INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team TEXMEX

Efficient Exploitation of Multimedia Documents: exploring, indexing and searching in very large databases

Rennes

THEME 3A

Table of contents

1. Team **1**
2. Overall Objectives **1**
2.1.1. Our first field of competence 2
2.1.2. Our second field of competence 2
2.1.3. First topic of research: searching in large image bases. 2
2.1.4. Second topic of research: towards more semantic search engines. 3
2.1.5. Third subject of research: multimedia and coupling between media 3
3. Scientific Foundations **3**
3.1. Introduction 3
3.2. Metadata and Document Description 3
3.2.1. Image Description 4
3.2.2. Video Description 5
3.2.3. Text Description 5
3.2.3.1. Acquisition of lexicons based on Rastier's differential semantics. 5
3.2.3.2. Acquisition of elements of Pustejovsky's Generative lexicon. 6
3.2.3.3. Characterization of huge sets of thematically homogeneous texts. 6
3.2.4. Evaluation of Descriptors 6
3.2.5. Metadata 7
3.3. Efficient Exploitation of Descriptors 7
3.3.1. Statistics for Huge Datasets 7
3.3.2. Multidimensional Indexing Techniques 8
3.3.2.1. Traditional Approaches, Cells and Filtering Rules. 8
3.3.2.2. Approximate NN-Searches. 9
4. Application Domains **10**
4.1. Still Image Database Management 10
4.2. Video Database Management 10
4.3. Textual Database Management 11
4.4. Robotics and Visual Servoing 11
5. Software **12**
5.1.1. I-Description : 12
6. New Results **12**
6.1. Image Retrieval in Large Databases 12
6.1.1. Image Description, Compression and Watermarking 12
6.1.2. Combination of Descriptors by Association Rules and Multiple Correspondence Analysis 13
6.1.3. Approximate Searches: k-Neighbors + Precision 13
6.1.4. Coupling Action and Perception by Image Indexing and Visual Servoing 14
6.2. Text Retrieval in Large Databases 15
6.2.1. Natural Language Processing and Machine Learning 15
6.2.2. Visualization and Web Mining 16
6.2.3. Knowledge Extraction and Visualization from Textual Databases 16
6.3. Multimedia Document Description 16
6.3.1. Video Description 17
6.3.2. Image and Text Joint Description 18
7. Contracts and Grants with Industry **18**
7.1. Industrial Contracts 18
7.1.1. Contract with Thomson 18

7.1.2. Contract with France Télécom 18
7.1.3. Contract with Socio-logiciel 18
7.2. Contracts in the frame of National Networks of Technological Research 18
7.2.1. PRIAM Médiaworks Project 18
7.2.2. RNRT Diphonet Project: Photo Diffusion on Internet 19
7.2.3. RIAM Annapurna Project 19
7.2.4. RIAM FERIA Project 19
7.3. European Contracts 20
7.3.1. European IST Project BUSMAN : Bringing User Satisfaction to Media Access Networks 20
8. Other Grants and Activities 20
8.1. Regional Contracts 20
8.2. National Contracts 20
8.2.1. ACI Grid GénoGRID 20
8.2.2. Action Bio-info inter-EPST: Parallel and Reconfigurable Architectures for Genomic Data Extraction 20
8.2.3. Inter-EPST Bio-informatic Action Caderige-2: Automatic Document Categorization for the Extraction of Gene Interaction Networks 20
8.2.4. R & D INRIA Action SYNTAX 21
8.2.5. ACI Masse de données MDP2P 21
8.2.6. ACI Masse de données REMIX 21
8.2.7. ACI Jeunes Chercheurs TEXMEX 21
8.2.8. Participation to National Working Groups 21
8.3. International Collaborations 22
8.3.1. ERCIM Working Group on Image Undertanding 22
8.3.2. Collaboration with NII - Japan 22
8.3.3. PAI Jules Verne with University of Reykjavik 22
9. Dissemination 22
9.1. Conference, Workshop and Seminar Organization 22
9.2. Animation of the Scientific Community 23
9.3. Teaching Activities 23
9.4. Participation to Seminars, Workshops, Invitations 23
10. Bibliography 24

1. Team

TEXMEX is a joint project with CNRS and University of Rennes 1. The team has been created on January 1st, 2002 and became an INRIA project on November 1st, 2002.

Head of Project-team

Patrick Gros [research scientist CNRS]

Administrative assistant

Maryse Auffray [AJT INRIA, partial position in the project-team]

Faculty members (University of Rennes 1)

Laure Berti-Équille [associate professor]
Annie Morin [associate professor]
Pascale Sébillot [associate professor]

Faculty member (CNRS)

Laurent Amsaleg [research scientist]

Ph.D. students

Sid-Ahmed Berrani [CIFRE with Thomson]
Nicolas Bonnel [CIFRE with France Télécom R&D]
Vincent Claveau [MJENR grant, teaching assistant since Oct 10th]
Manolis Delakis [MJENR grant, since Oct 1st]
Ewa Kijak [CIFRE with Thomson until Sep 30th]
Anicet Kouomou-Choupo [MJENR grant]
Fabienne Moreau [INRIA grant, since Oct 1st]
Anthony Remazeilles [MJENR grant, also with VISTA]
Hervé Renault [since Oct 1st, 2003]
Mathias Rossignol [MJENR grant]
François Tonnin [INRIA - Britany grant, also with TEMICS]

Project technical staff

David Bonnaud [ANNAPURNA project]
Christophe Garcia [FERIA project, since Feb. 1st until Nov 30th]
Sophie Le Delliou [DIPHONET project, also with TEMICS, since Nov 1st]

2. Overall Objectives

Key words: *document content-based access*, *exploration*, *indexing*, *search*, *databases*, *multimedia*, *natural language processing*, *image recognition*, *machine learning*.

The explosion of the quantity of numerical documents raises the problem of the management of these documents. Beyond the problem of storage, we are interested in problems related with the management of the contents: how to exploit large bases of documents, how to classify them, how to index them to be able to search documents, how to visualize their contents? To solve these problems, we propose a multi-field work gathering within the same team specialists of the various media: image, video, text, and specialists in exploitation techniques of the data and metadata extracted from these data: databases, statistics, information retrieval. Our work is at the intersection of these fields and relates more particularly to 3 points: searching in large image databases, adding semantics to search engines, coupling media for multimedia document description.

The exploitation of the contents of large databases of digital multimedia documents is a problem with multiple facets, and the construction of a system exploiting such a database calls upon many techniques: study and description of documents, organization of the bases, search algorithms, classification, visualization, but also adapted management of the primary and secondary memories, interfaces and interaction with the user.

The five major challenges of the field appear to us to be the following ones:

- it is necessary, first of all, to be able **to treat large sets of documents**: it is important to develop techniques which scale up gracefully with respect to the quantity of documents they handle, and to evaluate their results according to quality and speed;
- the multimedia documents are not a simple juxtaposition of independent media, and it is important **to better exploit the existing links between the various media** present in the same document;
- **multimedia document databases are evolutionary**: the sets of documents evolve, as do the document description techniques and the modes of interrogation, which modifies in turn the way in which the bases are used;
- towards queries of a semantic nature, description techniques have only access to the document syntax; it is thus necessary to find means for **reducing this difference between semantic needs and syntactic description tools**;
- **the user-system interaction** is a central point: the user must be able to translate his needs efficiently and simply but with shades, to guide the system or to evaluate the results; he must be the one who controls the system.

We have adopted a matricial organization. On the one hand, we have competences in two main fields, automatic document description and exploitation of these descriptions, and on the other hand, we defined three transversal topics of research. The main idea is to concentrate on the questions where the team multidisciplinarity appears to be an asset to obtain original results. Most graduate students of the team work at the intersection of two domains and are thus advised by two persons.

2.1.1. Our first field of competence

is thus document description. The documents are generally not exploitable directly for search or indexing tasks: it is necessary to use intermediate descriptions which must carry the maximum information on document semantics, while being also computable automatically. To the documents and their descriptors, one can add metadata, which we define here as all the information (other than the descriptors) which inform, supplement or qualify the data (and the descriptors) to which they are associated.

2.1.2. Our second field of competence

relates to description exploitation. The question is to define the techniques which make it possible to handle and exploit large volumes of data, metadata and descriptors, which have been extracted from the documents: **organization and management of the bases**, logical and temporal consistency, selection and strategies of computation of descriptors and metadata; **statistical techniques** for the exploration of great volumes of data; **indexing techniques** aiming at confining in the smallest possible volume the exploitation of the data and thus avoiding an exhaustive examination whose cost is certainly controlled but crippling; **system problems** related to the physical organization of large volumes of data, like disc access management or cache memory management requiring new techniques which are adapted to the characteristics of the descriptors and to the way of using them.

2.1.3. First topic of research: searching in large image bases.

Going from corpora of a few thousands of images to corpora containing a few millions remains a research challenge today. The solution can come neither from the only descriptors nor from new indexing techniques, but necessitates to take into account all the various components of the system and their articulation. We thus propose to work on:

- data description, especially in the case of compressed or watermarked images,
- indexing and search algorithms,
- database organization and use of the metadata,
- system and hardware support,

and on the coupling between these various techniques to improve the performances of the current systems in terms of speed as well as of quality of recognition.

2.1.4. Second topic of research: towards more semantic search engines.

Search engines are extensively used tools, but they appear to be disappointing most of the time, due to their syntactic approach which is based on keywords. Natural language processing tools could however provide them more semantic capabilities, by allowing word sense disambiguation or the possibility to recognize the various formulations of the same concept. It is thus advisable to combine these two techniques.

This union is, however, not so simple. On the one hand, it requires to provide query extension strategies to search engines and then to translate these extensions in terms of similarity. On the other hand, natural language processing tools must work in much broader environments than the ones in which they are usually used. The contribution of such a modification of the engines must also be established, which requires a precise evaluation of the obtained results.

2.1.5. Third subject of research: multimedia and coupling between media

Studying media coupling is undertaken in two manners. Within the framework of video, we are interested in descriptions which jointly use the sound and image tracks of the video. Such techniques can be applied to automatic video structuring, but also to improve people detection and recognition techniques, whether by their face or their voice.

In addition, we study the coupling between text and image in the documents where these two media are strongly coupled, a common case in scientific bibliographical databases, on the web, in newspapers, in art books or technical documents. The goal is to connect, in the same document, the image and the text which is referred to it. This should make it possible to obtain an automatic and semantic description of the images, to connect different documents, either by the search for images visually similar, or by the search for texts treating the same subject, and thus to improve the description of the images and to remove possible ambiguities in the understanding of the text.

3. Scientific Foundations

3.1. Introduction

The work within the team needs two kinds of competencies: to exploit the content of documents, one should first be able to access this content, i.e. to characterize or describe this content. One should also be able to use this description in order to fulfill a task related to these documents. Finally, both description and exploitation techniques must satisfy the needs of the user (and the proof of this simple fact is not trivial).

Finding a solution requires the use of document description techniques based on text, image or video processing (sound and speech processing are studied by the METISS team with which we closely collaborate.) It is also necessary to exploit the correlation and complementarity between the different media, since they do not bring the same information and do not share the same limitations.

