal, while the second ones focuses on local symmetries. The context of this work is more described in the part on the spatio-temporal description of video contents (next section).

Secondly, at CNAM with Bruno Lameyre during his Master thesis (2003, Master STIC CNAM) and his CNAM engineer thesis (2004-2005), we began to investigate the combination local and global descriptors for object description with application to tracking [Lameyre and Gouet, 2004]. This study was performed with the French company Survi-

sion, whose domain is video surveillance and where Bruno works as senior engineer, within the scope of an industrial contract (2003-2004) that I supervised. For this work, coupled with the study on models of variability for local descriptors (work on similarity), Bruno obtained the award of the best CNAM engineer thesis in 2005. Because such a combination of features provides powerful synergies for describing object's content, we went further into this approach with object model-free recognition as objective. This work was initiated by the following observation: all the state-of-the-art approaches cited in section 1.1.3 rely on local features, with the aim of taking advantage of their interesting properties particularly relevant for object recognition in difficult conditions involving occlusions and cluttering. Unfortunately, by definition these features are not able to provide a global description of the object appearance. On the other hand, a more global description, such as for example a global shape, could be highly informative during the recognition process. But in general, global features are difficult to exploit when objects are mixed with background clutter. A preliminary step of image segmentation or of object detection is required before being able to exploit it. Moreover, object recognition with local features usually provide a poor localization of the recognized object. Here, one underlying objective was to exploit global features also with the aim of being able to obtain a very precise localization of the object. We have investigated these ideas during the PhD thesis of Bruno Lameyre (September 2005 - early 2009, Survision funding, CNAM), also within the scope of an industrial contract with Survision (2006-2009) under my supervision. We have proposed a solution that combines the advantages of local descriptors with those of more global descriptors to perform object recognition and precise localization without considering any prior step of segmentation nor of detection [Lameyre and Gouet-Brunet, 2006b; Gouet-Brunet and Lameyre, 2008]. At Survision, this work was implemented efficiently and is currently applied to surveillance of truck traffic on parking areas of French motorways.

Spatio-temporal description of video content *(Chapter 3)*

As stated in section 1.1.3, since few years video contents have become widely available, and are of great interest for researchers. I orientated my research in this direction because compared to images, such contents provide a lot of new relevant information, new challenges and applications, that we explored at INRIA with local descriptors. In 2004, we started a collaboration with the INA (Institut National de l'Audiovisuel) which is a public French organization whose main objective is to preserve the national audiovisual heritage by collecting and storing audiovisual programmes. At present, INA stores more than 300,000 hours of digital videos and has several needs such as the protection and the traceability of these contents. Through the PhD thesis of Julien Law-To (2004-2007, CIFRE funding, Versailles Saint-Quentin University, defense December 14, 2007), co-supervised by Nozha Boujemaa, Olivier Buisson (researcher at INA) and myself, we focussed on the problem of video copy detection [Law-To, 2007]. These three years of research led to the proposal of the system ViCopT (for Video Copy Tracking), which is the object of chapter 3. The main characteristic of this system is to describe the spatio-temporal behavior of local descriptors along video sequences with bottom-up and top-down cues, to provide a very distinctive description of the video content [Law-To et al., 2006d; Law-To et al., 2006a; Law-To et al., 2006c]. ViCopT was developed with the aim of being able to make the distinction between copies and near duplicates, to be robust to severe image and video transformations, as well

as dealing with large volume of video collections. It is important to note that, while being driven by the application - video copy detection, the solutions proposed during the thesis of Julien were designed with the aim of being sufficiently generic to allow addressing other applications of video content indexing (see the discussion of section 3.5 and publication [Law-To et al., 2007b]).

At this period, I took part in the European Network of Excellence MUSCLE on "Multimedia Understanding through Semantics, Computation and Learning" (FP6, 2004-2007) where I was deputy leader of the JPA2 work package on "Content-Based Description" (leader Nozha Boujemaa) and leader of the task on "Image and video processing". Julien's thesis was partly supported by MUSCLE, which allowed us collaborating with other European teams on the problem of visual saliency, to evaluate ViCopT against other systems dedicated to the same task (publication [Law-To et al., 2007a]) and to promote the system at industrial events such as the Cebit 2007 (international professional show dedicated to digital technologies).

Spatial relationships between objects in images *(Chapter 5)*

I started to investigate the problem of spatial layout description in images since 2007 in collaboration with Maude Manouvrier and Marta Rukoz of the LAMSADE laboratory of Paris-Dauphine University, within the scope of the federation Wisdom (2007-2010). Describing the spatial layout of images represents a very distinctive information that can drastically improve CBIR systems, but a lot of challenging research remains to do on this topic, see the publication [Gouet-Brunet et al., 2008] and chapter 5. In particular, our activities on this topic rely on the design of image analysis tools for representing the spatial layout of images as well as on the design of efficient multidimensional access methods for retrieval of images according to this representation. I have focussed more deeply on this topic during my sabbatical leave ("CRCT", February - July 2008). During this period, we have co-supervised the Master thesis of Nguyen Vu hoang (Master ISI, Paris-Dauphine University, Wisdom funding), who proposed a new approach of spatial layout representation, called δ-TSR, which is an amelioration of the state-of-the-art approach TSR. In addition, we have proposed two new research directions, concretized by the hiring of two PhD students that currently start (2008-2011): one concerns the proposal of a hierarchical description of the spatial layout of image contents, while the other addresses the definition of a spatial context for object retrieval/recognition, described with spatial relationships models.

1.2.2 Structuring of feature spaces for visual contents *(Chapter 4)*

In 2002, discussions with Michel Scholl and then my hiring as assistant professor at CNAM marked the beginning of my activity on access methods with multidimensional indexes. Literature on multidimensional index structures is very abundant (see for example the book [Samet, 2006] for an exhaustive survey). Research on this topic has drawn considerable attention and is still very active because of the production of more and more larger volumes of multimedia data and of the proposal of more and more sophisticated techniques of description, while several important problems persist when one is interested in accelerating

similarity search in feature spaces having a high dimension and containing many data. The most known is probably the problem of the curse of dimensionality: it gathers nonintuitive phenomena observed when the dimension of data increases and then makes reasoning on such spaces difficult and indexes less efficient than a sequential scan from a given dimension. As researchers on image descriptors try to bridge the semantic gap, researchers studying multidimensional access methods have the ambition of pushing back the limits of the curse of dimensionality. My research activity on this topic has been also motivated by this goal, but by considering data sets belonging to CBIR and by exploiting their specificities. We have addressed it according to the three directions, introduced in the following and fully described in chapter 4. They have been mainly investigated at CNAM since 2004 within the scope of the PhD thesis of Nouha Bouteldja (co-supervised with Michel Scholl, MENRT funding, defense December 2008).

Evaluation of multidimensional index structures *(Chapter 4)*

By studying literature on this topic, we observed that there is still no common framework for the evaluation of indexing structures, making them hardly comparable. In addition, many assumptions, on the curse of dimensionality phenomena as well as on the data sets used for evaluation, are made, such as the hypothesis of uniformity of the data set distribution. These difficulties and lacks drove us to re-evaluate the curse of dimensionality phenomena and the behavior of some state-of-the-art approaches, here on several data sets coming from CBIR and differently distributed [Bouteldja et al., 2006b; Bouteldja et al., 2008a].

Multiple queries *(Chapter 4)*

When considering feature vectors associated to image descriptors, there exist a lot of configurations where the entities of interest (image, part of image, object, video, video segment, etc) are described with descriptors that produce several feature vectors. For example, the description based on interest points can involve several hundreds of feature vectors associated to the points extracted, see for example figure 1.1(b). Consequently, performing a similarity search of these entities in a database involves a query defined by several feature vectors of the same family, often called a *multiple query*. One part of our work consisted in studying new strategies for accelerating retrieval of such configurations of feature vectors. The objective was to perform nearest neighbors retrieval of a multiple query more efficiently than performing several consecutive simple queries, by exploiting simultaneously all the feature vectors of the query during search. This work was published in [Bouteldja et al., 2006b] and supported by the French project BIOTIM (ACI "Masses of Data", 2003-2006) where it was applied to the fast search of similar images described with local descriptors in the context of large volumes of images dedicated to biodiversity.

HiPeR *(Chapter 4)*

Based on our previous studies and experiences on multidimensional access methods, we have proposed HiPeR, a Hierarchical and Progressive Retrieval model developed for speed-

ing up retrieval of nearest neighbors in high-dimensional spaces. This proposal represents the core of our work on multidimensional index structures. Here, it is important to notice that our aspiration was not to propose another index structure that performs retrieval more efficiently than other ones, but rather to define a model that allows integrating the most efficient state-of-the-art techniques into a framework, with the objective of making them more efficient in terms of retrieval time. The proposed model is said "index independent" in the sense that it can be applied to the index that gives the best performance with a given data set, to improve the performance of this index. In our experiments, HiPeR was in particular instantiated and evaluated on two different competitive access methods: the VA-File, which performs sequential scan more rapidly by the way of approximations, and the GC*tree, which is a tree-structured and space partitioning approach. For both these structures, we demonstrated the contribution of the model. HiPeR was designed to address three strategies of similarity search: exact search (published in [Bouteldja et al., 2008a]), approximate search (published in [Bouteldja et al., 2008b; Bouteldja et al., 2009]) and progressive search (published in [Bouteldja and Gouet-Brunet, 2008; Bouteldja et al., 2008c]).

During the period of this activity, I have also supervised two master theses: Arnaud Tournier (DEA SIR, Paris 6, 2002), Sami Stitou (with Nouha Bouteldja, Paris-Dauphine, 2006); and several engineer theses: Nizar Grira (with Nozha Boujemaa, SupCom Tunis, 2002), Akram Hentati (with Nozha Boujemaa, SupCom Tunis, 2003), Rochdi Bouchiha (with Michel Scholl, ENIS Tunis, 2004), Michel Martinez (with Nouha Bouteldja, CNAM, 2007). All of them contributed to this activity by developing or adapting state-of-the-art index structures, and evaluating them against other methods in a common framework (SR-tree, VA-File, X-tree, Pyramid-technique, GC*tree).

1.2.3 List of publications

This section gathers the list of my publications in journals, book chapters, conferences and workshops directly related to the research activities since 2001 previously described. Other publications (work done during my PhD thesis and publications concerning more general or other topics) as well as demonstrations, research reports and web tutorials are not reported here but are listed in my CV. The publications joined to this report (one per activity) are marked with symbol ★.

▶ Publications on visual description of image contents

[Gouet and Boujemaa, 2001] Gouet, V. and Boujemaa, N. (2001). Object-based queries using color points of interest. In *IEEE Workshop on Content-Based Access of Image and Video Libraries (CVPR/CBAIVL 2001)*, pages 30–36, Kauai, Hawaii, USA.

[Gouet and Boujemaa, 2002] Gouet, V. and Boujemaa, N. (2002). On the robustness of color points of interest for image retrieval. In IEEE *International Conference on Image Processing (ICIP'02)*, pages 377–380, Rochester, New York, USA.

[Boujemaa et al., 2003] Boujemaa, N., Fauqueur, J., and Gouet, V. (2003). What's beyond query by example? In *IAPR International Conference on Image and Signal Processing (ICISP'2003)*, Agadir, Morocco.

[Boujemaa et al., 2004a] Boujemaa, N., Fauqueur, J., and Gouet, V. (2004a). *Trends and Advances in Content-Based Image and Video Retrieval*, book chapter What's beyond query by example? L. Shapiro, H.P. Kriegel, R. Veltkamp (eds.), LNCS, Springer Verlag.

[Gouet and Lameyre, 2004] Gouet, V. and Lameyre, B. (2004). SAP: a robust approach to track objects in video streams with snakes and points. In *British Machine Vision Conference (BMVC'04)*, pages 737–746, Kingston University, London, UK.

[Lameyre and Gouet, 2004] Lameyre, B. and Gouet, V. (2004). Object tracking and identification in video streams with snakes and points. In *5th Pacific-Rim Conference on Multimedia (PCM'04), LNCS 3333 Springer-Verlag*, pages 61–68, Tokyo, Japan.

[Boujemaa et al., 2004b] Boujemaa, N., Fleuret, F., Gouet, V., and Sahbi, H. (2004b). Automatic textual annotation of video news based on semantic visual object extraction. In *IS&T/SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia*, San Jose CA, USA.

[Law-To et al., 2006b] Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2006b). Labeling complementary local descriptors behavior for video copy detection. In *IAPR and EURASIP International Workshop on Multimedia Concept Representation, Classification and Security (MCRCS'06)*, pages 290–297.

[Lameyre and Gouet-Brunet, 2006a] Lameyre, B. and Gouet-Brunet, V. (2006a). Connecting local and global descriptors for generic object recognition in videos. In *6th IEEE International Workshop on Visual Surveillance (VS'06, in conjunction with ECCV'06*, pages 57–64, Graz, Austria.

[Lameyre and Gouet-Brunet, 2006b] Lameyre, B. and Gouet-Brunet, V. (2006b). Connexions entre descripteurs locaux et globaux pour la reconnaissance d'objets dans les vidéos. In *Compression et Représentation des Signaux Audiovisuels (Coresa'06)*, pages 207–212, Caen, France.

★ **[Gouet-Brunet and Lameyre, 2008]** Gouet-Brunet, V. and Lameyre, B. (2008). Object recognition and segmentation in videos by connecting heterogeneous visual features. *Computer Vision and Image Understanding (CVIU)*, 111(1):86–109.

[Gouet-Brunet et al., 2008] Gouet-Brunet, V., Manouvrier, M., and Rukoz, M. (2008). Synthèse sur les modèles de représentation des relations spatiales dans les images symboliques. *Revue des Nouvelles Technologies de l'Information (RNTI).*

[Gouet-Brunet, 2008a] Gouet-Brunet, V. (2008a). *Encyclopedia of Database Systems: Multimedia Databases*, book chapter Image. L. Liu and T. Özsu (eds.), Springer Verlag.

[Gouet-Brunet, 2008b] Gouet-Brunet, V. (2008b). *Encyclopedia of Database Systems: Multimedia Databases*, book chapter Image representation. L. Liu and T. Özsu (eds.), Springer Verlag.

▶ Publications on spatio-temporal description of video contents

[Law-To et al., 2006d] Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2006d). Local behaviour labelling for content-based video copy detection. In *18th IAPR International Conference on Pattern Recognition (ICPR'06)*, pages 232–235.

★ **[Law-To et al., 2006a]** Law-To, J., Buisson, O., Gouet-Brunet, V., and Boujemaa, N. (2006a). Robust voting algorithm based on labels of behavior for video copy detection. In *14th ACM International Conference on Multimedia (ACM Multimedia'06)*, pages 835–844, Santa Barbara, USA.

[Law-To et al., 2006c] Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2006c). Labellisation du comportement de descripteurs locaux pour la détection de copies vidéo. In *Compression et Représentation des Signaux Audiovisuels (Coresa'06)*, pages 336–341, Caen, France.

[Law-To et al., 2007a] Law-To, J., Chen, L., Joly, A., Laptev, Y., Buisson, O., Gouet-Brunet, V., Boujemaa, N., and Stentiford, F. (2007a). Video copy detection: a comparative study. In *ACM International Conference on Image and Video Retrieval (CIVR'07)*, pages 371–378, Amsterdam, The Netherlands.

[Law-To et al., 2007b] Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2007b). Video copy detection on the internet: the challenges of copyright and multiplicity. In *IEEE International Conference on Multimedia & Expo (ICME'07)*, pages 2082 – 2085, Beijing.

▶ Publications on multidimensional index structures

[Gouet-Brunet, 2006] Gouet-Brunet, V. (2006). *Encyclopédie de l'Informatique et des systèmes d'information*, book chapter Recherche par contenu visuel dans les grandes collections d'images. J. Akoka and I. Commyn-Wattiau (eds.), Vuibert.

[Bouteldja et al., 2006b] Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2006b). Evaluation of strategies for multiple sphere queries with local image descriptors. In *IS&T/SPIE Conference on Multimedia Content Analysis, Management and Retrieval*, volume 6073, pages A1–12, San Jose CA, USA.

[Bouteldja et al., 2008a] Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2008a). HiPeR: Hierarchical progressive exact retrieval in multidimensional spaces. In *International Workshop on Similarity Search and Applications (SISAP'08, in conjunction with ICDE'08)*, pages 25–34, Cancún, Mexico.

[Bouteldja and Gouet-Brunet, 2008] Bouteldja, N. and Gouet-Brunet, V. (2008). Exact and progressive image retrieval with the HiPeR framework. In *IEEE International Conference on Multimedia & Expo (ICME'08)*, pages 1257–1260, Hannover, Germany.

★ **[Bouteldja et al., 2008b]** Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2008b). HiPeR: Hierarchies for approximate and exact retrieval in multidimensional spaces. In *Journées Bases de Données avancées (BDA'08)*, Guilherand-Granges, France. To appear.

[Bouteldja et al., 2008c] Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2008c). The many facets of progressive retrieval for CBIR. In *Pacific-Rim Conference on Multimedia (PCM'08)*, Tainan, Taiwan. To appear.

[Bouteldja et al., 2009] Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2009). Approximate retrieval with HiPeR: Application to VA-Hierarchies. In *ACM International Multimedia Modeling Conference (MMM'09)*, Sophia-Antipolis, France. To appear.

Chapter 2

Local description of image contents

This chapter gathers all the research work I have realized on the description of the image contents since 2001. My contributions to this topic concerned the extraction of visual information to describe finely image contents with local descriptors as well as the study of fine similarity measures dedicated to them. During period 2001-2004, one part of my research work at INRIA and at CEDRIC concerned the study of similarity measures for improving matching of images described with local descriptors and is described in section 2.1. This activity was supported by the European program "STOP" (2000-2001) and by the French project MediaWorks (RIAM, 2001-2004). With the recent proposal of new local features having different topological natures, one challenging idea has been to combine them judiciously to improve image content description. I have devoted a large part of my research on image description to the exploitation of this concept. During the PhD thesis of Julien Law-To (2004-2007) at INRIA, we proposed to combine different categories of interest points to improve video frames description, with application to video copy detection. This work is described in section 2.2. In parallel, during the Master thesis (2003), engineer thesis (2004-2005) and PhD thesis of Bruno Lameyre (2005-2009) at CNAM, we chose to combine local and global features, with the aim of enriching image description for object tracking, recognition and segmentation, with application to visual surveillance. This work was supported by industrial contracts with the French company Survision (2003-2004 and 2006-2009) and is described in section 2.3. Note that part of the activities presented here has taken video sequences as input, with applications to video copy detection and to object recognition for video surveillance. The main contributions described apply on video frames and then mainly consider videos as sets of images. The combination of local features as well as local and global features are applicable on pure collections of images, justifying their presentation in this chapter.

Contents

2.1 Similarity between images

We tackled the problem of the similarity between images though local descriptors at two levels of description: "interest point" level and "groups of interest points" level. Because local descriptors are very discriminant and then can be very sensitive to image transformations and noise, I began a study on the modeling of their variability for these transformations, with the main goal of integrating it in the similarity measure associated with local descriptors. This part is described in section 2.1.1. In parallel, in the continuity of the work realized during my PhD thesis on robust stereo image matching with geometrical information [Montesinos et al., 2000], I focussed on the study of solutions for integrating the spatial distribution of sets of interest points into the image description, adapted to the matching of large sets of images. This part is described in section 2.1.2.

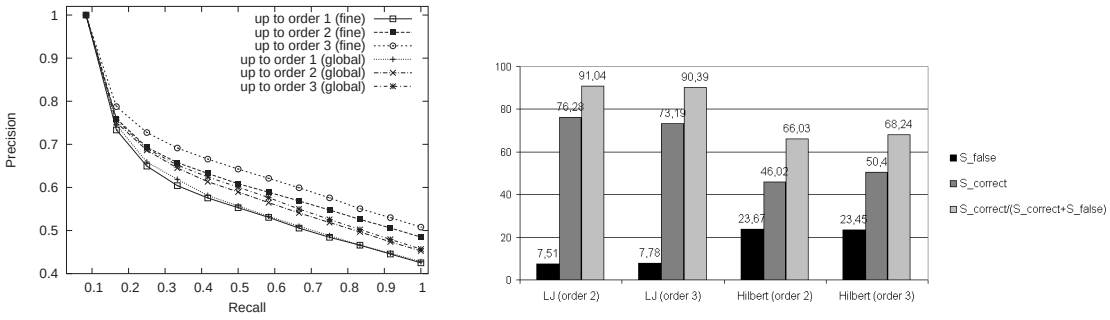
2.1.1 Modeling the variability of local features

Local descriptors are subject to different kinds of noise and consequently may vary from an image to another: by definition, they are at best only quasi-invariant to any point of view and in practice, they are sensitive to image acquisition (sensors and sampling errors may be important for images coming from video sequences), to numerical errors, to interest point extractor precision in localization, etc. These considerations show the importance of the similarity measure which must be carefully chosen to achieve best performances. In particular, an optimal similarity measure is directly related to the shape of the variability of the involved features. Under the gaussian assumption, a similarity measure commonly employed to compare two multidimensional feature vectors $\vec{v}_1$ and $\vec{v}_2$ is the Mahalanobis distance $\delta^2(\vec{v}_1, \vec{v}_2) = (\vec{v}_1 - \vec{v}_2)^T \Lambda^{-1} (\vec{v}_1 - \vec{v}_2)$. The involved covariance matrix Λ can take into account the different magnitudes, possible correlations and variability of the feature components. When a model of the variability of the components cannot be theoretically specified, the way to estimate Λ comes down to different empiric solutions:

- Estimating Λ from all the available data. This simple solution generates weights that are not discriminant, since representing a rough model of variability. However, this is the most common model encountered in literature to compare features with the Mahalanobis distance, because it is easy to compute. We call it Λ_{global}.
- Estimating Λ from points of interest whose local neighborhood is submitted to typical synthetic photometric and geometric transformations and perturbations that usually apply to images. Another solution is to estimate Λ from training sequences of real images: several points on different images with representative perturbations are tracked and a combination of the covariance matrices obtained can be used as the model of features variability; this solution was adopted in [Schmid and Mohr, 1997] for image retrieval. The Mahalanobis distance obtained from these approaches is noted Λ_{fine}.

In 2004 with Jean-Philippe Tarel at INRIA, we performed experiences to estimate the variability of local features components by using simulated perturbations in images [Imedia, 2004]. The model Λ_{fine} obtained was applied to the color local jet which is a descriptor based on the partial derivatives of the signal, and evaluated on the Columbia

coil-100 database [Nene et al., 1996]. It enabled us to improve retrieval results comparing to retrieval performed with model Λ_{global}, whatever the order of derivatives employed, as illustrated in figure 2.1(a). In particular, the best improvements were obtained on jets involving higher orders of derivatives, that are reputed to be very discriminant but sensitive to noise; in comparison, with model Λ_{global}, the use of order 3 does not clearly improve description . These results confirmed us that a model of variability is important to finely exhibit the relevance of the features employed.



(a) On the database of images coil-100 (b) On video surveillance frames

Figure 2.1: *Retrieval with/without modeling the variability of the descriptor. (a) Average precision vs. recall obtained for local jets up to order 3 according to models Λ_{global} and Λ_{fine} ; (b) Point matching scores obtained with model Λ_{fine} and two point characterizations: the local jet (LJ) and the Hilbert's differential invariants (up to orders 2 and 3). $S_{correct}$ is the number of correct matches, while S_{false} is the number of uncorrect matches.*

However, this analysis has to be qualified: additional experiments demonstrated that the improvements obtained with the very sensitive orders of the jet strongly depend on the support where there are computed. Even with model Λ_{fine}, these features do not resist to images of poor quality, as we demonstrated it in [Gouet and Boujemaa, 2002] for image retrieval facing JPEG compression. During the Master thesis of Bruno Lameyre, these results were confirmed on video frames issued from surveillance cameras, where Λ_{fine} was estimated by tracking interest points along sequences: here, local jet up to order 2 provides better results than up to order 3; see figure 2.1(b) and publication [Gouet and Lameyre, 2004]. An important conclusion of these studies is that such features must be used with attention: adding very discriminant but sensitive information into the feature vectors does not necessarily improve the description. Even so, we observed that a lot of work in literature exploits these features while comparing them according to a similarity measure based on the global model Λ_{global}, whatever the quality of the images considered.

Another interesting result is that such models of variability can be also employed to specify more precisely the range where descriptor components are allowed to vary, without enforcing complete invariance, as it is usually performed. For instance, it is possible to constrain invariance only to a range of rotation angles. in [Gouet and Lameyre, 2004], we estimated Λ_{fine} for the local jet components that are only invariant to translation on training sequences differing from small rotations. As illustrated in figure 2.1(b), the best matching results were obtained with this model, facing Hilbert's descriptors that are invariant to rotation. Similar results were obtained with the coil-100 database, where objects differ only from a small range of image rotation [Imedia, 2004]. These experiments

lead to the following conclusion we generalize to other usual image transformations: when images or objects differ from small transformations (it is typical in consecutive frames of video sequences), the way to develop the most distinctive and robust descriptor is to keep it the less invariant possible and to make it tolerant to these transformations by learning its variability facing them. Developing features with larger degrees of freedom naturally leads to a less selective description.

