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 SYM

Table of contents

1. Team

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

Team Leader

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]
Vincent Claveau [Assistant Lecturer until January, 31st]

Faculty Member (CNRS)

Laurent Amsaleg [Research Scientist]

Ph.D. Students

Sid-Ahmed Berrani [CIFRE with Thomson, until February, 29th]
Nicolas Bonnel [CIFRE with France Télécom R&D]
Manolis Delakis [MJENR grant, also with METISS]
Stéphane Huet [MJENR grant, since October, 1st, also with METISS]
Zied Jemai [INRIA grant, since October, 1st]
Anicet Kouomou-Choupo [MJENR grant]
Fabienne Moreau [INRIA grant]
Xavier Naturel [INRIA grant]
Anthony Remazeilles [MJENR grant, also with LAGADIC, INSA assistant lecturer since October, 1st]
Hervé Renault [until September, 30th]
Mathias Rossignol [MJENR grant]
François Tonnin [INRIA - Britany grant, also with TEMICS]

Post-Doctoral Fellow

Panagiotis Hadjidoukas [MDP2P project]

Project Technical Staff

Philippe Daubias [FERIA project, since November, 1st]
Claire-Hélène Demarty [FERIA project, until September, 30th]
Sophie Le Delliou [DIPHONET project, also with TEMICS Project, until September, 30th]
Brigitte Fauvet [FERIA project, also with VISTA, since December, 1st]
Boris Rousseau [Enthrone Project, since October 1st]

2. Overall Objectives

Keywords: *Databases*, *Document Content-Based Access*, *Exploration*, *Image Recognition*, *Indexing*, *Machine Learning*, *Multimedia*, *Natural Language Processing*, *Search*.

The explosion of the quantity of numerical documents raises the problem of the management of these documents. Beyond the storage, we are interested in the problems linked to the management of the contents: how to exploit the large bases of documents, how to classify them, to index them to be able to search efficiently 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 data and related metadata exploitation techniques such as the database techniques, statistics, and information retrieval. Our work is at the intersection of these fields and relates more particularly to 3 points: i) searching in large

image databases, ii) adding semantics to search engines, and iii) coupling media for multimedia document description.

Content exploitation 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 taken into account (millions of images, months of videos), and to evaluate their results as well in quality as in speed;
- multimedia documents are not a simple juxtaposition of independent media, and it is important of **to better exploit the existing links between the various media** present in a same document;
- **multimedia document databases are evolutionary**: the sets of documents evolve, as do the document description techniques and the modes of questioning, which modifies in return the way the bases are used;
- towards queries of a semantic nature for their majority, 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 documents description and exploitation of these descriptions, and on the other hand, we defined three transverse axis of research. The underlying idea is to concentrate on the questions where the team multidisciplinarity appears an asset to obtain original results.

Our First Field of Competence: Document Description. 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, but must also be computable automatically. To the documents and their descriptors, one can add metadata, which we define here as all additional information which inform, supplement or qualify the data with which they are associated.

Our Second Field of Competence: Description Exploitation. The question is to define the techniques which make it possible to apprehend, handle and exploit large volumes of data, metadata and descriptors, which have been extracted from the documents: i) **organization and management of the multimedia data bases**, including the control of logical and temporal consistency, strategies of computation and selection of descriptors and metadata; ii) **statistical techniques** for the exploration of great volumes of data; iii) **indexing techniques** aiming at confining in the smallest possible volume the exploitation of the data and thus avoiding an exhaustive processing whose cost is certainly controlled but crippling; iv) **system problems** related to the physical organization of large volumes of data, like disk access management or cache memory management requiring new techniques which are adapted to the characteristics of the descriptors and to the way they are used.

First Axis 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 neither come from the only descriptors nor from new indexing schemes, but it requires to take into account all the various components of the system and their articulations. 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 speed as well as in quality of recognition.

Second Axis 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 based on keywords searching. Natural language processing tools could however provide them more semantic capabilities, by allowing word sense disambiguation and the possibility to recognize the various formulations of a same concept. It is thus advisable to marry these two techniques.
This union is, however, not so simple. On the one hand, it requires to provide query and document 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 environment 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 work on the evaluation of information retrieval systems (IRS).

Third Axis of Research: Multimedia and Coupling Between the Media Studying media coupling is undertaken in three 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 it is by their face or their voice. Another interesting direction consists of using natural language processing techniques to the result of speech transcription. As a matter of fact, speech carries a lot of semantic information and NLP techniques are among the most efficient ones to extract semantic from textual data.
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 refers 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 about a same subject, and thus to improve the description of the images and to remove possible ambiguities in the comprehension of the text.
Moreover since October 1st, we have also begun the study of the coupling between speech and text together with the METISS Team.

3. Scientific Foundations

3.1. Background

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 the tasks related to these documents. Finally, both the descriptors and exploitation techniques must satisfy the needs of the user (and proving this simple fact is not so 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, that are two constraints usually opposite.

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

3.2. Document Description and Metadata

Keywords: *Low-level 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 (sets of characters or of pixels). More concise and informative descriptions are needed in order to handle these documents.

3.2.1. Image Description

Keywords: *Image Indexing*, *Image Matching*, *Image Recognition*, *Invariants*.

Computing image descriptors has been studied for about thirty 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 which judges this similarity, on the problem this user tries to solve, and on the set of images he uses at this moment. 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 (*e.g.*, 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 has been proposed in the literature. The most frequent frame of use considered is that of 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 (with the major drawback that is to require an automatic image segmentation). This field of research is still active: color histograms provide too poor information to solve any problem as soon as the size of the database increases [105] and several solutions have been proposed to remedy this problem: correlograms [78], weighted histograms [46]...

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

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 opens the possibility of partial recognitions like object recognitions independently of the background [103].

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 too.) The most widely used extractor is the Harris [75] point detector which provides not very precise but "repeatable" points. Other detectors exist, even for points [88].

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 [45] or by properties established only experimentally [60][59].

In the case of points, the classical technique consists of characterizing the signal around each point by its convolution with the Gaussian 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 obtained respectively by Florack [61], Dufournaud [54] and Mikolajczyk [92], photometric invariance was demonstrated for grey-levels by Schmid [103] and for color by Gros [67]. 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 databases, especially when these are quite small. In this case, a simple syntactic criterion can give the feeling of a good semantic description, but this does not tell anything about what would happen with a larger database.

3.2.2. Video Description

Keywords: *Key-Events*, *Structuring*, *Video Indexing*.

Professional and domestic video collections are usually much bigger than the corresponding still image collections: a common factor is 1000 between the two. If the images often have a weaker quality (motion, fuzzy images...), they present a temporal redundancy which can be exploited 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 [57], or still image-based descriptors, but which can use the image temporal redundancy [74].

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 [47][62]. Nevertheless, shots appear to be a too low-level segment 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 his topic.

3.2.3. Text Description

Keywords: *Corpus-Based Acquisition of Linguistic Resources*, *Exploratory Data Analysis*, *Lexical Semantics*, *Machine Learning*, *Natural Language Processing*.

Automating indexing of textual documents [102] 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 questioned; second, dealing with the fact that the representation *is* a word-based one and not a concept-based one. Therefore information retrieval has to be able to overcome two semantic problems: various possibilities to formulate the same idea (how to match a concept in a text and a query 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 another. One solution is to make use of domain-specific linguistic resources, both to disambiguate words and to expand user queries with synonyms, hyponyms, etc. 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, though symbolic approaches also present a growing interest [37]. 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 [98]. We also acquire through symbolic machine learning (inductive logic programming [93]) noun-verb relations defined within J. Pustejovsky's Generative lexicon theory [97]; 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 [98] 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 from corpora lexicons based on Rastier's principles. First, with the help of a hierarchical classification method (Linkage Likelihood Analysis, LLA [86]) 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 [97], 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 [93], 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.

3.2.4. Retrieval and Description Evaluation

Keywords: *Discriminating Power*, *Evaluation*, *Performance*.

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 [101].

Setting protocols of evaluation that compare different content-based information systems (CBIR) is a very hard task, especially when considering the relevance feedback from users who submit an image or a video as query-by-example to the CBIR system. In this context, our idea is to automatically learn user profiles during the searching scenarii and to correlate some feedback indicators (non-intrusively collected) with the sets of descriptors used in the query to compute the results. Finally, the objective is to adapt the next query execution or the image/video browsing, with taking into account dynamically the last feedback.

3.2.5. Metadata Integrated Management, Selection and Mining

Keywords: *Automatic Selection*, *Integration*, *MPEG-21*, *MPEG-7*, *Metadata*, *Metadata Management*, *Standard*, *TV-Anytime*.

To improve the data organization or to define the strategies to compute some descriptors, it may be advisable to use additional information, called metadata. Metadata (data about the data) must describe the data sufficiently well as to be used as a surrogate for the data when making decisions regarding description and use of the data. Metadata can give complex information concerning structure description, semantics and contents of data items, their associated processes and –more widely– the respective domains of this various information.

Metadata are: i) data describing and documenting data, ii) data about datasets and usage aspects of them, iii) the content, quality, constraints, and other characteristics of data.

The documenting role of metadata is fundamental. This information can provide decision elements in order to choose the most appropriate dataset or processing techniques and also, the most appropriate data presentation mode. In the case of large amounts of data, it is difficult to analyze data content in a straight way. Metadata then give appreciation or description informative elements of the dataset.

However, metadata role is not restricted to documenting information. Metadata must also allow:

Data acquisition and transformation that are complex steps for data producers. Metadata can, on one hand, represent the production memory by describing operations carried out during data acquisition and transformation process, and it can, on the other hand, prevent a data producer from repeating the production step of an already existing dataset,

Description of structure and role of data, in order to allow its interpretation and treatment by a user, especially during transfer steps.