After this description stage, it is necessary to exploit the descriptions to satisfy the user's query. At this second stage are needed sorting, indexing, retrieving algorithms which must provide good and fast results, two constraints usually opposite.

These two aspects are not independent and any solution with only one of the two aspects can not solve any real problem. The combination of the two in the context of large databases raises many difficult, but interesting, questions, and their solution may only come from a confrontation of people and ideas coming from both sides.

3.2. Metadata and Document Description

Key words: *descriptor*, *metadata*.

All the multimedia documents have the ambivalent characteristic to be, on the one hand, very rich semantically and, on the other hand, very poor, especially when considering the elementary components which constitute them (continuations of letters or pixels). More concise and informative descriptions are needed in order to handle these documents.

3.2.1. Image Description

Key words: *image matching*, *image recognition*, *image indexing*, *invariants*.

Computing image descriptors has been studied for about ten years. The aim of such a description is to extract indices called descriptors whose distance reflects those of the images they are computed from. This problem can be seen as a coding problem: how images should be coded such that the similarity between the codes reflects the similarity between the original images?

The first difficulty of the problem is that image similarity is not a well defined concept. Images are polysemic, and their level of similarity will depend on the user who judges this similarity, on the problem this user tries to solve, and on the set of images he is using. As a consequence, there does not exist a single descriptor which can solve every problem.

The problem can be specialized with respect to the different kinds of users, databases and needs. As an example, the problem of professional users is usually very specific, when domestic users need more generic solutions. The same difference occurs between databases composed of very dissimilar images and those composed only of images of one kind (fingerprints or X ray images). Finally, retrieving one particular image from an excerpt or browsing in a database to choose a set of images may require very different descriptors.

To solve these problems, many descriptors have been proposed in the literature. The most frequent framework is image retrieval from a large database of dissimilar images using the query by example paradigm. In this case, the descriptors integrate the information of the whole image : color histograms in various color spaces, texture descriptors, shape descriptors (their major drawback is to require automatic image segmentation). This field of research is still active: Color histograms provide a too poor information to solve any problem as soon as the size of the database increases [75] and several solutions have been proposed to remedy this problem: correlograms [57], weighted histograms [36]...

Texture histograms are usually useful for one kind of texture, but they fail to describe all the possible textures, and no technique exist to decide in which category a given texture falls, and thus which descriptor should be used to describe it properly. Shape descriptors suffer from the lack of robustness of shape extractors.

Many other works have been done in the case of specific databases. Face detection and recognition is the most classical and important case, but other works concern medical images for example.

In the team, we work with a different paradigm based on local descriptors: one image is described by a set of descriptors. This solution offers the possibility of partial recognitions like object recognitions independently of the background [74].

The main stages of the method are the following. First, simple features are extracted from each image (interest point in our case, but edges and regions can be used as well). The most widely used extractor is the Harris [55] point detector which provide not very precise but "repeatable" points. Other detectors exist, even for points [63].

The similarity between images are then translated into the concept of invariance: Measurements of the image invariants to some geometric (rotation, translations, scalings) or photometric (intensity variations) transformations are searched for. In practice, this concept of invariance is usually replaced by the weaker concept of quasi-invariance [35] or by properties established only experimentally [46][45].

In the case of points, the classical technique consists of characterizing the signal around each point by its convolution with a Gaussian kernel and its first derivatives and by mixing these measurements in order to obtain the invariance properties. The invariance with respect to rotations, scalings and affine transformations was obtain respectively by Florack [47], Dufournaud [41] and Mikolajczyk [65], photometric invariance was demonstrated for grey-levels by Schmid [74] and for color by Gros [52]. The difficult point is that not only invariant quantities have to be computed, but that the feature extractor has to be invariant itself to the same set of transformations.

One of the main difficulties of the domain is the evaluation and the comparison of the methods. Each one corresponds to a slightly different problem and comparing them is difficult and usually unfair: The results depend on the used database, especially when they are quite small. In this case, a simple syntactic criteria can

give the impression of a good semantic description, but this does not tell anything about what would happen with a larger database.

3.2.2. Video Description

Key words: *video indexing, structuring, key-events.*

Professional and domestic video collections are usually much bigger than the corresponding still image collections: there is a common factor of 1000 in between the two. If the images often have a weaker quality (motion, fuzzy images...), they present a temporal redundancy which can be exploited in order to gain some robustness.

Video indexing is a large concept which covers different topics of research. Video structuring consists of finding the temporal units of a video (shots, scenes) and is a first step to compute a table of contents of a video. Key-event detection is more oriented to the creation of an index of the video. Finally, all the extracted elements can be characterized with various descriptors: motion descriptors [43], or still image based descriptors, which can also use the image temporal redundancy [54].

Many contributions have been proposed in the literature in order to compute a temporal segmentation of videos, and especially to detect shot boundaries and transitions [37][48]. Nevertheless, shots appear to be too low-level segments for many applications since a video can contain more than 3000 of them. Scene segmentation, or what is called macro-segmentation is a solution, but it remains an open problem. The combination of media is probably an important axis of research to progress on this topic.

3.2.3. Text Description

Key words: *natural language processing, lexical semantics, machine learning, corpus based acquisition of linguistic resources, exploratory data analysis.*

Automating indexing of textual documents [73] has to tackle two main problems: first choosing indexing terms, *i.e.* simple or complex words automatically extracted from a document, that "represent" its semantic content and make its detection possible when the document database is considered; second, dealing with the fact that the representation *is* a word one and not a conceptual one. Therefore information retrieval has to be able to overcome two semantic problems: various possibilities to formulate the same idea (how to match a same concept in a text and a query but expressed with different words); word ambiguity (a same word –graphical chain– can cover different concepts). In addition to these difficulties, the meaning of a word, and thus the semantic relations that link it to other words, varies from one domain to the other. One solution is to make use of domain-specific linguistic resources, both to disambiguate words and to expand user queries with synonyms, hyponyms... These domain-specific resources are however not pre-existing and must be automatically extracted from corpora (collections of texts) using machine learning techniques.

Lots of works have been done during the last decade in the domain of automatic corpus-based acquisition of lexical resources, essentially based on statistical methods, even though symbolic approaches also present a growing interest [78]. We focus on these two kinds of methods and aim at developing machine learning solutions that are generic and fully automatic to give the possibility to extract from a corpus the kind of lexical elements required by a given application. We specifically extract semantic relations between words (especially noun-noun relations) using hierarchical classification techniques and implementing principles of F. Rastier's differential semantic theory [70]. We also acquire through symbolic machine learning (inductive logic programming [66]) noun-verb relations defined within J. Pustejovsky's Generative lexicon theory [69]; those peculiar links give access to interesting reformulations of terms (*disk shop - to sell disks*) that are up to now not often used in information retrieval systems. Our research both concerns the machine learning algorithms developed to extract lexical elements from corpora, and the linguistic and applicative interests of the learnt elements.

3.2.3.1. Acquisition of lexicons based on Rastier's differential semantics.

Differential (or interpretative) semantics [70] is a linguistic theory in which the meaning of a word is defined through the differences that it presents with the other meanings in the lexicon. A lexicon is thus a

network of words, structured in classes, in which differences between meanings are represented by *semes* (*i.e.* semantic features). Within a given semantic class –group of words that can be exchanged in some contexts–, words share *generic semes* that characterize their common points and are used to build the class (*e.g.* /*to seat*/ is associated with {*chair, armchair, stool...*}), and *specific* ones that explicit their differences (/*has arms*/ differentiates *armchair* from the two others). Following Rastier, two kinds of linguistic contexts are fundamental to characterize relations of lexical meaning: the topic of the text unit in which a word occurrence is found, and its neighborhood. Differential semantics states that valid semantic classes, in which specific semes can be determined, can only be defined within specific topic. And a topic can be recognized within a text by the presence of a semantic isotopy, *i.e.* the copresence within the sets of semes (named *sememes*) representing some of its words of some recurrent semes. For example, a *war* topic can be detected in a text unit that contains the words *soldier, offensive, general...* by the presence of the same seme */war/* within the sememes of all these words.

We have developed a 3-level method to extract lexicons based on Rastier's principles. First, with the help of a hierarchical classification method (Linkage Likelihood Analysis, LLA [61]) applied on the distribution of nouns and adjectives among the paragraphs, we automatically learn sets of keywords that characterize the main topics of the studied corpus. These sets are then used to split the corpus into topic-specific corpora, in which semantic classes are built using LLA technique on shared contexts. Finally we try to characterize similarity and dissimilarity links between words within each semantic class.

3.2.3.2. Acquisition of elements of Pustejovsky's Generative lexicon.

In one of the components of this lexical model [69], called the *qualia structure*, words are described in terms of semantic roles. For example, the *telic* role indicates the purpose or function of an item (*cut* for *knife*), the agentive role its creation mode (*build* for *house*)... The qualia structure of a noun is mainly made up of verbal associations, encoding relational information.

We have developed a learning method, using inductive logic programming [66], that enables us to automatically extract, from a morpho-syntactically and semantically tagged corpus, noun-verb pairs whose elements are linked by one of the semantic relations defined in the qualia structure in the Generative lexicon. This method also infers rules explaining what in the surrounding context distinguishes such pairs from others also found in sentences of the corpus but which are not relevant. In our work, stress is put on the learning efficiency that is required to be able to deal with all the available contextual information, and to produce linguistically meaningful rules. And the obtained method and system, named ASARES, is generic enough to be applied to the extraction of other kinds of semantic lexical information.

3.2.3.3. Characterization of huge sets of thematically homogeneous texts.

A collection of texts is said to be thematically homogeneous if the texts share some domains of interest. We are concerned by the indexing and analysis of such texts. The research of relevant keywords is not trivial : even in thematically homogeneous sets, there is a high variability in the used words and even in the concerned sub-fields. Apart from the indexing of the texts, it is valuable to detect thematic evolutions in the underlying corpus.

Generally, textual data are not structured and we must suppose that the files we are concerned with have either a minimal structure, or a general common thema. The method we use is the factorial correspondence analysis. We get clusters of documents and their characteristic words. Recently, R Priam defended a thesis where he proposes methods very close to the Kohonen maps to visualize words and documents local proximities.

3.2.4. Evaluation of Descriptors

Key words: *evaluation, performance, discriminating power*.

The situation on this subject is very different according to the concerned media. Reference test bases exist for text, sound or speech, and regular evaluations campaign are organized (NIST for sound and speech

recognition, TREC for text in English, AMARYLLIS for text in French, SENSEVAL or ROMANSEVAL for text disambiguation)[1].

In the domain of images and videos, The BENCHATLON provides a database to evaluate image retrieval systems while TREC provides test database for video indexing. A system to evaluate shot transition algorithms has been developed by G. Quenot and P. Joly [72].

3.2.5. Metadata

Key words: *metadata*.

To improve the data organization or to define the strategies to choose or compute some descriptors, it may be advisable to use contextual informations. These informations, called metadata (or data about the data) can provide some information about how the data were produced, obtained, or used, they can provide information about the users or describe the content of a document.