2.1.2 Spatial consistency of groups of interest points

Retrieving similar images, parts of images or objects that are described with sets of local descriptors ideally consists in comparing sets of points, i.e. checking if some of them are similar according to their local description and eventually if the topologies of the matched (sub)sets are similar. Theoretically, this problem involves graphs comparison, where vertices are the interest points and edges can be the spatial relationships existing between them. In practice, using graph matching is possible when comparing two images, for example for stereovision purposes as we did during my PhD thesis by proposing an iterative relaxation algorithm dedicated to the matching of two large sets of interest points [Montesinos et al., 2000]. But it is not usable in general when dealing with sets of images because of the well-known complexity of graph matching algorithms. In CBIR, checking the spatial configuration of sets of points is usually a work done *a posteriori* when similar points have been found in candidate images. How this problem is handled depends on the geometric transformations existing between the images. If the data or the application make the type of transformation known, then usually this problem consists in estimating the existing transformation and in checking if most of the similar points agree it. Otherwise (unknown transformation, global model not existing such as with viewpoint change, non rigid scenes, etc), more general constraints can be employed, such as semi-local geometric constraints that only suppose the semi rigidity of the scenes involved in the images.

During my postdoc at INRIA, I addressed this problem under these two configurations. Section 2.1.2.1 revisits the work performed for logo detection, where geometric transformations are known, while section 2.1.2.2 presents the solution we proposed to deal with complex transformations and non rigid scenes.

2.1.2.1 Specific constraints for logo detection

This work was realized within the French project MediaWorks (2001-2004, RIAMM), that consisted in the proposal of an hybrid text-image indexing and retrieval platform for video news. One objective of this project was developing tools for helping archivists of the French TV channel TF1 in the automatic annotation (by suggesting textual metadata) of video news, based on visual content analysis. In particular, for determining the copyrights of the incoming videos, one need was the automatic detection and identification of logos, from a thesaurus of TV channels logos made available by TF1.

Local descriptors were employed because of their ability to detect particular objects in an image without requiring a delicate step of segmentation. The thesaurus of logos was indexed with local descriptors, that were also extracted from key frames of the incoming

video to study. Then, the comparison of the two sets of features allowed coming to light from one side particular points of the key frame where a logo can be, and on the other side, particular points in the thesaurus associated with candidate logos. Here, because of the specificity of the 2D objects involved in this application, the geometrical transformations existing between the logos of the thesaurus and the one of the key frame are translation and scale. We developed a specific similarity measure that integrates a registration process, by estimating the parameters of this transformation with the RANdom SAmple Consensus approach (RANSAC) from the set of matched points. Then, from this estimation, the similarity measure consisted in quantifying the reliability, in the tested key frame, of the interest points that describe the candidate logo in the thesaurus. This work is described in details in publication [Boujemaa et al., 2004b].

2.1.2.2 Semi-local constraints for complex transformations

By considering semi-local constraints in the description of the geometry of a set of interest points, it is possible to deal with complex transformations where a model is not reasonably possible, such as with non rigid scenes.

A first classical constraint, called *constraint on neighborhood*, considers that a couple of candidate points (m_1, m_2) as a good match if in the spatial neighborhood of m_1 there are enough points matched with spatial neighbors of m_2. Some heuristics based on geometrical constraints can be added to this one. It depends on the transformations existing between the images to match. In [Gouet and Boujemaa, 2001], we proposed geometrical constraints invariant to similitude group, in concordance with the local descriptors used (color Hilbert's differential invariants). The proposed constraint consisted in considering the property of angles conservation between neighbors. Some authors [Schmid and Mohr, 1997; Marie-Julie and Essafi, 1998] propose to consider the angle defined between couples of neighbors of the point to match, which must be constant from an image to another. For our part, we defined an angular constraint between the point to match and only one of its neighbors. In comparison to the state-of-the-art approaches, this angular constraint leads to a complexity of minor importance, seeing that it implies only one neighbor at once in the computation.

During the comparison process of two interest points, the comparison of the angles involved in this constraint is done for all the points inside the semi-local neighborhood of these points. Then the coherence of the angles comes to reenforce the visual similarity computed from the associated local descriptors; see [Gouet and Boujemaa, 2001] for details on the whole similarity measure. This work was implemented in IKONA, the CBIR system of Imedia for image retrieval; a demo of sub-image retrieval with these features is available here[1]. It was also applied on a concrete scenario of use, described in the following.

Application to Police investigation aids. The present research was supported by the European program "STOP" (2000-2001) related to fight against organized crime, in collaboration with the French Judicial Police where it was exploited as an investigation aids. The French Judicial Police created a department for the fight against children's pornography. This department collects images via the Internet and seizures during searches at suspect's houses. At present, the volume of accumulated images does not enable Police

[1] IKONA website: `http://www-rocq.inria.fr/cgi-bin/imedia/circario.cgi/demos`.

officers classifying these data manually. In particular, one important need of this department was the possibility to gather images involving the adult responsible of them. Because most of the time this adult is not visible in the image, the chosen criterion for classification was visual details of the crime scene, judged as representative of the adult, such as clothes or elements of the decor (painting, sofa's patterns, etc). As illustrated with the query of figure 2.2 performed on generic images from INA, it is possible to employ local descriptors in order to gather images involving the same scene, under difficult conditions such as occlusions, different viewpoints and characters. In this example, we focussed on the upper left part of the image, which shows partially a wine storeroom, and we interrogated the database with a partial query build on this area. The retrieved images are presented on the right of the figure. We show that the query area was retrieved in five images, which imply the same room but with different characters. Such a scenario precisely corresponds to the scenario of use that Police officers apply on their particular images to classify them. Because of these images differ from complex transformations and the queries can involve deformable objects (clothes), the proposal of semi-local geometrical constraints was necessary to improve sub-image/object retrieval with local descriptors.

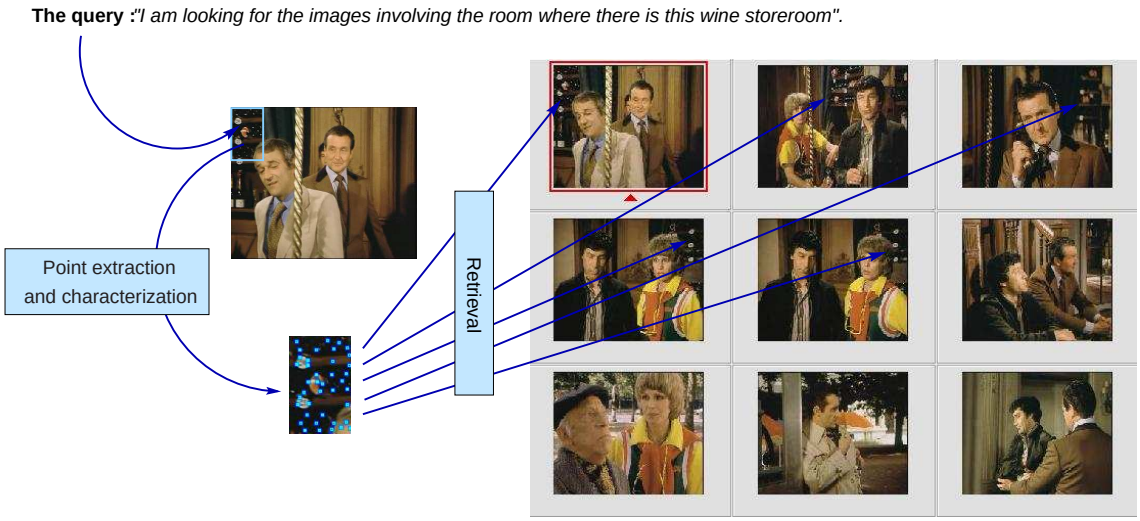


Figure 2.2: *Partial query on a particular area of the scene. On the right, the nine best responses are presented, they are sorted by decreasing order of the image score.*

2.2 Combination of different categories of local features

As stated in section 1.1.3, the recent proposal of local descriptors involving different topological natures of image support has given the possibility to study some combinations of them to improve image description. For example, in [Sivic and Zisserman, 2003], the authors use two types of viewpoint covariant regions to describe each frame of a video: the first type is constructed by elliptical shape adaptation around a Harris interest point and the region extracted is called Shape Adapted (SA), while the second type is called Maximally Stable (MS) regions and is based on an intensity watershed segmentation. They use these two types of regions of interest because each represents a different visual content: SA

regions are centered on corners and MS regions are blobs of high contrast.

2.2.1 Harris points and symmetry centers

During the PhD thesis of Julien Law-To dedicated to video copy detection, we chose to describe the frames content of video sequences by considering the precise version of the Harris detector of points of interest [Harris and Stephens, 1988; Schmid and Mohr, 1997], because this detector is well-known for its good repeatability and has a competitive computational time of extraction. While this detector extracts points on sites of the image characterized with high variations of intensity in several directions, we decided to exploit *points of symmetry* jointly to this detector. Such points are extracted on sites of images whose local neighborhood contains symmetrical edges. Several implementations of detectors of symmetry points exist: Reisfeld et al. developed an attention operator based on the intuitive notion of symmetry which is thought closely related to human focalization of attention [Reisfeld et al., 1994], followed by [Heidemann, 2004] who proposed a color version of this detector. Loy [Loy and Zelinsky, 2003] developed a fast algorithm for detecting symmetry points of interest, that uses image gradients.

The decision to combine these two kinds of local detectors to index the visual content of videos follows from the fact that, on the one hand, the two supports involved clearly do not describe the same sites, as illustrated in the images of figure 2.3.

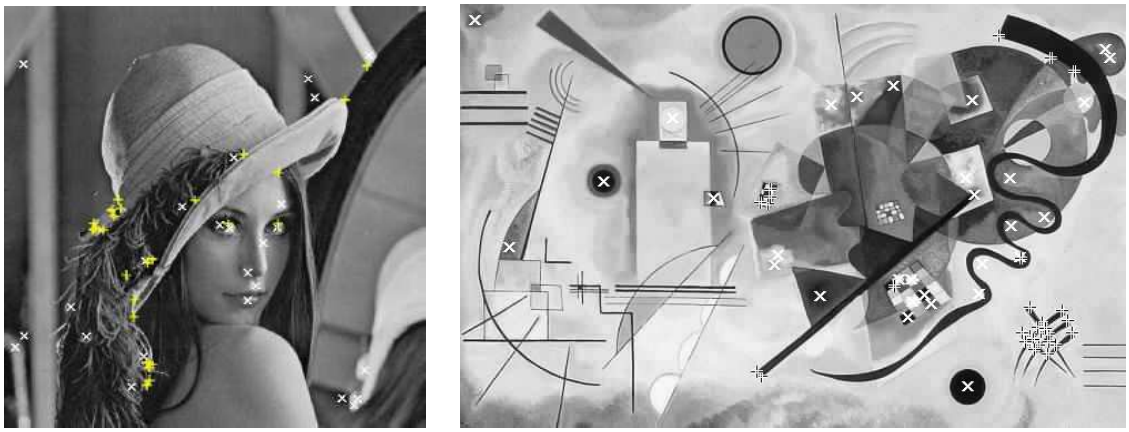


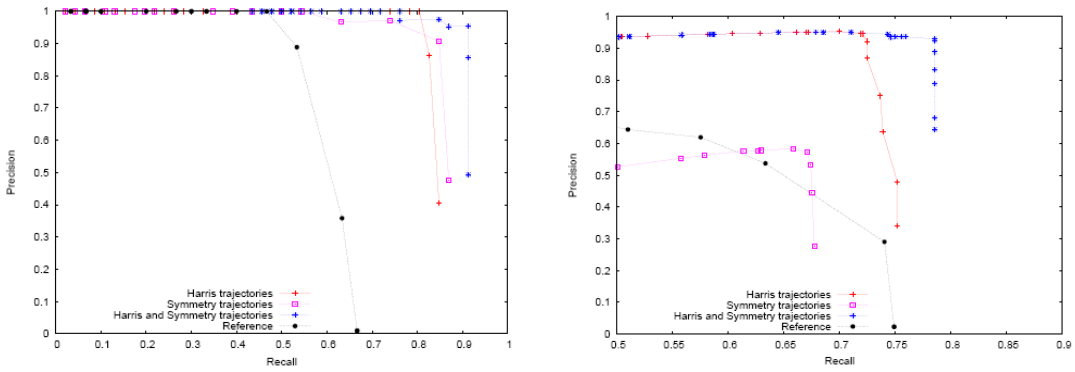
Figure 2.3: *Extraction of two categories of interest points: Harris points (+) and symmetry points (×). W. Kandinsky's painting: "Yellow, Red, Blue", 1925.*

On the other hand, symmetry points have shown some interesting semantic properties: because they correspond to sites of visual attention [Locher and Nodine, 1987], we feel they have the ability to index areas of interest with a strong semantic content, that should be reasonably less damaged by human post-production modifications that can occur when considering video copy detection. This affirmation is strengthened by several works that connect symmetry with visual attention and then with semantic, such as [Lin and Lin, 1996] who demonstrated that symmetry is an advantage in applications of copy detection for TV sequences, and Privitera and Stark [Privitera and Stark, 2000] who showed by a comparison with an eye-tracking system that a local symmetry algorithm is very efficient

for finding human regions of interest in generic images.

2.2.2 Evaluation dedicated to video copy detection

By jointly integrating Harris points and symmetry corners into the video frames description, these hypotheses were confirmed by the experiments we performed for video copy detection on a database of 1,000 hours of heterogeneous video contents taken from the archives of INA. In particular, we compared the retrieval results obtained using only one nature of points of interest and those using both natures. Figure 2.4 presents the associated precision and recall curves, for retrieval of video segments and retrieval of frames (that measures the temporal precision of the detected segments and then indicates if the boundaries of the detected segment are precisely retrieved). The reference technique compared in the one proposed in [Joly et al., 2007] that exploits only Harris points and not trajectories of Harris and/or symmetry points as we did.



Precision/Recall by segment Precision/Recall by frame

Figure 2.4: *Complementarity of the natures of interest points.*

Whatever the scenario of retrieval, we observed that description based on symmetry centers is globally less distinctive than description based on Harris points, even if it shows a slightly better maximum recall (but less precision) than Harris points for segment retrieval. Using them jointly is clearly profitable to improve the number of video segments correctly detected, as well as the temporal precision of the detection: for retrieval of segments, their complementarity allows an increase of 4% in the recall of Harris points and of 11% in the recall of symmetry centers for a precision of 95%. When considering retrieval of frames, using both natures of trajectories really increases the recall of more than 4%, while keeping a high precision.

This work was published in [Law-To et al., 2006b] and additional experiments are available in the PhD thesis of Julien Law-To [Law-To, 2007].

2.3 Synergies between local and global features

Improving the local description of visual contents with other features, as explained in section 1.1.3 mainly relies on local features. Unfortunately, by definition these features are not

able to provide a *global* description of the object appearance. On the other hand, a generic and more global description of the object appearance, such as for example global shape, dominant colors of the object or global texture description, could be highly informative during the recognition process, as demonstrated in [Yan et al., 2004] and in [Lisin et al., 2005a] where recognition on objects is improved by combining local and global features. But in general, global features are difficult to exploit when objects are mixed with background clutter. A preliminary step of image segmentation or of object detection is required before being able to exhibit such a global description. Unfortunately, these pre-processings are incompatible with pure model-free object recognition from ordinary images. Indeed, segmentation can be reasonably achieved with specific images (as in [Lisin et al., 2005a] where images are marine organisms) or with specific objects that can be detected according to models (as in [Yan et al., 2004] where the objects are faces), but it generally remains a delicate processing requiring some a priori knowledge about the image, to build a segmentation related to the high-level features that objects are, as explained in [Borenstein and Ullman, 2002] where object segmentation is performed by the use of top-down processes.

From these observations, we proposed an approach for improving object content description in ordinary still images or video frames. Our main objective was to combine the potentiality and advantages of local descriptors, i.e. their *robustness* to complex images, with those of more global descriptors, i.e. their *richness*, to perform object recognition without considering any prior step of segmentation nor of detection. The underlying objective was also to be able to obtain a precise localization and segmentation of the recognized object, which is usually lacking with approaches only based on local descriptors.

As global descriptor, our choice fell on active contours and is discussed in section 2.3.1. Then I present in section 2.3.2 a preliminary work on the use of such descriptor with local features to enhance object tracking in video sequences. This work was extended to object recognition and segmentation and is described in section 2.3.3.

2.3.1 Active contours as global descriptors

Characterizing the appearance of an object with features of higher level than local descriptors can be done with several image primitives: local shapes, contours, regions, etc. To evaluate the validity of our approach, we have chosen to describe the global shape of objects with active contours, also called snakes [Kass et al., 1988]. There were several a priori reasons for this choice: they are quite easy to track in video sequences, because their deformation is small between two consecutive frames. In addition, we have experimented that they can help during points tracking and vice versa, as presented in section 2.3.2. But moreover, snakes describe the global shape of objects, providing a high-level visual description of them. Used alone, this information can be sufficient to recognize most of the objects, as illustrated with the examples of figure 2.5. Because computed locally, interest points do not describe this global high-level information at all. Moreover, they do not characterize the same parts of the object, since the solution chosen to describe its shape is contour-based (and not region-based). We thought that their complementarity may produce a profitable synergy.

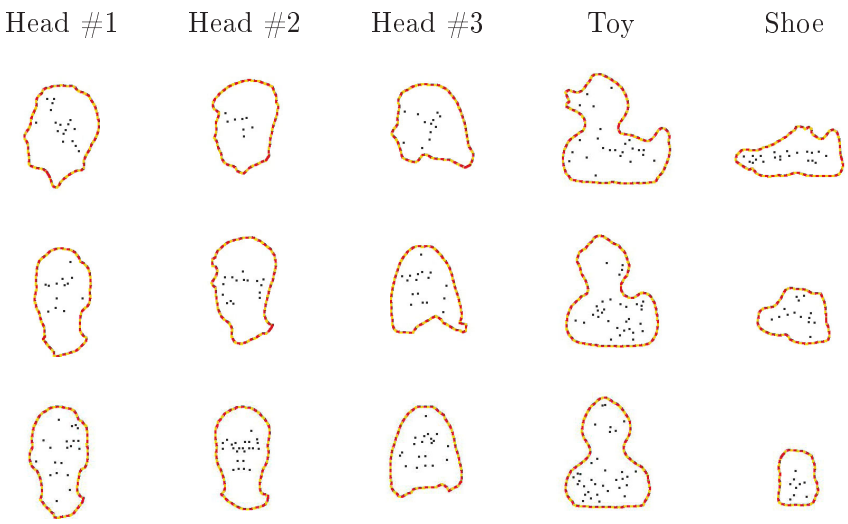


Figure 2.5: *Examples of local and global descriptors associated with five moving objects.*

2.3.2 Combination for robust tracking in video sequences

Combining complementary local and global features can be exploited for enhancing object tracking in video sequences. We built on such a combination according to two directions: first, the snake allows reducing the points tracking to a limited area in each frame, and secondly the spatial point description can be exploited to improve snake tracking along frames. The process was made robust to wide occlusions by the exploitation of a *short term memory* that temporally stores the visual appearance of the object before occlusion and helps in tracking it again when it reappears entirely.

This work is fully described in publication [Lameyre and Gouet, 2004] where it was experimented on video sequences involving severe occlusions. It was done during the Master thesis (2003) and CNAM engineer thesis (2004-2005) of Bruno Lameyre, who obtained the award of the best CNAM engineer thesis in 2005 for this work combined with the study on models of variability for local descriptors (see section 2.1.1).

2.3.3 Combination for recognition and precise segmentation

The work initialized with Bruno Lameyre during his engineer and Master thesis has been deepened during his PhD thesis (2005-2009) where we focus on object recognition and segmentation. Before presenting the proposed approach in the following sections, it is necessary to remind what the term "object recognition" covers and what are our objectives during this thesis. Recognition of 3D objects from 2D views has been an active research area for a long time. It can be presented according to the three following problems:

Categorization: also called *generic object recognition* or *category-level recognition*, it concerns the identification of a class of objects (e.g. a bicycle) in a given image, among several classes of objects. When no prior knowledge of the classes of objects can be used, the solutions encountered are *model-free* and involve bottom-up features.

According to the approach, categorization can provide a label to the image, and can also go further by providing a more or less localization of it. Obviously, the additional task of localization of the objects increases the complexity of the problem.

Recognition: also called *specific object recognition* or *instance-level recognition*, it differs from categorization in the sense that recognition would distinguish between two structurally distinct objects of the same category (e.g. "Eric's bicycle").

Detection: this refers to deciding whether or not *one class of object* is present in the image. Classically, the encountered approaches are *model-based*, insofar as they suppose a prior knowledge of the class (e.g. faces) and exploit its specificity with top-down and bottom-up features. By scanning a window over the image at all positions, detectors provide a localization of the object. In theory, it would be possible to perform categorization or recognition by applying a detector for each category or instance to the image, but in practice, this solution becomes inefficient given a large number of objects.

The work done during the PhD thesis of Bruno Lameyre intends to better answer the problem of specific object recognition, with application to visual surveillance of video sequences. While being able to provide a temporal localization of the recognized object in the video sequence, we are also interested in the spatial localization problem. Surveillance imposes near real-time performances: these requirements have orientated our choices in the techniques and algorithms proposed. This research is also done with the more general objective of video database annotation.

The main concept of the proposed approach is presented in section 2.3.3.1. The challenge consists in efficiently structuring the feature spaces involved, in order to perform recognition efficiently; this part is described in section 2.3.3.2. Then, we present the algorithms for object recognition in a *still-to-video scenario* (recognition in a given image) in section 2.3.3.3 and its generalization to *video-to-video scenario* (recognition in a video sequence) in section 2.3.3.4. As illustrated in section 2.3.3.5, the approach was also designed to deal with multiple-object recognition and to be robust to occlusions and to several object deformations. Because this work is done in collaboration with a French company on visual surveillance, I finish this part by presenting the application of our approach to surveillance of truck traffic on motorways in section 2.3.4.

2.3.3.1 Overview of the approach

Our first contribution consisted in building two feature spaces that describe the visual appearance of objects: one dedicated to the local description with interest points, describing the visual content of the objects, and the other dedicated to a global description of the objects. Here, interest points are used as *primary source* in a first and classical process of recognition. Then, points that matched are seen as *anchors* that allow going deeper towards recognition. During training, we associate some interest points with some snakes, because they belong to the same view of an object. More precisely, we will see that descriptors of the two categories are linked according to a many-to-many relationship: one snake is usually associated to several interest points, and one interest point can be associated to

several snakes. These relations reflect the intuitive idea that a given local feature cannot be associated with any object and then probably appears in the image in the neighborhood of a limited number of particular shapes, and that a given shape associated with an object probably refers to a limited number of local features describing the object content. Then, during a second step of recognition, anchor points *index* relevant snakes via the connections established. Then, these snakes are back-projected into the image, to confirm or not the first decision. In order to achieve this, our second contribution addressed the structuring of the feature spaces and more precisely the definition of *connections* between *elementary local descriptors* and *elementary global descriptors*. Besides object identification, anchor points give a first idea of the object localization, while back-projecting the snake into the image provide very good segmentation of it as a result of the recognition process. Figure 2.6 summarizes the whole approach proposed.

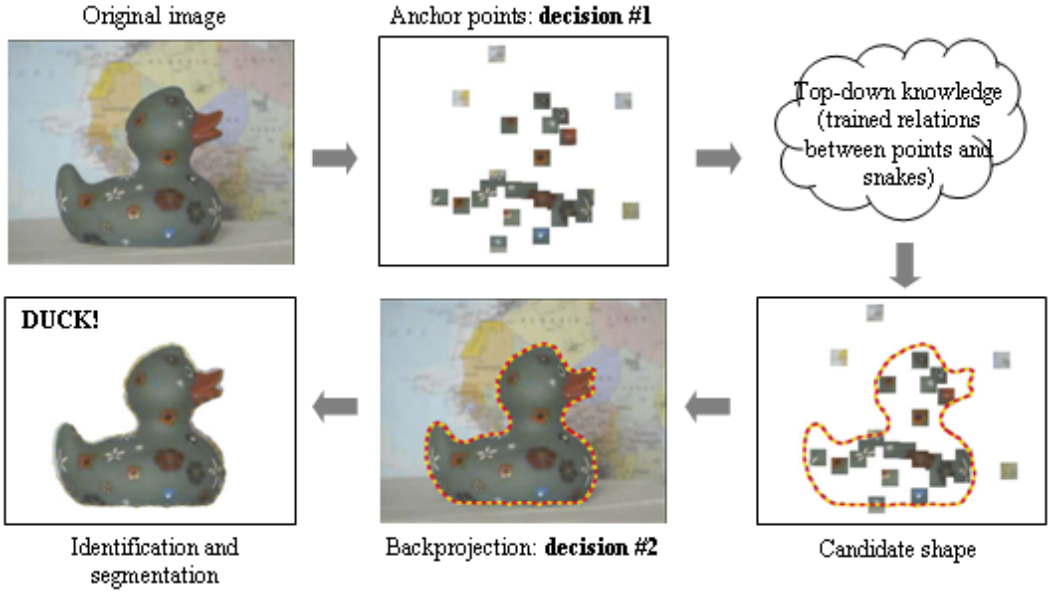


Figure 2.6: *Outline of the recognition and segmentation procedure.*

Using local and global features separately and in sequence, instead of combining them into the same descriptor, has several advantages: this solution allows combining the strong points of local descriptors, i.e. their robustness to complex images, with those of more global descriptors, i.e. their richness, to perform object recognition without considering any prior step of segmentation nor of detection. In addition, it allows keeping feature spaces of moderate dimensions (20 dimensions for each descriptor in our experiments), which becomes crucial with large volumes of data and real-time applications.

Global descriptors as top-down knowledge

Because it is directly computed from the signal without any prior knowledge about the objects to recognize, the set of anchor points is considered as a bottom-up information. Such points play the part of a coarse model-free detector, since they eliminate areas of the image where no point is matched; moreover, in the remaining area(s), the local descriptors

associated limit the number of candidate objects. They define a suitable context that makes it possible to go further by exploiting other descriptors, more informative but that could not directly be applied alone. These last descriptors, in this case a snake describing the shape of objects, can be seen as a top-down knowledge that drives final recognition and localization. Indeed, focusing on a particular snake cannot be deduced from the low-level information available at this stage, i.e. the signal and the anchor points. To decide which global information applies, this task supposes a prior knowledge that guides attention to relevant shapes: in our approach, it is the anchor points context *plus* the connections established during training. This statement is interesting because this simple top-down visual cue is learnt during training and consequently does not impose to provide to the system any goal-oriented knowledge or constraints on the objects to recognize.

