INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team TEXMEX

Efficient Exploitation of Multimedia Documents: Exploring, Indexing and Searching in Very Large Databases

Rennes

THEME SYM

Activity Report

2005

Table of contents

1. Team **1**
2. Overall Objectives **1**
2.1. Overall Objectives 1
3. Scientific Foundations **4**
3.1. Background 4
3.2. Document Description and Metadata 4
3.2.1. Image Description 4
3.2.2. Video Description 5
3.2.3. Text Description 5
3.2.3.1. Characterization of Huge Sets of Thematically Homogeneous Texts 6
3.2.4. Retrieval and Description Evaluation 6
3.2.5. Metadata Integrated Management, Selection and Mining 7
3.3. Efficient Exploitation of Descriptors and Metadata 8
3.3.1. Statistics and Data Quality over Huge Datasets 8
3.3.2. Multidimensional Indexing Techniques 9
3.3.2.1. Traditional Approaches, Cells and Filtering Rules 9
3.3.2.2. Approximate NN-Searches 10
4. Application Domains **11**
4.1. Still Image Database Management 11
4.2. Video Database Management 12
4.3. Textual Database Management 12
4.4. Robotics and Visual Servoing 13
5. Software **13**
5.1. Softwares 13
5.1.1. I-Description 13
5.1.2. Asares 13
5.1.3. Faestos 14
5.1.4. 2PAC 14
5.1.5. M-Tool 14
5.2. Experimental Platforms 14
6. New Results **15**
6.1. Image Retrieval for Large Databases 15
6.1.1. Image Description, Compression and Watermarking 15
6.1.2. PvS: Projections versus Segmentations 16
6.1.3. Management of Visual Features for Content-Based Image Retrieval 17
6.1.4. Coupling Action and Perception by Image Indexing and Visual Servoing 18
6.2. Text Retrieval for Large Databases 19
6.2.1. Natural Language Processing and Machine Learning 19
6.2.2. Intensive Use of Factorial Analysis for Text Mining: Indicators and Displays 20
6.3. Data Mining and Data Quality for Large Databases 21
6.3.1. Visualization and Web Mining 21
6.3.2. Quality-Aware Data Management and Mining 21
6.4. Multimedia Document Description 22
6.4.1. Segments Models for Video Description 22
6.4.2. Automatic Structuring and Labeling of Television Streams 23
6.4.3. Image and Text Joint Description 24
6.4.4. Text and Speech Joint Description 25

7. Contracts and Grants with Industry **25**
7.1. Contracts, Initiatives and Participation to Networks of Technological Research 25
7.1.1. Pôle de Compétitivité 25
7.1.2. Quaero 26
7.1.3. GIS Æternum Multimedia 26
7.1.4. Contract with Thomson 26
7.1.5. Contract with France Télécom 26
7.1.6. RIAM FERIA Project 26
7.2. European Initiatives 27
7.2.1. European Network of Excellence MUSCLE: Multimedia Understanding through Semantics, Computation, and Learning 27
7.2.2. European Integrated Project aceMedia: Integrating Knowlege, Semantics and Content for User-Centered Intelligent Media Services 27
7.2.3. European Integrated Project ENTHRONE Phase 1: End-to-End QoS through Integrated Management of Content, Networks and Terminals 27
8. Other Grants and Activities **28**
8.1. National Initiatives 28
8.1.1. ACI masses de données Remix: Indexing Large Amounts of Data with Reconfigurable Chips 28
8.1.2. ACI masses de données MDP2P: Managing Large Amounts of Data in P2P Systems 28
8.1.3. ACI masses de données DEMI-TON: Multimodal Description for Automatic Structuring of TV Streams 28
8.1.4. Action Bio-Info Inter-EPST: Parallel and Reconfigurable Architectures for Genomic Data Extraction 29
8.1.5. INRIA R&D action SYNTAX 29
8.1.6. Participation to National Working Groups 29
8.2. International Collaborations 30
8.2.1. Working Group Image Undertanding of ERCIM 30
8.2.2. Collaboration with Reykjavík University, Iceland 30
8.2.3. Collaboration with Croatia and Slovenia 30
8.2.4. Collaboration with Japan 30
8.2.5. Collaboration with University of Geneva, Switzerland 31
8.2.6. Collaboration with Dublin City University, Ireland 31
8.2.7. Collaboration with University of Montreal, Quebec, Canada 31
9. Dissemination **31**
9.1. Conference, Workshop and Seminar Organization 31
9.2. Involvement with the Scientific Community 32
9.3. Teaching Activities 34
9.4. Participation to Seminars, Workshops, Invited Conferences 35
10. Bibliography **35**