V. Kashyap and A. Sheth [58] proposed a classification of the metadata for multimedia in two main classes: metadata which contain external information (date, localization, author...) and metadata which contain internal information linked to the content or the way this content is represented within the document. The latter are called descriptors:

1. Some of them are computed directly from the documents;
2. The others are provided by a human, like keywords.

The organization of the metadata can be completed and become matricial when considering the various objectives of the metadata: easy data access, data abstract, interoperability, media or content representation...), or the way the metadata were obtained. Metadata are a privileged way to keep information relative to a document or its descriptors in order to facilitate future processing. They appear to be a key point in a coherent exploitation of large multimedia databases.

3.3. Efficient Exploitation of Descriptors

Key words: *Statistics, Data Analysis, Indexing*.

Even if the description of the documents can be done automatically, this is not enough to build a complete indexing and retrieval system usable in practice. As a matter of fact, the system must be able to answer a query in a reasonable amount of time, and thus requires tools in order to guarantee this aspect. The section is devoted to some of these tools.

On-line and off-line processing define the two main categories of exploitation. Off-line processing correspond usually to all techniques which need to consider all the data, and the complexity in time is thus not the main issue. On the other hand, on-line processing need to go really fast. To gain such a performance, these procedures use the result of the off-line processing to limit the treatment to the smallest data subset necessary to answer the query.

3.3.1. Statistics for Huge Datasets

Key words: *exploratory data analysis, statistics, sampling*.

The situation where we have few available data has been well studied but a huge amount of data generates different kinds of problems : for instance, the use of classical inferential statistics results in hypothesis testing conclude rather often to reject the null hypothesis. Besides, the methods of models identification fail very often or we overestimate the quality of the model. The question is : how can we set a representative sampling in such datasets? We must also add that some clustering algorithms are unusable with such large datasets. Therefore, it is clear that working with huge datasets is difficult because of their computational complexity, because of the data quality and because of the scaling problem in inferential statistics.

[1] It should be noted that within TREC, the aim is now not only to retrieve pertinent documents, but to find in these documents the parts which provide the answer to a precise query.

However, statistical methods can be used with caution if the data quality is good. So the first step is the cleaning and the checking of data to be sure of their coherence. The second step depends on our goal. Either we want to build a global model, either we are looking for hidden structures in the data. In the first case, we can work on a sample of the data and use methods such as clustering, segmentation, regression models. In case we are looking for hidden structures, sampling is not appropriate and we need to use other heuristics.

Exploratory data analysis (EDA) is an essential tool to deal with huge amount of data. EDA describes data in an interactive way, without a priori hypothesis and provides useful graphical representations. Visualization methods when the dimension of the data is greater than three is also indispensable: for instance, parallel coordinates. All these previous methods analyse the data to discover their properties.

We can add that most of the available data mining programs are very expensive, and that their contents are very disappointing and poor for most of them.

3.3.2. Multidimensional Indexing Techniques

Key words: *Multidimensional Indexing Techniques, Databases, Curse of Dimensionality, Approximate Searches, Nearest-Neighbors.*

This section gives an overview of the techniques used in databases for indexing multimedia data (often focusing on still images). Database indexing techniques are needed as soon as the space required to store all the descriptors gets too big to fit in main memory. Database indexing techniques are therefore used for storing descriptors on disks and for accelerating the search process by using multi-dimensional index structures. Their goal is mainly to minimize the resulting number of I/Os. This section first gives an overview of traditional multidimensional indexing approaches achieving exact NN-searches. We especially focus on the filtering rules theses techniques use to dramatically reduce their response times. We then move to approximate NN-search schemes.

3.3.2.1. Traditional Approaches, Cells and Filtering Rules.

Traditional database multidimensional indexing techniques typically divide the data space into cells containing vectors. Cell construction strategies can be classified in two broad categories: *data-partitioning* indexing methods [33][79] that divide the data space according to the distribution of data and *space-partitioning* [56][77] indexing methods that divide the data space along predefined lines regardless of the actual values of data and store each descriptors in the appropriate cell.

Data-partitioning index methods all derive from the seminal R-Tree [53], originally designed for indexing bi-dimensional data used in Geographical Information Systems. The R-tree was latter extended to cope with multi-dimensional data. The SS-Tree [79] is an extension that rely on spheres instead of rectangles. The SR-Tree [59] specifies its cells as being the intersection of a bounding sphere and a bounding rectangle.

Space-partitioning techniques like grid-file [67], K-D-B-Tree [71], LSD$^{\text{h}}$-Tree [56] typically divide the data space along predetermined lines regardless of data clusters. Actual data are subsequently stored in the appropriate cells.

NN-algorithms typically use the geometrical properties of cells to eliminate those cells that can niether have any impact on the result of the current query [38]. Eliminating irrelevant cells avoids having to subsequently analyze all the vectors they contain, which, in turn, reduces response times. Eliminating irrelevant cells is often enforced at run-time by applying two rather similar *filtering rules*. The first rule is applied at the very beginning of the search process and identifies irrelevant cells as follows:

$$\text{if } dmin(q, C_i) \geqslant dmax(q, C_j) \text{ then } C_i \text{ is irrelevant,} \tag{1}$$

where $dmin(q, C_i)$ is the minimum distance between the query point q and the cell C_i and $dmax(q, C_j)$ the maximum distance between q and cell C_j.

The search process then ranks the remaining cells on their increasing distances to q. It then accesses the cells, one after the other, fetches all the vectors each cell contains, and computes the distance between q and each vector of the cell. This may possibly update the current set of the k best neighbors found so far.

The second filtering rule is applied to stop the search as soon as it is detected that none of the vectors in any remaining cell can possibly impact the current set of neighbors ; all remaining cells are skipped. This second rule is:

$$\text{if } \; dmin(q, C_i) \geqslant d(q, nn_k) \; \text{ then } \; \text{stop}, \tag{2}$$

where C_i is the cell to process next, $d(q, nn_k)$ is the distance between q and the current k^{th}-NN.

The "curse of dimensionality" phenomenon makes these filtering rules ineffective in high-dimensional spaces [77][34][68][38][60].

3.3.2.2. Approximate NN-Searches.

This phenomenon is particularly prevalent when performing *exact* NN-searches. There is therefore an increasing interest in performing *approximate* NN-searches, where result quality is traded for reduced query execution time. Many approaches to approximate NN-searches have been published.

3.3.2.2.1. Dimensionality Reduction Approaches.

Dimension reduction techniques have been used to overcome the "curse of dimensionality" phenomenon. These techniques, such as PCA, SVD or DFT (see [49]), exploit the underlying correlation of vectors and/or their self similarity [60], frequent with real datasets. NN-search schemes using dimension reduction techniques are approximate because the reduction only coarsely preserves the distances between vectors. Therefore, the neighbors of query points found in the transformed feature space might not be the ones that would be found using the original feature space. These techniques introduce imprecision on the results of NN-searches which can not be controlled nor precisely measured. In addition, such techniques are effective only when the number of dimensions of the transformed space become very small, otherwise the "curse of dimensionality" phenomenon remains. This makes their use problematic when facing very high-dimensional datasets.

3.3.2.2.2. Early Stopping Approaches.

Weber and Böhm with their approximate version of theVA-File [76] and Li et al. with Clindex [62] perform approximate NN-searches by interrupting the search after having accessed an arbitrary, predetermined and fixed number of cells. These two techniques are efficient in terms of response times, but give no clue on the quality of the result returned to the user. Ferhatosmanoglu et al., in [44], combine this with a dimensionality reduction technique: it is possible to improve the quality of an approximate result by either reading more cells or by increasing the number of dimensions for distance calculations. This scheme suffers from the drawbacks mentioned here and above.

3.3.2.2.3. Geometrical Approaches.

Geometrical approaches typically consider an approximation of the sizes of cells instead of considering their exact sizes. They typically account for an additional ε value when computing the minimum and maximum distances to cells, making somehow cells "smaller". Shrunk cells make the filtering rules more effective, which, in turn, increases the number of irrelevant cells. However, cells containing interesting vectors might be filtered out.

In [76], the VA-BND scheme empirically estimates ε by sampling database vectors. It is shown that this ε is big enough to increase the filtering power of the rules while small enough in the majority of cases to avoid missing the true nearest-neighbors. The main drawback of this approach is that the same ε is applied to all existing cells. This does not account for the possibly very different data distributions in cells.

The AC-NN scheme for M-Trees presented in [40] also relies on a single value ε set by the user. Here, ε represents the maximum relative error allowed between the distance from q and its exact NN and the distance from q and its approximate NN. In this scheme, setting ε is far from being intuitive. Their experiments showed that, in general, the actual relative error is always much smaller than ε. Ciaccia and Patella also present an extension to AC-NN called PAC-NN which uses a probabilistic technique to determine an estimation of the distance between q and its NN. It then stops the search as soon as it founds a vector closer than this estimated distance. Unfortunately, AC-NN and PAC-NN can not search for k neighbors.

3.3.2.2.4. Hashing-based Approaches.

Approximate NN-searches using locality sensitive hashing (LSH) techniques are described in [50]. These schemes project the vectors into the Hamming cube and then use several hash functions such that co-located vectors are likely to collide in buckets. LSH techniques tune the hash functions based on a value for ε which drives the precision of searches. As for the above schemes, setting the right value for ε is key and tricky. The maximum distance between any query point and its NN is also key for tuning the hash functions. While finding the appropriate setting is, in general, very hard, [50] observes that choosing only one value for this maximum distance gives good results in practice. This, however, makes more difficult any assessment on the quality of the returned result. Finally, the LSH scheme presented in [50] might, in certain cases, return less than k vectors in the result.

3.3.2.2.5. Probabilistic Approaches.

DBIN [32] clusters data using the EM (Expectation Maximization) algorithm. It aborts the search when the estimated probability for a remaining database vector to be a better neighbor than the one currently known falls below a predetermined threshold. DBIN bases its computations on the assumption that the points are IID samples from the estimated mixture-of-Gaussians probability density function. Unfortunately, DBIN can not search for k neighbors.

P-Sphere Trees [51] investigate the trading of (disk) space for time when searching for the approximate NN of query points. In this scheme, some vectors are first picked from a sample of the DB, and each picked vector becomes the center of one hypersphere. Then, the DB is scanned and all the vectors that have one particular center as nearest neighbor go into the corresponding hypersphere. Vectors belonging to overlapping hyperspheres are replicated. Hyperspheres are built in such a manner that the probability of finding the true NN can be enforced at run time by solely scanning the sphere whose center is the closest to the query point. P-Sphere Trees can also not search for k neighbors.

To our knowledge, no technique linking the precision of the search to a probability of improving the result can search for k neighbors.

4. Application Domains

4.1. Still Image Database Management

Key words: *Image databases, photo agencies, digital pictures, medical imagery.*

We are particularly interested in large image bases, like those managed by photo agencies. These agencies have between five hundred thousands and twelve millions of images. The Andia Press agency has a million of them, Sigma twelve millions, the Corbis agency which gathers the whole of acquisitions of Bill Gates has thirty six millions of them. These agencies work according to two modes. In the first one, they respond to a customer query by sending him a set of images. The customer pays for the images that he publishes. In the second mode, the customers are subscribers at the agencies which send their new photographs systematically to them, the mode of payment being the same one. This working method is that of the AFP or Reuters.