The idea of using top-down knowledge combined with bottom-up features to improve recognition and localization was already proposed. Close in spirit to our method, the approach of recognition and localization of [Leibe et al., 2004] first involves local features and then an implicit model of the object shape back-projected in the tested image. But, where it differs from our model which is explicit, the shape model employed is implicit, because learnt from configurations of points; according to the results presented in the paper, the segmentation obtained is quite coarse. Similarly, in [Russell et al., 2006], multiple segmentations are combined with a bag-of-features representation to segment objects during recognition; the approach can deal with multiple instances of objects but not with occlusions, while our approach can. In [Borenstein and Ullman, 2004], image segmentation and recognition is performed by using the prior knowledge of fragments of object regions that are back-projected into the image along shapes of objects, in a jigsaw puzzle fashion. But the approach does not address the problem of object multi-occurrences nor occlusion, as in [Winn and Jojic, 2005] where the approach uses a generative probabilistic model to combine bottom-up cues of color and edge with top-down cues of shape and pose. In the same spirit, [Fussenegger et al., 2006] addresses these problems but with boundary fragments, as well as [Levin and Weiss, 2006] but with a different training process involving conditional random fields. In [Kumar et al., 2005], bottom-up and top-down features are also combined, here with the objective of object category segmentation, and not recognition.

2.3.3.2 Construction and structuring of the feature spaces

For recognition, the local description employed is the $n = 20$ first coefficients of the Discrete Cosine Transform, producing a feature space called V_n^{point}. The similarity measure associated is the Mahalanobis distance, trained to learn the descriptors variability as explained in section 2.1.1. We describe the visual appearance of the snake with a distance centroid signature converted to 1D time series with Fourier descriptors, as in [Zhang and Lu., 2001] where it was evaluated as very efficient for shape retrieval. The feature space produced is V_m^{snake}, where m is its dimension and $m = 20$ in our experiments.

Feature spaces clustering

During training, all the collected local features are gathered in V_n^{point}. This feature space is then structured in clusters called *Elementary Local Patterns* (ELPs). Clustering feature

spaces of local descriptors was applied in several works involving recognition with local descriptors from still images, in order to build a visual vocabulary in a bag-of-features representation [Sivic and Zisserman, 2003; Willamowski et al., 2004] or from videos [Grabner and Bischof, 2005; Opelt et al., 2005; Sivic et al., 2005a] in order to reduce temporal redundancy. Usually, the related approaches use k-means algorithms that fix the number of clusters to obtain. To be independent of the k parameter and because we do not need a fixed number of clusters, in our experiments we used the unsupervised clustering approach CA (Competitive Agglomeration [Frigui and Krishnapuram, 1998]) where the number of clusters is automatically determined during the processing. In a similar way, all the features associated to snakes are collected and stored in V_m^{snake}. All those shape features are clustered with the Competitive Agglomeration approach, as for the local descriptors. It provides several clusters that represent what we call *Elementary Global Shapes* (EGSs).

Feature spaces V_n^{point} and V_m^{snake} are now respectively represented by ELP and EGS clusters. Such a structuring has some advantages: firstly, it allows reducing the redundancy of the features along images and to share features inter/intra object classes, making V_n^{point} and V_m^{snake} more compact. Consequently, retrieval time may be efficiently reduced during recognition. Secondly, it will allow dynamical improvements of the objects description as recognition proceeds in new images, with a minimal growth of the catalogue: only new relevant features will be added, by generating new clusters.

Establishing many-to-many relationships between elementary features

The second level of structuring consists in *connecting* the feature spaces V_n^{point} and V_m^{snake}, and more precisely in associating ELPs with EGSs and vice versa, according to a many-to-many relationship model. Indeed, each feature point of V_n^{point} belongs to a ELP cluster and is linked to an image. This image contains a snake that has been extracted, stored in V_m^{snake} and that belongs to an EGS. Therefore, an ELP can be connected to one or several EGSs. Conversely, an EGS can be connected to several ELPs, since a snake usually surrounds several points in a frame. Such connections represent an important contribution of our work. They have the two following interesting properties:

- They integrate some semantics because they express the intuitive idea that, in a set of given objects, there probably exists a relationship between an object content and its shape (usually, human eyes and mouth do not occur with a car- or duck-like shape). The connections established between ELPs and EGSs allow modeling this high-level knowledge. Associated to them, ELPs can be seen as a context suitable for indexing particular shapes.
- Connections from ELPs towards EGSs are the only way to properly exploit a global description of the object for recognition in cluttered scenes. Other alternatives would require a time expensive step of global primitive registration in the image, or at least would require a preliminary stage of segmentation or of object detection.

In this work, the connections are established between points and snakes as a proof of concept, but such a structuring has been designed with the aim of applying with other couples of categories of descriptors satisfying a general many-to-many relationship. To be

relevant, one of the descriptors must be robust, easy to extract and match in the image (local descriptors are), while the other one comes to bring a complementary and high-level description of the object (snakes are, and we think that several other global or semi-global descriptors can apply).

2.3.3.3 Still-to-video object recognition

From the visual features extracted and structured in previous section, the algorithm of recognition and segmentation, developed for one-object recognition in an input image, can be decomposed in the following steps, fully described in [Gouet-Brunet and Lameyre, 2008]:

1. **Searching for candidate feature points.** Let $\{P_1 \ ... \ P_L\}$ be the set of local descriptors extracted in the whole image I tested. Object recognition begins by searching in V_n^{point} if these points are similar to some ELPs of the catalogue, by the way of a nearest neighbor search. If sufficiently points are similar to existing ELPs, then we suppose that I potentially contains a trained object (the candidate object) and we continue. We note these points P_i^* and the ELPs that matched are the *anchors*, noted A_j in the following. A confidence weight is associated to each of them, in function of the similarity of the points P_i^* to the ELPs.
2. **Searching for candidate shapes.** Here, the aim is to find in V_m^{snake} the best EGS clusters corresponding to the candidate object. To be considered as candidate, an EGS cluster must be connected to enough anchors. A confidence vote is attributed to each EGS, according to the proportion of connections with the anchors obtained and their associated confidence weight. At this point, only the EGS clusters associated with the best scores are selected. In the following, such EGSs are noted S_v and their prototypes s_v (medoid) are considered as the best candidate shapes for image I, with corresponding confidence votes $CR_{Points}(I, S_v)$.
3. **Checking candidate shape reliability.** This stage consists in checking if one candidate s_v has a reality in the test image I, which should confirm that the object associated with the s_v elected is the appropriate one. It is first necessary to back-project s_v in I by estimating the geometric transformation that exists between the real shape of the candidate object in I and s_v in the catalogue. In our experiments, we considered a transformation of 5 parameters (translation, rotation and scale), estimated with a RANSAC algorithm. Then, to check the coherence of the back-projected shape, the local shape of the snake around each control point is compared to the local shapes encountered in image I, according to a procedure illustrated in figure 2.7, producing a confidence score $CR_{Snake}(I, S_v)$ that quantifies the coherence of S_v according to the image content.
4. **Final decision.** The final decision combines scores CR_{Points} and CR_{Snake} into a global confidence score CR computed for each candidate shape S_v. A candidate is definitely elected if the associated score is the highest among all the scores computed and is higher than a predefined threshold. If such a candidate exists, then we assert that the object associated with it is present in I. In addition, the elected shape S_v can give an indication of its 3D pose, while the back-projection in I of its cluster prototype s_v gives its location, by segmenting it precisely in I.

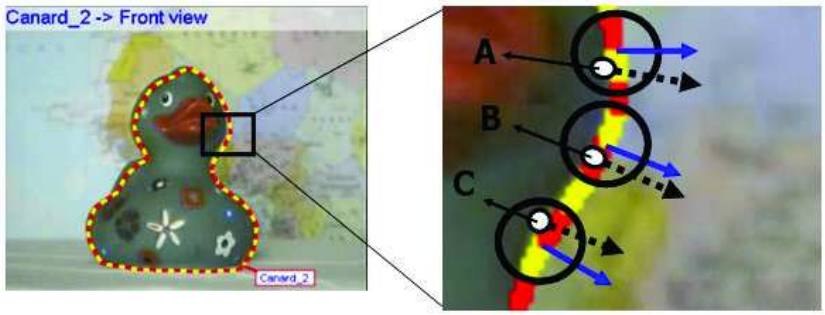


Figure 2.7: *Checking the candidate shape reliability. In the left image, the snake (dotted lines) is the shape candidate back-projected in I. A part of it is zoomed in the right image, where the three solid arrows represent the gradient direction* $\vec{\nabla}_{C_k^I}$ *of 3 control points* C_k^I *of the snake. The circles represent the searching area* $W_{C_k^I}$ *around each control point. Here, there is one pixel in each* $W_{C_k^I}$ *(points A, B and C) that owns a gradient direction (dotted arrows) close to* $\vec{\nabla}_{C_k^I}$*, conducing to the conclusion that the candidate shape is in coherence with this part of the image.*

Evaluation in terms of recognition rate and of spatial segmentation

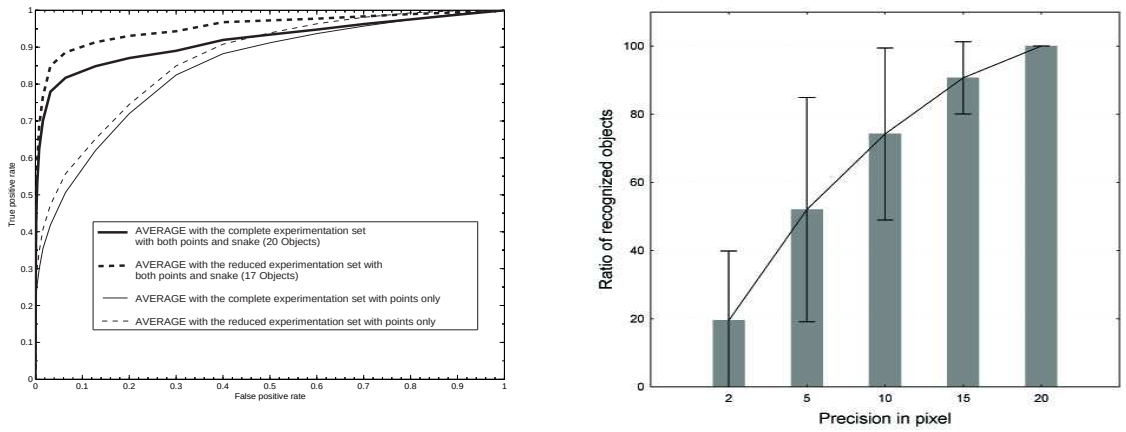
The whole approach proposed was evaluated on 20 classes of objects, 16 of them with different inter-class visual appearances in terms of content and shape, and 4 others (heads) which are more similar. It was compared to a state-of-the-art approach considering only local features as descriptor of the object content. In order to perform the most objective evaluation, each experiment took as inputs the whole set of objects and also a subset of 17 objects where 3 objects were not considered because very poor in extracted points (because of a surface too uniform); obviously, recognition results were better on the second set. Figure 2.8(A) presents the results of recognition obtained under the form of Receiver Operating Characteristic (ROC) curves and averaged ROC Equal Error Rates (EER).

Clearly, the results obtained are much better when the global description is added. In particular, the shape of the curves associated with the proposed approach shows that the use of the informative global descriptor allows reducing the rate of false positives drastically, for a given rate of false negatives. This improvement is very important in the case of video surveillance applications, as discussed in section 2.3.3.4.

To give an idea on the quality of the segmentation obtained after recognition, we defined its precision as the maximum distance between the control points of the back-projected shape and the border of the real objet in the image. Figure 2.8(B) shows the average repartition of this precision over 10 recognized objects. To illustrate, image (f) of figure 2.9 shows an example of head recognition with a segmentation of precision 20 pixels.

2.3.3.4 Video-to-video object recognition

The main algorithm of recognition presented in section 2.3.3.3 considers one image as input of the recognition process. Considering video sequences and then the temporal continuity



(A) Quality of the recognition process (B) Quality of the spatial segmentation

Figure 2.8: *Evaluation of the approach: (A) Recognition with and without the contribution of the global descriptors, over the reduced set (17 objects) and over the complete one (20 objects). The associated EER is 79.4% (resp. 81.8%) for the state-of-the-art approach with the whole set of objects (resp. with the reduced one), while it is 87.2% (resp. 91.1%) with the proposed approach that combines local and global features. (B) Average ratio of recognized objects over 10 recognized objects, according to the segmentation precision chosen. For example, 100% of the recognized objects are segmented with a precision of 20 pixels, on average 52% of them are segmented with a precision of 5 pixels and then 48% of the recognized objects are segmented with a precision between 6 and 20 pixels.*

of object presence over several consecutive frames would probably help in recognition, in terms of reduction of the false alarms number as well as in processing time. To keep a temporal precision during detection and to benefit from this temporal continuity, we chose to proceed recognition from several consecutive frames, instead of for example analyzing a sub-sampled video sequence. The adaptation to video of the initial algorithm was realized through several improvements, resumed as follows: in the absence of occlusion, because point extraction in each frame is an expensive operation, interest points are tracked along the sequence; another improvement consisted in benefiting from the spatial and precise segmentation obtained in the previous frame to restrict search in the current frame to close locations; the last improvement consisted in integrating decisions taken concerning frames before the current one F_t, over the temporal window $[F_{t-w}..F_{t-1}]$ of size w, in addition with the local decision at F_t. These adaptations to video are described and experimented fully in [Gouet-Brunet and Lameyre, 2008].

On the choice of the ROC operating point for surveillance scenarios

As done previously for recognition in images, it is fairly common to compare classifiers with the ROC Equal Error Rate EER, that is a satisfactory method with still images. More generally, in a recognition system, the cost of misclassification can be considered as the sum of the costs of misclassifying positive and negative cases. This cost can be written $(1-p)C_1x + pC_2(1-y)$, where C_1 is the cost of false alarm, C_2 the cost of

missing a positive and p the proportion of positive cases ($C_1 = C_2$ and $p = 0.5$ with EER). The question is how to choose the costs of misclassifying positive and negative cases. It probably does not have a universal answer, because it depends on the specific criteria imposed by the application. When considering video-to-video scenarios, it is necessary to take into account the significant temporal redundancy present in the sequence, as well as the significant number of images to be processed. In this context, we think that the cost C_1 of raising a false alarm must be much larger than the cost C_2 of missing a detection. A low rate of false positives comes to counterbalance, in a very effective way, a high rate of false negatives. The reason is that when an object appears in a sequence, it is certainly present in a large number of successive frames. The effect of false negatives is mitigated by the high number of frames containing the object. On the other hand, the false positives generate alarms that have to be verified by a human intervention (in particular in video surveillance applications). Because of the large number of frames, a high rate of false positives can lead to frequent human intervention. Therefore, a good detector, dedicated to video applications, must have a very low rate of false alarms and a satisfactory rate of false negatives.

The approach proposed in this work was developed with such an objective. We chose as operating point the following parameters: $C_1 = 200$, $C_2 = 1$ and $p = 0.5$. The corresponding averaged error rates obtained for the proposed approach in a still-to-video scenario are shown on the first row of table 2.1. We observed that the global descriptor allows reducing the rate of false positives drastically, while only multiplying the false negatives rate by an amount 10 times lower. In practice in a video sequence, these results imply that, when the object is present in the sequence, it is recognized half-time. Since we imposed a very low rate of false positives, it makes highly improbable that such detections are bad. For comparison, second row of table 2.1 also presents false alarm rates obtained with the state-of-the-art approach, according to the false alarms obtained with our method.

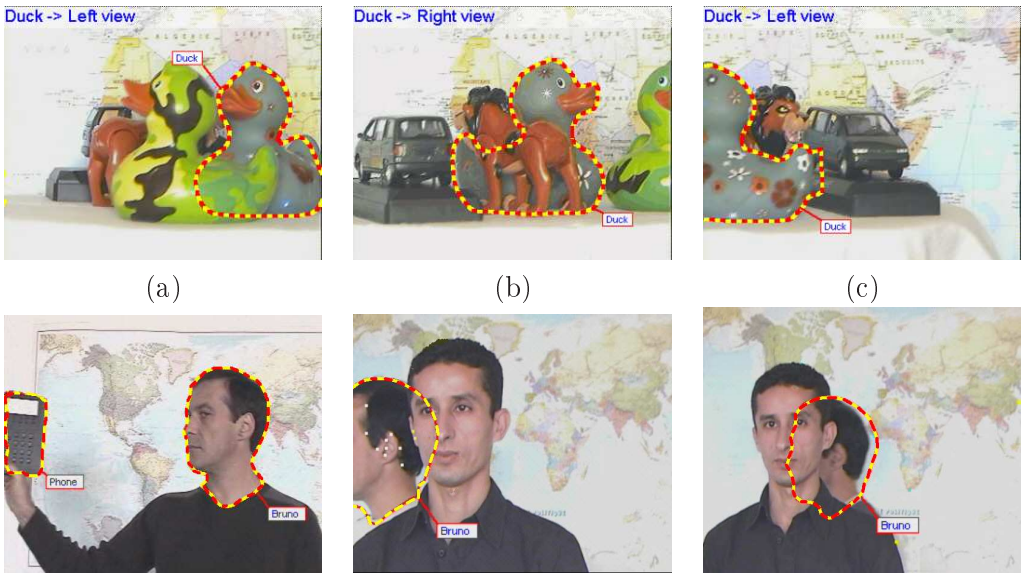
Method	**Scenario**	**False Neg. rate**	**False Pos. rate**
Points + Snake	Still-to-video	52%	0.26%
Only points	Still-to-video	76%	0.26%
		52%	5.4%
Points + Snake	Video-to-video	13%	0.26%
		52%	0.02%

Table 2.1: *Recognition on the complete experimentation set with parameters imposing low false positives rate, according to the approach (our approach and the state-of-the-art one) and the scenario (still-to-video and video-to-video).*

Integrating temporal information into the video-to-video recognition scenario also contributed to reduce the number of isolated false detections, under the hypothesis that it is not probable that an object appears in only one frame: results are improved, as shown on the last row of table 2.1. These results were confirmed by a fine analysis on the distribution of missed and correct detections, with transition probabilities modeling responses along a sequence as a Markov process and published in [Gouet-Brunet and Lameyre, 2008].

2.3.3.5 Robustness according to complex scenes

The complete approach can be easily customized to deal with more complex scenes involving occlusions and several objects to recognize. It was also evaluated under these scenarios, as illustrated in figure 2.9 with the results of recognition and segmentations in videos presenting occlusions and several objects to recognize.



(a) (b) (c)

(d) (e) (f)

Figure 2.9: *Recognition and segmentation in presence of large occlusions and multiple objects. On all the sequences, the camera is moving (translation, small rotation and zoom).*

Multiple-object recognition. The three first steps of the algorithm of recognition (section 2.3.3.3) do not impose the presence in the image of only one object. To adapt it to the presence of several objects, we can easily modify the final decision step of the algorithm by enabling to elect all the best candidates S_v satisfying a global score CR higher than threshold, instead of electing the best candidate only. An example of recognition with two objects is presented in image (d) of figure 2.9.

Occlusions. The step of section 2.3.3.3 that checks the reliability of the back-projected shape in the image is parameterized, among others, by a factor T_{snake} that imposes the minimal proportion of control points of the snake that should be locally in visual coherence with the image content. For the evaluation of the approach in presence of potential occlusions, we took $T_{snake} = 1/3$, that indicates that third of the snake has to be precisely in coherence with the image content to be considered as a relevant shape. The results of figure 2.9 were obtained under these conditions.

Object deformations. In theory, the approach can apply with any object, rigid as well as deformable, as soon as the different visual appearances involved were learnt and stored in the catalogue (the only condition being that sufficient interest points can be extracted). Because described locally, interest points remain robust even with deformable objects. For the global description based on a snake, several cases must be distinguished:

- Small global deformations of the object are processed with our approach, because the shape descriptor chosen is tolerant to such transformations, as well as to different 3D poses that also lead to shape deformations.
- As soon as the different global shapes involved are learnt and stored in the catalogue, every configuration of the deformable object can be recognized. Clearly in practice, the approach will fail with highly deformable objects because every configuration of the object cannot be learnt, or at least would probably generate a very large catalogue of EGSs. However, note that objects with few degrees of freedom, whatever their magnitude, could be processed (for example, a bird flapping wings), since it would not increase the catalogue so much.
- If the object shape is locally deformed in comparison to the EGS shapes stored (for example a car with a door opened), then the approach works, because the deformation is processed as an occlusion.

2.3.4 Application to surveillance of truck traffic

The work on object tracking, recognition and segmentation realized during the engineer thesis, Master thesis and PhD thesis of Bruno Lameyre is currently applied to surveillance of truck traffic on secured parking areas of motorways. It is done in collaboration with the French company Survision (two industrial contracts 2003-2004 and 2006-2009) where Bruno Lameyre works as senior engineer.

Here, the objective of the system is to increase the security of trucks and their trailers in the parking area. It has been developed with the two aims of controlling how many times trucks and trailers stay into the parking, and also of detecting swapping of trailers in the area, observable when a truck leaves the area either with a trailer different from the one it had when entering, or without trailer. Such events may indicate that a trailer is going to be stolen, or that a trailer has been warehoused in the parking area during a period that does not correspond (longer) to the arrival and leaving of the truck. Each time a truck goes in the parking area, two short sequences are taken: one of the front view of the truck and another one of the back view of its trailer. First row of figure 2.10 shows a back view of three different trucks and trailers entering a parking area. When the truck leaves the area, new front and back sequences are taken. Then, after having identified the leaving truck (from a database of front views), the system controls that it is always associated with the same trailer, by the help of the back views. The second row of the figure shows the trucks leaving the area, and the results of trailer recognition obtained with our approach. Note that, since the sequences are taken at the entrance and exit of the area, the backgrounds are different, justifying the use of local descriptors that are robust to clutter backgrounds. Moreover, most of the time, as seen in these examples, lighting conditions are drastically different, justifying the use of a shape descriptor as visual feature for recognition.

Automatic detection and recognition of truck license plates can be used as a primary source of pruning, but the experiences of Survision on this task demonstrated that they do not always represent a reliable source of information: they can be too dirty, partially occulted, unreadable because of extreme lighting conditions, ambiguous (several plates can appear: the truck one, the trailer one, those of cars carried by trailers), or because of

Figure 2.10: *Object recognition: application to surveillance of truck traffic.*

plates traffic. Then recognition through the analysis of truck visual appearance represents a complementary or alternative solution to surveillance of truck traffic. Figure 2.11 illustrates some of these difficult conditions: here, license plate detection failed, while our system matched the trucks and their trailers at their arrival and departure. Since September 2006, the system has been evaluated on parking areas of the French motorways "Autoroutes du Sud de la France". Statistics showed that in 30% of the cases, license plate detection on front and/or back views failed, making uncontrolled the corresponding vehicles. In such a case, the system switches to the visual recognition mechanism. Thanks to this mechanism, 70% of the missed trucks are recognized, i.e. 21% of all the filmed vehicles. To sum up, the complete approach allows reaching a percentage of detections of 91%, while it was 70% with license plates detection only.

(a) (b) (c)

Figure 2.11: *Three examples of successful recognition between truck arriving (first row) and leaving (second row). Here, license plate detection failed, because of: (a) plate saturated due to sidelight too strong; (b) and (c) very low sun and dirty plate.*

Chapter 3

Spatio-temporal description of video contents for copy detection

This chapter presents the main elements of my work on the adaptation of local image descriptors to video sequences, with application to content-based video copy detection. This research was done in the context of my collaboration with the group Imedia at INRIA Rocquencourt, through the PhD thesis of Julien Law-To [Law-To, 2007] with Nozha Boujemaa. This CIFRE thesis was done in collaboration with Olivier Buisson, researcher at the Institut National de l'Audiovisuel (INA). The work was co-funded by INA and also supported by the European Network of Excellence MUSCLE (2004-2007). After an introduction to content-based copy detection in section 3.1 and a brief state of the art on the domain in section 3.2, I present ViCopT (for Video Copy Tracking), the approach we proposed for the efficient retrieval by content of copies of videos. It is based on the analysis of trajectories of local descriptors along sequences and was developed with the aim of being robust to severe image transformations as well as dealing with large volume of video collections. In 2007, ViCopT was evaluated facing other academic and industrial techniques dedicated to video copy detection, where it proved its effectiveness, as resumed in section 3.4. Because this work was done with the objective of being able to address new future scenarios, I give perspectives to other applications in section 3.5.

Contents

3.1 The challenge of content-based video copy detection

This section is an introduction to the theme of video copy detection. Section 3.1.1 gives a panorama of the motivations that make this topic very challenging. A system of copy detection driven by the analysis of the visual content involves techniques of CBIR. But an efficient system dedicated to this task has specificities and must master some difficulties that are discussed in section 3.1.2.

3.1.1 Motivations

Over recent last years, the continuing drop in prices of computers and storage, the democratization of digital images and videos, the expansion of networking and communication bandwidth via the Internet or the HDTV (high-definition television) and associated services (user-generated-content web sites, blogs, video on demand, etc.) greatly contribute to the production and dissemination of larger and larger volumes of digital contents. Among them, videos are now involved in a large number of domains, they concern various categories of people ranging from video producers to video consumers; additionally, they are easily manipulable and transformable with consumers or professional softwares. Managing such a volume and a diversity of video contents has created new needs and new challenges for the scientific community: in particular, finding copies in a large video database has become a critical new issue.