V. Kashyap and A. Sheth [80] proposed a first classification of metadata for multimedia documents in two main classes: metadata which contain external information (date, localization, author...) and metadata which contain internal information directly dependent on the content (such as low-level descriptors) or describing the content independently (such as keywords annotations) [39], [49], [51], [65], [79]. Many standardized metadata

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

such as in MPEG-7, TV-Anytime, etc. and also *ad hoc* content-descriptive metadata can be included in this classification [89], [91].

The key elements of the metadata managed by TEXMEX include (but are not limited to):

- Media description metadata (such as global descriptors –color, texture, motion, shape, etc. or local descriptors)– extracted from the images or from the bitstream and that are eventually formated as MPEG-7 or TV-AnyTime metadata or MPEG-21 Digital Item Declaration,
- Media usage metadata (such as relevance feedback in searching scenarii, access rights, availability, encryption, conditional access, etc.),
- User metadata (such as user preferences, usage history, etc.) and natural environment characteristics metadata (such as location, audio environment, illumination characteristics, etc.).

The selection and organization of metadata is highly application-dependent and also depends on the various objectives of metadata consumption that can facilitate: data access, data summary, data interoperability, media or content presentation and adaptation, etc.

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. But the bulk of potentially available metadata raises two important problems: the coherent and integrated management of metadata (usually formated in XML files) and the adaptive selection of relevant metadata relatively to each application. Our work is here to use exploratory data mining techniques to propose generic solutions for these two problems.

3.3. Efficient Exploitation of Descriptors and Metadata

Keywords: *Data Analysis*, *Data Quality*, *Indexing*, *Statistics*.

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 needs 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. On one hand, off-line processing corresponds 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 needs 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 and Data Quality over Huge Datasets

Keywords: *Data Quality Metrics*, *Exploratory Data Analysis*, *Sampling*, *Statistics*.

The situation where we have few available data has been well studied but a huge amount of data generates different kinds of problem: for instance, the use of classical inferential statistics results in hypothesis testing concludes rather often to reject the null hypothesis. Besides, the methods of models identification fail very often or the quality of the model is overestimated. The question is: how can we set a representative sampling in such datasets? We must add also 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.

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, or 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 watch the data to discover their properties.

Let us 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.

Many data analysis applications, such as multimedia mining or text mining, require various forms of data preparation with several data processing techniques, because the input to the data mining algorithms is assumed to conform to "nice" data distributions, containing no missing, inconsistent or incorrect values. This leaves a large gap between the available data and the available machinery to process the data. In fine, the evaluation of results obtained from data analysis is usually made by specialists (experts, analysts, etc.). The cost of this task is often very high, and the way to reduce it is to help the specialists while giving them relevant decision criteria as quality indicators or interest measures of results.

These measures of knowledge quality have to be designed in order to combine two dimensions: the objective dimension related to data quality, and the subjective dimension related to the specialists' focus of interest. Our work deals specifically with data quality issues and related knowledge discovery techniques. It intends to address methods, techniques of massive data analysis, methodologies, new algorithmic approaches or approaches to developing data quality metrics in order to understand and to explore data, to find data glitches and to ensure both data quality and knowledge quality discovered from data. In the context of data quality, we focus on techniques of:

- Detection of contradictory data, outliers, duplicates, inconsistencies, and noise,
- Mining for patterns of non- or poor quality data,
- Data transformations, reconciliation, consolidation,
- Data cleaning techniques.

3.3.2. Multidimensional Indexing Techniques

Keywords: *Approximate Searches*, *Curse of Dimensionality*, *Databases*, *Multidimensional Indexing Techniques*, *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 indexing 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 nearest-neighbors searches. We especially focus on the filtering rules these 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 [41][109] that divide the data space according to the distribution of data and *space-partitioning* [76][108] indexing methods that divide the data space along predefined lines regardless of the actual values of data and store each descriptor in the appropriate cell.

Data-partitioning index methods all derive from the seminal R-Tree [69], 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 [109] is an extension that relies on spheres instead of rectangles. The SR-Tree [81] specifies its cells as being the intersection of a bounding sphere and a bounding rectangle.

Space-partitioning techniques like grid-file [94], K-D-B-Tree [100], LSD^h-Tree [76] 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 cells that cannot have any impact on the result of the current query [48]. 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:

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

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

The search process 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:

$$\text{if } dmin(q, C_i) \geq d(q, nn_k) \text{ then stop,} \quad (2)$$

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 [108][44][96][48][82].

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.

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 [63]), exploit the underlying correlation of vectors and/or their self similarity [82], frequent with real datasets. NN-search schemes using dimension reduction techniques are approximated 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 cannot 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.

Early Stopping Approaches. Weber and Böhm with their approximate version of the VA-File [107] and Li et al. with Clindex [87] 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 [58], 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.

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. Cells containing interesting vectors might be filtered out, however.

In [107], 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 very different data distributions possible in cells.

The AC-NN scheme for M-Trees presented in [52] also relies on a single value ε set by the user. Here, ε represents the maximum relative error allowed between the distance from q to its exact NN and the distance from q to its approximate NN. In this scheme, setting ε is far from being intuitive. The 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 finds a vector closer than this estimated distance. Unfortunately, AC-NN and PAC-NN cannot search for k neighbors.

Hashing-based Approaches. Approximate NN-searches using locality sensitive hashing (LSH) techniques are described in [64]. 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, [64] 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 [64] might, in certain cases, return less than k vectors in the result.

Probabilistic Approaches. DBIN [40] 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 [66] 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 neither 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

Keywords: *Digital Pictures*, *Image Databases*, *Medical Imagery*, *Photo Agencies*.

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 images, Sigma twelve millions, the Corbis agency which gathers the whole of acquisitions of Bill Gates has thirty six millions of images. 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 subscribed 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 by indexing that the image is not a fragment of an image of the agency base.

Another important field where the management of the images acquires an increasing importance is that of the medical images. The access to the medically interesting contents of the image is a true difficulty, so is the level of quality imposed by this field to the recognition system. The applications of content-based methods are thus still to come in this field.

4.2. Video Database Management

Keywords: *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: private and professional files. On one hand, the files of private individuals include 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 have 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, are professional files: TV channels archives, registration of copyright, cineclubs, producers... These files are of a much larger size, but benefit from the attentive care of professionals of documentation and archiving. In this field, the 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

Keywords: *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 relating more on concepts than 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 makes 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 to provide the list of the most discriminating keywords for this or these documents. The index validation is carried out by searching information in a database more general than that used to build 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 on the automatic extraction, from texts of bioinformatics coming from bases such as Medline, of the zones of texts describing the interactions between genes and to model 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 binds between words are exploited to determine these relevant zones of texts. To a corpus of summaries extracted from Medline, we apply a training by inductive logic programming to try to learn what distinguishes the sentences containing interaction 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 free 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

Keywords: *Planning*, *Robotics*, *Visual Memory*, *Visual Servoing*.

If collaboration between robotics and vision is an already old subject, it underwent 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 it 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 [56][50][90]. The results are completely 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.

The control problem drives 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.)

5.1.2. Asares

is a symbolic machine learning system (based on inductive logic programming) that automatically infers, from descriptions of pairs of linguistic elements (noun-noun, noun-verb...) found in a corpus in which the components are linked by a given semantic relation (synonymy, hyperonumy, qualia...), corpus-specific morpho-syntactic and semantic patterns that convey the target relation. The patterns are explanatory and linguistically motivated, and can be applied to a corpus to efficiently extract resources and populate semantic lexicons. Two semi-supervised versions of Asares also exist, that rely on a combination of the supervised symbolic pattern learner and a statistical extraction technique. They both rival Asares's supervised version.

5.1.3. Caractopics

is a tool composed of a sequence of statistical treatments that extracts from a morpho-syntactically tagged corpus sets of keywords that characterize the main topics that it contains. The system exploits the distribution of words of the corpus over its paragraphs, and requires neither human intervention nor given knowledge about the number or nature of the topics of the corpus. The extracted lists are employed in order to detect the presence of a topic in a paragraph, revealed by a keyword cooccurrence.

6. New Results

6.1. Image Retrieval for 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? Without sticking too tightly to the JPEG'2000 format, we try to find new description schemes based on wavelet decomposition of images. 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 by using data mining techniques applied to visual descriptors.

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 LAGADIC 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 appears 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

Keywords: *Image Compression*, *Image Description*, *Image Indexing*.

Participants: Patrick Gros, François Tonnin.

This is a joint work with the TEMICS team (C. Guillemot).

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.

The main difficulty comes from the fact that the wavelets used for compression are critical and are very sensitive to small translations in the image. To gain invariance towards translations, we chose to use redundant wavelets which offer much better performances. The average robustness obtained is just 10 % less than the multiresolution Harris point extractor. We now investigate how to describe, in the wavelet space, the neighborhood of these points by means of vectors invariant to allowable image manipulations. Once more, the main difficulty is to find a wavelet description of the image which is locally invariant to geometric transforms. Our approach is currently based on D. Lowe's ideas. Another point we consider is the comparison of robustness and speed between classical and wavelet-based local signatures.

6.1.2. Approximate Searches: k-Neighbors + Precision

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

Participants: Laurent Amsaleg, Sid-Ahmed Berrani, Patrick Gros, Zied Jemai, Annie Morin, Panagiotis Hadjidoukas.

This is a joint work with Thomson R&D France. S.A. received the SPECIF thesis prize for his work on this topic.

We designed an approximate search-scheme for high-dimensional databases 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. We also present 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 $5\,106$ vectors of 24 dimensions, even when the probability of missing one of the true nearest-neighbors is below 0.01.

This work was done in the frame of Sid-Ahmed Berrani's thesis in collaboration with Thomson. We began to study a new approach in collaboration with the University of Reykjavik. The aim of this new method is to tend to a constant-time retrieval method which would be independent from the number of data in the database. At least this dependance should be as small as possible. This new method should also allow to revisit the problem of clustering. This is the subject of Zied Jemai's thesis.