One of the concerns of the agencies is of course the digital rights management, and the fact that they are not unduly used by people or institutions while not having discharged the rights. Watermarking and indexing are two techniques planned to control image diffusion, either by seeking a watermark of property in the images, or by checking, using indexing techniques, that the image is not a fragment of an image of the agency base.

4.2. Video Database Management

Key words: *video bases, video structuring.*

The existing video databases are generally little digitized. The progressive passage to digital television should quickly change this point. As a matter of fact, TF1 passed to an entirely digitized production, the cameras remaining the only analogical stage of the production. Treatment, assembly and diffusion are digital. In

addition, domestic digital decoders can, from now on, be equipped with hard disks allowing a storage initially modest, of ten hours of video, but larger in the long term, of a thousand of hours.

One can then distinguish two types of digital files. First of all those of the private individuals, including recordings of diffused programs and films taken using digital camcorders. If the effort of management of such bases will be probably weak, without rigorous method, there is a great need for tools to help the user: automatic creation of summaries and synopses to allow to find information easily, or to gain, in a few minutes, a general idea of a program. Even if the service is rustic, it is initially evaluated according to the appreciation which it brings to a system (video tape recorder, decoder), will have to remain not very expensive, but will benefit from a large diffusion.

On the other hand professional files (TV channels archives, registration of copyright, cineclubs, producers...) are of a much larger size, but benefit from the attentive care of professionals of documentation and archiving. In this field, systems can be much more expensive and are judged according to the profits of productivity and the assistance which they bring to documentalists, journalists and users.

4.3. Textual Database Management

Key words: *bibliography, indexing.*

Searching in large textual corpora has already been the topic of many researches. The current stakes are the management of very large volumes of data, the possibility to answer requests relying more on concepts that on simple inclusions of words in the texts, and the characterization of sets of texts.

We work on the exploitation of scientific bibliographical bases. The explosion of the number of scientific publications make the retrieval of relevant data for a researcher a very difficult task. The generalization of document indexing in data banks did not solve the problem. The main difficulty is to choose the keywords which will encircle a domain of interest. The statistical method used, the factorial analysis of correspondences, makes it possible to index the documents or a whole set of documents and provides the list of the most discriminating keywords for this or these documents. The validation of the indexing is carried out by making a search for information in databases more general than the one which made it possible to work out the index and by studying the reported documents. That in general makes it possible to still reduce the subset of words characterizing a field.

Another difficulty is to find within a given document the parts which tackle a subject. We thus worked, from texts of bioinformatics coming from bases such as Medline, on the automatic extraction of the zones of texts describing the interactions between genes and on the modeling of the described interaction. Modeling requiring a fine and expensive analysis of sentences, it should be carried out only on zones of texts likely to contain an interaction indeed. Our methodologies of training of semantic bonds between words are exploited to determine these relevant zones of texts. To a corpus of summaries extracted from Medline, we apply a training by ILP to try to learn what distinguishes the sentences containing interactions description of the others.

We also explore scientific documentary corpora to solve two different problems: to index the publications by the way of meta-keys and to identify the relevant publications in a large textual database. For that, we use factorial data analysis which allows us to find the minimal sets of relevant words that we call meta-keys and to clear out the bibliographical search from the problems of noise and silence. The performances of factorial correspondence analysis are sharply greater than classic search by logical equation.

4.4. Robotics and Visual Servoing

Key words: *robotics, visual servoing, visual memory, planning.*

If collaboration between robotics and vision is an already old subject, it has undergone an important change of paradigm in the five last years. Hitherto, collaboration was considered on the level of planning: a camera observed the world around a robot to enable him to plan its displacements. The results appeared to be not so satisfactory.

The field of collaboration then moved towards control: the vision is not any more used to plan a movement, but to ensure its follow-up and good execution, by setting up a closed loop of control including vision [42][39][64]. The results are promising and many industrial applications already exist.

Some difficulties remain: the tasks to be achieved are specified using a target image that should be reached, but that assumes that the robot is able to establish a bond between this image and the current image provided by the camera. This is a classical image matching problem. If these two images do not have anything in common, it will be necessary to use a collection of intermediate images, which define intermediate positions of the robot before reaching the final position.

Therefore, the control problem corresponds to an image collection management problem, with dynamic collections to follow the evolution of the environment of the robot, and needs for fast access for recognition. This application appears important because it widely opens the experimental use conditions of visual servoing: once an environment collected in a base, the robot can start from any position to go towards any target. If this kind of approach presents little interest for articulated arm for which the articular co-ordinates can be read directly, an autonomous vehicle can benefit from it in restricted environments such as car parks. In this case, the systems of positioning as the GPS do not offer sufficient relative precision and do not give information of orientation.

5. Software

5.1.1. I-Description :

this software allows to compute local or global image descriptors: differential local invariants, global and local color histograms or weighed histograms. It was deposited to "Agence pour la Protection des Programmes" under the number
IDDN.FR.001.270047.000.S.P.2003.000.21000. (Correspondant: Patrick Gros.)

6. New Results

6.1. Image Retrieval in Large Databases

Our work on image description does not aim at finding new general descriptors. The IMEDIA and LEAR teams are very active in this field, and we use their results. The originality of our work comes from the size of the database we want to handle. In large databases, most images will be compressed. Is it possible to describe an image without decompressing it? In many databases, images will also be watermarked, and the influence of watermarking (and of the systems for breaking watermarks) on the content-bases description techniques is not clear. This is our first direction of research.

A second direction concerns the combination of descriptors: when documents are described by many descriptors, how a query should be processed in order to provide the fastest as possible answer? To answer this question, we study the information that each descriptor can provide about the other ones. The aim is to determine the order in which the descriptors should be considered.

The third direction is description indexing and retrieval. In the local description scheme, 1 million of images can give raise to 600 millions of descriptors, and retrieving any information in such an amount of data requires really fast access techniques, whatever the aim of this access may be.

A fourth direction is due to our collaboration with the roboticians of the VISTA team. They work on visual servoing and using a database is a good way to improve the applicability of their techniques to large displacements. Our description technique appear to be particularly well suited to such an application where a matching between images is required, and not only a global link of similarity between images.

6.1.1. Image Description, Compression and Watermarking

Participants: Patrick Gros, François Tonnin.

Key words: *image indexing*, *image description*, *image compression*.

This is a joint work with the TEMICS team (S. Pateux).
Image authentication is becoming very important for certifying image data integrity. A key issue in image authentication is the design of a compact signature being robust under allowable manipulations. Watermarking has been mostly investigated to deal with the problem of detection of illegal copies. But it provides only an assumption, not a proof, of illegacy. We believe that content based image description techniques may provide robust detection of illegal copies. Big databases are made of compressed images. In order to speed up the matching scheme, it is of interest to calculate signatures from the compressed images. Thanks to its wavelet analysis, JPEG2000 compression standard allows the design of multiresolution signatures. Inspired by classical content based local description techniques, we have developed a robust point extractor in the wavelet space. Its average robustness is 10 % less than multiresolution Harris point extractor reference. We will investigate how to describe (in the wavelet space) the neighborhood of these points by means of vectors invariant to allowable image manipulations. Another point we consider is the comparison of robustness and speed between classical local signatures and wavelet signatures.

6.1.2. Combination of Descriptors by Association Rules and Multiple Correspondence Analysis

Participants: Laure Berti, Anicet Kouomou-Choupo, Annie Morin.

Content-based image retrieval is not easy when image databases become very large. Fixed image database can be described in several ways by global visual descriptors of color, texture, or form (pixel level). Most frequent queries imply and combine results of several type of descriptors such as: "retrieve all images that have similar color and similar texture to the given example image". To retrieve more efficiently and more effectively an image of a large database, we exploited combinations of descriptors. Firstly we surveyed the state of the art of image mining and content-based image retrieval. Then, our objective was to study the interest of association rules between descriptors to accelerate response time of queries on large fixed image databases. We used 5 MPEG-7 descriptors to describe several thousands of fixed images. We initially used K-means based algorithm to compute clusters of images for each descriptor. We then generated relations between different clusters in form of association rules. Multiple correspondence analysis was used to study the relevance of found associations and to validate our approach. We are now exploiting association rules between clusters of descriptors to optimize content based retrieval.

6.1.3. Approximate Searches: k-Neighbors + Precision

Participants: Laurent Amsaleg, Sid-Ahmed Berrani, Patrick Gros.

Key words: *Multidimensional Indexing Techniques, Databases, Curse of Dimensionality, Approximate Searches, Nearest-Neighbors.*

6.1.3.1. References:
[11][12][18][19][20]

This is a joint work with Thomson R&D France (cf. 7.1.1).
We designed an approximate search-scheme for high-dimensional DB where the precision of the search can be stochastically controlled and where the search can retrieve the k nearest-neighbors of query points. It allows a fine and intuitive control over the precision by setting at run time the maximum probability for a vector that would be in the exact answer set to be missing in the approximate answer set. This off-line scheme computes controlled approximations shrinking each cluster within which feature vectors are enclosed. Those approximations are values for (approximate) radii of clusters, and they are computed for all the levels of precision defined beforehand. To answer a query, the search process considers the appropriate approximations corresponding to the desired level of precision. This may cause the actual nearest-neighbors of the query point to be ignored. Our method, however, bounds the probability for this to happen. This paper also presents a performance study of the implementation using real datasets. It shows, for example, that our method is 6.72 times faster than the sequential scan when it handles more than $5 \quad 10^6$ 24-dimensional vectors, even when the probability of missing one of the true nearest-neighbors is below 0.01.

This approach first clusters vectors. It encloses clusters in minimum bounding hyperspheres in an Euclidean space. All existing vectors might not be in clusters because the clustering isolates outliers. Outliers are stored

in a specific file that we treat separately. The clustering algorithm we use is derived from the first phase of Birch [80]. It has a couple of crucial differences, however. Birch ends its first phase when all the created micro-clusters can fit in the allowed main memory. Instead, we stop our clustering when the number of micro-clusters created falls below the maximum number of clusters that are allowed to exist. The variance of data points drives the radius of Birch' clusters. Instead, radii of clusters in our implementation are exact in the sense that each defines a minimum bounding hypersphere.

The output of the clustering phase is a set of minimum bounding hyperspheres defined by their center and their exact radius. As for Birch, clusters might overlap and outliers are treated separately. Data points are stored sequentially on disk on a per cluster basis. No specific data structure is used to index the clusters. Outliers are also stored in a separate data file, in a sequential manner.

Each cluster is analyzed off-line to derive several approximate radii given the exact radius, the volume and the distribution of vectors within each cluster. For each cluster, several approximate radii are determined, each corresponding to a predetermined level of precision. All the approximate radii of one cluster are always smaller than the exact radius of the same cluster. Approximate radii will ultimately be considered during the approximate NN-searches.

At query submission time, a user provides, along with the query, an imprecision level called α controlling the quality of the approximate NN-search. α is chosen among the set of predefined values, and it corresponds to the maximum probability for a vector that belongs to the exact answer set to be actually missing in the approximate set of answers eventually returned.

This imprecision level then determines which specific approximate radii must be taken into account by the filtering rules during the NN-search. Irrelevant clusters are thus filtered out and the remaining clusters are then ranked with respect to the distance of their centers to the query point. Clusters are then accessed one after the other. When a cluster is accessed, all the data points it contains (all points enclosed within its *exact* bounding hypersphere) are fetched in memory. The search then computes the distances between all points in the cluster and the query vector. This might in turn update the current set of neighbors. It might also filter out more clusters. The search stops when k neighbors have been found and when the approximate minimum distance to the next cluster is greater than the current distance to the k^{th} neighbor.