In this context, Content-Based Copy Detection (CBCD) is an interesting solution to find, identify and trace documents in collections of visual contents. It is a sub-field of content-based image/video retrieval, and represents an alternative to watermarking. The latter consists in embedding visible or non visible information into the original content in order to be able to establish its owner [Lin et al., 2005]. However, present watermarking algorithms are not sufficiently robust to the hard transformations that can appear during copy creation (filtering, crop, insertion of objects, etc.) and because insertion is done a priori, it cannot apply if copies of the original content were disseminated before the insertion of the mark, situation which appears with a large part of the disseminated content. CBCD refers to a a posteriori process that can address these two issues.

Two examples of a concrete application of CBCD are trace of uses of videos of archive professionals, such as the INA that has more than 300,000 hours of digital videos to protect (and 800,000 hours expected in 2015), and the automatic control of the copyrights of videos uploaded by users on user-generated-content web servers, such as YouTube which presents more than 100 million videos daily. More generally, by establishing links between the several versions of the same document (or parts of them), other promising applications are possible: data mining with semantic interpretation, for example by reasoning on statistics on the broadcast of a particular video on TV (period duration, number of channels, national/international scope, etc.); content-based database structuring, for example by factorizing or purging redundancy, by mutualizing or checking other modalities and by proposing new paths for browsing (e.g. focus on a particular event shown partially by jumping to another more detailed video). These last tools could be particularly useful for poorly structured media such as the internet and user-generated-content web sites.

3.1.2 Specificities and difficulties to address

According to [Hoad and Zobel, 2003], a copy of a document is a "co-derivative" document in the sense that it was derived from the content of the original document. More formally in [Joly et al., 2007], it is defined as a transformed version of this document, according to a set of tolerated transformations that keeps the new document recognizable.

Copies are very specific similar documents

As a sub-field of image and video retrieval, CBCD involves similarity search, whose requirements greatly depend on the application and scenario of retrieval considered. For CBCD, two copies are considered as similar documents with very specific characteristics: in particular, copies must be differentiated from near duplicates which are documents having very similar visual contents but are not obtained via a transformation of the original document. For example, two videos of the same scene taken at the same time by two cameras may contain near duplicates, as well as videos taken at different time but sharing a common visual design like some TV shows or specific events as the one of figure 3.1(a).

(a) Two similar videos which are not copies (the ties are different)

(b) Two copies (one video is used to make the other). Source: Gala du Midem. G. Ulmer 1970 (c) INA

Figure 3.1: *Two examples exhibiting the difference between copy and near duplicate.*

Moreover, two documents can be copies while they look visually less similar than other documents, because of the transformations applied; figure 3.1(b) illustrates this idea with an extreme transformation (compositing, i.e. the combining of visual elements from separate sources into single images). More commonly, a large palette of transformations may occur, like change in resolution, frame and bit rate, compression codec and signal quality. Photometric or geometric transformations (including gamma and contrast transformations, crop, shift) can also greatly modify the signal, as well as the insertion of objects (logos, text). Figure 3.2 presents common examples of transformations, found on TV shows.

The specificities of similarity retrieval dedicated to CBCD impose the proposal of *discriminant* indexing techniques, i.e. that the description of the video should be unique to the video and its copies. The possible image and video transformations that can occur impose to index the content with visual features sufficiently *robust* to deal with these transformations. Note that this property is in contradiction with the property of discriminance: a description that is robust to many transformations would be less discriminating.

Figure 3.2: *Two examples of re-encoding and post-production effects on videos.*

Scenarios of retrieval

When considering video sequences, finding copies in a database of videos can cover several scenarios of retrieval by example. The simplest case is when the query is a single video: here, the video to retrieve has almost the same length and there is no boundary to determine. A more complicated case is when only a segment of the query video is a copy of only a segment of a video in the database. In this case, it is also necessary to determine the boundaries of the copy in the query and/or in the reference database. In applications like TV monitoring, a query video does not have boundaries, it can be an infinite stream. Consequently, the indexing technique developed (visual features as well as similarity measure) should be sufficiently *flexible* to deal with these last scenarios.

Volume of data and future applications

The needs and applications listed in section 3.1.1 deal with very large volumes of data, often going beyond several hundreds thousands hours of videos. Developing efficient techniques to manage such volumes require a specific structuring of the feature spaces involved by using dedicated multidimensional index structures (see chapter 4 for a solution to this problem). Moreover, the time spent to index and structure such volumes makes very hard and even impossible the re-indexing in case of new application. For example, at present the INA has more than 300,000 hours of digitalized videos and expects to have digitalized and indexed 800,000 hours in 2015, but what will be the needs and applications in 2015? From this constatation, it might be relevant to consider the computation of visual features as indexes that are sufficiently *generic* to be exploited and customized in future applications.

3.2 Related work

CBCD generally consists in extracting a small number of pertinent signatures from all the frames or particular frames of the video stream. According to the survey presented in [Yang et al., 2003], the palette of encountered approaches share several characteristics such as the kind of extracted features and their robustness to categories of transformations, the matching function, the level of identification (video, shot, key frame, frame, spatial window, etc.) and the type of query (a single video or a stream with segment detection). Several kinds of techniques were proposed in the literature for video retrieval, copy or near duplicate detection. In the following, we revisit the main trends encountered and give one

example at least for each of them. A more detailed state of the art can be found in [Law-To et al., 2007a].

Coming from the watermarking community, the authors of [Oostveen et al., 2001] were the first to propose the concept of *visual hashing* as a tool for video identification based on a hashing function. The general idea is to transform the data into a small set of features invariant to desired transformations, the hash values, that as used as signature [Oostveen et al., 2001; De Roover et al., 2005; Coskun et al., 2006]. In [Oostveen et al., 2001], the whole video is represented by a string sequence where each item is obtained by differences of mean luminances computed on partitioned grids in spatial and temporal consecutive regions. In general, the hashing functions have been designed to be robust to several transformations, but they cannot deal with the case where only a part of the query video is a copy of only an extract of a video in the reference database.

While the hashing functions apply at a very low-level by considering all the pixels equally, other approaches belonging to CBIR aim at analyzing the video content in order to exhibit spatial, temporal or spatio-temporal properties relevant for video identification. Classically, the features extracted by these approaches can be classified into *global* or *local* approaches:

A large number of global approaches are based on *color histograms* as signatures to describe the video content [Naphade et al., 2000; Ferman et al., 2002; Cheung and Zakhor, 2003; Li et al., 2005]. For example, in [Naphade et al., 2000], the authors use histogram intersection of the YUV histograms from the DCT sequence of the MPEG video; they also propose an efficient compression technique for the histograms based on polynomial approximations, making compact the set of signatures obtained and allowing supporting queries at multiple temporal resolutions. We compared this approach against other ones and our solution; results are presented in section 3.4.2. The main drawback of global methods based on histograms is that distinctiveness is not guaranteed and consequently the solutions proposed may mix up near duplicate detection and copy detection.

Some global approaches are based on *ordinal measures* [Hampapur and Bolle, 2002; Hua et al., 2004; Kim and Vasudev, 2005; Chen and Stentiford, 2008]. Originally proposed in [Bhat and Nayar, 1998] for computing image correspondence and then adapted to video by [Mohan, 1998] for retrieving video clips that depict similar actions, it is adapted to CBCD purposes in [Hampapur and Bolle, 2002]: each frame of the video is partitioned into N systematic blocks and the average gray level in each block is computed. The set of average intensities is sorted in ascending order, the rank is assigned to each block and the frame signature consists in the list of N ranks. In this approach, two other signatures are added, one involves motion direction computed on blocks of each frame while the other rests on a color histogram similar to the approach of [Naphade et al., 2000] (without the compression part). We also evaluated the approaches of [Hampapur and Bolle, 2002] and [Chen and Stentiford, 2008] (see section 3.4.2). Globally, ordinal measure has proved to be robust to changes in the frame rate and resolution, to color shifting and are generally characterized by a high-speed matching step. Its drawbacks is its lack of robustness as regards logo insertion, spatial shifting or cropping.

Image transformations like occlusion, insertion and cropping are very present in post-production effects (insertion of logos or text, superimposing of a border, etc). Oppositely

to global approaches that are not robust to such transformations, local signatures based on interest points have demonstrated their effectiveness in these conditions. While literature on local descriptors is abundant for object recognition, few literature exists on their use for image copy detection [Berrani et al., 2003; Ke et al., 2004; Lejsek et al., 2006] and for video copy detection purposes. In [Massoudi et al., 2006], multi-scale interest points based on differences of gaussian are extracted on particular key frames obtained using the visual hash function of [De Roover et al., 2005] and that correspond to frames with the smallest content variation along the shot. For each point, the associated local signature is the concatenation of histograms of the normalized pixels orientation computed for nine local slices around the point; it is invariant to image rotation and has 144 dimensions. In [Joly et al., 2007], the authors present a method based on local features which decomposes the sequence into key frames based on the image activity, and Harris interest points are extracted on them. The associated local description computes local jet of the signal around the point, as well as around 3 other spatio-temporal positions near the point with the aim of limiting the correlation and redundancy between the detected features. The use of an index structure and an approximate similarity search allows the database to be very large (1,944,000,000 descriptors for 30,000 hours of video in the database). This method were used as a reference in our experimentations and then also evaluated in section 3.4.2.

3.3 Presentation and characteristics of ViCopT

The work done during the PhD thesis of Julien Law-To (2004-2007) was in the continuation of the PhD thesis of Alexis Joly (2001-2005) done at INA, who proposed an approach based on local descriptors for video copy detection as well as an index structure to perform search quickly in large volumes of videos with these descriptors [Joly et al., 2007]. The main objectives of this new thesis entitled "From genericity to distinctiveness of video content description, application to Video Copy Detection" were to study a new approach of content description more discriminant with the ambition of reducing the false alarms remaining with the previous approach, while keeping in mind its ability of genericity with customization to copy detection as well as to future scenarios. The solutions proposed were integrated into a system of copy detection called ViCopT. The first description of its characteristics was published in [Law-To et al., 2006d] and [Law-To et al., 2006c].

Section 3.3.1 presents the basis of the choices we made for describing the video content. The features chosen are low-level in order to characterize the signal the most robustly and discriminately possible. From such features, then we defined more high-level features obtained through top-down processes, thus customizable according to a specific application; they are presented in section 3.3.2 where they are dedicated to copy detection. Then, the main steps of the algorithm of online video retrieval are explained in section 3.3.3.

3.3.1 A low-level description of the video content

Because they are able to describe the image content precisely and are robustly, interest points and local descriptors are very popular to deal with object recognition purposes. We thought that these properties make them also particularly relevant for copy detec-

tion purposes where severe transformations of the image can occur, for instance cropping and insertion of objects. Besides, they were already developed and experimented for this purpose with success in several works, such as in [Massoudi et al., 2006; Joly et al., 2007].

This section revisits the choices we made to describe the video content with local features. In this work, we did not propose a new interest point detector or a new local descriptor, but we rather focussed on the improvement of existing solutions for indexing large volumes of video contents dedicated to copy detection, according to the two axes described in sections 3.3.1.1 and 3.3.1.2. Note that the whole strategy is generic in the sense that it can apply with different kinds of local features.

3.3.1.1 Spatial description with interest points

As already mentioned in section 1.1.3, the recent proposal of local descriptors involving different natures of image support has given the possibility to study some combinations of them to improve image description. We chose to describe the frames content by considering the precise version of the Harris detector of points of interest [Harris and Stephens, 1988; Schmid and Mohr, 1997], combined with points of symmetry with the fats implementation of [Loy and Zelinsky, 2003]. Because related to the description of image content and not specific to video, this part of our work is described in the section 2.2 of chapter 2.

In the continuity of the work done by Alexis Joly [Joly et al., 2007], we employed the local jet of the signal as descriptor around the interest points. The signature obtained per point is a 20-dimensional vector which is invariant to translation and is normalized to be invariant to affine illumination transformations. The associated feature space is associated with distance L_2 and is called S_{Points} in the following.

3.3.1.2 Spatio-temporal description of the interest points

Differently to several approaches of video indexing, like in [Joly et al., 2007], that index only the content of key frames (see the state of the art of section 3.2), we chose to extract interest points *in every frame* of the video sequence. This choice was motivated by two arguments: by describing every frame, first it will be possible to detect copies of segments of videos with a more high temporal precision than by only integrating particular frames into the description. Secondly and above all, it gives the possibility to track the extracted points along the sequence and then to characterize their *kinematic behavior*. We think that such an information is of great importance in the precise characterization of video contents regarding to copy detection. As an illustration, by considering the motion of local features in the example (a) of figure 3.1, it is obvious that the two near-duplicates found would not be considered as copies (the body movements of Kofi Annan in the two sequences are not the same).

One consequence of this choice is that, in a whole sequence, many features are extracted and a lot of them are similar from one frame to consecutive frames. But, tracking points may also help in reducing the volume of produced features drastically and in limiting their redundancy, while enhancing their robustness by modeling their variability along the trajectory. The final description chosen for the interest points along their trajectory,

described in the following, takes advantage of these considerations.

During the off-line indexing step, each trajectory is represented by a set of trajectory parameters called S_{Traj}, defined according to the four sets of components listed in table 3.1. Only a global and compact representation of the trajectory is kept, in the form of a spatio-temporal box (T_1 and T_2) which corresponds to the variation of the interest point location in space and time. As the trajectory is computed from frame to frame, the local signature of the point may vary along the trajectory. The parameters of the point tracking algorithm were chosen in order to obtain trajectories where the local signature does not vary so much, making relevant the consideration of the average local signatures of the point along the trajectory $\vec{S}_{mean}$.

T_1:	Time code of the beginning and the end: $[tc_{in}, tc_{out}]$
T_2:	Bounds of the spatial position of the point: $[x_{min}, x_{max}]$ and $[y_{min}, y_{max}]$
T_3:	Standard deviation of the spatial position of the point: (σ_x, σ_y)
T_4:	Average local signature of the point along the trajectory: $\vec{S}_{mean}$ (from S_{Points})

Table 3.1: *S_{Traj} set: characteristics of the trajectories of interest points that are indexed.*

Note that, in order to describe the kinematic behavior of local features in video sequences, spatio-temporal solutions were studied, like [Laptev and Lindeberg, 2003; Dollár et al., 2005] who propose a generalization of 2D interest points to 3D (space-time) interest points for motion interpretation and behavior recognition. Unfortunately, experiments performed on the approach of [Laptev and Lindeberg, 2003] demonstrated that the produced features characterize too specific areas of the video to be efficient for copy detection purposes (STIP, see section 3.4.2.1 and publication [Law-To et al., 2007a]). The solution we proposed is more general and more representative of the video content, because it allows taking into consideration all the behaviors of interest points, including those which are salient in 3D.

3.3.2 A higher level of description: trends of points behavior

From the generic description of the trajectories described in previous section, we can go further by interpreting them according to the considered application. As illustrated with the example of figure 3.3, one relevant angle of view is to focus on the trends of behaviors of the interest points along the sequence. In the example, the different trends of behaviors observed exhibit several visual and semantic components of the image (the speaker, his attitude, the background and the frame insert), very specific of the sequence.

Defining categories of spatio-temporal behaviors provides a specific interpretation of the video content, dedicated to the goal considered. It supposes a top-down knowledge on what kind of behavior is relevant for the considered application. A lot of categories can be imagined. When considering video copy detection purposes, we studied two particular labels: the first label, called L_{Still}, characterizes motionless and persistent points, that are particularly *robust* because stable by definition; frequently, it can be associated to features of the background of the scene, often typical of a TV show. The second label represents a very different behavior: it is called L_{Motion} and is associated to the moving and persistent points of the video; such points bring *discriminance* to the video sequence

Figure 3.3: *Categorization of interest points, according to their spatio-temporal behavior: the boxes represent the amplitude of moving points along their trajectory (motionless points do not have a box). The "+" corresponds to the mean position of such points. The "×" shows a non persistent point (the eye blinking).*

description. When the cameras are moving, the interpretation of these labels can become less informative semantically. But when applied to CBCD, such labels still remain relevant because the motion of the background due to camera motion is also a very distinctive visual cue.

Classifying interest points into such categories enriches the local description with a temporal context associated to each point. To determine the categories of trends of behavior and classify all the interest points into them, we proposed two alternatives that manipulate the characteristics T_1, T_2 and T_3 of the trajectories:

Heuristic thresholds. The simplest solution is to consider heuristic thresholds that constraint the size of the 3D box determined by intervals of T_1 and T_2 for finding the motionless points, the moving points and the persistent points along the video.

Clustering. By considering an unsupervised algorithm of classification, it is possible to determine clusters of trajectories and then trends of behavior, automatically deduced from the data distribution without a priori information. We chose the ARC algorithm that stands for "Adaptive Robust Competition" introduced by [Le Saux and Boujemaa, 2002], because of its capacity (1) of fine adaptation to the density of each cluster and (2) of collecting feature space outliers into a noise cluster. In particular, this last property allows detecting and eliminate abnormal trajectories. Labels of behavior L_{Still} and L_{Motion} are then assigned to the cluster prototypes by using the heuristic thresholds and then propagated to all the points of the clusters. Note that by using a clustering algorithm, classifying all the trajectories on the whole video sequence would not provide a precise classification, because of the high variability of trajectory parameters in the whole sequence. To obtain trends of behavior finely adapted to the video content, we performed the classification locally in time, by only running ARC on trajectories of the same shot. The cutting of the video into shots was performed simply by analyzing the distribution of ends and beginnings of trajectories along the sequence. This simple solution can over-detect some cuts but we experimented that it does not damage the relevance of the ensuing classification.

Table 3.2 gives an idea of the volume of features involved on 1,000 hours of video, for labels L_{Still} and L_{Motion} obtained with thresholds or with ARC. Here, a third category of labels was also considered, that consists of the random classification of trajectories according to labels L_{Rand1} and L_{Rand2}. As a reference, we selected the technique of [Joly et al., 2007] which is also based on a local description but does not integrate trajectories

of points nor labels of behavior.

Total number of Harris Trajectories		316,946,586	
Labeling Technique	First label	Second label	Total
Thresholds	$L_{Still} : 34,983,888$	$L_{Motion} : 32,118,506$	67,103,394
ARC	$L_{Still} : 107,468,852$	$L_{Motion} : 34,289,601$	141,758,453
Random	$L_{Rand1} : 66,081,802$	$L_{Rand2} : 65,930,010$	132,011,812
Reference technique	No labels		67,488,552

Table 3.2: *Number of features for 1,000 hours of video.*

A qualitative evaluation of the approach was performed in terms of number of segments found and of the number of frames found. Whatever the scenario considered, the construction of trajectories and the use of labels of behavior (even the random ones) improves recall and precision, facing the reference technique. The classification of labels L_{Still} and L_{Motion} with ARC allows a large improvement compared to a random labeling algorithm, while involving approximately the same number of features. With a smaller number of features, the method which uses thresholds also presents a good performance and is even better than the reference technique that involves a similar quantity of features. More details can be found in [Law-To et al., 2006d; Law-To et al., 2006a] and in the thesis [Law-To, 2007].

3.3.3 Online retrieval of copies

This section summarizes the main steps of the online algorithm of video copy retrieval, which is fully described in publication [Law-To et al., 2006a]. It is based on the high-level description of the video content presented in previous section.

As the off-line indexing step needs long computation time, the proposed approach of retrieval is *asymmetrical* in order to proceed queries in real-time: the same features are not extracted from the query videos and from the reference videos. The features extracted from the query video are Harris and symmetry points characterized with low-level descriptors, called $\vec{S}_{query}$ and belonging to feature space S_{Points} and do not take trajectories and labels of behavior into account. This description is extracted on a selection of frames of the query, according to the two following parameters:

- period p of the selected frame in the video stream;
- number n of extracted points per selected frame.

These parameters can be chosen online and then highly contribute to the flexibility of the approach. This flexibility is an important advantage for dealing with all the situations where copies have to be found: for case where the full-length video needs to be identified, p can be chosen larger in order to speed up the detection. Oppositely, the scenario of segment detection in video streams may require shorter values of p in order to find the boundaries of a detected segment precisely, particularly with very short segments.

A fast similarity search. Retrieval begins by a first step of frame filtering that consists in finding in the database the frames most similar to the ones constituting the query,

according to the local descriptors involved. The objective is to find the nearest neighbors of the feature vectors $\vec{S}_{query}$ in the multidimensional space S_{Points}. In the database, the solutions are taken in the set of features $\vec{S}_{mean}$ associated with trajectory characteristics S_{Traj}, which are of same nature as the $\vec{S}_{query}$. To be efficient, such similarity search in a multidimensional space is performed with an index structure that helps accelerating retrieval in S_{Points}. The structure used performs approximate retrieval by the way of a distortion-based probabilistic similarity search and was proposed in [Joly et al., 2007] and improved in [Poullot et al., 2007].

Spatio-temporal registration. From the set of nearest neighbors returned, a spatio-temporal registration can be performed in order to determine the most similar frames of the database with precision, i.e. with the aim of taking into consideration the transformations authorized by scenarios of copy detection and of eliminating possible outliers. As transformation, we focussed on a simple linear model $P' = P + B$ with $P = (x, y, t_c)$ (that means that zoom, slow-motion and rotation were not considered). The estimation of the best parameters of B provides the possible spatial translation existing between corresponding interest points, plus a temporal offset between the query video and the corresponding reference video in the database. Here, the asymmetrical approach of description defined imposed one difficulty: the local features of the database that matched with the ones of the query are geometrically described with a trajectory within a 3D box (parameters T_1 and T_2 in table 3.1), while the features of the query are associated with a simple 3D location in the video query (time code and spatial coordinates in the frame). The algorithm used for estimating the transformation parameters is an adapted version of the RANSAC [Fischler and Bolles, 1981], modified to take as input 3D data (the queries) and interval-valued 3D data (the features of the database), and to provide as output an interval-valued solution. This registration is applied to each video of the database having frames where nearest neighbors were found at previous step, and the decision algorithm consists in computing a vote that considers the number of local features in coherence with the transformation. At the end of this step, the interval-valued offset $B^{\sqcap}$ is also estimated.

Combination of labels of behavior and natures of points. The previous step of registration is applied separately on each label of behavior, because the information brought by them might not be relevant in the same way: label L_{Still} is associated with features presenting an accurate spatial position but great temporal imprecision because of the persistence of the points along frames, while label L_{Motion} is associated with features which are more accurate in the temporal domain but less spatially. Then, the votes obtained for each category of points are combined by a simple multiplication in case of compatibility (i.e. intersection of offsets $B^{\sqcap} = B^{\sqcap}_{L_{Still}}$ and $B^{\sqcap}_{L_{Motion}}$ not empty). Integrating the two natures of the interest points into the decision process is done similarly, by adapting the fusion function for enhancing the vote if two natures of points of interest are found with the same registration estimation and if the two labels of behavior are found.

Aggregation and propagation of the detected segments. This last step aggregates the results of the previous step, in order to provide a global vote for the whole matched sequence and also to determine the boundaries of the detected video segment with the highest accuracy. Labels of behavior are also exploited to determine a starting frame for this aggregation: in order to start from a very confident frame, the beginning of the aggregation must contain all the labels. In the case of copy detection, starting from a

frame which only shares the same background as the query (supposed to be represented by label L_{Still}) could conduct to temporal imprecision in detection; a good example is TV shows where the same background is present during a long time.

3.4 Performances of ViCopT

For evaluating the relevance of the choices made during the design of ViCopT, extensive performance evaluations were carried out on a database of 1,000 hours of heterogeneous video contents taken from the archives of INA, where queries were subjected to several basic (change in contrast, resizing, re-encoding, blur, etc.) and more complicated transformations (crop, insertion, combination of basic transformations, etc.). Several experiments were performed: evaluation of the relevance of labels of behavior in general and for copy detection (see section 3.3.2); evaluation of the complementarity of the different natures of points (see section 2.2 of chapter 2); measure of the computational time elapsed to off-line indexing and online retrieval; the influence of the query parameters p and n, particularly on the temporal precision of detected segments. All these experiments are fully reported in the PhD thesis of Julien Law-To [Law-To, 2007]. In the following, we illustrate the relevance of the whole system with a very hard scenario of copy detection encountered on a TV show, while section 3.4.2 gives main results of two comparative evaluations performed on academic and industrial systems dedicated to video copy detection, including ViCopT.

3.4.1 Evaluation on a hard TV case

A French TV show named "Les duos de l'impossible" provides a very hard scenario of copy detection: several post-production effects are used, including compositing, i.e. the combining of visual elements from separate sources into single images. Here, video archives of singers from the 60s are mixed with videos of current singers; see figure 3.4 for examples.

Copy detection evaluation with ViCopT and the reference system [Joly et al., 2007] was carried out by including 20 hours of the original videos used in the compositing into the 1,000-hour video database. Then three hours of the TV show were considered as video queries. Table 3.3 plots the number of segments retrieved according to the length of the segments that a perfect system should find.