In addition, another set of related issues is currently under investigation between researchers from IRISA and colleagues from the University of Reykjavik. These issues can be divided into the following main streams:

Multidimensional Indexing and Searching. Some work at IRISA clearly showed the benefits in using local description schemes for the automatic enforcement of the copyright on still images. To speed-up the retrieval process, a pre-clustering of images is required. In this case, it is possible to improve the response time of copyright enforcement queries by 90% without any significant loss in effectiveness. Yet, this improvement is not sufficient when dealing with realistic collection sizes of few millions of images. In addition, pre-clustering data does not scale due to its quadratic nature. We designed a very different indexing algorithm which goes 300 times faster than the sequential scan during the retrieval, and it can even be made faster by restricting resource consumption to only three I/O per query descriptor. In this case, the response time is flat and independent of the size of the image collection. We are currently in the process of evaluating our solution against the largest ever made database of images described with local descriptors and composed of more than 220 millions descriptors computed over 192.000 images occupying more than 22 Gigabytes on disk [85].

Clustering for Indexing. In parallel to the study mentionned above, we also investigated further issues related to the use of clustering algorithms for accelerating the indexing process and the subsequent retrieval of similar images. We first analysed existing clustering algorithms and investigated their adequacy to be used in the specific context of CBIR systems [55]. We then conducted an extensive performance study where we were comparing the advanced clustering schemed developed at IRISA and other more traditional schemes in order to understand the tradeoffs between the size of such clusters and the cost to process them [104]. This study shows that clusters based solely on proximity of data points are very effective in concentrating the neighbors but fail to provide any balance on their population, *i.e.*, they might frequently induce a large processing overhead due to the large number of points they each host. On the other hand, clusters defined solely on their population tend to overlap and therefore exhibit poor effectiveness – they are very efficiently processed however [104]. We are currently designing a clustering scheme mixing both approaches.

Another work with Panagiotis Hadjidoukas, as a post-doctoral fellow, aims at the development of parallel data clustering algorithms able to handle very large databases by exploiting the processing power of shared memory multiprocessors and clusters of compute nodes (this work is done in collaboration with the PARIS and ATLAS project-teams).

For this purpose, we developed PCURE, a Parallel implementation of CURE (Clustering Using REpresentatives) [68], a well-known and efficient hierarchical data clustering algorithm, using the OpenMP programming model [95]. Despite its efficiency, the worst-case time complexity of CURE is O(n^2logn), where n is the number of points to be clustered. The parallelization of the algorithm aims at enabling the direct use CURE on very large data sets and this is the first time that OpenMP has been applied to such an application.

The asymmetry and non-determinism of this algorithm necessitate the exploitation of its nested loop-level parallelism. Therefore, we also developed an alternative implementation of the NANOS OpenMP runtime library (NthLib) [70] that targets portability and efficient support of multiple levels of parallelism. In addition, we integrated this library into available open-source OpenMP compilers, reducing thus their overhead and providing them with inherent support for nested parallelism.

We are currently working on our OpenMP implementation on Software Distributed Shared Memory [73][106], in order to run PCURE on clusters of compute nodes. Furthermore, an extension of the OpenMP model for master-slave message passing computing [72][71] will enable us to use PCURE on heterogeneous clusters with respect to hardware and operating systems.

6.1.3. Visual Features Mining for Improving Content-Based Image Retrieval Performance

Keywords: *Association Rules*, *Content-Based Image Retrieval*, *Progressive Query*, *Query-by-example Execution Plans*, *Visual Features Mining*.

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

Still images can be retrieved by similarity searching on global visual features such as color, texture, layout or shape at the pixel level. Content-based retrieval systems then use and combine all the available low-level features whose computing cost can be prohibitive and they rank the images according to how well they match the submitted query-by-example. Finally, they return the best few matches to the user in a ranked result list. But, a subset of features could be sufficient enough to answer very quickly while offering an acceptable quality of results. Moreover, the administration of very large collections of images accentuates the classical problems of indexing and efficiently querying information. Our research focuses on the elaboration of fully automatic and generic strategies of visual global features usage for content-based retrieval on very large still image databases.

During 2004, our work, in the frame of the Ph.D. Thesis of Anicet Kouomou-Choupo, has especially consisted in the design of a method for automatic selection and scheduling of visual features as search criteria that are relevant to the query-by-example on a large still image database [83][84].This method is composed of two main steps: the indexing and the retrieval processes. The indexing step uses K-mean clustering and association rule discovery as dimensionality reduction techniques on global MPEG-7 descriptor values. The retrieval step is designed to order the query execution plans for speeding up the content-based retrieval over

the image database. At this step, query-by-example processing is adapted in order to propose instantaneous and intermediate results that are progressively merged together with the advantage, for the users, on one hand, not to wait until the whole database has been processed by similarity search and, on the other hand, to allow them to stop the current execution of the query without losing the first partial results. We have evaluated our method by comparing query execution time and result quality with the sequential search. Our experiments show that we can get a result similar to the final one in less than half the time of the sequential search, which is a promising result for optimizing content-based image retrieval.

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

Keywords: *Robot Motion control*, *Visual Servoing*.

Participants: Patrick Gros, Anthony Remazeilles.

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

We are working on automatic robot motion control, using visual information provided by an on-board camera, and an image database 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 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.

Our navigation scheme is based on the potential field theory. The robot moves in order to make the features defined on the image path and initially out of the camera field of view, become visible. The command law is defined inline, by using suitable potential functions, depending on the feature projections on the current image plane, and the image path. Explicit $3D$ reconstruction is not necessary any more.

During 2004, we have extended our method, in order to be able to deal with general motions and $3D$ environments. We have defined new potential functions which give the robotic system the ability to control its motions of translation along the optical axis, and of rotation around objects of the environment. Some simulations have been performed and confirm that the control law is correct for a six-degree of freedom robot.

We have performed first simulations in road-like or corridor-like environments. As expected, as long as the robotic system does not have any notion of obstacles, our method in its current state cannot avoid wall collisions. The next step should be to find a way to prevent those collisions, either by using the camera or another sensor to detect obstacles or by adding more constraints on the motions the robot can perform.

As long as the navigation is a large displacement, landmarks that are belonging to the camera field of view before the motion start are likely to disappear during the navigation. Visual landmark update is therefore a key problem that has to be considered. This year, we have proposed an automatic update of visible points, based on image transfer theory. Once more, the $3D$ model is not used; epipolar geometry, computed between the current view and the image path, is a sufficient information for making points projections. However, even if we have obtained some promising results on image sequences, this method should be improved to be used into for controlling a robotic system. Taking into account the structure of urban scenes, possible motions and using some statistical outlier rejection laws should be some ideas that could improve this landmark update.

At last, we have proposed during this year new methods for improving the image data base structure. In a first time, the graph structure has been changed in order to take directly in account motions that can perform a mobile robotic system. In a sense, the idea is that the image path provided by the graph corresponds to motions

that are easy to perform by a non-holomic system. In a second time, we have proposed a hierarchical graph definition, in order to reduce the search dimension for an image path.

In future works, it should be interesting to combine this visual description of the environment with information provided by other sensors, like GPS for exemple, and to define an hybrid GIS (Geographic Infornation System).

6.2. Text Retrieval for Large Databases

6.2.1. Natural Language Processing and Machine Learning

Keywords: *Corpus-Based Acquisition of Lexical Relations*, *Hierarchical Classification*, *Inductive Logic Programming*, *Lexical Semantics*, *Machine Learning*, *Natural Language Processing*, *Semi-Supervised Learning*.

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

Our research focuses on the elaboration of fully automatic and generic machine learning solutions –using both symbolic and statistical approaches– to extract from textual corpora linguistic resources needed by a given application. We mainly apply these solutions to the acquisition of semantic lexical relations that enrich the description of nouns, in order to both disambiguate them, and give access to semantic variants. Linguistic theories are used to determine the relevant lexical links, and to validate what we acquire. We particularly learn two kinds of relations: within F. Rastier's differential semantics framework, we acquire inter-categorial links (synonyms, but finer-grained ones too) with the help of hierarchical classification. We also focus on noun-verb relations defined in J. Pustejovsky's Generative lexicon theory, with the help of ASARES, our acquisition system based on inductive logic programming. We evaluate the interest of the semantic lexical resources that we get when inserted into an information retrieval system, trying to discover accurate solutions to the problem of their integration into the system.

During 2004, our work has especially concerned the 2 following points:

Acquisition of Semantic Lexicons Based on Rastier's Differential Semantics: Automatic generation of semantic classes within domain-specific corpora

In a previous work described in 2003 activity report, we have proposed a sequence of statistical data analysis treatments that enables 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 a (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, and thus to split the initial non-specialized corpus into several topic-specific ones. This gives us 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 domain-specific semantic classes from these subcorpora [16], work on which we have focused our attention during 2004. The principle of our method is to bring together words used in a similar way, and the main problem to address is the small amount of data available (about 400,000 words per topical corpus). We have developed a two-stage process to overcome that issue: a first classification tree is built with data collected on the whole corpus. An ultrametrics defined on that tree is then used as an assumed semantic distance between words when computing similarities on topical subcorpora. An ultrametrics is a metric that strictly verifies the triangular inequality. For that task, the similarity between two words is inferred from a similarity measure defined between their contexts, using a random sampling technique that makes the measure independent of the numbers of occurrences of the compared words. The classification trees obtained using that measure create relevant classes for about 2/3 of their words, but must be read manually to extract final semantic classes. Those constitute the starting point of our last task, the extraction of specific semantic information distinguishing the meanings of words within a semantic class. Current experiments aim at automatically detecting " revealing contexts " typical of a precise specific semantic marker.