Before returning the result to the user, a sequential scan of the file where outliers are stored is performed. This might also update the current set of neighbors.

6.1.4. Coupling Action and Perception by Image Indexing and Visual Servoing

Participants: Patrick Gros, Anthony Remazeilles.

This is a joint work with the VISTA team (F. Chaumette).

We are working on automatic robot motion control, using visual information provided by an on-board camera, and an image data base of the navigation space. The image base describes the environment in which the robotic system moves. More exactly, it describes features that can observe the robot camera. Thanks to this base, the robot localization is nothing but a k nearest-neighbor search [18] of the initial image given by the camera before the motion. The localization stage therefore avoids reconstructing the entire scene, which is a time consuming and complex process.

The definition of the path the robot has to follow is also defined in terms of images : the desired position corresponds to the image the camera should obtain at the end of the motion. The same image retrieval method presented before enables to localize the desired position. By translating the image base into a valuated graph (corresponding to the feasibility to go from on image to an other), and using graph theory, the shortest image path can be easily found between the initial image and the desired one. Those images extracted from the database describe in a continuous way the space the robot has to pass through in order to reach the desired position.

During this year, we have defined a formalism that enables to control robot motions, given this image sequence, the features matched between each consecutive couple of images, and the images acquired by the camera. 3D reconstruction is not necessary yet. Furthermore, robot motion is not defined during an off-line

stage; motion are determined for each image acquired by the camera. This will permit us to take easily in account within our scheme unexpected exterior events, like occlusion, obstacles, ...

Our method is based on potential field theory. The robot moves in order to make features defined on the image path, initially out of the camera field of view, become visible. Furthermore, the obtained trajectory is independent of the intermediate image positions. This work has been validated through experiments with a planar environment, and planar motions, with an articulated arm. We are trying to relax those constraints in order to be able to deal with more general motions, and on a 3D scene. Then non-holonomic constraints will be added in order to manage mobile robots in real environments.

Furthermore, we want to improve the data base management, which could accelerate the retrieval process. For example grabbing conditions (moment of the day, weather conditions, ...) are criteria that can be extracted automatically from the image signal. Those information could help to categorize the images of the base, and also to provide to the robot images that best correspond to the current exterior condition (which can be very useful as long as those images are used in the feature tracking stage). At last, a protocol for the autonomous image base acquisition should be defined, in order to be able to make experiments with the robot Cycab that owns Irisa.

6.2. Text Retrieval in Large Databases

6.2.1. Natural Language Processing and Machine Learning

Participants: Vincent Claveau, Fabienne Moreau, Mathias Rossignol, Pascale Sébillot.

Key words: *natural language processing, machine learning, lexical semantics, corpus-based acquisition of lexical relations, semi-supervised learning, inductive logic programming, hierarchical classification.*

6.2.1.1. References:

[8][22][23][15]

The general frame of our work is explained in 3.2.3.

During 2003, our work has especially concerned the 3 following points:

1. ASARES: two semi-supervised versions combining symbolic and statistical approaches.
   ASARES is our inductive logic programming (ILP) based system (using ALEPH algorithm) that automatically infers from a set of positive and negative examples of elements in a given relation (*e.g.* noun-verb (N-V) pairs in which the V plays for the N one of the roles defined in the qualia structure in Pustejovsky's Generative Lexicon model, or do not play such a role; we shall refer respectively to these two cases as N-V qualia (resp. non-qualia) pairs) morpho-syntactic and semantic patterns that characterize this relation and can be applied to a corpus to get new elements in this same relation. [15] fully describes this system and its application to the extraction of N-V qualia pairs; it also explains the refinement operator well-adapted to the hierarchical knowledge we deal with that we have built, that allows us to travel efficiently through our hypothesis search space. However the automation of the system and its easy application to a new corpus is limited by the supervised nature of ILP. We have thus proposed two semi-supervised versions of ASARES that combine in two different ways a statistical approach (N-V qualia pairs are considered as one special kind of cooccurrences) and the ILP approach. The first and sequential combination is presented in [22], whereas the second one, that more deeply integrates the two techniques, is described in [23]. The two solutions lead to the same results as ASARES supervised version, but keep advantages of the two approaches they mix: robustness and automation of the statistical method, quality of the results and expressiveness of the symbolic one.
2. Acquisition of semantic lexicons based on Rastier's differential semantics: automatic generation of sets of keywords for topic characterization and detection.
   We have ended the elaboration of a sequence of statistical data analysis treatments, that refines and enriches the results of an initial LLA (linkage likelihood analysis) classification of the words of

a given corpus based on their distribution over its paragraphs, and allows us to obtain in a fully automatic way and with the use of no prior information sets of keywords that characterize the main topics of the (morpho-syntactically tagged) corpus. Each class can then be used to detect the presence of *its* topic in any paragraph of the corpus, by a simple keyword cooccurrence criterion. The obtained sets enable us to split an initial non-specialized corpus into several topic-specific ones and to get the linguistic material necessary to carry on the second step of the elaboration of semantic lexicons based on Rastier's principles, *i.e.* the automatic constitution of semantic classes within homogeneous topics. We have begun a work in this direction. In order to achieve this goal, once again without any human intervention or external data, thus eventually favoring precision over recall, we consider statistical techniques that group words appearing in similar contexts. We plan to take into account different lengths of contexts for different word categories, and to consider the relative positions of contextual elements. This idea leads us to the need of the definition of a non-symmetric similarity measure to automatically build the semantic classes.

3. Linguistic resources and information retrieval (IR).
   We have evaluated the interest of N-V qualia relations for the expansion of queries in an information retrieval system (IRS). More precisely, we have used Salton's IRS SMART and the data of the IRS evaluation campaign Amaryllis, and ASARES has learnt qualia pairs from one Amaryllis corpus. Our experiments have shown [8] that expanding a query with verbs that play one qualia role for nouns that it contains locally but significantly improves the results of SMART. More precisely, the relevance of the first ten documents is increased, and these results are particularly interesting if we consider the way search engines are commonly used. Moreover, Fabienne Moreau has begun in October a PhD thesis which aims at exploring methods of extending Salton's vector space model (VSM) to improve its ability to capture the semantics of natural language texts. Currently, under Salton's theory, documents are represented as a set of features, without regard for the relationship between individual terms. The goal of this thesis is to adapt the VSM to allow information gained from natural language processing to inform IR.

6.2.2. Visualization and Web Mining

Participants: Nicolas Bonnel, Annie Morin.

This is joint work with France Telecom R & D (cf. 7.1.2).
Nicolas Bonnel, a second year PHD student, is currently working on the dynamic generation of 2D and 3D multimedia interactive presentations, that aims at representing the results of a search in a database. N. Bonnel has a Cifre contract with France Telecom and his thesis is done in cooperation with FT. For that, he uses metaphors developed by France Telecom and works on the relevance of descriptors for the documents and the improvement of the graphical representation. We need therefore to perform quality evaluation and we have to take into account the user profile to optimize the results of a query.

6.2.3. Knowledge Extraction and Visualization from Textual Databases

Participant: Annie Morin.

Key words: *Exploratory data analysis*.

Knowledge extraction from textual databases is not obvious. Among the used methods, we find factorial analysis, neural networks or Kohonen maps. R Priam 's PhD thesis [10] proposes an adaptation of Kohonen maps to discrete data and develops a new algorithm called CASOM for correspondence analysis and Self organizing maps. CASOM is a non-linear extension of correspondence analysis which allows a graphical representation of words and documents.

6.3. Multimedia Document Description

Participants: Manolis Delakis, Patrick Gros, Ewa Kijak, Hervé Renault, Pascale Sébillot.

Key words: *multimedia*.

The term multimedia documents is broadly used and covers in fact most documents. It is in fact more and more appropriate since any document are now truly multimedia and contain several media: sound, image, video text. The description of these documents, videos for example, remains quite difficult. Research groups are often monodisciplinary and specialist of only one of these media, and the interaction between the different media of a same document is not taken into account. Nevertheless, it is clear that this interaction is a very rich source of information and allows to avoid the limitations of the techniques devoted to a single media since the limits vary according to the concerned media.

We propose to investigate, in conjunction with other teams like METISS and VISTA, this new aspect of multimedia document description. We propose to follow two directions: the first one concerns the case of video, where image and sound are closely related and provide complementary information. The sound track also opens the possibility of speech recognition, and requires the use of natural language processing in order to use this new modality. In this case, one of the problem is to handle the dynamics proper to each media. The second direction concerns the documents which mix text and still images like journals, technical manuals, or most of web pages.

6.3.1. Video Description

Participants: Manolis Delakis, Patrick Gros, Ewa Kijak.

Key words: *image - sound interaction, video structuring, Hidden Markov Models*.

6.3.1.1. References:

[9][28][25][29][26][27][30]

Our work on this topic is done in close collaboration with the METISS and VISTA teams of IRISA and with the Thomson company where E. Kijak has done most of her thesis (cf. 7.1.1).

The aim of this work is to define a general method allowing to describe all the media of a video, as well as their interaction. Another constraint is that this method should also allow a user to formulate a task or a query concerning videos. This problem was first studied during the thesis of E. Kijak in the frame of a limited case, the structuring of sport (and particularly tennis) reports. In such documents, there are four main sources of information: the tennis rules which explain how a tennis game is organized, the production rules which explain how the producer works, what tools it uses and how he tries to reflect what's going on by formal techniques, the image and the sound tracks.

It is clear that none of these sources can explain the video alone. The goal is thus to integrate these sources such that their complementarity can be used to obtain the most complete description as possible. Three ways of integration are possible. In the late integration frame, the processing is done independently of each media, and the results are merged in a second time. The main difficulty of such a method comes from the merging operation since there is not coherence between the results provided by each media, and usually no satisfying way to solve this point.

The second way is to give a leading role to one of the modality. We experimented such a solution with image as the leading media. To help the characterization of the shots of a tennis reports, each shot was characterized by the presence, during this shot, of speech, applauds or ball noise. Sound is seen as a complementary source of information and allows to improve the results obtained previously with images only. Such a solution appears to be not fully satisfactory since only a small portion of the information carried by the sound track can be taken into account.

The third way we plan to study is early integration, where the different media are mixed from the beginning and all decision are taken based of the whole stream of information.

Such integration frames must be supported by a foundation technique which allow to handle them. Hidden Markov Models where chosen first, due to their nice properties to represent temporal streams and their ability to represent a priori knowledge about tennis and production rules. A hierarchical model was used to represent the complete structure of a tennis report, ans the Viterbi algorithm was used in order to identify this structure

from the video stream. A problem with this model is that the time model it used is not flexible enough to handle properly the different time granularity of the various media.

We propose to use another kind of models, segment models, to circumvent this problem. In these models, each state does not correspond to a unique observation as it is the case in HMM, but can correspond to a variable number of observations. It is also possible to use several streams of observations. On the other hand, the models are more complicated and their use is more costly. The use of these models is the subject of the thesis of M. Delakis.