Retrieved segments with the reference technique					43
Retrieved segments with our technique					82
Correct detections with reference technique					7min 53s
Correct detections with our technique					10min 44s
Segments length	$< 1s$	$1s - 5s$	$5s - 10s$	$10s - 20s$	$> 20s$
Reference Technique	0	11	10	17	5
ViCopT	7	35	13	21	6

Table 3.3: *Results of copy detection on the videos of figure 3.4.*

The results obtained lead to some observations: based on local features and temporal context, ViCopT is robust to large transformations made by a professional post-production

(a) Left: *Les duos de l'impossible* 2005, Right: *Palmarès des Chansons* R. Pradines 1966(c)INA

(b) Left: *Les duos de l'impossible* 2005, Right: *Vient de Paraître.* J. Guyon 1965 (c)INA

(c) Left: *Les duos de l'impossible* 2005, Right: *Sacha Show.* JP. Marchand 1963 (c)INA

(d) Left: *Les duos de l'impossible* 2005, Right: *Système deux.* C. Fayard 1975 (c)INA

Figure 3.4: *Copy detection on a hard TV case: video queries are on the left and videos detected by ViCopT are on the right.*

and it presents an increase in the number of correct detections facing a state-of-the art reference technique. In addition, ViCopT is efficient even when searching for short video sequences in a video stream: in table 3.3, 7 very short segments (lower than 1 second) were retrieved while the reference technique did not find any of them. The improvement (36% in terms of video length detected) is very significant for a copyright management application. This test which used a real TV stream, illustrates the accuracy and the performances of our method for video copy detection in a real-life situation. These results were confirmed by additional experiments performed on on a large set of videos (5,600 videos) downloaded on the Internet on popular web sites which propose videos, see [Law-To, 2007].

3.4.2 Comparison with other copy detection techniques

This section presents two types of comparative evaluations of systems and techniques dedicated to video copy detection (including ViCopT) performed in 2007. In section 3.4.2.1, the work presented is an evaluation, supported by the Network Of Excellence MUSCLE, performed during the PhD thesis of Julien Law-To, in collaboration with research groups from the academic field involved in the e-team of MUSCLE concerned by Visual saliency. The second evaluation, presented in section 3.4.2.2, is a live benchmark which took place in Amsterdam during the ACM International Conference on Image and Video Retrieval (CIVR'07) and was part of an evaluation showcase funded by MUSCLE.

3.4.2.1 A comparative study of state-of-the-arts techniques

This evaluation was a joint work with Alexis Joly (INRIA, Imedia group), Ivan Laptev (INRIA, Vista group), Li Chen and Fred Stentiford (UCL Adastral Park Campus). Different state-of-the-art techniques, using various categories of descriptors and voting functions, were compared: ViCopT, Harris points computed on key frames [Joly et al., 2007], Temporal Ordinal Measurement [Chen and Stentiford, 2008], Spatio-temporal interest points (STIP) [Laptev and Lindeberg, 2003], Ordinal Measurement and motion direction [Hampapur and Bolle, 2002] and Color histogram [Naphade et al., 2000]. All the techniques were tested and compared within the same framework, by evaluating their robustness under single and mixed transformations, as well as for different lengths of video segments. All the experiments were carried out on the BBC open news archives (3.1 hours of videos).

The results of these experiments were published in [Law-To et al., 2007a] and can be summarized as follows: for all the transformations tested alone, the Temporal Ordinal Measurement approach provides the best results (all the segments were found without false alarms). When considering mixed random transformations, the local approaches of description present the best results in term of robustness, while having more computational costs. The best results were obtained with ViCopT, with an average precision of 0.86.

These results lead to the observation that no single technique and no universal description is optimal for all the applications involving video copy detection. Choosing a copy detection system strongly depends on the required scenario (retrieval of full-length movies, retrieval of segments, temporal precision of the retrieved segments), on the encountered transformations, from none (exact copy) to severe post-production effects, and also on the size of the videos/segments to retrieve. To find full-length movies affected by small modifications, methods with global features like Temporal Ordinal Measurement are probably faster with the same efficiency as methods with local features that would probably be expensive as the video length is considerable. When considering more complicated cases, as in the audiovisual domain where post-production takes an important place, local approaches seem to be more relevant. Among them, ViCopT has good performances, even with short segments.

3.4.2.2 The live benchmark of CIVR'07

The live benchmark organized during CIVR'07 gave the opportunity to several academic and industrial teams to evaluate their system of video copy detection exactly in the same conditions (same framework, same video database and queries) and in limited time. Two tasks, corresponding to two levels of difficulty, were proposed: (ST1) the queries were videos which are copies of whole long videos (from 5 minutes to 1 hour), with possible re-encoding, noise or slight retouch; here copy retrieval consisted in being able to detect the copy in the database. Task (ST2) involved sequences post-processed by professional audiovisual archivists (including cropping, fade cuts, insertion of logos, borders, static and moving texts, contrast, gamma) and video streams as queries with length from 1 second to 1 minute; here, sequences belonging to the database had to be identified and located by their start and end. The video database consisted of 101 videos (total of 80 hours) of various contents downloaded on different web sites. The results of these tasks are presented

in table 3.4.

Team - run	Mean precision	Search time
Advestigo	0.86	64 min
CAS - 1	0.46	41 min
CAS - 2	0.53	14 min
City Univ. of Hong Kong	0.66	45 min
IBM - 1	0.86	44 min
IBM - 2	0.86	68 min
IBM - 3	0.86	99 min
ViCopT (large p)	**0.93**	**25 min**

Task ST1

Team - run	Mean precision	Search time
Segment-level granularity		
Advestigo	0.33	33 min
City Univ. of Hong Kong	0.86	35 min
ViCopT ($p = n = 20$)	**0.95**	**7 min**
Frame-level granularity		
Advestigo	0.17	33 min
City Univ. of Hong Kong	0.76	35 min
ViCopT ($p = n = 20$)	**0.84**	**7 min**

Task ST2

Table 3.4: *Results of the CIVR'07 live benchmark for tasks ST1 and ST2.*

Because Julien Law-To was one of the organizers of this event, ViCopT did not officially take part to the evaluation. But, in order to evaluate the system in a real challenging situation, the queries were also submitted to the system during the showcase in exactly the same conditions as the other systems. Here, ViCopT obtained the best recall, the best precision with the fastest technique whatever the task. In order to perform the queries quickly during task ST1, a large value of parameter p was considered. Oppositely, to obtain precise temporal detections of segments in task ST2, this parameter was taken shorter.

3.5 Generalization to other applications

With the huge production and broadcasting of visual contents, many applications of video retrieval have to manipulate larger and larger volumes of videos. Unfortunately, such quantities cannot anymore be managed manually and then indexing by content will certainly take an increasing part in the future solutions to manage multimedia data. But at present, the time spent to index by content and structure such volumes is far to be negligible; in addition, it may be increased because of the complexity of the more powerful future techniques of content description. For example, the whole description involved in ViCopT requires 1,200 hours to be computed on 1,000 hours of video. Therefore, it appears very hard to re-index such a volume with other descriptors in case of new application. The research performed during the PhD thesis of Julien Law-To was oriented to cope with this problem. During the off-line indexing step, most of the time is spent computing the low-level features, but this purely bottom-up process is entirely generic and then performed only one time, independently of the application. The step of high-level features extraction follows from a top-down process that is specific to the application considered, but has a very negligible computational cost: on 1,000 hours of video, associating labels of behavior to interest points only takes 5 minutes by using the heuristic thresholds. This property makes ViCopT relevant to deal with future applications or scenarios of video retrieval. In particular, we think that the three following tasks can be addressed without recomputing the low-level features, only by customizing the high-level ones:

Thematic linkage of videos. The combined use of labels of behavior L_{Still} and L_{Motion} has proved to be relevant for copy detection. Changing these features allows computing another similarity measures dedicated to other scenarios. When considering videos like TV shows, news or weather forecast programs which are recognizable because of a specific visual identity design (background, banner, logo, etc), most often recorded with motionless cameras, label L_{Still} can be employed alone to describe these visual identities, as illustrated in figure 3.5 with a particular French TV show. In this scenario, this is the stability of interest points along frames and the robustness of local descriptors facing viewpoint changes that are exploited. Experiments on videos from the Internet were performed according to this scenario of application, and were published in [Law-To et al., 2007b].

Figure 3.5: *Similar videos found by taking into consideration label L_{Still}. The videos retrieved belong to the same TV show.*

Object segmentation. Spatio-temporal segmentation of video sequences attempts to extract independent objects or background in video sequences. A lot of work has been done to address this task, among them are the approaches of segmentation based on trajectory grouping. Differently to the previous scenario, ViCopT can be customized by only focusing on moving local features with the help of dedicated labels of behavior assigned to interest points with the categorization algorithm employed in section 3.3.2. Here, the spatial position of the trajectory (center) has to be used as input of the clustering, in addition with the sizes of the 3D box ; it was not the case for copy detection purposes. Examples of segmentation obtained with this approach are given in [Law-To, 2007].

Event-based segmentation of the video. Another potential application is to assist the segmentation of the video. Indeed, detecting specific motions is possible with ViCopT by defining dedicated labels of behavior: for example, focusing on clusters of stable points having a long vertical motion at the same time allows detecting events like the credits at the end of programs. None of these events are at a semantic level, no interpretation of the video is involved but this can help for the semi-supervised indexing of video sequences.

Chapter 4

Structuring of feature spaces for visual contents

This chapter presents a synthetic view of my activity in the field of access methods with multidimensional index structures for accelerating similarity search in high-dimensional spaces. The major part of this research has been performed at CNAM during the PhD thesis of Nouha Bouteldja (2004-2008) and part of it was supported by the French project BIOTIM (ACI "Masses of Data", 2003-2006) where it was applied to images dedicated to biodiversity. Section 4.1 revisits the literature of databases and content-based image retrieval related to multidimensional index structures. Here I discuss the difficulties and lacks encountered in the literature for this topic, that motivated the first part of my work on this domain: the re-evaluation of classical existing structures for multidimensional features, in particular features extracted from images. This work is described in section 4.2. In section 4.3, I present a preliminary work done for accelerating retrieval with multiple queries as input, i.e. queries that are composed of several multidimensional vectors; this study was applied to local image descriptors where queries are exactly defined on this model. Then section 4.4 is dedicated to the description of HiPeR, a new model proposed for improving retrieval in high-dimensional spaces, based on a hierarchy of feature spaces and indexes. HiPeR was designed to address exact, approximate and progressive nearest neighbors retrieval. This work was applied to CBIR by evaluating our proposals on data sets of image descriptors (data sets of local descriptors according to the case) but is not dedicated to.

Contents

4.1 Scientific context

This section has for objective to briefly present the literature of databases and CBIR related to multidimensional index structures. Section 4.1.1 revisits the main new strategies that were proposed to accelerate similarity search in high dimensional spaces; in particular, we will focus on approximate and progressive retrieval. Index structure traditionally suffer from effects of the well-known problem of the curse of dimensionality. The new proposals of index structures try to limit these effects by the way of different strategies, but the curse of dimensionality remains a fundamental difficulty for access methods in high-dimensional spaces. We go further into this problem in section 4.1.2, where we also tackle other important problems such as the definition of a "high dimensional space" and the evaluation of index structures.

4.1.1 Strategies of retrieval and associated index structures

During the last years, similarity search has drawn considerable attention in multimedia systems, decision support and data mining. In these systems, there is the need for finding a small set of objects which are similar to a given query object. As shown in chapters 2 and 3, content-based image/video retrieval is one example of application where the objects to be retrieved are images, videos or parts of them. Usually, similarity search is implemented abstractions of these entities, the signatures, and consists in the nearest neighbors search of a given query, under the form of a k-NN search (i.e. retrieving the k objects closest to the query) or of a ϵ-*sphere query* search (i.e. retrieving all the objects at a distance lower then a given threshold ϵ).

Because the feature spaces associated with the indexed entities can be high-dimensional and can contain many points, a large number of multidimensional access methods have been proposed in the literature to accelerate the retrieval process, see for example [Samet, 2006] for an exhaustive survey on these structures. Among them, we can cite the classical ones that divide the feature space recursively with trees such as the SR-tree [Katayama and Satoh, 1997], those that filter the data by approximating them such as the VA-File [Weber et al., 1998], those that apply on metric spaces such as the M-tree [Ciaccia et al., 1997], and those that define a mapping between the d-dimensional feature space to one dimension such as the Pyramid-technique [Berchtold et al., 1998] or approaches based on clustering by random projections [Urruty et al., 2007]. The VA-File was one of the first approaches which was proposed to face the problems due to dimensionality with image descriptors. Specifically with local image descriptors, work [Sivic and Zisserman, 2003] was the first to structure the local features differently, by using a bag-of-features representation that only involves one feature vector per image, thus generating a very sparse feature space that can be accessed quickly with simple inverted files.

All the aforementioned access methods perform *exact* retrieval, in contrast with other more recent techniques studied for large databases and that perform *approximate* retrieval, where the retrieval process as accelerated at the cost of a loss in precision: the set of neighbors retrieved is approximate insofar as it is a subset of the real answers. Generally, the system can be tuned according to a precision parameter that controls the proportion of missed nearest neighbors. According to the manner of controlling this precision, ap-

proximate techniques are *c-approximate* or *probabilistic*: c-approximate methods retrieve approximate points with a distance from the query that at most equals c times the distance between the query and its exact nearest neighbors, like Local Sensitive Hashing methods (see the comparison [Andoni and Indyk, 2006]) and random projections [Kleinberg, 1997]. Probabilistic methods, also known as δ-NN, define probabilistic filtering rules which select only the regions of the feature space having the highest probability to contain a correct answer. They make some assumptions on the database distribution in order to guarantee the retrieval of exact neighbors with a given probability, such as [Ciaccia and Patella, 2000; Berrani et al., 2003; Joly et al., 2007]. The VA-BND algorithm is also a probabilistic method that adapts the VA-File to approximate search [Weber and Böhm, 2000]. Note that the above classification is not exclusive and different techniques are hybrid. One can find in [Patella and Ciaccia, 2008] a more complete classification of these approaches.

As an alternate solution, *progressive* retrieval has been advocated more recently. The principle of progressive similarity search is to offer intermediate answers to the user during retrieval: coarse responses are quickly returned, and then refined progressively until reaching exact responses or user satisfaction. Several kinds of solutions exist to deal with progressive retrieval, each of them tackles the problem under very different viewpoints; we reported the different proposals in publication [Bouteldja et al., 2008c]. A part of them exploits the specificity of some image descriptors such as [Lisin et al., 2005b] who combine complementary descriptors involving different capacities of filtering, or [Landré et al., 2001] who use an image descriptor based on wavelets that can be computed at several levels of resolution; while other approaches develop dedicated index structures like in [Jomier et al., 2005] where a tree-based structure describes the image content spatially and allows recursive retrieval. For further details, the reader can consult publications [Bouteldja and Gouet-Brunet, 2008; Bouteldja et al., 2008c].

4.1.2 Difficulties and lacks

High-dimensional spaces show surprising properties that are counter intuitive with respect to the behavior of low-dimensional data. Theses properties were known and have been studied since a long time in several disciplines such as economics or statistics. They are traditionally gathered under the term of *Curse of Dimensionality* (CoD), denomination coined by Richard Bellman to describe the problem caused by the exponential increase of the volume with the augmentation of the space dimension when addressing the problem of optimizing functions with several variables [Bellman, 1961]. Later, the term was used to indicate, more generally, nonintuitive phenomena observed when the dimension of data increases. Among these phenomena, the most important ones observed are (under the assumption of a uniform distribution): the "*exponential increase of the space volume*" phenomenon, whose main effect is that the number of partitions generated by structures based on space partitioning grow dramatically and has for a consequence a drastic performance decrease of these structures; the "*measures concentration*" phenomenon, which asserts that in high dimensions, distances between objects become almost identical, with one consequence that the noise coming from data acquisition may disturb too much the usual metrics. For details, the reader can see [Beyer et al., 1999; Weber et al., 1998; Böhm et al., 2001; Verleysen and François, 2005] where these phenomena are fully described.

As the dimension increases, the aforementioned effects increase enough to have a negative impact on the performance of the index structure. One may then wonder from which dimension the CoD is reached, or what is a "high-dimensional space". Unfortunately in literature, the response is not unique. Besides those that consider that the CoD is reached after three dimensions, because data cannot be visualized, several responses to these questions were given: Verleysen and al. for neuronal networks [Verleysen et al., 2003], Weber and al. for multidimensional access methods [Weber et al., 1998], show that an exhaustive scan is faster than traversing a tree-structured index from a dimension of about 10, but under the assumption of uniformity of the vectors composing the data set; Beyer and al [Beyer et al., 1999] demonstrate that k-NN search becomes meaningless from dimension 20 (under some assumptions such as data uniform distribution), because of the measures concentration phenomenon; Verleysen in [Verleysen, 2003] links the curse of dimensionality problem to the number of points composing the database, by assuming that a space is high-dimensional when the size of its population does not grow exponentially with its dimension; similarly in [Verleysen and François, 2005] for learning purposes, he defines a high-dimensional space as a space where learning mechanisms suffer from the lack of sufficient training samples w.r.t. the dimensionality of the space.

These different responses clearly indicate that it is still not possible to know *a priori* if a given index structure will be efficient in a given feature space. Moreover, all the demonstrations on the CoD phenomena as well as the evaluation of many index structures suppose some assumptions on the data distribution, most of the time a uniform or a gaussian distribution. Consequently, are these definitions of a high-dimensional space still relevant on data sets differently distributed? Experiments we performed on real image data sets (see section 4.2) conduced to the conclusion that the response is no. In addition, while the existing research efforts produced a large number visual content descriptors, the characteristics of these descriptors were mainly studied from a quality of description point of view, and the links between these characteristics and the index structures remain largely unexplored. These links should allow selecting existing (or put forward new) index structures that can provide the scalability required for content-based search in very large databases. Last but not least, most the proposed access methods are not compared to each others in the same work and then with the same evaluation framework (same data sets, same queries, same material, etc), making difficult their evaluation. As an illustration, experiments we conducted on our data sets for evaluating strategies of multiple-query retrieval did not lead to the same conclusions as experiments previously done by other researchers on these strategies but with different data sets (see section 4.3). Unfortunately, today there is no benchmark initiative to compare multidimensional access structures dedicated to similarity search. In the community of CBIR, some discussions are in progress but it is important to notice that such evaluations will remain challenging due to the large range of existing approaches and the different parameters to consider.

4.2 Evaluation of index structures for image databases

According to the problems described in section 4.1.2 on the evaluation of access methods w.r.t. the CoD phenomena, I started my activity on multidimensional access methods by studying the behavior of some state-of-the-art indexing approaches for similarity search on

several data sets coming from CBIR and differently distributed. This work is described in section 4.2.1 for the experiments performed according to the CoD phenomena, while an extensive evaluation of classical access methods is presented in section 4.2.2.

4.2.1 Revisiting the curse of dimensionality phenomena

Several demonstrations and experiences have been done in literature on the curse of dimensionality according to tree structures. In particular, under the assumptions of uniformity, [Weber et al., 1998] showed that an exhaustive scan is faster than traversing a tree-structured index from a dimension of about 10. Indeed, from this dimension, nodes overlap increases so drastically that most of the tree tends to be visited. To reduce the problem of dimensionality, in the same article he proposed the VA-File structure, that performs a sequential scan on approximations of the data to accelerate retrieval. Because this demonstration was done on uniform distributions, as for several other comparisons, we decided to perform extensive experimental evaluations to study the performance of tree structures on several categories of distributions of data. As tree-structured index, we chose to use the SR-tree [Katayama and Satoh, 1997], because from one hand, its code was available on the line at the time when the experiment started, and from another its performance is known to be far better than that of the regular R*tree [Beckmann et al., 1990] for high dimensions. All our experiments were performed on ϵ-sphere queries. When dealing with local image descriptors, we are convinced that the k-NN strategy is not well suited because k may be very different for each of the points composing the query, and cannot be determined intuitively as with global descriptors where it can be fixed by the user as the number of images to return.

The experiments were performed over different data sets, split into three categories: 1 data set obtained from local descriptors extracted from image and video contents, 1 synthetic data set uniformly distributed and 3 data sets involving synthetic clusters following a gaussian distribution with 3 different standard deviations and centers randomly generated. These distributions were computed for two sizes of population with the aim of obtaining a "small" experimentation set (218,000 points) and a "big" experimentation set (1 million of points). In addition, each data set was generated for different dimensions (8, 17 and 29 for the small database and 9, 18 and 30 for the big one). We obtained a total of 30 data sets. We evaluated the gain in CPU time obtained for ϵ-sphere queries with the index structure against a sequential scan, and also the gain in I/O access (or in number of acceded nodes). Some of the results of this study were published in [Bouteldja et al., 2006b] and additional results are detailed in the research report [Bouteldja et al., 2006a]. In the following, we report the most interesting conclusions, that clearly nuance the state-of-the-art demonstrations performed on uniform data sets.

Impact of the data distribution. The experiments demonstrated that, for the real data sets as well as for the most clustered distributions, the curse of dimensionality is clearly not reached! The speed up for the largest dimensions ($d = 29, 30$) is significant even for large values of the number of neighbors. In contrast with the uniform data set, there is a very small gain at dimensions $d = 17, 18$ and no gain at dimensions $d = 29, 30$. This trend was confirmed by the measurement of the ratio of nodes acceded during the tree traversal: this ratio remains small ($\leq 27\%$) for the real and clustered distributions, while

it quickly reaches 100% for the large dimensions. These results confirm the conclusions of Weber in [Weber et al., 1998] on the behavior of tree-structured indexes facing the dimensionality, but we observed that with differently distributed data, the conclusion are very different, tree structures remain still effective.

Impact of the dimension. Additional experiments on the influence of the dimension of the CoD, conducted on a database of size of 200,000 points obtained with CBIR descriptors computed on images, led to the following conclusions: unsurprisingly, the performance of the index structures decrease while the dimension increases. More precisely, the SR-tree remains more efficient than a sequential scan up to 120 dimensions. For comparison, the performances obtained with a VA-File structure [Weber et al., 1998] also get worse, but more slowly that with the SR-tree: it is affected by the CoD after 512 dimensions. These experiences demonstrated that the dimension from which the CoD affects indexing techniques depends on several parameters such as the data distribution and the indexing technique.

Impact of the database size. Other experiments with the same data sets consisted in fixing the dimension of the feature vectors (64) and in varying the database size from 200,000 to 1 million of vectors. With these experiments, we observed that the dimension from which the CoD affects indexing techniques also depends on the database size: here, the gain increases with the database size (over or sub-linearly according to the distribution and the index structure). Consequently, the larger the database size the higher the dimension where the CoD appears. These results confirmed the experimented performed by Verleysen in [Verleysen, 2003] who links the CoD problem to the number of points composing the database.

This series of experiments led us to the conclusion that we can not assert that an index is inefficient for a given dimension or is better than another index, since there are many parameters affecting the indexes behavior, such as the distribution, the dimension and the size of the data set.

4.2.2 A study on state-of-the-art indexes' behavior

In this work, our objective was to compare different categories of state-of-the-art multi-dimensional indexes that perform exact retrieval in exactly the same framework, i.e. with the same data sets, the same material and the same queries. The underlying objective was to study the characteristics of each category of approach that enables reaching good performances. Because they are representative of the database literature, the four following indexing schemes were evaluated:

- *Pyramid-technique:* this approach [Berchtold et al., 1998] transforms each high-dimensional vector into a one-dimensional vector. B+tree search on them is then performed as a filtering step. By using all the initial representation of the resulting vectors, a sequential step returns the correct answers.
- *SR-tree :* this approach [Katayama and Satoh, 1997] structures data in a hierarchy of clusters represented by the intersection of bounding sphere and rectangles, with the aim of minimizing overlap in high-dimensional spaces.

- *VA-File:* The VA-File [Weber et al., 1998] accelerates sequential scan by using vector approximations as a first filtering step. Then, the original vectors selected in the first step are processed to get the exact answers.

- *GC*Tree :* we use a clustered variant of the GC-tree [Cha and Chung, 2002], developed by Nouha Bouteldja during her PhD thesis. The latter structure is based on a space partitioning strategy which is optimized for a clustered high-dimensional image data set. It uses a recursive hierarchical decomposition of the multidimensional space into cells so as to reach leaves whose points hold in a disk page. The GC*Tree uses adaptive space decomposition and clustering; it has some differences with the GC-tree that are reported in [Bouteldja et al., 2008a].

The data sets were taken in the domain of CBIR: 4 millions of 128-dimensional feature vectors where extracted from images of heterogeneous contents. Three very differently distributed data sets were considered by using three techniques of description: two global image descriptors (noted $DFTT$ and RGB), and one local descriptor (noted $LOWE$). The first experiment evaluated CPU time as a function of the number of neighbors found in the sphere query for 500,000 points and for each data set; see the results in figure 4.1.

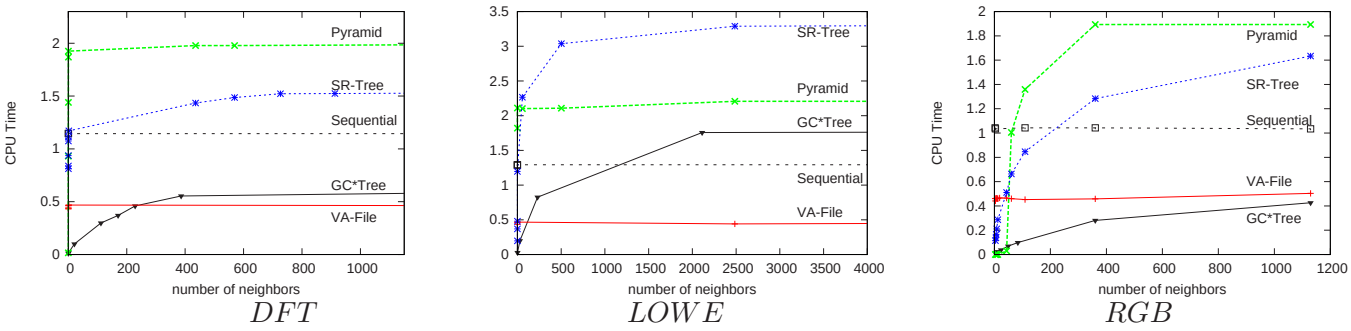


Figure 4.1: *CPU Time versus numbers of neighbors with 500,000 feature vectors.*

The pyramid-technique is outperformed by sequential scan for all the data sets. This is due to the intersection of the sphere query with too many pyramids in high dimensional spaces, leading to a high percentage of neighbors returned after the B+Tree search; this observation confirms the results of Amsaleg and al.[Amsaleg et al., 2004]. As the Pyramid-technique, the SR-Tree is affected by the CoD for almost all the data sets. The behavior is better with the DFT and RGB databases; this is due to the their distributions which are more clustered than the $LOWE$ data set. The GC*Tree is also affected by the CoD, but only with the $LOWE$ data set, most likely because of the distribution of this data set, which is more uniform than the ones of the two other data sets. For $LOWE$, the GC*Tree is outperformed by the VA-File even for relatively small values of ϵ. In contrast, it performs well for the two other data sets: for the DFT set, is is even better than the VA-File until roughly 200 neighbors in average. For the RGB set, it outperforms the VA-File for all the tested values of ϵ.