Linguistic Resources and Information Retrieval (IR)

Our aim in this domain is to explore methods that enable information retrieval systems (IRS) to capture the semantics of natural language (NL) texts, and to exploit the semantic information that natural language

processing (NLP) techniques can automatically extract from textual documents. Currently, for example, systems based on Salton's vector space model (VSM) represent documents and queries as sets of words, without regard for the relationship (or even linear order) between them. Fabienne Moreau's Ph.D. thesis aims at adapting current IR models to allow information gleaned from NLP to inform IR. During 2004, we have realized a state-of-the-art of the wide domain of NLP+IR, that is, of the attempts that have already been conducted to partially insert linguistic resources into IRS. We have thus reviewed and compiled systems that integrate some morphological, syntactic or semantic information, at the indexing stage or for query expansion. We have also begun a study of the possibilities that the various kinds of IR models (VSM, boolean model, probabilistic model, neural network model, latent semantic indexing model, language modeling, etc.) offer to take linguistic resources into account.

6.3. Data Mining and Data Quality for Large Databases

6.3.1. Visualization and Web Mining

Keywords: *Exploratory Data Analysis*.

Participants: Nicolas Bonnel, Annie Morin.

Knowledge extraction from textual databases is not obvious. Among the used methods, we find factorial analysis, neural networks or Kohonen maps. We are currently working on the dynamic generation of 2D and 3D multimedia interactive presentations. That means how to represent the results of a search in a database. This work is done in cooperation with France Telecom R & D (in the context of a CIFRE contract). In this thesis subject, we can distinguish two main steps. The first one is how to organize efficiently the results of a web query. A good solution for that are the Kohonen self-organizing maps (which are an unsupersived clustering algorithm). Indeed, this approach enables to make groups of results and to have an organization of these groups. We study the distance used in this algorithm and compare it to others distances or dissimilarities. Another important point is the quality evaluation of the proposed classification. The second step concerns the visualization metaphors. The aim is to improve the graphical representation with the use of hybrid interfaces (interfaces composed by a 2D java applet and a 3D scene). We have many metaphors and an adaptive interface. The choice of the metaphor can be automatically or manually done. Finally, this work is integrated in a prototype developed by France Telecom R & D.

6.3.2. Data Quality: Measurement, Control and Correction

Keywords: *Data Cleaning Techniques*, *Data Quality Metrics*, *Quality-Adaptive Query Processing*.

Participant: Laure Berti-Équille.

The main challenges of exploratory data mining and of data quality in large databases are:

- Heterogeneity of data gathered from different sources: analyzing such data sets using a single method or a black box approach can produce misleading, if not totally incorrect results, because the combined information presented as a single large data set usually contains already a superposition of several statistical processes.
- Data quality: gathering data from different sources and scraping data off the web makes the information rich in content but poor in quality. It is hard to correlate data across sources since they are often no common keys to match on. For example, creating huge image databases or searching in several distributed multimedia databases usually dodge the problem of duplicates, truncated data, incomplete processing, errors, outliers or missing data detection.
- Scale: aside from the issues of collection, storage and retrieval of the sheer volume of the data, the analyst's focus is now to summarize the data meaningfully and accurately for efficient exploitation.

Data quality is a notoriously messy problem that refuses to be put into a neat container, and therefore is often viewed as technically intractable [53], [77], [99]. But data quality problems can be adressed with several methods from many disciplines such as statistics, database techniques and metadata [42].

For non-collaborative distributed data sources, both cost estimate-based query optimization and quality-driven query processing are difficult to achieve because the information sources do not export cost information nor data quality indicators. In our recent work [43], we propose an expressive query language extension using QML (QoS Modeling Language) syntax for defining in a flexible way dimensions, metrics of data quality and data source quality. We present a new framework for adaptive query processing on quality-extended query declarations. This processing includes the negotiation of quality contracts between the distributed data sources. The principle is to find dynamically the best trade-off between the cost of the query and the quality of the result retrieved from several distributed sources.

6.4. Multimedia Document Description

Keywords: *Multimedia Description*.

Participants: Laure Berti-Équille, Manolis Delakis, Patrick Gros, Ewa Kijak, Hervé Renault, Pascale Sébillot.

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.

Due to our close collaboration with teams like METISS and VISTA, and with external partners like Thomson and INA, we propose to focus on video description and TV in particular. We follow two approaches. On the one hand, we study extensions of Hidden Markov Models to integrate the information from the various media while respecting their various temporal granularity. These models also allow to integrate some *a priori* information in the structure of the models and are well suited for a fine description of the video. On the other hand, we study the problems which arise when indexing continuous streams of TV. In this case, a first challenge is to segment the various programs and to match the signal with a TV guide. This approach is much more coarse-grain oriented.

6.4.1. Segments Models for Video Description

Keywords: *Hidden Markov Models*, *Image-Sound Interaction*, *Video Structuring*.

Participants: Manolis Delakis, Patrick Gros, Pascale Sébillot.

Our work on this topic is done in close collaboration with the METISS and VISTA teams of IRISA and INA and Thomson as external partners.

In this work we aim to describe and to decode a video document in a human meaningful way, and thus, to construct its table of contents. Having analyzed the video in a such a way, we can access any of its particular points of interest in non-linear time, thus extremely facilitating tasks like video archiving or video querying. This work also presents theoretical values by applying machine learning algorithms for multimedia document understanding, a research field that has attracted the interest of scientists that last few years. We have focused at first on tennis broadcast videos, where the game rules as well as the work of the producer on it result in an organized document. Due to this set of rules, some patterns emerge that modern computer vision is charged to modelize and detect.

This work attempts to extend the work of E. Kijak on this topic. A Hidden Markov Model framework was used to modelize the temporal structure of a tennis video. In this way, the problem of understanding and decoding of the structure is translated into an optimization problem, which is solved efficiently by the Viterbi algorithm. It was confirmed out intuition that none of the video modalities (images and sound were concerned)

can solve efficiently the problem alone. The fact that these modalities operate in different time scales as well as their asynchronicity make an efficient multimodal fusion a really challenging problem.

A possible solution to this problem is to segment the stream of a reference modality (images, in our case) and to collect the observations of the other modalities in this fixed time period. This approach can improve the results, but a clear indication was found that we should extend our search towards a different time scale for each modality. For doing so, we opted for Segment Models, which provide a generalization of Hidden Markov Models. The former are based on the association of each state with an set of observations, while the latter are based on a single observation for each state. Segment models also introduce the notion of duration modeling for a given state.

We first modelized a scene using smaller (*i.e.*, with fewer states) and simpler Hidden Markov Models, trained via the Baum-Welch algorithm. The underlying game and production rules are now learned in a more implicit way. Furthermore, we used a Viterbi algorithm that uncovers not only the most likely sequence of labels, but also provides the most likely segmentation of the video document into scenes. This algorithm uses the Hidden Markov Models as a scorer for a complete scene and also a duration modeling of each scene, learned by the training data. Experiments conducted so far using only the visual modality demonstrated that this modeling can improve the results, so we can conclude that it is a more realistic modeling of a video document.

Segment models facilitate the incorporation of the audio modality since a scene contains all the information needed for sound classification and Viterbi decoding. As audio observations do not need any more to be studied at the video rate, various ways of modeling the audio observations are to be explored. Furthermore, as the underlying structure of the video document is learned more implicitly and with less engineering effort, we believe that the application of Segment Models to other types of structured video will be more easy than that of Hidden Markov Models.

6.4.2. TV Streams Indexing

Keywords: *Digital TV*, *Video Mining*, *Video Stream*.

Participants: Laure Berti-Équille, Patrick Gros, Xavier Naturel.

One of the benefits of the upcoming digital television is to allow random access and easy manipulation of the video stream. One desirable feature is to segment the video into meaningful sequences of interest to the user, for a late viewing, storage or annotation. One of the first step is to segment carefully the television broadcat into programs. This is the subject of the Ph.D. thesis of Xavier Naturel, and one important axis of the DEMI-TON project, in which Texmex and INA are involved.

To detect the programs embedded in a television broadcast, we focus on the localization and structuring of the transition between 2 programs, transition that we call the "deadzone". The deadzones can be roughly localized by the program schedule but this is highly unaccurate. A first phase, rather classical, is to segment the video stream into shots, and then try to classify the shots into program or non-programs according to various low-level properties. Non-programs are usually more colorful, louder, and exhibit a higher cut activity than traditional programs. We use a statistical framework to classify the shots into programs/non-programs, keeping in mind that this classification cannot be accurate due to the high variability of the low-level features, and that it is rather used as a first indication of the non-program presence and boundaries.

The main idea to detect a deadzone is to recognize the video sequences that are typical of a non-program segment and that are repeateadly broadcasted. We build a signature for each shot and then try to match it to one of the signatures stored in the database. This work is in progress.

6.4.3. Image and Text Joint Description

Keywords: *Image-Text Interaction*.

Participants: Patrick Gros, Hervé Renault, Pascale Sébillot.

In text retrieval engines, images are not taken into account; conversely, 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 reach this goal is to determine the parts of the texts which are related to images. This could lead to textual descriptions of images, and thus to the possibility of textual queries to retrieve them, in a much richer way than what is currently offered by systems using simple keywords associated with the images. Moreover, it is possible to find two documents containing a same image and to use both associated relevant notions of texts to disambiguate or improve the understanding of the textual material.

Our first work was to develop methods to find markers indicating a physical reference to an image within a text. This allows us to isolate portions of interesting text. We now plan to study more implicit relations, based on the text alone, without explicit references.

6.4.4. Text and Speech Joint Description

Keywords: *Text-Speech Interaction*.

Participants: Stéphane Huet, Pascale Sébillot.

A lot of sound documents contain speech. In order to be indexed and exploited, they have to be structured, and relevant knowledge such as topics or named entities (people, places, ...) have to be extracted from their contents. Currently, this extraction phase is a two-step one: in a first step, a speech recognition system produces a textual transcription of the documents; natural language processing (NLP) techniques are then used on the transcription in order to extract the relevant information. However, the linear vision of this two-step system is conceptually limited: NLP techniques can suffer from non grammatical transcriptions; transcriptions are conducted without taking into account useful clues provided by the text itself. Stéphane Huet has begun in October a Ph.D. thesis (under the double supervision of Guillaume Gravier from the METISS Project-Team, and Pascale Sébillot) which aims at exploring ways to deeply integrate knowledge coming from NLP and speech recognition and transcription domains in order to both offer a better transcription of speech documents, and extract information necessary to their indexing, structuring and exploitation.