6.3.2. Image and Text Joint Description

Participants: Patrick Gros, Hervé Renault, Pascale Sébillot.

Key words: *image - text interaction.*

In text retrieval engines, images are not taken into account. When an image retrieval engine exists aside the previous one, it treats the images independently of the text surrounding them. Of course, it should be better to couple these two engines or, at least, to couple the information that both media can provide.

The first way to do this is to determine the parts of the text which are related to images. This should allow to get a textual description of the images, and to make textual query to retrieve images in a much richer context than that if systems using simple keywords linked with the images.

In the other way, it is possible to find documents containing a same image and to use both texts to disambiguate or improve the understanding of the text. These two points are the thesis subject of H. Renault.

7. Contracts and Grants with Industry

7.1. Industrial Contracts

7.1.1. Contract with Thomson

Participants: Patrick Gros, Ewa Kijak, Sid-Ahmed Berrani.

The thesis of S.A. Berrani and E. Kijak were supported by a CIFRE grant in the frame of a contract between Thomson and TEXMEX.

7.1.2. Contract with France Télécom

Participants: Annie Morin, Nicolas Bonnel.

The thesis of N. Bonnel was supported by a CIFRE grant in the frame of a contract between Thomson and TEXMEX.

7.1.3. Contract with Socio-logiciel

Participant: Annie Morin.

For one year, we collaborate with a service company Socio logiciel working in statistics and data analysis in the marketing area. We were consultants on some statistical problems and we have to report on new statistical methods. This contract was conducted with the laboratory ERIC of the university of Lyon 2.

7.2. Contracts in the frame of National Networks of Technological Research

7.2.1. PRIAM Médiaworks Project

Participant: Patrick Gros.

Key words: *TV archives, video databases.*

This is a joint project with the VISTA team of IRISA. Duration 46 months, Start date: September 2000. Partners: LIMSI - CNRS, AEGIS, INRIA (VISTA, TEXMEX, and IMEDIA projects), TF1.

The Mediaworks project was created in the frame of the PRIAM program and the French information society program, financed by the Ministry of industry. It began on September 1st 2000. This project gathers the TF1

TV channel, the LIMSI lab, AEGIS (a SME) and INRIA (projects IMEDIA of the INRIA Rocquencourt and VISTA). It concerns the development of a system to assist documentalists who index TV archives. Its principal features are the cooperation between the text and image media, and the development of a semantic search engine. TEXMEX works together with VISTA to develop tools for automatic structuring of video in plans and for computing an iconic representation of these plans.

7.2.2. RNRT Diphonet Project: Photo Diffusion on Internet

Participants: Laurent Amsaleg, Patrick Gros, Sophie Le Delliou.

Key words: *copyright protection, image databases, image recognition, piracy.*

Duration : 30 months, starting January 2002. Partners: Canon, L2S, Andia Presse, INRIA (projects CODES, TEMICS and TEXMEX).
Copyright protection is a key component to allow photo holders like photo agencies to diffuse their collections through the Web. It seems today impossible to avoid skilled hackers to thrust their way in Web sites and rob images. Therefore, legal image holders need a way to check whether the images available on a third party site do not originate from their own database (DB) of images. This is particularly crucial if that third party is making money by selling the images it pretends to possess. Watermarking is a first solution to this problem, but it requires a complex organization to become a legal argument. Moreover, pirated images are often washed out in order to remove the inserted marks.

This project addresses the problem enforcing the protection of copyright by relying on a content-based image retrieval (CBIR) scheme for that. The idea is provide a tool allowing professionals (e.g. photo agencies) to check if a published image comes from their DBs using only visual similarity. Its goal is therefore to detect matches between a set of doubtful images (e.g. downloaded from the Web) and the ones stored in the DB of the legal holders of photographies. If an image was indeed stolen and used to create a pirated copy, it tries to identify from which original image the pirated copy comes from.

So far, we performed extensive experiments showing that our image description scheme was useful in this context.

7.2.3. RIAM Annapurna Project

Participants: Patrick Gros, David Bonnaud.

Key words: *Personal photo archives, image indexing.*

Duration 12 months, Start date: January 2003. Partners: Thomson, LTU, CLIPS-IMAG.
The ANAPURNA project is a small project focused on demonstrating the feasibility of a management system of pictures on a digital set-top box in a familial context of use. The projects concerns the definition of the possible ergonomy of such a system, and brings together three technologies: image description, image indexing and search, and tuning of the system on an experimental platform developed by Thomson.

TEXMEX was in charge of the indexing and retrieval part of the project, which allowed to transfer to Thomson the results of Sid-Ahmed Berrani's thesis.

7.2.4. RIAM FERIA Project

Participants: Patrick Gros, Christophe Garcia, Manolis Delakis.

Key words: *TV archives, video databases.*

Duration 24 months, Start date: october 2003. Partners: Communications et Systèmes, INA, IRIT, Canal+ Technologie, Vecsys, Arte France.
The FERIA project aims at developing a framework for the development of multimedia applications in the domain of archive diffusion and valorization. This framework should allow to develop easily applications in the domain of multimedia production. These applications, in a second stage, will be used to produces DVD, web sites or other products.

Within this project, TEXMEX is in charge of still image analysis (logo and text detection, face detection and recognition), and of coordinating a research group on multimedia description of video documents.

7.3. European Contracts

7.3.1. European IST Project BUSMAN : Bringing User Satisfaction to Media Access Networks

Participant: Laurent Amsaleg.

Key words: *video, indexing videos.*

Duration : 30 months. Partners: IRISA (projet TEMICS et TEXMEX), Motorola, Telefonica, Technical University Munich, Queen Mary University of London, BTexact Technologies, Heinrich-Hertz Institute Berlin, FramePOOL.

This project is concerned with the design of new algorithms for indexing and watermarking video streams in order to create new multimedia-related services. Our contributions are focused on the architecture of the indexing and search schemes involved in the design of the database of videos.

8. Other Grants and Activities

8.1. Regional Contracts

- L. Berti-Équille collaborates with INSERM U522 on the problem of management and cleaning of data issued from hepatic transcriptome experiments.
- L. Berti-Équille collaborates with INRA Animal Genetics on the problem of management and sharing of data on gene interaction.

8.2. National Contracts

8.2.1. ACI Grid GénoGRID

Participant: Laurent Amsaleg.

Key words: *reconfigurable architecture, FPGA, genomics.*

Joint work with SYMBIOSE, ADEPT and R2D2.

The goal of this ACI is to provide a portal allowing the access to shared computing resources geographically distributed in order to boost bio-related algorithms (such as DNA alignments for instance). Our team is involved in the design of the architecture of the grid.

8.2.2. Action Bio-info inter-EPST: Parallel and Reconfigurable Architectures for Genomic Data Extraction

Participant: Laurent Amsaleg.

Key words: *architecture reconfigurable, FPGA, genomics.*

Joint work with SYMBIOSE and R2D2.

This works aims at defining a specialized highly-parallel architecture devoted to process large amount of data such as genomics sequences. This architecture is based on FPGA. We are involved in its design.

8.2.3. Inter-EPST Bio-informatic Action Caderige-2: Automatic Document Categorization for the Extraction of Gene Interaction Networks

Participant: Pascale Sébillot.

Key words: *Information extraction, bio-informatics, genomic data.*

Inter-EPST bio-informatic action from CNRS, INSERM, INRA, INRIA, Ministère de la Recherche, grouping people from Symbiose and TexMex teams at Inria, plus the following laboratories: Leibniz/Imag, LIPN, LRI, MIG/INRA and Inra-Ensar. This action, which has stopped at the end of October 2003, aimed at discovering within bio-informatic textual databases (MedLine), texts dealing with gene interactions, and at extracting the

interaction networks from those texts. Our participation concerned the automatic detection of text segments that describe the interactions.

8.2.4. R & D INRIA Action SYNTAX

Participant: Pascale Sébillot.

Key words: *Document analysis, information retrieval.*

This national action, coordinated by L. Romary (Loria), concerns information retrieval within electronic textual databases. Together with industrial firms, it aims at developing a software chain able to capture, analyze, and search through textual documents, using and grouping research solutions proposed by different INRIA teams.

8.2.5. ACI Masse de données MDP2P

Participants: Laurent Amsaleg, Patrick Gros.

Key words: *peer-to-peer, multimedia.*

Joint work with Atlas, Gemo and Paris.
We investigate, in this 3 years project, how peer-to-peer (P2P) distributed architectures can be used to cope with the fast increasing volumes of numeric data (such as text and multimedia data) often stored in autonomous, heterogeneous and distributed equipments. The potential advantages of P2P systems are node autonomy, scale up to large numbers of nodes, high availability (through replication) and performance (through parallelism). Thus, the main objective of the project is to provide high-level services for managing text and multimedia data in large-scale P2P systems.

Our role in this project is focused on the management (querying and retrieval) of multimedia data.

8.2.6. ACI Masse de données REMIX

Participant: Laurent Amsaleg.

Key words: *architecture, very large main memory, image retrieval.*

Joint work with Symbiose, R2D2 and Equipage (Valoria, Vannes).
The REMIX project proposes the design of a dedicated and very large RAM index memory (several hundreds of Giga bytes), big enough to entirely store huge indexes in main memory, avoiding the use of any disk. The use of an almost unlimited main memory raises completely new issues when designing indexes and allows to entirely revisit the principles that are at the root of almost all existing indexing strategies. Here, within this scheme, direct access to data, massive parallel processing, huge data redundancy, precomputed structures, etc., can be advantageously promoted to speed-up the search.

Our role in this project is focused on the design of main-memory oriented index structures dedicated to boost content-based retrievals.

8.2.7. ACI Jeunes Chercheurs TEXMEX

Participants: Laurent Amsaleg, Laure Berti-Équille, Patrick Gros.

This program of the French ministry of Research aims at helping the creation of new research teams by young researchers. In the frame of this program, Texmex received 103 kEUR for 3 years.

8.2.8. Participation to National Working Groups

- L. Amsaleg participates to the AS of the STIC department of CNRS: multimedia data, interrogation and storage.
- L. Berti-Équille participates to the GafoQualité group of AS GafoDonnées of the STIC department of CNRS.
- L. Berti-Équille participates to the "Documents Mutimédia" and "Médiation" working groups of GDR I3.

- L. Berti-Équille participates to the AS "Médiation d'informations via les méta-données" of the STIC department of CNRS.
- P. Gros is a member of the stearing committees of the RTP 25 (Computer Vision) and RTP33 (Documents and Contents: creation, indexing, browsing) of the STIC department of CNRS.
- P. Gros and L. Amsaleg participate to the AS image mining of STIC department of CNRS.
- P. Sébillot is a member of the thematic network "Information and knowledge: discovering and abstracting" of the STIC department of CNRS.
- P. Sébillot is member of the AS "Semantic Web" of the STIC department of CNRS.
- P. Sébillot is a member of AFIA Café (Collège apprentissage, fouille et extraction).
- P. Sébillot is a member of the working group A3CTE: Application, Learning and Knowledge Acquisition from Electronic Texts of GDR I3.

8.3. International Collaborations

8.3.1. ERCIM Working Group on Image Undertanding

Participant: Patrick Gros.