Other experiments were carried out, among them the evaluation of the cost in terms of I/O and of the scalability that measures the impact of the data set size on the index

performance. By varying the size from 500,000 to 4 millions of points, we observed that the indexes that do not suffer from the CoD at 500,000 points (VA-File and GC*Tree) increase their performances with the database size, in accordance with the experiences on the impact of the database size already done with the SR-tree (section 4.2.1).

This work has been performed during the PhD thesis of Nouha Bouteldja and was published in [Bouteldja et al., 2008a]. More experiments, that confirm the aforementioned conclusions in addition with the evaluation of index X-tree [Berchtold et al., 1996], were also done during the CNAM engineer thesis of Michel Martinez [Martinez, 2007].

4.3 Nearest neighbors search with multiple queries

When considering feature vectors associated to image descriptors, there exist a lot of configurations where the entities of interest (image, part of image, object, video, video segment, etc) are described with descriptors that produce several feature vectors. For example, the description based on interest points can involve several hundreds of feature vectors associated to the points extracted, see figure 1.2(c). Image description based on regions of interest can involve several tens of regions of interest and as many feature vectors, see figure 1.2(b) and for example the work [Fauqueur and Boujemaa, 2006] on image description with such descriptors. Even with global visual descriptors, such as color histograms that describe globally the content of images and produce one feature vector per image, describing a video content by extracting such a descriptor on every frames produces a large number of high-dimensional feature vectors, like in [Cheung and Zakhor, 2003] for copy detection. Consequently, performing a similarity search of these entities in a database involves a query defined by several feature vectors of the same family, often called a *multiple query*. During her PhD thesis, early work of Nouha Bouteldja consisted in studying new strategies for accelerating retrieval of such configurations of feature vectors. The objective was to perform nearest neighbors retrieval of a multiple query more efficiently than performing several consecutive simple queries, by exploiting simultaneously all the feature vectors of the query during search.

To our knowledge, Braunmüller et al. [Braunmüller et al., 2000] were the first to study multiple similarity queries, for mining in metric databases. They proposed a general framework for answering simultaneously a set of similarity queries (k-NN and sphere queries) and give algorithms both for reducing I/O cost and CPU time. In particular, they proposed CPU saving by using the triangle inequality and two lemmas. Experiments were provided for multiple k-NN queries that evaluates the speed-up for a linear scan of the collection of points as well as for an X-tree traversal.

4.3.1 Strategies for tree-based structures and metric spaces

For reducing I/O cost and CPU time, we used similar ideas as in [Braunmüller et al., 2000] but with adaptation to ϵ-sphere queries instead of k-NN queries. We investigated two orthogonal strategies to speed-up the processing of multiple similarity queries in metric databases and tree-based index structures: reducing of I/O cost (i.e. the number of disk accesses) and reducing of the CPU cost (i.e. the number of distance calculations). For

improving CPU time, we also exploited the distance triangle inequality relying on the lemmas proposed by Braunmüller and al. as well on three other novel lemmas. In our experiments, the tree structure is an SR-tree, but it is important to note that the proposed or customized strategies are easily adaptable to other tree structures.

Strategy #1: This strategy saves I/O time by reading a single page (node or leaf of the tree) only once and processing it for the whole set points in the query.

Strategy #2: This one attempts to save CPU time by avoiding distance computations as much as possible in applying the distance triangle inequality in the metric space. It supposes that, prior to the sphere query, the distances $d_{p_i p_j}$ between any two points p_i and p_j of the query, were pre-computed. Such a strategy is motivated by the fact that tests implied by the triangle inequality are significantly cheaper than computing a distance. For this strategy, we used five lemmas. Lemmas 1 and 2 were already proposed in [Braunmüller et al., 2000], while Lemmas 2a, 3 and 3a were new contributions we proposed for ϵ-sphere queries. In particular, lemma 2a (lemma 3a) takes advantage of a successful test with lemma 2 (lemma 3) on a couple of points (q_1,q_2) for other couples (q_1,q_3). The proposed nearest neighbors retrieval algorithm simultaneously uses lemmas 1, 2 (or 2a) and 3 (or 3a).

The two strategies, the five lemmas and the corresponding algorithms are described in publication [Bouteldja et al., 2006b].

4.3.2 Comparison of the strategies

The two strategies were evaluated against the simple strategy that consists in sequentially running a classical nearest neighbor search for each of the points in the query. Here, no optimization is done for limiting distance computations as well as I/O access. The data sets used are the ones presented in section 4.2.1 for experiments on curse of dimensionality. Several configurations of queries were considered according to their size and distributions: queries uniformly distributed, gaussian distribution and real queries extracted from parts of images. Figure 4.2 plots the gain in acceded nodes versus the number of neighbors found for various sizes of multiple queries m. The gain is the ratio of the average number of nodes acceded with the basic strategy times m over the average number of nodes acceded by using strategy #1. Note that for all values of m, the larger the dimension, the higher the gain.

Strategy #2 was evaluated by comparing the performances of the five lemmas independently and then by combining the most performing ones. These experiments showed that combination of lemmas 1, 2a and 3 provide the best gain in CPU time, as illustrated in figure 4.3. Such a gain was confirmed on another data set of 1 million of feature points and on a real scenario of query by example performed on a database of heterogeneous visual contents (one of the queries was the one of figure 1.1(b)).

The experiments and associated results are fully described in publication [Bouteldja et al., 2006b]. This work was supported by the French project BIOTIM (ACI "Masses of Data", 2003-2006) where it was applied to the fast search of similar images described with local descriptors in the context of large volumes of images dedicated to biodiversity.

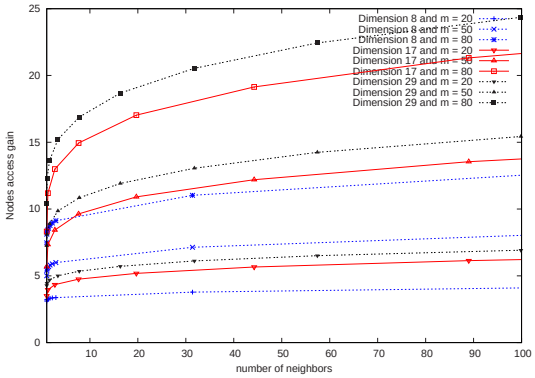


Figure 4.2: *Nodes' access gain for m in {20, 50, 80}.*

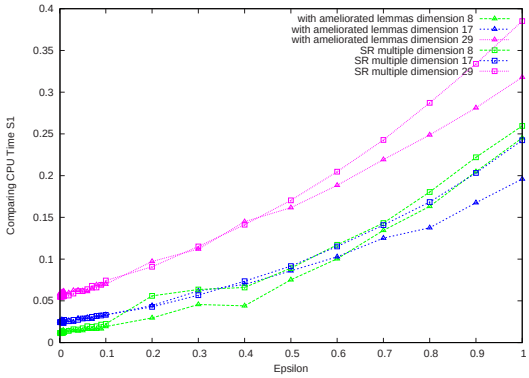


Figure 4.3: *CPU time with and without Lemmas 1, 2a and 3 used jointly.*

4.4 HiPeR: a hierarchical model for accelerating retrieval in high-dimensional spaces

HiPeR is a Hierarchical and Progressive Retrieval model we have proposed for speeding up retrieval of nearest neighbors in high-dimensional spaces. Its main principle is presented in section 4.4.1. The model was designed to address three scenarios of similarity search: exact search that is described in section 4.4.2, approximate search, described in section 4.4.3 and progressive search, described in section 4.4.4. As a conclusion, the main characteristics and results of the experiments on HiPeR are summarized in section 4.4.5.

4.4.1 General concept

HiPeR is based on a hierarchy of feature vectors, having the aim of minimizing the CoD phenomena by integrating low-dimensional representations of the feature vectors. Given a high-dimensional feature vector describing an entity, the hierarchy associated with it is composed of this vector as well as one or several lower-dimensional vectors describing the same entity. Retrieval through such a hierarchy consists in progressively scanning it from the smallest dimensional vectors up to the larger ones as long as closeness to the query vector is verified or as long as the user requires more precise responses. According to the case, the similarity search performed at each level of the hierarchy can be exact or approximate. HiPeR involves specific dimension reduction methods to build the hierarchy of vectors. Oppositely to classical feature selection techniques, it is able to perform exact search and can deal with dynamic databases. In this work, HiPeR was developed for similarity search with ϵ-sphere queries.

Let E be an entity described by v^E, a d-dimensional point in feature space V. Defining a hierarchy H_n for v^E consists in building an ordered set of n ($n \geq 2$) vectors with increasing dimensions:

$$H_n(v^E) = \{v_1^E, v_2^E, ..., v_{n-1}^E, v_n^E = v^E\} \tag{4.1}$$

where $v_i^E, i < n$ is a d_i-dimensional vector of space V_i, obtained from v_{i+1} by a mapping

m_i. Any two vectors in H_n can be linked by combining the mapping functions m_i; for instance, each v_i^E is related with v^E by a mapping M_i were $M_i = \odot_{k=i}^{n-1} m_k$. According to this definition, entities are now described into several feature spaces $\{V_1, V_2, ..., V_n = V\}$ of increasing dimensions ($d_i < d_{i+1}, \forall i < n$). The spaces V_i with $i < n$ are called intermediate feature spaces.

Let $\mathcal{N}_V(E)$ be the set of entities, which are the nearest neighbors of entity E, retrieved by similarity search in multidimensional feature space V. In order to accelerate the retrieval of $\mathcal{N}_V(E)$, we propose to exploit the hierarchy defined above by performing search successively in subsets of the different feature spaces: search is done in space V_i, prior to be performed in space V_{i+1}, $i < n$, as follows: the initial sphere query takes as an entry the whole set of features V_1. The subset of neighbors $\mathcal{N}_{V_i}(E)$ resulting form the search in V_i is the subset of entities on which sphere query in space V_{i+1} is done. Obviously, the relations between the intermediate feature spaces and high-dimensional space V have some constraints. In particular, when considering exact nearest neighbors retrieval, the following relation has to be satisfied:

$$\mathcal{N}_{V_i}(E) \supseteq \mathcal{N}_{V_{i+1}}(E) \quad \forall\ 1 \leq i < n \tag{4.2}$$

Intermediate feature spaces V_i ($i < n$) may characterize less precisely the entities and then needlessly provide larger sets of neighbors that contain non relevant entities. But on the other hand, it is well-known that similarity search is much faster at low dimension: indeed the CoD effects are not yet significant.

Here, we do not make any assumption on the multidimensional index structure employed to accelerate nearest neighbor retrieval at each level of the hierarchy. Let Idx_i denote the index defined in space V_i, $i \in [1..n]$. We only suppose that each entity E is described by a vector having the same identifier in the different spaces or equivalently there is a one-to-one mapping between the vector describing entity E in index Idx_i and the one describing the same entity in index Idx_{i+1} ($i < n$). This allows exploiting the index of V_{i+1} on the subset $\mathcal{N}_{V_i}(E)$ obtained in the previous filtering step. Such a correspondence depends on the implementation of the index structures. We proposed an implementation for indexes VA-File and GC*tree; the one for VA-File is described in section 4.4.1.3.

Let *dist* be the distance associated with space V, and E' another entity characterized by the hierarchy $H_n(v^{E'}) = \{v_1^{E'}, ..., v_n^{E'}\}$ involving the same mappings $m_i, i \in [1..n[$ defined for E, with $v_n^{E'} \in V$. We suppose that *dist* is sufficiently generic to apply or to be adapted to all the intermediate spaces of the hierarchy. In the case of sphere queries, we further assume that the sphere radius ϵ is the same at all levels in the hierarchy. If the following implication is satisfied for all $1 \leq i < j \leq n$, then relation 4.2 is satisfied too:

$$dist(v_i^E, v_i^{E'}) \geq \epsilon \Rightarrow dist(v_j^E, v_j^{E'}) \geq \epsilon \tag{4.3}$$

Condition 4.3 expresses that vectors that are not neighbors in V_i cannot be neighbors in V_j with a larger dimension. Conversely, two neighbors in V_i are not necessarily neighbors in V_j with a larger dimension.

4.4.1.1 Mappings for intermediate feature spaces construction

In a hierarchy H_n, one important parameter is the mapping m_i that provides a feature space V_i of lower dimension ($1 \leq i < n$). Mapping m_i must provide vectors of V_i that preserve rule 4.2. In publication [Bouteldja et al., 2008a], we proposed and demonstrated the validity of two families of mappings for any of the distances L_p: the first one simply consists in truncating the feature vectors of V to obtain those of V_i, while the second one considers partial p-norms of the vector components. When experimenting HiPeR on image descriptors based on color histograms, we used L_1 as distance for all levels of the hierarchy, L_1 being traditionally employed with this family of descriptors; the mappings m_i considered were partial p-norms.

4.4.1.2 Heuristics for optimal navigation in the hierarchy

By default, HiPeR suggests a *systematic scan* of the whole hierarchy of feature vectors from lowest dimension (V_1) to the final one (V), i.e. searching neighbors of query E in space V_{i+1} from the set $\mathcal{N}_{V_i}(E)$ previously obtained ($i < n$). The challenge for efficiency is the appropriate choice of intermediate feature spaces that have a high filtering capacity. Unless the distribution of vectors in each space is uniform, such a criteria is not easy to determine a priori. In the presence of skewed data, the filtering capacity of a level clearly depends on the spatial location of E. We proposed an alternate strategy, called *selective scan* in which we might skip search at level i if we can decide a priori for a given query that this search has low filtering capacity. We can assert that scanning space V_i ($i < n$) is relevant at query location E if:

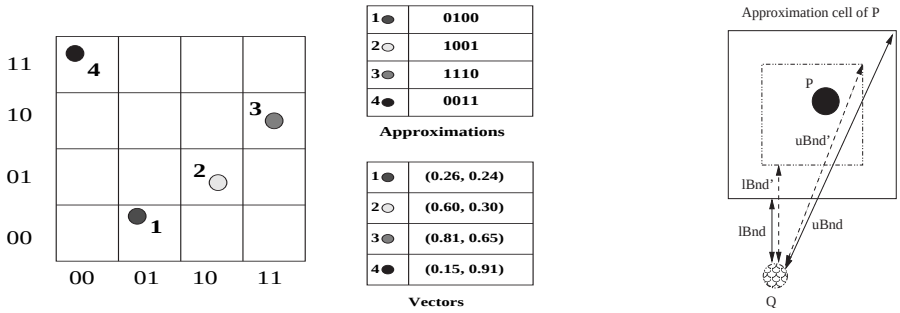
$$N_{i-1}(E).t_i^E + N_i(E).t_{i+1}^E < N_{i-1}(E).t_{i+1}^E \tag{4.4}$$

where $N_j(E)$ is the number of entities filtered at level j for query E (i.e the cardinality of $\mathcal{N}_{V_j}(E)$ or of the whole data set if $j = 0$) and t_j the average time of vector processing in V_j (i.e. sphere query through the index structure). This equation can be used to determine the relevance of exploiting feature space V_i. In publication [Bouteldja et al., 2008a], we proposed a criterion that estimates the selectivity of feature space of level i according to equation 4.4, for the instantiation of HiPeR on structures VA-File and the GC*Tree. If this criterion is not satisfied at level i, it implies that this level does not eliminate enough vectors to accelerate retrieval efficiently. Therefore, it is preferable to skip it and to jump to the following level $i+1$. Otherwise search can be performed in V_i, from the intermediate set of NN $\mathcal{N}_{V_{i-1}}(E)$ obtained at previous level or from the entire data set if $i = 1$. Note that, the estimation is never performed at the last level, since here the sphere query has to be run anyway in order to get the exact answers to a given query.

4.4.1.3 Illustration with the implementation of a hierarchy of VA-Files

The VA-File [Weber et al., 1998] was an index structure proposed to accelerate sequential scan in high-dimensional spaces by using vector approximations (stored in a Vector Approximation File, called the VA-File) instead of the features themselves, as a first filtering

step. Figure 4.4(a) gives an example of VA File with its approximation cells.



(a) Vectors representation (b) Similarity search

Figure 4.4: *Illustration of a VA File with two dimensions.*

When searching points similar to a query one, the VA-File is first sequentially scanned entirely to exclude a part of vectors (filtering step). For each approximation, a lower and an upper bound ($lBnd$, $uBnd$) are computed between the query and the approximation (see figure 4.4(b)), allowing to eliminate irrelevant vectors. After the filtering step, a superset of the solution is produced and only a small subset of the real features are loaded from second memory in order to check for similarity on the vectors themselves.

To exploit the HiPeR hierarchy of size n on index structures, any two descriptors v_j^E and v_i^E $(1 \leq j < i \leq n)$ associated with the same entity E must have the same identifier in the hierarchy. In order to avoid I/O increase with the implementation of HiPeR on VA-Files, we create a hierarchy of VA-Files, called a VA-hierarchy, that only utilizes the approximations of each VA-File in the intermediate levels which reside in main memory: I/O are necessary only at the last level to fetch the features. This insures that I/O access with exact retrieval is at worst equal to the one given by the simple index VA-File. Approximations as well as data vectors are identified by their rank in each VA-File. Because they are inserted in the same order in the respective approximations files, they have the same identifier in the hierarchy. The details of this implementation are described in [Bouteldja et al., 2008a], as well as the implementation of HiPeR with the GC*tree structure.

4.4.2 Performances for exact similarity search

The evaluation of HiPeR for exact similarity search was performed on the same data sets DFT, RGB and $LOWE$ as the ones exploited in section 4.2.2 for comparing state-of-the-art index structures. With these data sets, the most efficient index structures were the VA-File and the GC*tree (see figure 4.1). Consequently, HiPeR was instantiated on these structured, conducting to a hierarchy of VA-Files (a VA-hierarchy) and a hierarchy of GC*tree called GC*hierarchy. The performances of these implementations were compared to the ones of the simple index structures VA-File and GC*tree, in terms of CPU time and I/O access. The complete evaluation was published in [Bouteldja et al., 2008a] and the most important conclusions are reported here:

Impact of the hierarchy depth. For each hierarchy, the number of levels n was augmented until performance stagnation or deterioration, under a systematic scan scenario.

The configurations of dimensions considered were: 128-64 (2 levels), 128-64-32 (3 levels), 128-64-32-16 (4 levels) and 128-64-32-16-8 (5 levels). Whatever the configuration and the number of neighbors considered, the CPU time obtained was much more better than using the simple index VA-File or GC*tree, justifying the relevance of the model (see table 4.1 for an idea of the gains obtained). The second observation was that the optimal number of levels in the hierarchy is distribution dependent: the 4-level configuration was the best for DFT, the 2-level configuration provided good performance for RGB, while it was three levels with $LOWE$. From these observations, we concluded that an a priori study is necessary to choose the best number of levels, before the use of the systematic strategy on a given distribution. If not, a compromise for these data sets is to use configuration 32-64-128 that provides good results in average. Concerning the I/O access, because the implementations of the VA-hierarchy and the GC*hierarchy exploit vector approximations in the intermediate levels, I/O are only acceded at the last one, which ensured that I/O access with them is at worst equal to the one given by the simple indexes.

Scalability. The size of each data set was varied from 500,000 to 4 millions of points. We observed that HiPeR is robust when scaling the data set size.

Data set	**Best index configuration**	**Data set size (in millions)**					
		0.5	1	1.5	2	3	4
DFT	4-level VA-hierarchy	4.89	3.56	3.46	3.45	3.37	3.24
$LOWE$	3-level VA-hierarchy	2.59	2.03	2.03	1.99	2.31	1.90
RGB	2-level GC*hierarchy	1.52	2.07	1.58	1.48	2.19	2.02

Table 4.1: *CPU gain obtained with the best configurations of hierarchical indexes VA-hierarchy and GC*hierarchy, w.r.t. their simple version VA-File and GC*tree. Here, the gain is computed as* $CPU_{Simple\ version}/CPU_{Hierarchy}$.

Systematic versus selective strategy. CPU time was measured for the two strategies of hierarchy traversal: systematic and selective scan. From the experiments, it is clear that whatever the data set, selective scan strategy has similar performances than the best configuration of systematic scan strategy. Even though skipping one or more levels of the hierarchy does not bring a significant improvement (the estimator employed does not penalize retrieval time either), selective scan strategy presents the advantage of adapting to every data set since it does not require an a priori study as with the systematic one.

4.4.3 Approximate similarity search

Besides exact similarity search, we extended HiPeR to deal with approximate similarity search. The model was implemented for hierarchies of VA-Files. The algorithm proposed for each level of the hierarchy is presented in section 4.4.3.1. Then, section 4.4.3.2 presents the complete algorithm of approximate similarity search for the whole hierarchy of VA-Files. In this study, the difficulty was to propose relevant models for the estimation of the precision loss implied by approximate search, for the VA-File (at each level) as well as for the whole hierarchy VA-hierarchy.

4.4.3.1 Approximate VA-File

In [Weber and Böhm, 2000], two methods for approximate k-NN search with the VA-File were proposed: first, the VA-LOW strategy that simply returns as neighbors the approximations instead of the features, thus saving I/O access as well as distances computation; second, the VA-BND strategy that modifies the bounds of the approximation cells to accelerate retrieval.

We chose to focus on the VA-BND strategy, because VA-LOW can be easily incorporated into any similarity search algorithm. VA-BND is based on the following observation: reducing the size of the vector approximation cells saves retrieval time while modifying the exact answers set only slightly. The filtering strategy of VA-BND only differ with exact k-NN retrieval by the introduction of operation $Alter(lBnd, uBnd, \alpha)$ which modifies the lower bound $lBnd$ and upper bound $uBnd$ of the approximation cells as following (see also the illustration of figure 4.4(b)):

$$lBnd' = lBnd + \alpha \quad \text{and} \quad uBnd' = uBnd - \alpha \tag{4.5}$$

According to this modification of the cells, a model for the estimation of the errors introduced in the nearest neighbors set was proposed in [Weber and Böhm, 2000] for k-NN search, under the assumption of uniformity of the data set distribution. In the following, I present the algorithm of retrieval we proposed for ϵ-sphere queries and the associated model of precision loss, which is probabilistic and more general than the k-NN one because supposing few assumptions on the data set distribution.

Approximate VA-File for ϵ-sphere queries: ϵ-VA-BND algorithm

ϵ-VA-BND is the algorithm of approximate similarity search we proposed for ϵ-sphere queries; it is described in algorithm 1. As VA-BND, it modifies the bounds of the approximation cells with parameter α. Differently, filtering rule #1 allows loading only the vectors associated with the cells which intersect sphere query Q. The real features are read from the vector file (VF) using function $GetVector()$ which loads the page including the vector only if not already loaded. Filtering rule #2 avoids distance computation when the approximation cell is inside the query.

Model for precision loss estimation

The estimation of the precision loss during similarity retrieval using ϵ-VA-BND differs from the one of VA-BND: it is not based on definition a theoretic probability density but on cumulated histograms of distances. Therefore ϵ-VA-BND does not need a function that approximates well f^{δ_l} and few assumptions are made on the data distribution. In addition, for a fixed shift α, the precision loss also depends on ϵ. In [Bouteldja et al., 2008b; Bouteldja et al., 2009], we proposed to define the probability P_l of having an effective precision loss due to replacing $lBnd$ by $lBnd'$ as a function of the number of points inside a query (n_d) and the number of approximation cells that intersect the same query when using $lBnd$ (resp. $lBnd'$) (n_{lBnd} and $n_{lBnd'}$). Since these values are not

Algorithm 1: ϵ-VA-BND(F,Q)

```
Input  : F a Vector Approximation File of size N, Q a sphere query with center q and
         radius ε, α the shift parameter of Equation 4.5
Output: V the set of NN vectors
//  Retrieves the set V of vectors inside Q
V := empty;
for i = 1 , N do
    a := F.GetApproximation(i);
    GetBounds(a, q, lBnd, uBnd);
    Alter(lBnd, uBnd, α);
    //  Filtering rule #1
    if lBnd < ε then
        v := GetVector(a);
        //  Filtering rule #2
        if uBnd < ε then  V+={v};
        else
            d := GetDistance(v, q);
            if d < ε then V+={v};
        end if;
    end if;
end for;
return V;
```

known in advance, they are approximated from samples of the data set: for each query sample, the distance (and the $lBnd$ bound) between it and each data set vector (resp. approximation cell) is computed, providing a cumulated histogram CH^d (CH^{lBnd} resp.) of the frequency function f^d (f^{lBnd} resp.). For a given ϵ value, CH^d gives the ratio n_d of the number of neighbors inside a query, while CH^{lBnd} provides an approximation of n_{lBnd}. Given a shift α, an approximation of $n_{lBnd'}$ is deduced by shifting histogram CH^{lBnd} with α. Estimation of $n_{lBnd'}$ from CH^{lBnd} and a given α allows computing P_l quickly, given precomputed CH^d and CH^{lBnd}. Consequently, the computation of the precision loss can be done on line, and then quality of answers can be selected by user at query time.