7. Contracts and Grants with Industry

7.1. Contracts, Initiatives and Participation to Networks of Technological Research

7.1.1. PRIAM Médiaworks Project

Keywords: *TV Archives*, *Video Databases*.

Participant: Patrick Gros.

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

The Mediaworks project was labeled by the PRIAMM program and the French information society program, financed by the Ministry of industry. 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 of automatic structuring of video in plans and to compute an iconic representation of these plans.

7.1.2. RNRT Diphonet project: Photo Diffusion on Internet

Keywords: *Copyright Protection*, *Image Recognition*, *Image Databases*, *Piracy*.

Participants: Laurent Amsaleg, Patrick Gros.

Duration : 30 months, starting in January 2002. Partners: IRISA, Canon, L2S, Andia Presse, INRIA (projects CODES, TEMICS and TEXMEX).

Copyright protection is a key component to allow photo holders like photo agencies to provide their collections on 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 to 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.

7.1.3. Contract with Thomson

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

Duration : 36 months, starting in March 2001.

The Ph.D. thesis of S.A. Berrani was supported by a CIFRE grant in the framework of a contract between Thomson and TEXMEX.

7.1.4. Contract with France Télécom

Participants: Annie Morin, Nicolas Bonnel.

Duration : 36 months, starting in December 2002.

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

7.1.5. RIAM FERIA Project

Keywords: *TV Archives*, *Video Databases*.

Participants: Patrick Gros, Claire Hélène Demarty, Philippe Daubias, Manolis Delakis.

Duration 21 months, starting in October 2003. Partners: Communications et Systèmes, INA, IRIT, NDS, 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 produce 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.

7.2. European Initiatives

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

Keywords: *Indexing Videos*, *Video*.

Participant: Laurent Amsaleg.

Duration : 30 months, starting in April 2002. Partners: IRISA (TEMICS and TEXMEX teams), 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.

7.2.2. European Network of Excellence MUSCLE : Multimedia Understanding through Semantics, Computation, and Learning

Keywords: *Images*, *Multimedia*, *Naturel Language Processing*, *Video*.

Participants: Patrick Gros, Laurent Amsaleg, Pascale Sébillot.

Duration: 4 years, starting in April 2004. 42 partners. Prime: ERCIM, scientific coordinator: Eric Pawels, CWI – Amsterdam.

This project aims at developing the collaboration in the domain of automatic multimedia document analysis, in particular to be able to manipulate their meaning. The project is thus concerned by all content-based analysis tools available for every media (text, sound and speech, image and video), but also by the techniques which allow to combine the information extracted from each media, and by the common techniques needed to handle such data (optimization, classification, intensive computation).

The TEXMEX group, as one of the partners, will contribute to monomedia aspects (text, images, video) and to multimedia combination techniques.

7.2.3. European Integrated Project aceMedia : Integrating Knowlege, Semantics and Content for User-Centered Intelligent Media Services

Keywords: *Multimedia*, *Video*, *Video indexing*.

Participants: Patrick Gros, Laurent Amsaleg, Zied Jemai.

Duration: 4 years, starting in January 2004. 15 partners. Prime: Motorola Ltd.

The goal of the project is to develop a new way to encode multimedia document for their diffusion on networks like Internet, telecommunication networks or broadcasting systems. In this new encoding scheme, data, metadata and an intelligence layer containing code will be embedded in a single autonomous entity called an ACE (autonomous content entity).

Within the project, we are in charge of developing algorithms to index and retrieve numerical data. The documents will be described by both conceptual and content-based descriptors. The latter are usually vectors of real numbers and require special algorithms not available in traditional DBMS.

7.2.4. European Integrated Project ENTHRONE : End-to-End QoS through Integrated Management of Content, Networks and Terminals

Keywords: *Content Generation*, *Digital Item*, *Heterogeneous Networks*, *MPEG-21*, *Metadata*, *QoS Adaptation*.

Participants: Laure Berti-Équille, Patrick Gros, Boris Rousseau.

Duration: 4 years, starting in December 2003. 25 partners. Prime : Thales Broadcast & Multimedia

The ENTHRONE project proposes an integrated management solution which covers an entire audio-visual service distribution chain, including content generation and protection, distribution across networks and reception at user terminals. The aim is not to unify or impose a strategy on each individual entity of the chain, but to harmonise their functionality, in order to support an end-to-end QoS architecture over heterogeneous networks, applied to a variety of audio-visuals services, which are delivered at various user terminals. To meet its objectives, the project will rely on an efficient, distributed and open management architecture for the end-to-end delivery chain.

The availability and access to resources will be clearly identified, described and controlled all the way along the content distribution chain. The MPEG-21 data model is used to provide the common support for implementing and managing the resources functionalities and to contribute to implement the *Universal Multimedia Access*. Our work under the workpackage entitled "Metadata definition and specification" is to provide a study and a characterization of relevant metadata that is necessary for the requirements and the usage scenarios of the ENTHRONE system previously defined (such as VoD, pay-TV, TV Broadcast, e-learning). The choice and the detailed description of the relevant metadata have been made in conformance with the MPEG-21 data model. In the frame of ENTHRONE, we gave an overview of metadata standards (such as

Dublin Core, MPEG-7, TV-Anytime, etc.) and standardized frameworks and tools (such as MPEG-21) to declare, use, and modify metadata for the multimedia delivery chain. We also identified the required metadata and how this data is/has to be managed by the ENTHRONE project. We described the metadata generation, storage, usage and exchange processes through the ENTHRONE system architecture, focusing in particular the description of metadata flows and we proposed a conceptual model of the ENTHRONE metadata that corresponds to our metadata specification and includes the required metadata packages and data types with the UML formalism.

8. Other Grants and Activities

8.1. National Initiatives

8.1.1. ACI Grid GénoGRID

Keywords: *FPGA*, *Genomics*, *Reconfigurable Architecture*.

Participant: Laurent Amsaleg.

Joint work with SYMBIOSE, ADEPT and R2D2. Duration: 3 years, starting in December 2001.

The goal of this Ministry grant 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.

8.1.2. ACI masses de données Remix: Mémoire reconfigurable pour l'indexation de masses de données

Participant: Laurent Amsaleg.

Joint work with Symbiose and R2D2. Duration 3 years starting in September 2003.

This project aims at developing new technologies to access very large databases like DNA or image databases. The core technology used in the project is based on FPGA chips.

8.1.3. ACI masses de données M2PDP: Gestion de masses de données dans les systèmes P2P

Participants: Laurent Amsaleg, Panagiotis Hadjidoukas.

Joint work with PARIS and ATLAS. Duration: 3 years, starting in September 2004.

This project aims at studying the problem arising when managing lots of data on P2P systems like PC clusters. These problems are studied from a system point of view (memory management) and from an algorithmic point of view (parallelization of algorithms).

8.1.4. ACI masses de données DEMI-TON: Multimodale description for automatic structuring of TV streams

Participants: Manolis Delakis, Stéphane Huet, Patrick Gros, Xavier Naturel, Pascale Sébillot.

Other partners : INA, METISS project-teams. Duration 3 years, starting in December 2004.

This project concerns the development of new techniques to index large collections of TV programs. INA is supposed to record and index more than 50 channels 24 hours a day. As the number of available documentalists did not increase as fast as the number of channels to be indexed, they has to rely on a more automatic process. The first need is to verify that the programs present is the stream correspond effectively to what was announced in the TV program guide and to synchronize the stream with this program guide. In a second stage, some programs like news reports have to be indexed to the topics that were tackled by the program and which, of course, could not be announced in the program guide.

8.1.5. Action Bio-Info Inter-EPST: Parallel and Reconfigurable Architectures for Genomic Data Extraction

Keywords: *Architecture Reconfigurable*, *FPGA*, *Genomics*.

Participant: Laurent Amsaleg.

Joint work with SYMBIOSE and R2D2. Duration 2 years, starting in April 2004.

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.

8.1.6. R&D INRIA action SYNTAX

Keywords: *Document Analysis*, *Information Retrieval*.

Participant: Pascale Sébillot.

21 partners. Prime : INRIA. Started in December 2002.

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.

8.1.7. 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 kEuros for 3 years.

8.1.8. Participation to National Working Groups

- L. Amsaleg is a member of the CNRS AS "Données Multimédias : Interrogation et Stockage" of the RTP9.
- L. Amsaleg was invited on Mai 19, 2003 to give a seminar entitled "indexation d'images – le point de vue des bases de données" in the context of the CNRS specific action "fouille d'images", RTP 25.
- L. Berti-Équille participates to the "Documents Mutimédia" and "Médiation" working groups of GDR I3.
- 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. 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 participates to AS "Text Mining" 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.
- P. Sébillot is a member of the working group PRC I3-AFIA TIA (terminologie et intelligence artificielle)

8.2. International Collaborations

8.2.1. Working Group Image Undertanding of ERCIM

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.

8.2.2. Collaboration with Reykjavik University - Iceland

Keywords: *Caching*, *Disks*, *Memory Management*.

Participant: Laurent Amsaleg.

Björn Jónsson's group of Reykjavik University is recognized by INRIA as an associate group to TEXMEX since 2004.

Laurent Amsaleg and Björn Jónsson boosted their cooperation that takes place between France and Iceland. B. Þ. Jónsson, *Associate Professor* at the University of Reykjavík in Iceland, visited IRISA three times (1 week in March 2002, 4 weeks in June 2003, 3 days in September 2004). Another visit is scheduled for 1 week starting November 14th. L. Amsaleg visited Reykjavík University twice (1 week in December 2003 and 2 months in summer 2004). Another visit is scheduled for 1 week starting December 6. Their work is to develop techniques that integrate efficiency and effectiveness in CBIRS (Content-based image retrieval systems). The long-term benefits of this work are expected to be improved image retrieval systems that are key for emerging applications. In order to achieve this goal, we have to look at each aspect of content-based image retrieval systems. This mainly includes multidimensional indexing and searching, user interaction, interactions with the operating system and implementation of internal data structures.