This working group aims at encouraging research activities in video and image analysis and understanding among the members of ERCIM. Its main action was to organize the MUSCLE consortium which has been accepted as a Network of excellence in the 6th Framework program.

8.3.2. Collaboration with NII - Japan

Participant: Laure Berti-Équille.

Key words: *metadata*, *culture*, *geography*.

A MOU (Memorandum Of Understanding) has been signed between IRISA (TEXMEX Team) and NII - National Institue of Informatics - of Tokyo - Japan in 2002, to provide a general framework to initiate an M4 project (metadata and multimedia management).

8.3.3. PAI Jules Verne with University of Reykjavik

Participants: Laurent Amsaleg, Patrick Gros.

This work is supported by Égide.
Image databases and therefore content-based retrieval systems have become increasingly important in many applications areas. While extremely effective (they return high quality results), these techniques are very inefficient (they answer very slowly) due to their complexity and because of the inadequacy of traditional operating system support.

The goal of this project is to develop techniques that integrate efficiency and effectiveness in content-based image retrieval systems. The long-term benefits of this work are expected to be much improved image retrieval systems that are key for emerging applications.

9. Dissemination

9.1. Conference, Workshop and Seminar Organization

- A. Morin organized an invited paper session during the meeting of the International statistical institute in Berlin, august 2003 on Impact of developments in information systems on statistics education.

- In the frame of the AS image mining of the STIC Deparment of CNRS, A. Morin coorganized with Thierry Denoeux a meeting on data analysis, statistics and learning for image mining on the 1st of April 2003. L. Amsaleg and P. Gros organized another meeting on image mining and databases.
- P. Sébillot was a member of the program committee and the organization committee of TALN'03 (Traitement automatique des langues naturelles) in June 2003 at Batz sur Mer.
- P. Sébillot was a member of the organization committee of RECITAL 2003 (Young researcher conference associated with TALN'03).
- P. Gros was a member of the organizing committee and of the program committee of the European Content-Based Multimedia Indexing workshop which took place in Rennes in September 2003.

9.2. Animation of the Scientific Community

- L. Berti-Équille was a member of the program committee of the conference INFORSID 2003.
- L. Berti-Équille was a member of the program committee of the conference SEMSOFT 2003.
- P. Gros is associate editor of the journal "Traitement du signal".
- P. Gros is a member of the scientific board of Universtiy of Rennes 1.
- P. Sébillot is associate editor of the journal In Cognito.
- P. Sébillot is associate editor of Jedai (Journal Électronique d'Intelligence Artificielle).
- P. Sébillot was a member of the programm committee of GL'03 (Generative approaches to the Lexicon, Genève, Suisse, 15-17 May 2003.)
- P. Sébillot was a member of the programm committee of the Workshop "Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel", plate-forme AFIA2003, Laval, 1-4 juillet 2003.
- P. Sébillot is a member of the board of ATALA (Association pour le traitement automatique des langues).
- P. Sébillot is a member of the scientific committee of the TCAN program od CNRS (Traitement des conaissances, apprentissage et NTIC).

9.3. Teaching Activities

- DEA Computer Science, Rennes. P. Sébillot, L. Amsaleg and P. Gros : Multimedia Indexing: Techniques and Applications.
- DESS Mitic, Ifsic, Rennes 1. L. Amsaleg, P. Gros et P. Sébillot: Digital documents indexing and retrieval. L. Amsaleg and P. Sébillot: data management. L. Amsaleg and P. Sébillot: Databases.
- INSA Rennes, 5th year. L. Berti-Équille : bioinformatics - biological data management.
- ENST Bretagne, 3rd year. L. Berti-Équille : dataware houses and data mining.
- Diic2, LSI, 2nd year. L. Amsaleg: databases.
- IUP Miage, 3rd year. L. Amsaleg: Datawarehouses and Datamining.

9.4. Participation to Seminars, Workshops, Invitations

- P. Sébillot gave an invited talk to the Workshop "Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel", plate-forme AFIA2003, Laval, 1-4 juillet 2003.
- P. Gros was invited to a France - Taïwan seminar of recent issues in Multimedia (March 2003)

10. Bibliography

Major publications by the team in recent years

[1] L. AMSALEG, P. GROS. *Content-based Retrieval Using Local Descriptors: Problems and Issues from a Database Perspective.* in « Pattern Analysis and Applications », number 4, volume 2001, 2001, pages 108-124.

[2] J. ANDRÉ, A. MORIN, H. RICHY. *Comparison of Literary Texts using Biological Sequence Comparison and Structured Documents Capabilities.* in « Proceedings of the ICCLSDP, Calcutta, Inde », February, 1998.

[3] L. BERTI-EQUILLE. *Annotation et recommandation collaboratives de documents selon leur qualité.* in « Revue ISI-NIS, Numéro spécial Recherche et filtrage d'information », number 1-2/2002, volume 7, 2002, pages 125-156.

[4] V. CLAVEAU, P. SÉBILLOT, P. BOUILLON, C. FABRE. *Acquérir des éléments du lexique génératif : quels résultats et à quels coûts ?.* in « Traitement automatique des langues, numéro spécial Lexiques sémantiques », number 3, volume 42, 2001, pages 729-753.

[5] B. LAMIROY, P. GROS. *Rapid Object Indexing and Recognition Using Enhanced Geometric Hashing.* in « Proceedings of the 4th European Conference on Computer Vision, Cambridge, Angleterre », volume 1, pages 59-70, April, 1996.

[6] R. PRIAM, A. MORIN. *Visualisation des corpus textuels par treillis de multinomiales auto-organisees - Généralisation de l'analyse factorielle des correspondances.* in « Revue Extraction des Connaissances et Apprentissage (Actes EGC'2002) », number 4, volume 1, 2002, pages 407-412.

[7] M. ROSSIGNOL, P. SÉBILLOT. *Automatic Generation of Sets of Keywords for Theme Characterization and Detection.* in « Sixièmes journées internationales d'analyse statistique des données textuelles », A. MORIN, P. SÉBILLOT, editors, Saint-Malo, France, 2002.

Doctoral dissertations and "Habilitation" theses

[8] V. CLAVEAU. *Acquisition automatique de lexiques sémantiques pour la recherche d'information.* Ph. D. Thesis, University of Rennes 1, France, December, 2003.

[9] E. KIJAK. *Structuration multimodale des vidéos de sport par modèles stochastiques.* Ph. D. Thesis, University of Rennes 1, France, December, 2003.

[10] R. PRIAM. *Méthodes de carte auto organisatrice par mélange de lois contraintes. Application à l'exploration dans les tableaux de contingence textuels.* Ph. D. Thesis, University of Rennes 1, France, October, 2003.

Articles in referred journals and book chapters

[11] L. AMSALEG, P. GROS, S.-A. BERRANI. *Robust Object Recognition in Images and the Related Database Problems.* in « Special issue of the Journal of Multimedia Tools and Applications », 2003.

[12] S.-A. BERRANI, L. AMSALEG, P. GROS. *Recherche approximative de plus proches voisins : application à*

la reconnaissance d'images par descripteurs locaux. in « Technique et Science Informatiques », number 9, volume 22, 2003, pages 1201–1230.

[13] L. BERTI. *Quality-based recommendation of XML documents.* in « Journal of Digital Information Management », number 3, volume 1, September, 2003, pages 117-128.

[14] L. BERTI. *Renseigner la qualité des connaissances par la fusion d'indicateurs sur la qualité des données.* in « Revue RSTI-ECA (Actes EGC'2003), Hermès », number 1-2-3, volume 17, 2003.

[15] V. CLAVEAU, P. SÉBILLOT, C. FABRE, P. BOUILLON. *Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming.* in « Journal of Machine Learning Research, special issue on Inductive Logic Programming », volume 4, August, 2003, pages 493–525.

[16] R. PRIAM, A. MORIN. *Visualisation des corpus textuels par Treillis de Multinomiales Auto-organisees - Propriété, Navigation Interactive Multicritère.* in « Revue RSTI-ECA (Actes EGC'2003), Hermès », number 1-2-3, volume 17, 2003, pages 549-550.

Publications in Conferences and Workshops

[17] J.-A. BENVENUTI, L. BERTI, ?. JACOPIN. *Ontological Parsing of XML Documents: A Use Case in the domain of training French military staff.* in « 21st ICDE World Conference on Open Learning and Distance Education », Hong-Kong, 2003/2004.

[18] S.-A. BERRANI, L. AMSALEG, P. GROS. *Approximate Searches: k-Neighbors + Precision.* in « Proc. of the 12th ACM International Conference on Information and Knowledge Management », pages 24–31, La Nouvelle Orléans, Louisiane, USA, November, 2003.

[19] S.-A. BERRANI, L. AMSALEG, P. GROS. *Probabilistically Controlling the Precision of Approximate Nearest-Neighbor Searches.* in « Actes des 19e journées Bases de Données Avancées, BDA'03 », pages 89–108, Lyon, France, October, 2003.

[20] S.-A. BERRANI, L. AMSALEG, P. GROS. *Robust Content-Based Image Searches for Copyright Protection.* in « Proc. of the ACM International Workshop on Multimedia Databases », pages 70–77, La Nouvelle Orléans, Louisiane, USA, November, 2003.

[21] L. BERTI. *Quality-extended query processing for distributed sources.* in « International Workshop on Data Quality in Cooperative Information Systems, DQCIS'2003 », Siena, Italy, January, 2003.

[22] V. CLAVEAU. *Extraction de couples nom-verbe : une technique symbolique automatique.* in « Actes de la 10e conférence de Traitement automatique des langues naturelles (TALN'03) », Batz-sur-Mer, France, June, 2003.

[23] V. CLAVEAU, P. SÉBILLOT. *Apprentissage symbolique pour l'acquisition de ressources linguistiques.* in « Actes de l'atelier "Acquisition, apprentissage et exploitation de connaissances sémantiques pour l'accès au contenu textuel" de la plateforme AFIA », Laval, France, July, 2003.

[24] M. DELAKIS, C. GARCIA. *Training Convolutional Filters for Robust Face Detection.* in « Proc. of the IEEE international Workshop of Neural Networks for Signal Processing (NNSP'03) », pages 739-748, Toulouse,

France, September, 2003.

[25] E. KIJAK, G. GRAVIER, P. GROS, L. OISEL, F. BIMBOT. *HMM based structuring of tennis videos using visual and audio cues.* in « IEEE Int. Conf. on Multimedia and Expo, ICME'03 », USA, July, 2003.

[26] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Audiovisual Integration for tennis broadcast structuring.* in « International Workshop on (CBMI'03) », Rennes, France, September, 2003.

[27] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Structuration multimodale d'une vidéo de tennis par modèles de Markov cachés.* in « Colloque GRETSI sur le traitement du signal et des images », Paris, France, September, 2003.

[28] E. KIJAK, P. GROS, L. OISEL. *Temporal structure analysis of broadcast tennis video using hidden Markov models.* in « SPIE Storage and Retrieval for Media Databases », Santa Clara, USA, January, 2003.

[29] E. KIJAK, L. OISEL, P. GROS. *Hierarchical structure analysis of sport videos using HMMS.* in « IEEE Int. Conf. on Image Processing, ICIP'03 », Barcelone, Espagne, September, 2003.