A similar study can be conducted for upper bounds when using $uBnd'$ to define P_u, the probability of precision loss PL_u. Note that taking $uBnd'$ into account instead of $uBnd$ implies a possible *add of false answers*, while shifting $lBnd$ may imply *loose of correct answers*. Consequently, altering upper bounds may save CPU time, whereas modifying lower bounds saves both I/O access and CPU time. From the definitions of P_l and P_u, we defined the global model of precision loss $P_{l,u}$ as following:

$$P_{l,u} = \frac{P_l + P_u}{1 + P_u} \tag{4.6}$$

ϵ-VA-BND was evaluated on the same data sets DFT, RGB and $LOWE$ employed for evaluating HiPeR for exact retrieval. The details of the experiments are reported in publications [Bouteldja et al., 2008b; Bouteldja et al., 2009].

4.4.3.2 Approximate VA-hierarchies

Let VA-H_n be a hierarchy of n VA-Files organized according to the HiPeR model. The algorithm for approximate similarity search at level i of this hierarchy ($i \leq n$), referred as ϵ-VA-BND$_i$, is largely inspired from the ϵ-VA-BND one described in algorithm 1. But it differs in two points, presented in the two following paragraphs:

Given a sphere query of size ϵ and shift α_i, the input/outputs are different: while ϵ-VA-BND searches candidate neighbors in the Vector Approximation File VAF and provides feature vectors as response, ϵ-VA-BND$_i$ takes as input the IDs of the approximations in the Vector Approximation File VAF$_k$ returned at level k (with $k = i - 1$ in case of systematic scan and $k < i$ for selective scan) or all the IDs of VAF$_1$ if $i = 1$. As output, if i is the last level ($i = n$ or fixed as the last level by the user), it provides the IDs of approximations in the current VAF$_i$ that verify the similarity criterion, otherwise it returns the feature vectors considered as nearest neighbors.

As already stated in section 4.4.1.3, the retrieval process in the intermediate spaces only exploits the approximation files. As a matter of fact, filtering using rule #2 of Algorithm 1 (associated to upper bounds) becomes useless since it was designed for diminishing vector distance computations, which are not computed at these levels. Consequently, the ϵ-VA-BND$_i$ algorithm designed at level i with $i < d$ does not exploit upper bounds nor rule #2 that is suppressed. Differently, retrieval at last level n exploits both upper and lower bounds. Here, the only difference with algorithm ϵ-VA-BND is the input which is a list of IDs in the Vector Approximation File of previous level.

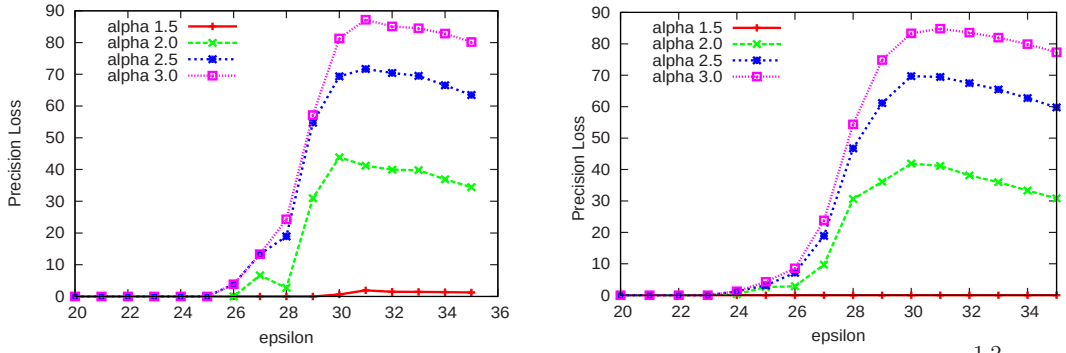
Model for estimation of the global precision loss over the HiPeR hierarchy

According to a level i of the hierarchy, the probability of equation 4.6 corresponds to the precision loss if retrieval is performed the first time with level i using all the approximations as if there were no prior filtering of approximations in steps 1 to $i - 1$. We rename it $P^i_{l,u}$. To define the global precision loss when all the levels of the hierarchy are used together, it is necessary to specify probabilities $P_l^{1,n}$ and $P_u^{1,n}$ that estimate proportions of respectively correct neighbors eliminated and false neighbors added to the answer set: since upper bounds are shifted only at the last level, $P_u^{1,n}$ is defined as $P_u^{1,n} = P_u^n$, whereas probability of loosing correct answers $P_l^{1,n}$ is defined as follows:

$$P_l^{1,n} = P_l^{1,n-1} + (1 - P_l^{1,n-1}).P_l^n \quad with \quad P_l^{1,1} = P_l^1 \tag{4.7}$$

Having $P_l^{1,n}$ and $P_u^{1,n}$, the global precision loss estimation $P_{l,u}^{1,n}$ is deduced from equation 4.6. Note that this model is independent of the index implementation chosen. Examples of effective and estimated losses in precision are displayed in figure 4.5, for a uniform distribution of 1 million of 128-dimensional points and several values of shift α. The hierarchy is composed of $n = 2$ levels (dim. 64-128). We observe that the curves fit well.

In the experiments, a three-level hierarchy VA-H_3 was evaluated (dim. 32-64-128). Because the approximate version of the VA-Hierarchies saves I/O access, while the hierarchy itself, in an exact or approximate retrieval scenario, allows saving CPU time significantly (see the experiments of section 4.4.2), here we focus on the performances of the



(a) Effective precision loss (%) (b) Estimated precision loss $P_{l,u}^{1,2}$ (%)

Figure 4.5: *Average precision loss vs. ϵ with approximate VA-hierarchy.*

VA-hierarchy on I/O access. Several experiments were performed and are reported in publications [Bouteldja et al., 2008b; Bouteldja et al., 2009]. Among them, table 4.2 shows interesting results on the performance of VA-H_3 in terms of I/O gain and scalability w.r.t. the VA-File technique computed in dimension 128, for fixed values of α_i and ϵ.

	Data set size (in millions)					
	0.5	1	1.5	2	3	4
I/O Gain	94.85	91.83	90.12	90.19	88.82	85.50
$P_{l,u}^{1,3}$	63.81	29.43	16.80	16.18	5.69	1.04
DFT ($\alpha_1 = \alpha_2 = \alpha_3 = 1200 = 0.1\epsilon$)						
I/O Gain	39.36	39.30	39.57	39.49	39.52	39.31
$P_{l,u}^{1,3}$	3.52	3.52	3.51	3.57	3.50	3.56
$LOWE$ ($\alpha_1 = \alpha_2 = \alpha_3 = 150 = 0.1\epsilon$)						
I/O Gain	49.94	50.66	53.04	53.03	52.44	51.88
$P_{l,u}^{1,3}$	3.50	3.65	4.83	4.87	4.44	4.11
RGB ($\alpha_1 = \alpha_2 = \alpha_3 = 2500 = 0.17\epsilon$)						

Table 4.2: *VA-hierarchy evaluation: I/O gain (%) and precision loss $P_{l,u}^{1,3}$ (%) when scaling the database size. Here, the gain obtained w.r.t. the VA-File technique is computed as 1-$IO_{VA-H_3}/IO_{VA-File}$.*

For both $LOWE$ and RGB data sets, one can observe that I/O gain is significative and almost constant when the database size is scaled, as well as the precision loss which remains reasonable. The stability of P_l for the same values of α_i across the data sets size shows that the cumulative histograms could be estimated from one of the samples of these data sets, and then be applied to larger data sets successfully. The behavior of DFT is different: we observed that the precision loss changes with the data set size, for fixed values of α_i. Obviously, this result does not mean that we can not achieve small precision loss with important I/O gains when this data set is small, but that the estimation of the α_i from a given precision loss cannot be made from cumulative histograms computed on a predefined sample of DFT, because data is skewed, but rather on samples taken from the

considered data set.

4.4.4 Progressive similarity search

In the previous sections, the hierarchy on which HiPeR is based allowed us proposing efficient strategies for exact and approximate retrieval. Whatever the heuristic used (systematic or selective scan), traversing this hierarchy can be performed until last level n so as to obtain exactly or approximatively the entities satisfying the query. But such a hierarchy also enables a third strategy we call *progressive retrieval*, i.e. the possibility to bring back to the user progressively all the answers obtained at each level i (the set $\mathcal{N}_{V_i}(E)$) and to stop retrieval at a given level, according to some criteria such as the user satisfaction. Such a strategy is particularly meaningful in CBIR scenarios: in this case, the system can provide intermediate results to the user iteratively, aiming at giving him/her fast answers with coarse similarity and the possibility to stop retrieval.

Under this scenario, one important parameter is the quality of the intermediate results provided to the user at each level i. This quality depends on the mapping function employed to construct the intermediate feature spaces. In general, the feature vector employed to describe an entity (i.e. the vector of level n of the hierarchy) has a signification, but the intermediate feature vector obtained at level i ($i < n$) may not have any sense in terms of descriptor of the entity, making potentially irrelevant the return of the responses to the user at this level. Among the mapping functions we proposed in [Bouteldja et al., 2008a], the partial p-norms mapping preserves the significance of the features for several image descriptors, such as those involving histograms: for example data sets RGB and $LOWE$, previously used in experiments, involve histograms that are also meaningful histograms in the intermediate spaces of the hierarchy with this mapping.

The quality of progressive retrieval obtained at each level of the hierarchy was formally evaluated on the ALOI database containing a ground-truth of $100,000$ images [Geusebroek et al., 2005], and again described with the same categories of descriptors to get the DFT, $LOWE$ and RGB data sets of 128-dimensional vectors. Since the selective scan strategy brought good performance with 3 levels (see section 4.4.2), a hierarchy composed of 3 levels was built for each data set (dim. 32-64-128). Precision and recall (P/R) curves were computed for each level of the hierarchy used during retrieval. Then, in order to quantify the loss in precision due to retrieval up to intermediate levels, with respect to the final dimension that provides exact retrieval, the precision loss PL^d was defined for each recall value r as: $PL^d(r) = 1 - P^d(r)/P^{128}(r)$, where P^d represents the precision obtained by exploiting levels up to dimension d ($d \leq 128$). Figure 4.6 gives an overview of the compromise between precision loss and CPU gain, for each data set and each level (Li) of the HiPeR hierarchy.

Whatever the data set, loss in precision is always lower than 25%, which shows the relevance of the used mapping function. We observe that precision loss is small compared to the CPU gain. This experiments shows that, by considering progressive retrieval and then by providing intermediate results to the user, it is possible to accelerate retrieval time many more, while keeping reasonable quality in the returned responses.

Details and other experiments are published in [Bouteldja and Gouet-Brunet, 2008].

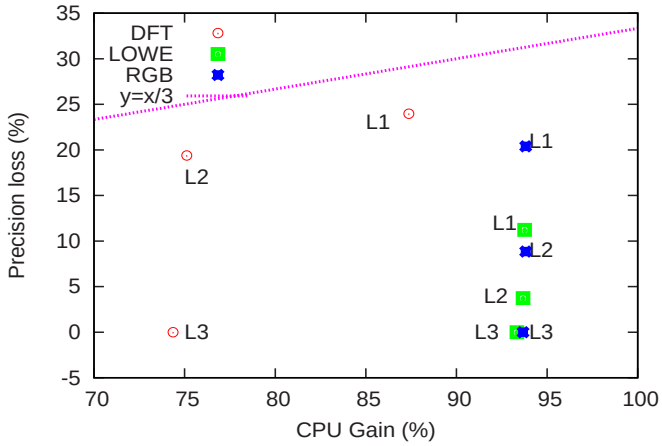


Figure 4.6: *Precision loss vs. CPU gain with all data sets. The gain is obtained w.r.t. sequential scan and is computed as 1-$CPU_{Hierarchy}/CPU_{Sequential}$.*

Additionally, we published a survey on the state-of-the-art approaches that enable progressive image retrieval in [Bouteldja et al., 2008c].

4.4.5 Main characteristics of HiPeR according to literature

HiPeR was instantiated and evaluated on VA-hierarchies and GC*hierarchies, and these implementations outperform the simple indexes VA-File and GC*tree, for the exact and approximate retrieval scenarios. We think that comparable improvements would be obtained by integrating into HiPeR any index structure that performs similarity search in multidimensional vector spaces. To do this, only few constraints are required: as stated in section 4.4.1, we only suppose that the chosen index enables a one-to-one mapping between the features describing entity E in index at level i and the ones describing it in index level $i+1$; such a mapping was easily implemented with VA-File and GC*tree. The algorithms and models of precision defined for approximate similarity search impose to use an index that enables approximate similarity search with a probabilistic control of the precision. The global model of precision loss proposed at equation 4.7 for the whole hierarchy is entirely independent of the index implemented.

According to the classification of algorithms for approximate retrieval of [Patella and Ciaccia, 2008], HiPeR has the followings characteristics: it was designed for vector spaces (**VS**). In the current implementations, it is $\mathbf{VS}_{L_p}$ because the families of mapping used for building the intermediate feature spaces are proposed for any of the L_p distances. Accelerating retrieval is done by changing the original space (**CS**). The current application to VA-File with algorithm VA-BND also puts HiPeR in category $\mathbf{RC}_{AP}$ where the approaches reduce the number of compared objects by pruning some parts of space. The model of precision loss defined for HiPeR is probabilistic and non parametric ($\mathbf{PG}_{npar}$) because few assumptions are made on the data distribution. Finally, HiPeR is an interactive approach (**IA**) in the sense that quality of answers is selected by user at query time.

Chapter 5

Conclusions, perspectives and new directions

This final chapter starts from the results obtained during the activities presented in chapters 2, 3 and 4, and gives the main perspectives that I consider in the continuation of these activities. In addition, I also present the current new work started within the scope of the federation Wisdom (PPF, 2007-2010) which gathers three parisian research groups (from laboratories CNAM/CEDRIC, Paris-Dauphine/LAMSADE and Paris-6/LIP6) that share common research interests in advanced databases. This new research direction concerns the description of the spatial relationships between objects in images. Part of these activities are supported by two French projects: DISCO (ANR MDCO, 2008-2010) whose main objective is to design and experiment generic and flexible techniques for content-based indexing and searching, dedicated to centralized and distributed sources of multimedia documents; and the project "Paris en images" funded by the "Mairie de Paris" (2007-2009) whose main objective is to propose indexing techniques to manage the visual contents of the different libraries of photographs and videos held by the city council.

Contents

All the methods proposed since 2001 and presented in the previous chapters raise some specific problems and require improvements that can be achieved in the short term. To have an idea of them, the reader can consult the conclusions and perspectives of the associated publications (see section 1.2.3 of chapter 1). The following sections rather focus on the new challenges I begin to investigate, in the continuation of my previous work on visual content description and on multidimensional access methods. Section 5.1 describes the new research topics I have just began to study, while section 5.2 presents some perspectives on my future research.

5.1 Ongoing new research

5.1.1 Multidimensional indexing of various media descriptors

The work I performed on multidimensional access methods and presented in chapter 4 was evaluated with image and video descriptors, but can apply with other media involving multidimensional features, as probably a large number of other access methods. Because non textual data (audio, image, video) are respectively 1D, 2D and 3D signals, the associated content descriptors involve high-dimensional feature spaces that probably share common characteristics. Consequently, one objective of my ongoing research activity concerning multidimensional indexing structures is to study how factorizing as largely as possible techniques applicable to all the types of multimedia documents. I address this problem within the scope of the DISCO project, in collaboration with the LAMSADE and also with the French institute IRCAM ("Institut de Recherche et Coordination Acoustique - Musique") that conducts multi-disciplinary research in the field of sciences and technologies of music and sound.

One of the key goal of this activity is the development of a framework which will allow a description of the audio media compatible with the way still image and video are described. A large part of the work will consist in finding a common formalism for the different types of media description, by proposing a typology for audio, image and video feature spaces. This study aspires to facilitate the use and development of future index structures for rapid access to data, the underlying objective being to mutualize the work for several modalities. I started to investigate this work with Geoffroy Peeters, researcher at IRCAM and Stanislav Barton, postdoc enrolled to work on these aspects within the DISCO project.

A second objective deals with the evaluation of index structures for large collections of multimedia signatures. The aim is to propose a framework for the evaluation of state-of-the-art index structures according to the proposed typology of descriptors. Such a study will also conduct to the definition of criteria allowing the dynamic selection of the most appropriate indexing technique according to a given descriptor, to a given query type as well as to the potential combination of several modalities available to build the query. This part of the work is done in collaboration with Stanislav Barton, Maude Manouvrier and Marta Rukoz of the LAMSADE laboratory (Wisdom partner).

5.1.2 The spatial layout of image contents

My past activities were focussed on the description of visual contents without considering spatial relationships models. During my postdoc at INRIA, I investigated the description of the spatial arrangement of group of interest points (see section 2.1.2 in chapter 2), but this work was dedicated to local descriptors and dealt with *ad hoc* methods. The spatial relationships existing between objects of interest in an image represent a very distinctive information that may improve CBIR systems, and a lot of challenging research remains to do on this topic. In particular, literature is very rich - see the survey we published in [Gouet-Brunet et al., 2008] - but many problems remain, among them: a large number of the existing approaches do not address concrete scenarios of retrieval relevant for actual real-world CBIR applications; many of the proposed models and associated similarity measures do not fit well the end-user expectations; retrieval time and real-time applications are rarely considered, while the description of the spatial arrangement of entities is often a problem having a high complexity; no rigorous evaluation exists, because most of the time the approaches are evaluated on different data sets and for different scenarios.

I started to investigate the problem of spatial layout description since 2007 in collaboration with Maude Manouvrier and Marta Rukoz of the LAMSADE laboratory of Paris-Dauphine University, within the scope of the federation Wisdom. In addition to the survey we published together in [Gouet-Brunet et al., 2008], we co-supervised the Master thesis of Mohamed Kechaou (Master ISI, Paris-Dauphine University, Wisdom funding) who studied the family of approaches describing the spatial layout of images based on 2D strings, and implemented the Z-string model. I have focussed more deeply on this activity during my sabbatical leave ("CRCT", February - July 2008). During this period, we have co-supervised the Master thesis of Nguyen Vu Hoang (Master ISI, Paris-Dauphine University, Wisdom funding), who proposed a new approach of spatial layout representation, called δ-TSR, which is an amelioration of the state-of-the-art approach TSR. In addition, we have proposed to investigate two axes of research on this topic, concretized by the hiring of two PhD students who are beginning their thesis; they are described in the two following sections. It is important to note that these activities rely on the design of image analysis tools for representing the spatial layout of images as well as on the design of efficient multidimensional access methods for retrieval of images according to the chosen representation.

5.1.2.1 Hierarchical description of the spatial layout of image contents

When considering the description of the spatial layout in images, one major difficulty remains in the comparison of two images described with these relations, most of the time represented with graphs involving a NP-hard comparison problem. We have chosen to study solutions to this problem by considering *hierarchies of spatial relationships*. We think that employing such a structuring to describe the spatial relationships in images should allow (1) the proposal of a new relevant approach of image content description and (2) the use of this description within the scope of the progressive retrieval paradigm aiming at accelerating retrieval in large volumes of data. More precisely, we focus on:

Hierarchical structuring of spatial layout for image matching. By considering such a structuring, our objective is triple: first, we want to give flexibility to the description according to the scenario of search and the user's interest, i.e. to be able to address several levels of semantic description online (e.g. a painting representing "a person at the left of a vase" versus "a person wearing a hat at the left of a vase decorated with birds"); second, we want to be able to characterize the spatial layout at several levels of descriptors, from high-level objects of interest (e.g. a car) to low-level objects of interest (e.g. interest points inside object "car"); and third, we want to exploit the hierarchy to accelerate image matching.

Progressive retrieval in large collections of images. The multi-level description previously described gives the possibility to deal with progressive retrieval, i.e. to return responses progressively from images respecting the spatial relationships coarsely to finer responses. Here, work will focus on the development of an efficient multidimensional index structure that embeds this description, as well as on the definition of a similarity measure updated online as iterations of retrieval proceed. Part of this work is the continuation of the PhD thesis of Nouha Bouteldja (see section 4.4.4 on progressive retrieval in chapter 4).

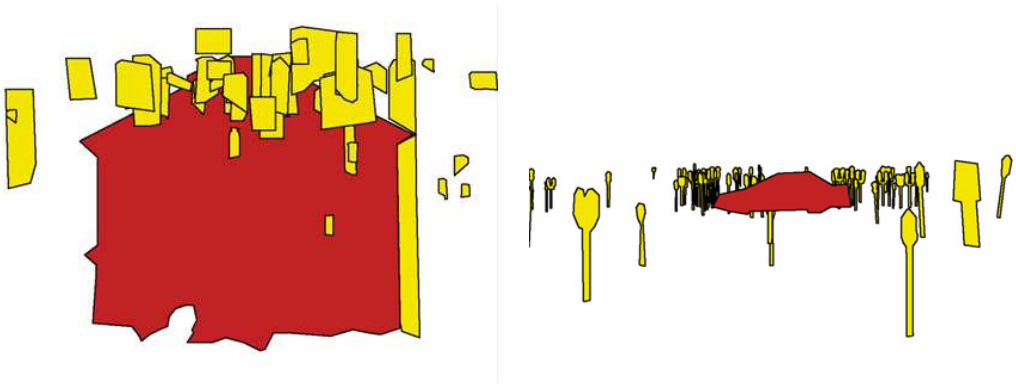
My past research on image descriptors and indexing structures, coupled with the ones of Maude and Marta particularly on quadtree-based image representation [Manouvrier et al., 2005] and on similarity measures for hierarchical representations [Rukoz et al., 2006] will allow addressing these two problems. This activity is performed within the scope of the PhD thesis of Radhwane Kissi (2008-2011) co-supervised by Maude, Marta and myself. According to my success to the "habilitation", I will be his thesis director. This activity is co-supported by two projects: the French projects DISCO where we apply it to image and video contents of online audiovisual collections/archives for consumers and professionals, provided by partners RMN (photo agency "Réunion des Musées Nationaux") and EWA (European Web Archives, a digital library of cultural contents on the Web); the project "Paris en images" where we consider the different libraries of photographs and videos held by the city council of Paris (artistic pictures, monuments pictures, pictures/videos related to the Mayor activities, to particular events, etc.).

5.1.2.2 Spatial context described with spatial relationships models

As stated in [Oliva and Torralba, 2007] and in other previous works, objects never occur in isolation in the real world. In general, they co-vary with other objects and particular environments, as illustrated with the two examples of figure 5.1.

These relationships provide a rich source of contextual associations to be exploited to improve object recognition or retrieval tasks. Recent work has shown that a statistical summary of the scene provides a complementary and effective source of information for contextual inference, which enables humans to quickly guide their attention and eyes to regions of interest in natural scenes [Rutishauser et al., 2004; Murphy et al., 2005; Zhang et al., 2006; Bonaiuto and Itti, 2006; Torralba et al., 2006].

Alternately, we think that another interesting and natural direction to address the problem of a priori object localization according to a spatial context consists in representing this context in terms of the spatial relationships that connect the object of interest with



House and chimney Car and parking meter

Figure 5.1: *Measure of the spatial dependencies existing among objects. These examples are extracted from article [Oliva and Torralba, 2007], they illustrate the dependencies between objects, by plotting the distribution of locations, sizes and shapes of a target object (shown in yellow) conditional on the presence of a reference object (shown in red). The objects belong to the data set of annotated objects LabelMe [Russell et al., 2008].*

other objects. The interesting challenge we propose to study is the proposal of a new paradigm of interrogation: given a query defined by an object plus a spatial relationship, we want to retrieve the images and define all the areas that may contain objects respecting the query relationship with the query object; here I present the problem in terms of binary relationships but it holds with relationships between tuples of objects. This problem is original because all the state-of-the-art approaches describing the spatial layout of images [Gouet-Brunet et al., 2008] only address the problem of similarity search via the query by example paradigm based on: either a query image (retrieval of images), or an image part query (retrieval of sub-images), or an object and a given relationship as query (retrieval of objects), or two objects as queries (retrieval of relationships). None of them address the problem of object recognition by considering spatial relationships to estimate the gist of the image for a given query. We are convinced that the studies of an appropriate model of representation of the spatial layout in images and of a dedicated index structure for this model that supports this new kind of query will provide an original solution to address the problem of the gist estimation for object retrieval, recognition or detection purposes. The proposed solutions will be evaluated with the aim of reducing the rate of false alarms observed during search as well as reducing the retrieval time of objects. According to the situation, the query relationship will be defined by an expert or, more interestingly, learnt from examples.

Since march 2008, I have co-supervised the Master thesis of Nguyen Vu Hoang (Master ISI, Paris-Dauphine University, Wisdom funding) with Maude and Marta on this topic. Because ranked first in the master, Nguyen obtained a PhD funding from the government (MENRT funding), leading to the beginning of his PhD thesis on the problem of the definition of context with spatial relationships, under the supervision of Maude, Marta and myself (2008-2011, Marta Rukoz thesis director). At present, we are studying application scenarios for this problem, that motivated my visit during three weeks in August 2008 to the LIS (Laboratory of Information Systems) of the UNICAMP University (São Paulo

state, Brazil), to set up a collaboration with Claudia Bauzer Medeiros and Ricardo Torres on the description of images with spatial relationships dedicated to remote sensing images (agriculture and biodiversity).