8.2.3. Collaboration with Croatia and Slovenia

Participant: Annie Morin.

Duration : 2 years. Partners: Andrija Stampar School of Public Health, Medical School, University of Zagreb, Zagreb, Croatia ; Faculty of Computer and Information Science, University of Ljubljana, Slovenia ; ERIC lab., University of Lyon ; Rudjer Boskovic Institute, Zagreb, Croatia.

A. Morin animates an Egide cooperation program with Croatia and Slovenia. Huge variety of medical problems and data characteristics, connected to such problems, as well as a large range of statistical methods and methods of intelligent data analysis of such data, bring out of the problem which method to apply for a given problem to get results of best quality and reliability. Another problem is how to conceptualise hidden relationship in dataset beside obvious relationship in dataset. During the project, the plan is to analyse three different medical data type, *i.e.*, four medical domains: epidemiological problems, clinical and laboratory problems (thyroid disease domain, HIV infection domain), genetic *i.e.* , genetic epidemiological domain. Medical problems and datasets will be collected and recorded in Croatia at the Andrija Stampar School of Public Health (epidemiological datasets, genetic-epidemiological datasets with all needed field research held on Adriatic islands connected to the isolated population of inhabitants of genetic interest), Dr. Fran Mihaljevic' - Clinic for infectious disease, AIDS referral centre (datasets on HIV infection in Croatia), Clinical hospital Sestre Milosrdnice, referral centre for thyroid disease (thyroid disease datasets). Application of different statistical methods and methods of intelligent data analysis will be performed at the Andrija Stampar School of Public Health (Department for medical statistics, epidemiology and medical informatics) and Institute Rugjer Boskovic (Laboratory for Information Systems) and in France: in IRISA, we are mainly concerned with datamining and multimedia mining with a special interest in the quality of data and quality of metadata used for the management of large databases. This problem of quality is becoming crucial in datamining.

Statistical methods will be applied as well as some specific algorithms of intelligent data analysis (visualization, artificial neural networks, Bayesian networks inductive learning etc.). Some algorithms like ILLM are developed in Croatia (Institute Rugjer Boskovic), and some analytical methods represent the school of analytical methods like French School which resulted in the development of application support such as SPAD-S,

SPAD-N. The differences in research approach, evaluation and result interpretation implies needs for cooperation. The goal of this project is to learn the differences and similarities in problem approach, analytical methods and evaluation of results depending on medical domain.

We organized a workshop in SRCE (university computing center) in Zagreb on june 12th 2004 . The goal was related to the design, analysis and implementation of data mining theory, and systems on hybrid data: numeric, textual or images. Various medical issues and datasets will be presented as a challenge to show how these paradigms can be used to improve medical decision-making. The workshop offered an opportunity for scientists with different backgrounds (medical doctors, clinicians, computer scientists, statisticians, mathematicians) to brainstorm on the subject. The purpose of this seminar was to explore the various approaches in data mining and intelligent data analysis and to choose several problem domains that will be analyzed during a 2-year project. The expected outcome of this project is the establishment of a standard approach to data analysis for the algorithm based on clinical decision rules and the development of guidelines for generating predictive medical decisions. A report is available on http://lis.irb.hr/IDADM/.

8.2.4. Collaboration with NII - Japan

Keywords: *Geomedia*, *Metadata*, *Statistics*.

Participants: Laure Berti-Équille, Annie Morin.

A MoU (Memorandum Of Understanding) has been signed in 2003 between IRISA (TEXMEX Project) and NII - National Institue of Informatics - of Tokyo (Japan) providing a frame for a starting collaboration on Multimedia Metadata Management intitled "M4" initially in the context of geomedia documents related to the Digital Silk Road Project in collaboration with UNESCO. An Exchange Research Program has been validated and founded this year by CNRS and JSPS and A. Morin spent two weeks in Japan in February 2004 and gave three seminars during her stay.

Further to some relationships with researchers at the RERF (Radiation Effects Research Foundation) in Hiroshima, Annie Morin first went to Hiroshima at the department of statistics of the RERF centre and met researchers from the department of epidemiology. The department of statistics analyzes interdisciplinary information collected to study radiation effects and assists research scientist in other RERF departments with data management. Annie Morin was particularly interested in their way to modelize risks using very advanced generalized linear models and also interested in their merging of different databases.

Her main contacts were Dr. Ichiro Ide and Pr. Shin Ichi Satoh, and Dr. Asanobu Kitamoto from the multimedia information research division.

8.2.5. Collaboration with University of Geneva

Participants: Laurent Amsaleg, Patrick Gros.

Following the CVBD'04 workshop we have organized, first contacts have been established with the VIPER group of University of Geneva headed by Stéphane Marchand-Maillet. We plan to work on video structuring and indexing in the case of very large collections.

8.2.6. Collaboration with Dublin City University

Participants: Laurent Amsaleg, Patrick Gros.

Our collaboration in the frame of aceMedia outlined the opportunity for a deeper collaboration on large image base indexing and retrieval. The first objective of this new collaboration is to make a first experience of indexing and retrieval with a database of one million images. DCU has a good knowledge on image descriptors and retrieval system interfaces which is complementary to ours. On the other hand, our indexing technology is more advanced.

9. Dissemination

9.1. Conference, Workshop and Seminar Organization

- A. Morin organized a workshop in Zagreb Croatia on Intelligent Analysis and Data Mining. The website is http://lis.irb.hr/IDADM/program.html

- L. Amsaleg and B. Jonsson (from the University of Reykjavick, Iceland) created, together with Vincent Oria from New Jersey Institute of Technology (USA) a new workshop entitled "First Computer Vision Meets Databases (CVDB) Workshop". This workshop was held in Paris, France, on June 13, 2004. The workshop was co-located with the 2004 ACM SIGMOD/PODS conference and was attended by forty-two participants from all over the world. They created this workshop because they wanted to understand why only few works in the computer vision community have adopted any of the indexing schemes that have been designed by database researchers. They discovered many valid scientific reasons but also that there was a great gap between the computer vision and the database communities. The goal was therefore to bridge that gap and to provide database researchers with a snapshot of what computer vision people are dealing with and vice-versa, with the aim of defining some research directions that can benefit both communities. The workshop was successful. Eight papers were selected for presentation and publication [36]. Additionally, we hand-picked two tutorialists to present their views of the research directions and contributions of the computer vision and database communities, respectively. Finally, we assembled a panel to focus on the applications of image databases in the near and distant future. For details of the papers, tutorials, and panel, including slides from all presentations, please visit the workshop web-site, which will remain open at http://cvdb04.irisa.fr. A workshop report will be published in Sigmod Record [38].
  Based on the observed need for a forum for exchanging ideas and results that are at the intersection of the computer vision and database research areas, we have decided to make the CVDB workshop an annual event and will apply again for co-location with SIGMOD/PODS in Baltimore, Maryland in June 2005.

9.2. Involvment with the Scientific Community

- L. Amsaleg:
  - was a Workshop Co-Chair, First Computer Vision meets Databases Workshop (CVDB), 2004, Paris, France,
  - was a program committee member of BDA 2004 (National Conference on Advanced Database Systems),
  - was a program committee member of PCM 2004 (Pacific-RIM Conference on Multimedia), Tokyo, Japan, December 2004,
  - was a program committee member of MIS 2004 (International Workshop on Multimedia Information Systems), College Park, MD, USA, August 2004,
  - was publicity chair and a member of the organizing committee of SIGMOD'04.
- L. Berti-Équille:
  - was a member of the program committee of CORIA 2004 (National Conference on Information Retrieval and Applications) in March 2004 at Toulouse, France,
  - was a member of the program committee of IQIS 2004 (International Workshop on Information Quality in Information System) co-located with the 2004 ACM SIGMOD/PODS conference, June, Paris, France,
  - was a member of the program committee of the Special Issue on Quality Metrics for Data Mining of the French Journal intitled "Revue Nationale des Technologies de l'Information (RNTI)", Cépaduès,

  - was a member of the program committee of the Special Issue on Quality of the French Journal intitled "Ingénierie des Systèmes d'Information", Hermès,
  - is a member of the editorial board of the International Journal intitled "Journal of Digital Information Management",
  - is a member of the editorial board of the International Journal intitled "Multimedia Tools and Application".

- P. Gros:

  - was a member of the program committee of the International Conference on Image and Video Retrieval CIVR'04 which was held in Dublin in July 2004,
  - was a member of the program committee of the member of the program committee of the COnference on Information Retrieval and Application CORIA'04 which was held in Grenoble in March 2004,
  - was a member of the program committee of the international workshop on multidisciplinary image, video, and audio retrieval and mining CORIMEDIA'04 which was held in Sherbrooke in October 2004,
  - was a member of the program committee of the First International Workshop on Computer Vision meets Databases which was held in Paris in June 2004,
  - was a member of the program committee of the Second International Conference on Intelligent Access of Multimedia Documents on Internet which was held in Tozeur (Tunisia) in November 2004,
  - was a member of the program committee of the Pacific Rim Conference on Multimedia which was held in Tokyo in December 2004,
  - is a member of the editorial board of the French journal intitled "Traitement du Signal".

- A. Morin:

  - was a member of the program committee of ITI 2004 (Information technology interfaces),
  - has been elected President of the GMF IIS (group of the French members of the international statistical institute) in may 2004 for 3 years,
  - is a member of the CNU (National Council of the University) in the computer science section.