Miscellaneous

[30] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Structuration multimodale d'une vidéo de tennis par modèles de Markov cachés.* in « Actes des journées francophones des jeunes chercheurs en analyse d'images et perception visuelle (ORASIS'03) », Gerardmer, France, May, 2003.

[31] R. PRIAM. *Vers de nouvelles représentations des cartes auto-organisatrices. Exemple de CASOM et des données textuelles.* in « Atelier Visualisation et Extraction Adaptative des Connaissances - EGC'2003 », 2003.

Bibliography in notes

[32] K. BENNETT, U. FAYYAD, D. GEIGER. *Density-Based Indexing for Approximate Nearest-Neighbor Queries.* in « Proc. of the 5th ACM Int. Conf. on Knowledge Discovery and Data Mining, San Diego, CA USA », August, 1999.

[33] S. BERCHTOLD, D. KEIM, H. KRIEGEL. *The X-tree : An Index Structure for High-Dimensional Data.* in « VLDB », 1996.

[34] K. BEYER, J. GOLDSTEIN, R. RAMAKRISHNAN, U. SHAFT. *When Is "Nearest Neighbor" Meaningful?.* in « Proc. of the 8th Int. Conf. on Database Theory, London, U. K. », January, 1999.

[35] T. BINFORD, T. LEVITT. *Quasi-Invariants: Theory and Exploitation.* in « Proceedings of DARPA Image Understanding Workshop », pages 819-829, 1993.

[36] N. BOUJEMAA, S. BOUGHORBEL, C. VERTAN. *Soft Color Signatures for Image Retrieval by Content.* in « Eusflat'2001 », volume 2, pages 394-401, 2001.

[37] P. BOUTHEMY, M. GELGON, F. GANANSIA. *A Unified Approach to Shot Change Detection and Camera Motion Characterization.* in « IEEE Transactions on Circuits and Video Technology », number 7, volume 9,

October, 1999, pages 1030-1044.

[38] C. BÖHM, S. BERCHTOLD, D. KEIM. *Searching in High-dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases.* in « ACM Computing Surveys », number 3, volume 33, September, 2001.

[39] F. CHAUMETTE. *De la perception à l'action : l'asservissement visuel, de l'action à la perception : la vision active.* Habilitation à diriger des recherches, Université de Rennes 1, January, 1998.

[40] P. CIACCIA, M. PATELLA. *PAC Nearest Neighbor Queries: Approximate and Controlled Search in High-Dimensional and Metric Spaces.* in « Proc. of the 16th Int. Conf. on Data Engineering, San Diego, California, USA », February, 2000.

[41] Y. DUFOURNAUD, C. SCHMID, R. HORAUD. *Appariement d'images à des échelles différentes.* in « Actes du 12e Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, Paris, France », volume 2, pages 327-336, February, 2000.

[42] B. ESPIAU, F. CHAUMETTE, P. RIVES. *A New Approach to Visual Servoing in Robotics.* in « IEEE Transactions on Robotics and Automation », number 3, volume 8, June, 1992, pages 313-326.

[43] R. FABLET. *Modélisation statistique non paramétrique et reconnaissance du mouvement dans des séquences d'images : application à l'indexation vidéo.* Ph. D. Thesis, Université de Rennes 1, July, 2001.

[44] H. FERHATOSMANOGLU, E. TUNCEL, D. AGRAWAL, A. E. ABBADI. *Approximate Nearest Neighbor Searching in Multimedia Databases.* in « Proc. of the 17th Int. Conf. on Data Engineering, Heidelberg, Germany », pages 503–511, April, 2001.

[45] G. FINLAYSON, S. CHATTERJEE, B. FUNT. *Color Angular Indexing.* in « Proceedings of the 4th European Conference on Computer Vision, Cambridge, Angleterre », pages 16-27, 1996.

[46] G. FINLAYSON, M. DREW, B. FUNT. *Color Constancy: Generalized Diagonal Transforms Suffice.* in « Journal of the Optical Society of America A », number 11, volume 11, November, 1994, pages 3011-3019.

[47] L. FLORACK, B. TER HAAR ROMENY, J. KOENDERINK, M. VIERGEVER. *General Intensity Transformation and Differential Invariants.* in « Journal of Mathematical Imaging and Vision », number 2, volume 4, 1994, pages 171-187.

[48] U. GARGI, R. KASTURI, S. ANTANI. *Performance characterization and comparison of video indexing algorithms.* in « Proceedings of the Conference on Computer Vision and Pattern Recognition, Santa Barbara, Californie, États-Unis », pages 559-565, June, 1998.

[49] J. GERBRANDS. *On The Relationships Between SVD, KLT and PCA.* in « Pattern Recognition », number 1-6, volume 14, 1981, pages 375–381.

[50] A. GIONIS, P. INDYK, R. MOTWANI. *Similarity Search in High Dimensions via Hashing.* in « Proc. of the 25th Int. Conf. on Very Large Data Bases, Edinburgh, Scotland, UK », pages 518–529, September, 1999.

[51] J. GOLDSTEIN, R. RAMAKRISHNAN. *Contrast Plots and P-Sphere Trees: Space vs. Time in Nearest Neighbor Searches.* in « Proc. of the 26th Int. Conf. on Very Large Data Bases, Cairo, Egypt », pages 429–440, September, 2000.

[52] P. GROS. *Experimental Evaluation of Color Illumination Models for Image Matching and Indexing.* in « Proceedings of the RIAO'2000 Conference on Content-Based Multimedia Information Access », pages 567-574, April, 2000.

[53] A. GUTTMAN. *R-Trees: A Dynamic Index Structure for Spatial Searching.* in « ACM SIGMOD », 1984.

[54] R. HAMMOUD, R. MOHR. *Mixture Densities for Video Objects Recognition.* in « Proceedings of the 15th International Conference on Pattern Recognition, Barcelone, Espagne », volume 2, IAPR, pages 71-75, September, 2000.

[55] C. HARRIS, M. STEPHENS. *A Combined Corner and Edge Detector.* in « Proceedings of the 4th Alvey Vision Conference », pages 147-151, 1988.

[56] A. HENRICH. *The LSD^h-Tree: An Access Structure for Feature Vectors.* in « ICDE », 1998.

[57] J. HUANG, S. R. KUMAR, M. MITRA, W. ZHU, R. ZABIH. *Image Indexing Using Color Correlograms.* in « Proceedings of the Conference on Computer Vision and Pattern Recognition, Puerto Rico, USA », pages 762-768, June, 1997.

[58] V. KASHYAP, A. SHETH. *Semantic Heterogeneity in Global Information Systems: The Role of Metadata, Context and Ontologies.* M. PAPAZOGLOU, G. SCHLAGETER, editors, in « Cooperative Information Systems », Academic Press, San Diego, Californie, États-Unis, 1998, pages 139-178.

[59] N. KATAYAMA, S. SATOH. *The SR-tree: An Index Structure for High-Dimensional Nearest Neighbor Queries.* in « ACM SIGMOD », 1997.

[60] F. KORN, B. PAGEL, C. FALOUTSOS. *On the 'Dimensionality Curse' and the 'Self-Similarity Blessing'.* in « IEEE Trans. on Knowledge and Data Engineering », number 1, volume 13, January, 2001, pages 96–111.

[61] I. LERMAN. *Foundations in the Likelihood Linkage Analysis Classification Method.* in « Applied Stochastic Models and Data Analysis », volume 7, 1991, pages 69–76.

[62] C. LI, E. CHANG, H. GARCIA-MOLINA, G. WIEDERHOLD. *Clustering for Approximate Similarity Search in High-Dimensional Spaces.* in « IEEE Trans. on Knowledge and Data Engineering », number 4, volume 14, July, 2002, pages 792–808.

[63] E. LOUPIAS, N. SEBE, S. BRES, J.-M. JOLION. *Wavelet-based Salient Points for Image Retrieval.* in « Proceedings of the IEEE International Conference on Image Processing, Vancouver, Canada », 2000.

[64] E. MALIS, F. CHAUMETTE, S. BOUDET. *2 1/2 D Visual Servoing.* in « IEEE Transactions on Robotics and Automation », number 2, volume 15, April, 1999, pages 238-250.

[65] K. MIKOLAJCZYK, C. SCHMID. *An Affine Invariant Interest Point Detector.* in « Proceedings of the 7th European Conference on Computer Vision, Copenhague, Danemark », 2002.

[66] S. MUGGLETON, L. DE-RAEDT. *Inductive Logic Programming: Theory and Methods.* in « Journal of Logic Programming », volume 19-20, 1994, pages 629-679.

[67] J. NIEVERGELT, H. HINTERBERGER, K. SEVCIK. *The Grid File: An Adaptable, Symmetric Multikey File Structure.* in « ACM TODS », number 1, volume 9, 1984.

[68] B. PAGEL, F. KORN, C. FALOUTSOS. *Deflating the Dimensionality Curse Using Multiple Fractal Dimensions.* in « Proc. of the 16th Int. Conf. on Data Engineering », San Diego, California, USA, March, 2000.

[69] J. PUSTEJOVSKY. *The Generative Lexicon.* MIT Press, Cambridge, 1995.

[70] F. RASTIER. *Sémantique Interprétative.* edition Second, Presses universitaires de France, 1996.

[71] J. ROBINSON. *The K-D-B-Tree: A Search Structure For Large Multidimensional Dynamic Indexes.* in « ACM SIGMOD », 1981.

[72] R. RUILOBA, P. JOLY, S. MARCHAND-MAILLET, G. QUENOT. *Towards a Standard Protocol for the Evaluation of Video-to-Shots Segmentation Algorithms.* in « Proceedings of the first European Workshop on Content Based Multimedia Indexing, Toulouse, France », October, 1999.

[73] G. SALTON. *Automatic Text Processing.* Addison-Wesley, 1989.

[74] C. SCHMID, R. MOHR. *Local Grayvalue Invariants for Image Retrieval.* in « IEEE Transactions on Pattern Analysis and Machine Intelligence », number 5, volume 19, May, 1997, pages 530-534, ftp://ftp.inrialpes.fr/pub/movi/publications/schmid_pami97.ps.gz.

[75] M. STRICKER, M. SWAIN. *The Capacity of Color Histogram Indexing.* in « Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, Washington, USA », 1994.

[76] R. WEBER, K. BÖHM. *Trading Quality for Time with Nearest Neighbor Search.* in « Proc. of the 7th Conf. on Extending Database Technology, Konstanz, Germany », March, 2000.

[77] R. WEBER, H. SCHEK, S. BLOTT. *A Quantitative Analysis of Performance Study for Similarity-Search Methods in High-Dimensional Spaces.* in « Proceedings of the 24th International Conference on Very Large Data Bases, New York City, New York, États-Unis », pages 194-205, August, 1998.

[78] S. WERMTER, E. RILOFF, G. SCHELER, editors, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing.* Lecture Notes in Computer Science, Vol. 1040, Springer Verlag, 1996.

[79] D. WHITE, R. JAIN. *Similarity Indexing with the SS-tree.* in « ICDE », 1996.

[80] T. ZHANG, R. RAMAKRISHNAN, M. LIVNY. *BIRCH: An Efficient Data Clustering Method for Very Large Databases.* pages 103–114, June, 1996.