5.2 Perspectives: human-centered combination of heterogeneous visual structures

A large part of my research activities on visual content description was dedicated to the combination of heterogeneous visual features (see sections 2.2 and 2.3 in chapter 2). The results obtained were satisfactory because they allowed improving rates of object recognition and of copy detection, but we can do better. Indeed, the combinations of feature chosen were defined a priori according to some criteria such as their pre-supposed complementarity: by definition, Harris points and symmetry centers do not describe the same sites of the image, while interest points and active contours do not represent the same information of the object visual appearance. We could consider more combinations of features to improve the richness of the description, but rapidly, we will need to define criteria that allow finding the best configuration for a given scenario or application. Because the knowledge of the behavior and needs of a human user are fundamental to define an efficient CBIR system [Jaimes and Sebe, 2007], one relevant criterion is based on the human visual perception. To exploit this information, my objective is to focus on two different directions:

Human-centered evaluation of local features. Most of the local detectors proposed in literature of Computer Vision were mainly concerned with the accuracy of point localization and repeatability of the detector. It is known that in pre-attentive vision, human explores specific salient areas in the image. How the classical detectors of interest points are in concordance with such areas? In contrast, other approaches, coming from psychology and neurosciences, were defined w.r.t. to the processes of the human brain, conducing to numerous models of human visual attention or saliency [Tuytelaars and Mikolajczyk, 2008]. However, the vast majority were only of theoretical interest and only few were implemented and tested on real images. Can these approaches be employed in a real-world CBIR system? Evaluating the relevance of these features according to the human visual attention is fundamental to build a CBIR system that exploits the power of local features optimally. My future research on point detectors and local descriptors will focus on these aspects.

Optimal combination of visual features. Exploiting different categories of visual (local or global) features simultaneously to improve visual content description raises several questions to better satisfy users' expectations: which categories of features should be used jointly? What is the minimal and optimal set of features to combine? What is the best similarity measure associated to this optimal set? Among all the techniques that deal with information fusion, I am interested in genetic programming, because such a framework allows addressing the combination of different features as well as the design of combined (linear and non linear) similarity functions in an uniform formalism. My visit to the LIS (Laboratory of Information Systems) of the UNICAMP University (São Paulo state, Brazil) allowed me considering a collaboration with Ricardo Torres on this aspect [da S. Torres et al., 2008].

References

Agarwal, S., Awan, A., and Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. IEEE *Transaction on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490.

Amores, J., Sebe, N., and Radeva, P. (2005). Fast spatial pattern discovery integrating boosting with constellations of contextual descriptors. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 769–774.

Amsaleg, L., Gros, P., and Berrani, S. (2004). Robust object recognition in images and the related database problems. *Multimedia Tools and Applications*, V23(3):221–235.

Andoni, A. and Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 459–468.

Beckmann, N., Kriegel, H.-P., Schneider, R., and Seeger, B. (1990). The R*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 322–331, Atlantic City, NJ. ACM Press.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press.

Berchtold, S., Böhm, C., and Kriegel, H. (1998). The pyramid-technique: towards breaking the curse of dimensionality. In *Proceedings of the ACM SIGMOD*, pages 142–153. ACM Press.

Berchtold, S., Keim, D. A., and peter Kriegel, H. (1996). The X-tree: An index structure for high-dimensional data. In *Proceedings of th 22nd International Conference on Very Large Databases (VLDB)*, pages 28–39, Mumbai (Bombay), India.

Berrani, S., Amsaleg, L., and Gros, P. (2003). Approximate searches: k-neighbors + precision. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 24–31, New Orleans, USA.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? *Lecture Notes in Computer Science*, 1540:217–235.

Bhat, D. and Nayar, S. (1998). Ordinal measures for image correspondence. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 20(4):415–423.

Böhm, C., Berchtold, S., and Keim, D. (2001). Searching in high-dimensional spaces - index structures for improving the performance of multimedia databases. *ACM Computing Survey*, 33(3):322–373.

Bonaiuto, J. and Itti, L. (2006). The use of attention and spatial information for rapid facial recognition in video. *Image and Vision Computing*, 24(5):557–563.

Borenstein, E. and Ullman, S. (2002). Class-specific, top-down segmentation. In *European Conference on Computer Vision*, pages 109–124.

Borenstein, E. and Ullman, S. (2004). Learning to segment. In *European Conference on Computer Vision*, pages 315–328.

Boujemaa, N., Fauqueur, J., and Gouet, V. (2003). What's beyond query by example? In *IAPR International Conference on Image and Signal Processing (ICISP'2003)*, Agadir, Morocco.

Boujemaa, N., Fauqueur, J., and Gouet, V. (2004a). *Trends and Advances in Content-Based Image and Video Retrieval*, book chapter What's beyond query by example? L. Shapiro, H.P. Kriegel, R. Veltkamp (eds.), LNCS, Springer Verlag.

Boujemaa, N., Fleuret, F., Gouet, V., and Sahbi, H. (2004b). Automatic textual annotation of video news based on semantic visual object extraction. In *IS&T/SPIE Conference on Storage and Retrieval Methods and Applications for Multimedia*, San Jose CA, USA.

Bouteldja, N. and Gouet-Brunet, V. (2008). Exact and progressive image retrieval with the HiPeR framework. In *IEEE International Conference on Multimedia & Expo (ICME'08)*, pages 1257–1260, Hannover, Germany.

Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2006a). Back to the curse of dimensionality with local image descriptors. Research report 1049, CNAM/CEDRIC.

Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2006b). Evaluation of strategies for multiple sphere queries with local image descriptors. In *IS&T/SPIE Conference on Multimedia Content Analysis, Management and Retrieval*, volume 6073, pages A1–12, San Jose CA, USA.

Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2008a). HiPeR: Hierarchical progressive exact retrieval in multidimensional spaces. In *International Workshop on Similarity Search and Applications (SISAP'08, in conjunction with ICDE'08)*, pages 25–34, Cancún, Mexico.

Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2008b). HiPeR: Hierarchies for approximate and exact retrieval in multidimensional spaces. In *Journées Bases de Données avancées (BDA'08)*, Guilherand-Granges, France. To appear.

Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2008c). The many facets of progressive retrieval for CBIR. In *Pacific-Rim Conference on Multimedia (PCM'08)*, Tainan, Taiwan. To appear.

Bouteldja, N., Gouet-Brunet, V., and Scholl, M. (2009). Approximate retrieval with HiPeR: Application to VA-Hierarchies. In *ACM International Multimedia Modeling Conference (MMM'09)*, Sophia-Antipolis, France. To appear.

Braunmüller, B., Ester, M., Kriegel, H.-P., and Sander, J. (2000). Efficiently supporting multiple similarity queries for mining in metric databases. In *ICDE*, pages 256–267.

Cao, L. and Fei-Fei, L. (2007). Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *IEEE 11th International Conference on Computer Vision*, pages 1–8.

Cha, G. and Chung, C. (2002). The GC-tree: a high-dimensional index structure for similarity search in image databases. *IEEE Transactions on Multimedia*, 4(2):235–247.

Chen, L. and Stentiford, F. (2008). Video sequence matching based on temporal ordinal measurement. *Pattern Recognition Letters, in press.*

Cheung, S. and Zakhor, A. (2003). Efficient video similarity measurement with video signature. *IEEE Transactions on Circuits Systems Video Techniques*, 13(1):59–74.

Ciaccia, P. and Patella, M. (2000). PAC nearest neighbor queries: Approximate and controlled search in high-dimensional and metric spaces. In *Proceedings of the 16th International Conference on Data Engineering*, pages 244–255, San Diego, California.

Ciaccia, P., Patella, M., and Zezula, P. (1997). M-tree: An efficient access method for similarity search in metric spaces. In *Proceedings of 23rd International Conference on Very Large Data Bases*, pages 426–435, Athens, Greece.

Coskun, B., Sankur, B., and Memon, N. (2006). Spatio-temporal transform-based video hashing. IEEE *Transactions on Multimedia*, 8(6):1190–1208.

da S. Torres, R., Falcão, A. X., Goncalves, M. A., Papa, J. P., Zhang, B., Fan, W., and Fox, E. A. (2008). A genetic programming framework for content-based image retrieval. *Pattern Recognition.* In press.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2).

De Roover, C., De Vleeschouwer, C., Lefèbvre, F., and Macq, B. (2005). Robust video hashing based on radial projections of key frames. *IEEE Transactions on Signal Processing, Supplement on Secure Media*, 10(53):4020–4037.

Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72.

Dorkó, G. and Schmid, C. (2003). Selection of scale-invariant parts for object class recognition. In *IEEE International Conference on Computer Vision.*

Everingham, M., Zisserman, A., Williams, C., Gool, L. V., Allan, M., Bishop, C., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Shawe-Taylor, J., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., and Zhang, J. (2006). The 2005 PASCAL visual object classes challenge. In *Selected Proceedings of the First PASCAL Challenges Workshop, LNAI, Springer-Verlag.*

Fauqueur, J. and Boujemaa, N. (2006). Mental image search by boolean composition of region categories. *Multimedia Tools and Applications*, 31(1):95–117.

Fergus, R., Perona, P., and Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271.

Ferman, A., Tekalp, A., and Mehrotra, R. (2002). Robust color histogram descriptors for video segment retrieval and identification. *IEEE Trans. Image Processing*, 11(5):497–508.

Ferrari, V., Tuytelaars, T., and Van Gool, L. (2006). Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, 67(2):159–188.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395.

Frigui, H. and Krishnapuram, R. (1998). A robust competitive clustering algorithm with applications in computer vision. IEEE *Transaction on Pattern Analysis and Machine Intelligence*, 21(5):450–465.

Fussenegger, M., Opelt, A., and Pinz, A. (2006). Object localization/segmentation using shape priors. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Geusebroek, J. M., Burghouts, G. J., and Smeulders, A. W. M. (2005). The Amsterdam library of object images. *Int. J. Comput. Vision*, 61(1):103–112.

Gouet, V. and Boujemaa, N. (2001). Object-based queries using color points of interest. In *IEEE Workshop on Content-Based Access of Image and Video Libraries (CVPR/CBAIVL 2001)*, pages 30–36, Kauai, Hawaii, USA.

Gouet, V. and Boujemaa, N. (2002). On the robustness of color points of interest for image retrieval. In IEEE *International Conference on Image Processing (ICIP'02)*, pages 377–380, Rochester, New York, USA.

Gouet, V. and Lameyre, B. (2004). SAP: a robust approach to track objects in video streams with snakes and points. In *British Machine Vision Conference (BMVC'04)*, pages 737–746, Kingston University, London, UK.

Gouet-Brunet, V. (2006). *Encyclopédie de l'Informatique et des systèmes d'information*, book chapter Recherche par contenu visuel dans les grandes collections d'images. J. Akoka and I. Commyn-Wattiau (eds.), Vuibert.

Gouet-Brunet, V. (2008a). *Encyclopedia of Database Systems: Multimedia Databases*, book chapter Image. L. Liu and T. Özsu (eds.), Springer Verlag.

Gouet-Brunet, V. (2008b). *Encyclopedia of Database Systems: Multimedia Databases*, book chapter Image representation. L. Liu and T. Özsu (eds.), Springer Verlag.

Gouet-Brunet, V. and Lameyre, B. (2008). Object recognition and segmentation in videos by connecting heterogeneous visual features. *Computer Vision and Image Understanding (CVIU)*, 111(1):86–109.

Gouet-Brunet, V., Manouvrier, M., and Rukoz, M. (2008). Synthèse sur les modèles de représentation des relations spatiales dans les images symboliques. *Revue des Nouvelles Technologies de l'Information (RNTI)*.

Grabner, M. and Bischof, H. (2005). Extracting object representation from local feature trajectories. In *1st Cognitive Vision Workshop*.

Hampapur, A. and Bolle, R. (2002). Comparison of sequence matching techniques for video copy detection. In *Conference on Storage and Retrieval for Media Databases*, pages 194–201.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. *Proceedings of the* 4^{th} *Alvey Vision Conference*, pages 147–151.

Heidemann, G. (2004). Focus-of-attention from local color symmetries. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 26(7):817–830.

Hoad, T. and Zobel, J. (2003). Video similarity detection for digital rights management. In *26th Australasian Computer Science Conference (ACSC'03)*, pages 237–245, Darlinghurst, Australia.

Hua, X.-S., Chen, X., and Zhang, H.-J. (2004). Robust video signature based on ordinal measure. In *International Conference on Image Processing*.

Imedia (2004). Images and multimedia : Indexing, retrieval and navigation. Research project activity report, INRIA Rocquencourt.

Jaimes, A. and Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*, 108(1-2):116–134.

Joly, A. (2007). New local descriptors based on dissociated dipoles. In *ACM International Conference on Image and Video Retrieval*, pages 573–580, Amsterdam, The Netherlands.

Joly, A., Buisson, O., and Frélicot, C. (2007). Content-based copy detection using distortion-based probabilistic similarity search. *IEEE Transactions on Multimedia*, 9(2):293–306.

Jomier, G., Manouvrier, M., Oria, V., and Rukoz, M. (2005). Multilevel index for global and partial content-based image retrieval. In *EMMA*, pages 66–75.

Jurie, F. and Schmid, C. (2004). Scale-invariant shape features for recognition of object categories. In *IEEE International Conference on Computer Vision and Pattern Recognition*.

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contours models. *International Journal of Computer Vision*, pages 321–331.

Katayama, N. and Satoh, S. (1997). The SR-tree: An index structure for high-dimensional nearest neighbor queries. In *Proceedings of the ACM SIGMOD*, pages 369–380.

Ke, Y., Sukthankar, R., and Huston, L. (2004). An efficient parts-based near-duplicate and sub-image retrieval system. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 869–876, New York, NY, USA.

Kim, C. and Vasudev, B. (2005). Spatiotemporal sequence matching techniques for video copy detection. IEEE *Transactions on Circuits and Systems for Video Technology*, 1(15):127–132.

Kleinberg, J. (1997). Two algorithms for nearest-neighbor search in high dimensions. In *Annual ACM Symposium on Theory of Computing*, pages 599–608, El Paso, Texas.

Kumar, M. P., Torr, P. H. S., and Zisserman, A. (2005). Obj cut. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 18–25, Washington, DC, USA. IEEE Computer Society.

Lameyre, B. and Gouet, V. (2004). Object tracking and identification in video streams with snakes and points. In *5th Pacific-Rim Conference on Multimedia (PCM'04), LNCS 3333 Springer-Verlag*, pages 61–68, Tokyo, Japan.

Lameyre, B. and Gouet-Brunet, V. (2006a). Connecting local and global descriptors for generic object recognition in videos. In *6th IEEE International Workshop on Visual Surveillance (VS'06, in conjunction with ECCV'06*, pages 57–64, Graz, Austria.

Lameyre, B. and Gouet-Brunet, V. (2006b). Connexions entre descripteurs locaux et globaux pour la reconnaissance d'objets dans les vidéos. In *Compression et Représentation des Signaux Audiovisuels (Coresa'06)*, pages 207–212, Caen, France.

Landré, J., Truchetet, F., S., M., and David, B. (2001). Automatic building of a visual interface for content-based multiresolution retrieval of paleontology images. *Journal of Electronic Imaging*, 10(4).

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *International Conference on Computer Vision*, pages 432–439, Nice, France.

Larlus, D. and Jurie, F. (2008). Combining appearance models and markov random fields for category level object segmentation. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 1–7.

Law-To, J. (2007). *From genericity to distinctiveness of video content description, application to Video Copy Detection, defense December 14, 2007*. PhD thesis, Versailles Saint-Quentin University, INRIA Rocquencourt, Imedia research group and INA.

Law-To, J., Buisson, O., Gouet-Brunet, V., and Boujemaa, N. (2006a). Robust voting algorithm based on labels of behavior for video copy detection. In *14th ACM International Conference on Multimedia (ACM Multimedia'06)*, pages 835–844, Santa Barbara, USA.

Law-To, J., Chen, L., Joly, A., Laptev, Y., Buisson, O., Gouet-Brunet, V., Boujemaa, N., and Stentiford, F. (2007a). Video copy detection: a comparative study. In *ACM International Conference on Image and Video Retrieval (CIVR'07)*, pages 371–378, Amsterdam, The Netherlands.

Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2006b). Labeling complementary local descriptors behavior for video copy detection. In *IAPR and EURASIP International Workshop on Multimedia Concept Representation, Classification and Security (MCRCS'06)*, pages 290–297.

Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2006c). Labellisation du comportement de descripteurs locaux pour la détection de copies vidéo. In *Compression et Représentation des Signaux Audiovisuels (Coresa'06)*, pages 336–341, Caen, France.

Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2006d). Local behaviour labelling for content-based video copy detection. In *18th IAPR International Conference on Pattern Recognition (ICPR'06)*, pages 232–235.

Law-To, J., Gouet-Brunet, V., Buisson, O., and Boujemaa, N. (2007b). Video copy detection on the internet: the challenges of copyright and multiplicity. In *IEEE International Conference on Multimedia & Expo (ICME'07)*, pages 2082 – 2085, Beijing.

Le Saux, B. and Boujemaa, N. (2002). Unsupervised robust clustering for image database categorization. In *IEEE-IAPR International Conference on Pattern Recognition*, pages 259– 262, Quebec, Canada.

Leibe, B., Leonardis, A., and Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *ECCV'04 Worskhop on Statistical Learning in Computer Vision*, pages 17–32, Prague, Czech Republic.

Lejsek, H., H.Ásmundsson, F., Thór-Jónsson, B., and Amsaleg, L. (2006). Blazingly fast image copyright enforcement. In *14th ACM International Conference on Multimedia*, Santa Barbara, USA.

Levin, A. and Weiss, Y. (2006). Learning to combine bottom-up and top-down segmentation. In *European Conference on Computer Vision*, pages 581–594.

Li, Y., Jin, L., and Zhou, X. (2005). Video matching using binary signature. In *Int. Symposium on Intelligent Signal Processing and Communication Systems*, pages 317–320.

Lin, C.-C. and Lin, W.-C. (1996). Extracting facial features by an inhibitory mechanism based on gradient distributions. In *Pattern Recognition*, volume 29.

Lin, E. T., Eskicioglu, A. M., Lagendijk, R. L., and Delp, E. J. (2005). Advances in digital video content protection. In *Proceedings of the IEEE*, pages 171–183.

Lisin, D., Mattar, M., and Blashcko, M. (2005a). Combining local and global image features for object class recognition. In *IEEE Workshop on Learning in Computer Vision and Pattern Recognition (in conjunction with CVPR)*, pages 47–54, San Diego, California.

Lisin, D. A., Mattar, M. A., Blaschko, M. B., Benfield, M. C., and Learned-Miller, E. G. (2005b). Combining local and global image features for object class recognition. In *CVPR*.

Locher, P. and Nodine, C. (1987). *Symmetry Catches the Eye*. Eye Movements: from Physiology to Cognition, J. O'Regan and A. Levy-Schoen. Elsevier Science Publishers B.V.

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

Loy, G. and Zelinsky, A. (2003). Fast radial symmetry for detecting points of interest. IEEE *Transactions on Pattern Analysis and Machine Intelligence*.

Manjunath, B., Salembier, P., and Sikora, T. (2002). *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley & Sons.

Manouvrier, M., Rukoz, M., and Jomier, G. (2005). *Spatial Databases: Technologies, Techniques and Trends*, chapter Chapter IV: Quadtree-Based Image Representation and Retrieval, pages 81–106. Idea Group Publishing.

Marie-Julie, J. and Essafi, H. (1998). Using ifs and moments to build a quasi invariant image index. In *European Conference on Computer Vision*.

Martinez, M. (2007). Une plateforme d'évaluation des strutures d'index pour la recherche d'images par contenu visuel. Mémoire d'ingénieur CNAM, Conservatoire National des Arts et Métiers (CNAM).

Massoudi, A., Lefebvre, F., Demarty, C.-H., Oisel, L., and Chupeau, B. (2006). A video fingerprint based on visual digest and local fingerprints. In *International Conference on Image Processing*, pages 2297–2300.

Mohan, R. (1998). Video sequence matching. In *Int. Conference on Audio, Speech and Signal Processing*.

Montesinos, P., Gouet, V., Deriche, R., and Pelé, D. (2000). Matching color uncalibrated images using differential invariants. *Image and Vision Computing*, 18(9):659–672.

Mortensen, E., Deng, H., and Shapiro, L. (2005). A SIFT descriptor with global context. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 184–190.

Murphy, K., Torralba, A., Eaton, D., and Freeman, W. (2005). Object detection and localization using local and global features. In *Sicily Workshop on Object Recognition*.

Naphade, M., Yeung, M., and Yeo, B. (2000). A novel scheme for fast and efficient video sequence matching using compact signatures. In *SPIE Storage and retrieval for media databases*, pages 564–572.

Nene, S. A., Nayar, S. K., and Murase, H. (1996). Columbia object image library (COIL-100). Technical report, Technical Report CUCS-006-96, Columbia University, New York.

Oliva, A. and Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527.

Oostveen, J., Kalker, T., and Haitsma, J. (2001). Visual hashing of digital video: Applications and techniques. In *SPIE Applications of Digital Image Processing XXIV*.

Opelt, A., Pinz, A., Fussenegger, M., and Auer, P. (2006). Generic object recognition with boosting. IEEE *Transaction on Pattern Analysis and Machine Intelligence*, 28(3):416–431.

Opelt, A., Sivic, J., and Pinz, A. (2005). Generic object recognition from video data. In *1st Cognitive Vision Workshop*.

Patella, M. and Ciaccia, P. (2008). The many facets of approximate similarity search. In *International Workshop on Similarity Search and Applications (in conjunction with ICDE'08)*, pages 10–21, Cancún, Mexico.

Poullot, S., Buisson, O., and Crucianu, M. (2007). Z-grid-based probabilistic retrieval for scaling up content-based copy detection. In *ACM International Conference on Image and Video Retrieval*, pages 348–355, Amsterdam.

Privitera, C. M. and Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982.

Reisfeld, D., Wolfson, H., and Yeshurun, Y. (1994). Context free attentional operators: the generalized symmetry transform. *International Journal of Computer Vision, Special Issue on Qualitative Vision*.

Rukoz, M., Manouvier, M., and Jomier, G. (2006). δ-distance: A family of dissimilarity metrics between images represented by multi-level feature vectors. *Information Retrieval*, 9(6):633–655.

Russell, B., Torralba, A., Murphy, K., and Freeman, W. T. (2008). LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173.

Russell, B. C., Efros, A. A., Sivic, J., Freeman, W. T., and Zisserman, A. (2006). Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of Computer Vision and Pattern Recognition*, volume 2, pages 1605–1614.

Rutishauser, U., Walther, D., Koch, C., and Perona, P. (2004). Is bottom-up attention useful for object recognition? In *IEEE International Conference on Pattern Recognition*.

Samet, H. (2006). *Foundations of Multidimensional and Metric Data Structures*. The Morgan Kaufmann Series in Computer Graphics.

Schmid, C. and Mohr, R. (1997). Local grayvalue invariants for image retrieval. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 19(5):530–534.

Sivic, J., Everingham, M., and Zisserman, A. (2005a). Person spotting: video shot retrieval for face sets. In *4th International Conference on Image and Video Retrieval*.

Sivic, J., Russell, B., Efros, A., Zisserman, A., and Freeman, W. (2005b). Discovering objects and their location in images. In *IEEE International Conference on Computer Vision*, pages 370– 377.

Sivic, J., Schaffalitzky, F., and Zisserman, A. (2004). Object level grouping for video shots. In *8th European Conference on Computer Vision*, pages 85–98.

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477.

Sivic, J. and Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 488–495, Washington, DC.

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. M. (2006). Contextual guidance of attention in natural scenes: The role of global features on object search. *Psychological Review*, 113(4):766–786.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280.

Urruty, T., Djeraba, C., and Simovici, D. (2007). Clustering by random projection. In *7th Industrial Conference on Data Mining*, pages 107–119.

Verleysen, M. (2003). *Limitations and future trends in neural computation*, chapter Learning high-dimensional data, pages 141–162. IOS Press.

Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In J. Cabestany, A. Prieto, F. S., editor, *Computational Intelligence and Bioinspired Systems*, Lecture Notes in Computer Science vol. 3512, pages 758–770.

Verleysen, M., François, D., Simon, G., and Wertz, V. (2003). On the effects of dimensionality on data analysis with neural networks. In *International work-conference on artificial and natural neural networks*, pages 105–112, Mao, Spain.

Wallraven, C., Caputo, B., and Graf, A. (2003). Recognition with local features: the kernel recipe. In *International Conference on Computer Vision*, pages 257–264.

Wang, J. Z., Boujemaa, N., Del Bimbo, A., Geman, D., Hauptmann, A., , and Tesic, J. (2006). Diversity in multimedia information retrieval research. In *8th ACM international workshop on Multimedia information retrieval*, pages 5 – 12.

Weber, R. and Böhm, K. (2000). Trading quality for time with nearest neighbor search. In *Proceedings of the 7th International Conference on Extending Database Technology: Advances in Database Technology*, pages 21–35.

Weber, R., Schek, H.-J., and Blott, S. (1998). A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of 24th International Conference on Very Large Data Bases*, pages 194–205, New York City, New York.

Willamowski, J., Arregui, D., Csurka, G., Dance, C., and Fan, L. (2004). Categorizing nine visual classes using local appearance descriptors. In *ICPR Workshop Learning for Adaptable Visual Systems*, Cambridge, United Kingdom.

Winn, J. and Jojic, N. (2005). LOCUS: Learning object classes with unsupervised segmentation. In *IEEE International Conference on Computer Vision*, volume 1, pages 756–763, Beijing.

Yan, S., He, X., Hu, Y., Zhang, H., Li, M., and Cheng, Q. (2004). Bayesian shape localization for face recognition using global and local textures. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):102–113.

Yang, X., Sun, Q., and Tian, Q. (2003). Content-based video identification: a survey. In *International Conference on Information Technology: Research and Education*, pages 50–54.

Zhang, D. and Lu., G. (2001). A comparative study on shape retrieval using fourier descriptors with different shape signatures. In *International Conference on Intelligent Multimedia and Distance Education*, pages 1–9, Fargo, ND, USA.

Zhang, H., Jia, W., He, X., and Wu, Q. (2006). Learning-based license plate detection using global and local features. In *IEEE International Conference on Pattern Recognition*.