- P. Sébillot:

  - is an associate editor of the journal In Cognito,
  - is an associate editor of Jedai (Journal Électronique d'Intelligence Artificielle),
  - was a member of the program committee of JADT'04 (7th International Conference on the Statistical Analysis of Textual Data) in March 2004 at Louvain-la-Neuve, Belgium,
  - was a member of the program committee of CIFT'04 (International Conference on Text Mining) in June 2004 at La Rochelle, France,
  - was a member of the program committee of CORIA 2004 (National Conference on Information Retrieval and Applications) in March 2004 at Toulouse, France,
  - was a member of the program committee of COLDOC' 2004 (La construction des observables en sciences du langage, National Workshop of Young Researchers in Linguistics), in April 2004 at Nanterre, France,
  - is a member of the board of ATALA (Association pour le traitement automatique des langues),
  - is a member of the scientific committee of the TCAN program of CNRS (Traitement des connaissances, apprentissage et NTIC).

9.3. Teaching Activities

- L. Amsaleg: Advanced Databases. DIIC2, LSI, 2nd year, IFSIC, University of Rennes I.
- L. Amsaleg: Datawarehouses and Data Mining. IUP Miage, 3rd year,IFSIC, University of Rennes I.
- L. Amsaleg and P. Gros,P. Sébillot, Multimedia Indexing: Techniques and Applications, Research Master in Computer Science, IFSIC, University of Rennes 1.
- L. Amsaleg, P. Gros and P. Sébillot: Digital documents indexing and retrieval. Professional Master in Computer Science, IFSIC, University of Rennes 1.
- L. Amsaleg and P. Sébillot: Advanced Databases.Professional Master in Computer Science, Ifsic, University of Rennes 1.

9.4. Participation to Seminars, Workshops, Invitated Conferences

- P. Gros gave an invited talk on at DCU, Dublin about TEXMEX work on video indexing, in October 2004.
- A. Morin gave three seminars at NII, Tokyo, Japan in February 2004 on : i) factorial correspondence analysis as a dual approach for semantic and indexing, ii) characterization of huge sets of thematically homogeneous texts and on iii) content-based research: use of the association rules how statistical data analysis methods can enrich or confirm the association rules.
- P. Sébillot gave an invited talk to the Workshop "La place des méthodes quantitatives dans le travail du linguiste" organized by ERSS, Toulouse, in June 2004.
- P. Sébillot gave an invited talk at OLST-RALI seminar, Montreal, in July 2004.

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", vol. 2001, n° 4, 2001, p. 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] 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", 2003, p. 24–31.

[4] L. BERTI-EQUILLE. *Quality-based recommendation of XML documents*, in "Journal of Digital Information Management", vol. 1, n° 3, 2003, p. 117-128.

[5] 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", vol. 4, august 2003, p. 493–525.

[6] E. KIJAK, G. GRAVIER, P. GROS, L. OISEL, F. BIMBOT. *HMM based structuring of tennis videos using visual and audio cues*, in "EEE Int. Conf. on Multimedia and Expo, ICME'03,", vol. 3, 2003, p. 309-312.

[7] 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)", vol. 1, n° 4, 2002, p. 407-412.

[8] M. ROSSIGNOL, P. SÉBILLOT. *Extraction statistique sur corpus de classes de mots-clés thématiques*, in "TAL (Traitement automatique des langues)", vol. 44, n° 3, 2003, p. 217-246.

Doctoral dissertations and Habilitation theses

[9] S.-A. BERRANI. *Recherche approximative de plus proches voisins avec contrôle probabiliste de la précision; application à la recherche d'images par le contenu*, Ph. D. Thesis, Université de Rennes1, February 2004, http://www.irisa.fr/texmex/publications/versionElect/these_SA.Berrani.ps.gz.

[10] A. REMAZEILLES. *Navigation à partir d'une mémoire d'images*, Ph. D. Thesis, Université de Rennes1, December 2004.

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", vol. 23, 2004, p. 221-235.

[12] L. BERTI-ÉQUILLE. *La qualité des données comme condition à la qualité des connaissances : un état de l'art*, in "Mesures de qualité pour la fouille de données. Numéro spécial, Revue Nationale des Technologies de l'Information (RNTI),Cépaduès", 2004.

[13] V. CLAVEAU, P. SÉBILLOT. *Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe*, in "TAL (Traitement automatique des langues)", vol. 45, n° 1, 2004, p. 153-182.

[14] V. CLAVEAU, P. SÉBILLOT. *Corpus-Based Qualia Element Acquisition by Inference of Extraction Patterns*, in "Generative Approaches to the Lexicon", J. PUSTEJOVSKY, P. BOUILLON, K. KANZAKI, I. ISAHARA, C. LEE (editors)., to appear, Kluwer Academic Publishers, 2004.

[15] A. KOUOMOU-CHOUPO, A. MORIN, L. BERTI-ÉQUILLE. *Recherche dans de grandes bases d'images fixes : une nouvelle approche guidée par les règles d'association*, in "Revue RNTI-E-2 (Actes EGC'2004), CÉPADUÈS", 2004, p. 65-70.

[16] M. ROSSIGNOL, P. SÉBILLOT. *Combining Statistical Data Analysis Techniques to Extract Topical Keyword Classes from Corpora*, in "IDA (Intelligent Data Analysis)", vol. 8, n° 6, 2004.

Publications in Conferences and Workshops

[17] J.-A. BENVENUTI, L. BERTI-ÉQUILLE, E. JACOPIN. *Le projet SABRE : de l'ontologie à l'inférence, (Poster)*, in "Actes des 15ièmes journées francophones d'Ingénierie des Connaissances, Lyon", May 2004.

[18] S.-A. BERRANI, L. AMSALEG, P. GROS. *Application de la recherche approximative de plus proches voisins à la recherche d'images par le contenu pour la détection des copies*, in "20e journées Bases de Données Avancées, BDA'04, Montpellier, France", October 2004.

[19] S.-A. BERRANI, L. AMSALEG, P. GROS. *Recherche d'images par le contenu pour la détection des copies*, in "14e Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, RFIA 2004, Toulouse, France", January 2004.

[20] L. BERTI-ÉQUILLE. *Quality-Adaptive Query Processing over distributed Sources*, in "International Conference on Information Quality, M.I.T., Boston, U.S.", November 2004.

[21] L. BERTI-ÉQUILLE, A. KOUOMOU-CHOUPO, A. MORIN. *Feature mining for multimedia indexing and retrieval (poster)*, in "5th International Workshop on Image Analysis for Multimedia Interactive Services, Lisbon, Portugal", 2004.

[22] V. CLAVEAU, P. SÉBILLOT. *Extension de requêtes par lien sémantique nom-verbe acquis sur corpus*, in "Traitement automatique des langues naturelles, TALN'04, Fes, Morocco", april 2004.

[23] V. CLAVEAU, P. SÉBILLOT. *From Efficiency to Portability: Acquisition of Semantic Relations by Semi-Supervised Machine Learning*, in "20th International Conference on Computational Linguistics, COLING'04, Geneva, Switzerland", august 2004.

[24] P. GROS, E. KIJAK, G. GRAVIER. *Automatic Video Structuring based on HMMs and Audio Visual Integration*, in "Proceedings of the Second International Symposium on Image/Video Communications over Fixed and Mobile Networks, Brest, France", jul 2004, p. 1–6.

[25] P. E. HADJIDOUKAS, L. AMSALEG. *Portable Support and Exploitation of Nested Parallelism in OpenMP*, in "Proceedings of the 6th European Workshop on OpenMP (EWOMP 2004), Stockholm, Sweden", October 2004.

[26] P. E. HADJIDOUKAS. *A Lightweight Framework for Executing Task Parallelism on top of MPI*, in "Proceedings of the 11th European PVM/MPI Users' Group Meeting (EuroPVM/MPI 2004), Budapest, Hungary", September 2004.

[27] P. E. HADJIDOUKAS. *A Unified Programming Approach to Master-Slave Computing*, in "Proceedings of the 4th International Workshop on Constructive Methods for Parallel Programming (CMPP 2004), Stirling, Scotland, UK", July 2004.

[28] A. KOUOMOU-CHOUPO, L. BERTI-ÉQUILLE. *Visual Feature Mining for Adapting Query-by-Example over Large Image Databases*, in "International Workshop on Multidisciplinary, Video, and Audio retrieval and Mining, Sherbrooke-Canada", October 25-26 2004.

[29] A. KOUOMOU-CHOUPO, L. BERTI-ÉQUILLE, A. MORIN. *Multimedia Indexing and Retrieval with Features Association Rules Mining*, in "IEEE International Conference on Multimedia and Expo (ICME'2004), Taipei, Taiwan", 2004.

[30] A. KOUOMOU-CHOUPO. *Sélection des critères de recherche par le contenu dans une base d'images : l'apport des règles d'associations entre les descripteurs visuels*, in "22 ème congrès francophone INFORSID, Biarritz, France", to appear, 2004.

[31] A. MORIN, A. KOUOMOU-CHOUPO, L. BERTI-ÉQUILLE. *Research in Image Databases: How Statistical*

Data Analysis Methods can enrich the Association Rules (poster), in "16th Symposium of International Association for Statistical Computing (COMPSTAT'2004), Prague, Czech Republic", 2004.

[32] A. REMAZEILLES, F. CHAUMETTE, P. GROS. *Contrôle des mouvements d'un robot à l'aide d'une mémoire visuelle*, in "14e Congrès Francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, RFIA 2004, Toulouse, France", January 2004.

[33] A. REMAZEILLES, F. CHAUMETTE, P. GROS. *Robot motion control from a visual memory*, in "IEEE Int. Conf. on Robotics and Automation, ICRA'04, La Nouvelle Orléans, LA, USA", April 2004.

[34] M. ROSSIGNOL, P. SÉBILLOT. *Extraction non supervisée sur corpus de classes de mots-clefs*, in "14e congrès francophone AFRIF-AFIA de reconnaissance des formes et intelligence artificielle, RFIA 2004, Toulouse, France", january 2004.

Miscellaneous

[35] A. KOUOMOU-CHOUPO, A. MORIN, L. BERTI-ÉQUILLE. *Recherche par le contenu dans une base d'images fixes : l'intérêt des règles d'associationAtelier Fouille de données complexes dans un processus d'extraction de connaissances - EGC'2004, Clermont Ferrand, France*, 2004.

Bibliography in notes

[36] L. AMSALEG, B. JONSSON, V. ORIA (editors). *Proc. of the First International Workshop on Computer Vision meets Databases (CVDB'04)*, June 2004.

[37] 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.

[38] L. AMSALEG, B. JONSSON, V. ORIA. *Report from the First International Workshop on Computer Vision meets Databases – CVDB 2004*, December 2004, ACM Sigmod Record.

[39] J. ANDERSON, M. STONEBRAKER. *Sequoia 2000 Metadata Schema for Satellite Images*, in "acm sigmod Record, Special issue on Metadata for Digital Media", vol. 23, nº 4, 1994.

[40] 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.

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

[42] L. BERTI-ÉQUILLE. *La qualité des données comme condition à la qualité des connaissances : un état de l'art*, in "Mesures de qualité pour la fouille de données. Numéro spécial, Revue Nationale des Technologies de l'Information (RNTI),Cépaduès", 2004.

[43] L. BERTI-ÉQUILLE. *Quality-Adaptive Query Processing over distributed Sources*, in "International Conference on Information Quality, M.I.T., Boston, U.S.", November 2004.

[44] 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.

[45] T. BINFORD, T. LEVITT. *Quasi-Invariants: Theory and Exploitation*, in "Proceedings of darpa Image Understanding Workshop", 1993, p. 819-829.

[46] N. BOUJEMAA, S. BOUGHORBEL, C. VERTAN. *Soft Color Signatures for Image Retrieval by Content*, in "Eusflat'2001", vol. 2, 2001, p. 394-401.

[47] 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", vol. 9, n° 7, October 1999, p. 1030-1044.

[48] C. BÖHM, S. BERCHTOLD, D. KEIM. *Searching in High-dimensional Spaces: Index Structures for Improving the Performance of Multimedia Databases*, in "ACM Computing Surveys", vol. 33, n° 3, September 2001.

[49] K. BÖHM, T. RAKOW. *Metadata for Multimedia Documents*, in "acm sigmod Record Special Issue on Metadata for Digital Media", vol. 23, n° 4, 1994, p. 21-26.

[50] 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.

[51] F. CHEN, M. HEARST, J. KUPIEC, J. PEDERSON, L. WILCOX. *Metadata for Mixed-Media Access*, in "acm sigmod Record, Special issue on Metadata for Digital Media", vol. 23, n° 4, 1994.

[52] 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.

[53] T. DASU, T. JOHNSON. *Exploratory Data Mining and Data Cleaning*, John Wiley and Sons, 2003.

[54] 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", vol. 2, February 2000, p. 327-336.

[55] I.-S. EINARSSON, B. ERLINGSSON, A.-G. VALGEIRSSON, B. JONSSON, L. AMSALEG. *Using clustering to index image descriptors: A performance evaluation*, Technical report, Reykjavik University, 2003.

[56] B. ESPIAU, F. CHAUMETTE, P. RIVES. *A New Approach to Visual Servoing in Robotics*, in "ieee Transactions on Robotics and Automation", vol. 8, n° 3, June 1992, p. 313-326.

[57] 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.

[58] 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", April 2001, p. 503–511.

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

[60] G. FINLAYSON, M. DREW, B. FUNT. *Color Constancy: Generalized Diagonal Transforms Suffice*, in "Journal of the Optical Society of America A", vol. 11, n° 11, November 1994, p. 3011-3019.

[61] L. FLORACK, B. TER HAAR ROMENY, J. KOENDERINK, M. VIERGEVER. *General Intensity Transformation and Differential Invariants*, in "Journal of Mathematical Imaging and Vision", vol. 4, n° 2, 1994, p. 171-187.

[62] 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", June 1998, p. 559-565.

[63] J. GERBRANDS. *On The Relationships Between SVD, KLT and PCA*, in "Pattern Recognition", vol. 14, n° 1-6, 1981, p. 375–381.

[64] 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", September 1999, p. 518–529.

[65] U. GLAVITSCH, P. SCHAUBLE, M. WECHSLER. *Metadata for Integrating Speech Documents in a Text Retrieval System*, in "acm sigmod Record, Special issue on Metadata for Digital Media", vol. 23, n° 4, 1994.

[66] 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", September 2000, p. 429–440.

[67] 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", April 2000, p. 567-574.

[68] S. GUHA, R. RASTOGI, K. SHIM. *CURE: An Efficient Clustering Algorithm for Large DataBases*, in "Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, Washington, US", June 1998.

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

[70] P.-E. HADJIDOUKAS, L. AMSALEG. *Portable Support and Exploitation of Nested Parallelism in OpenMP*, in "Proceedings of the 6th European Workshop on OpenMP (EWOMP 2004), Stockholm, Sweden", October 2004.

[71] P.-E. HADJIDOUKAS. *A Lightweight Framework for Executing Task Parallelism on top of MPI*, in "Proceedings of the 11th European PVM/MPI Users' Group Meeting (EuroPVM/MPI 2004), Budapest, Hungary",

September 2004.

[72] P.-E. HADJIDOUKAS. *A Unified Programming Approach to Master-Slave Computing*, in "Proceedings of the 4th International Workshop on Constructive Methods for Parallel Programming (CMPP 2004), Stirling, Scotland, UK", July 2004.

[73] P. E. HADJIDOUKAS, E. POLYCHRONOPOULOS, T. PAPATHEODOROU. *OpenMP Runtime Support for Clusters of Multiprocessors*, in "Proceedings of the International Workshop on OpenMP Applications and Tools (WOMPAT 2003), Toronto, Canada", June 2003.

[74] R. HAMMOUD, R. MOHR. *Mixture Densities for Video Objects Recognition*, in "Proceedings of the 15th International Conference on Pattern Recognition, Barcelone, Espagne", vol. 2, iapr, September 2000, p. 71-75.

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

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

[77] M. HERNANDEZ, S. STOLFO. *Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem*, in "Journal of Data Mining and Knowledge Discovery", vol. 2, n° 1, 1998, p. 9-37.

[78] 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", June 1997, p. 762-768.

[79] R. JAIN, A. HAMPAPURAM. *Representations of Video Databases*, in "acm sigmod Record, Special issue on Metadata for Digital Media", vol. 23, n° 4, 1994.

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

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

[82] F. KORN, B. PAGEL, C. FALOUTSOS. *On the 'Dimensionality Curse' and the 'Self-Similarity Blessing'*, in "IEEE Trans. on Knowledge and Data Engineering", vol. 13, n° 1, January 2001, p. 96–111.

[83] A. KOUOMOU-CHOUPO, L. BERTI-ÉQUILLE. *Visual Feature Mining for Adapting Query-by-Example over Large Image Databases*, in "International Workshop on Multidisciplinary, Video, and Audio retrieval and Mining, Sherbrooke-Canada", October 25-26 2004.

[84] A. KOUOMOU-CHOUPO, L. BERTI-ÉQUILLE, A. MORIN. *Multimedia Indexing and Retrieval with Features Association Rules Mining*, in "IEEE International Conference on Multimedia and Expo (ICME'2004), Taipei, Taiwan", 2004.

[85] H. LEJSEK, F.-H. ASMUNDSSON, B. JONSSON, L. AMSALEG. *The Application of the MEDRANK Algorithm to Content-Based Image Retrieval using Local Descriptors*, Technical report, Reykjavik University, August 2004.

[86] I. LERMAN. *Foundations in the Likelihood Linkage Analysis Classification Method*, in "Applied Stochastic Models and Data Analysis", vol. 7, 1991, p. 69–76.

[87] 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", vol. 14, n° 4, July 2002, p. 792–808.

[88] 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.

[89] A. LUGMAYR, S. NIIRANEN, S. S. KALLI. *Digital Interactive TV and Metadata, Future Broadcast Multimedia*, Signals and Communication Technology, Springer, 2004.

[90] E. MALIS, F. CHAUMETTE, S. BOUDET. *2 1/2 D Visual Servoing*, in "ieee Transactions on Robotics and Automation", vol. 15, n° 2, April 1999, p. 238-250.

[91] S. MARCUS, V. SUBRAHMANIAN. *Foundations of Multimedia Database Systems*, in "Journal of the acm", vol. 43, n° 3, 1996, p. 474-523.

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

[93] S. MUGGLETON, L. DE-RAEDT. *Inductive Logic Programming: Theory and Methods*, in "Journal of Logic Programming", vol. 19-20, 1994, p. 629-679.

[94] J. NIEVERGELT, H. HINTERBERGER, K. SEVCIK. *The Grid File: An Adaptable, Symmetric Multikey File Structure*, in "ACM TODS", vol. 9, n° 1, 1984.

[95] OPENMP ARCHITECTURE REVIEW BOARD. *OpenMP Specifications*, 2004, http://www.openmp.org.

[96] 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.

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

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

[99] T. REDMAN. *Data Quality for the Information Age*, ISBN 0-89006-8836, Artech House, 1996.

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

[101] 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.

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

[103] C. SCHMID, R. MOHR. *Local Grayvalue Invariants for Image Retrieval*, in "ieee Transactions on Pattern Analysis and Machine Intelligence", vol. 19, n° 5, May 1997, p. 530-534, ftp://ftp.inrialpes.fr/pub/movi/publications/schmid_pami97.ps.gz.

[104] R. SIGURDARDOTTIR, H. HAUKSSON, B. JONSSON, L. AMSALEG. *The Effect of Cluster Size on Image Descriptor Retrieval Performance*, Technical report, Reykjavik University, August 2004.

[105] 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.

[106] THE POP PROJECT CONSORTIUM. *POP (Performance Portability of OpenMP) IST/FET project (IST-2001-33071)*, 2001.

[107] 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.

[108] 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", August 1998, p. 194-205.

[109] D. WHITE, R. JAIN. *Similarity Indexing with the SS-tree*, in "ICDE", 1996.