1. Team

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

Team Leader
Patrick Gros [Research Scientist (CNRS)]

Administrative Assistant
Maryse Auffray [TR INRIA, partial position in the project-team]

Faculty Members (University of Rennes 1)
Laure Berti-Équille [Associate Professor, in the team until August 31st]
Annie Morin [Associate Professor]
Pascale Sébillot [Associate Professor, until August 31st]

Faculty Members (National Institute of Applied Sciences of Rennes)
Pascale Sébillot [Full Professor since September 1st]
Anthony Remazeilles [Teaching Assistant (ATER)]

CNRS Staff Members
Laurent Amsaleg [Research Scientist]
Vincent Claveau [Research Scientist since October 1st]

Ph.D. Students
Siwar Baghdadi [CIFRE grant with Thompson, since December 1st]
Nicolas Bonnel [CIFRE grant with France Télécom R&D]
Manolis Delakis [MJENR grant, also with METISS]
Stéphane Huet [MJENR grant, also with METISS]
Zied Jemai [INRIA grant]
Anicet Kouomou-Choupo [MJENR grant, until September 1st]
Fabienne Moreau [INRIA grant]
Xavier Naturel [INRIA grant]
Mathias Rossignol [until November 1st]
François Tonnin [INRIA - Brittany grant, also with TEMICS]

Post-Doctoral Fellows
Son Doan [INRIA grant, since December 15th]
Arnaud Dupuis [DEMI-TON project]

Technical Staff
Philippe Daubias [FERIA project, until June 30th]
Cédric Dufouil [Junior technical staff, since September 1st]
Brigitte Fauvet [FERIA project, also with VISTA, until June 30th]
Pierre-Hugues Joalland [AceMedia project, since March 15th]
Boris Rousseau [Enthrone Project, until August 31st]

2. Overall Objectives

2.1. 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, how to index them in order 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.

Exploiting the content of large databases of digital multimedia documents is a problem with multiple facets. Moreover, 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 needs an efficient management of the primary and secondary memories, as well as nice interfaces and interactions 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 process 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 in quality as well as in speed;
- multimedia documents are not a simple juxtaposition of independent media, and it is important **to better exploit the existing links between the various media** composing a unique 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 turn 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 very precisely, to guide the system or to evaluate the results; he must be the one who controls the system.

We have adopted a matricial organization for laying out our research. On the one hand, we have expertise 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 focus our work on the questions where the team's multidisciplinarity appears to be 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 databases**, including the control of logical and temporal consistency, strategies of computation and selection of descriptors and metadata; ii) **statistical techniques** for the exploration of large volumes of data; iii) **indexing techniques** aiming at confining in the smallest possible area the exploitation of the data and thus avoiding an exhaustive processing whose cost is certainly controlled but prohibitive; 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 solely come from new description schemes nor new indexing schemes, but it requires to take into account all the various components of the system and their articulations. Thus, we 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 merging of 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 (NLP) tools could however offer 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 merge NLP and traditional keyword-based approaches.

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 environments than the ones in which they are usually used. The contribution of such a modification of the engines must also be established, which requires a precise work on the evaluation of information retrieval systems.

Third Axis of Research: Multimedia and Coupling Between the Media We study media coupling along in three directions. 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 NLP techniques on texts produced by speech transcriptions. As a matter of fact, speech carries a lot of semantic information and NLP techniques are among the most efficient ones to extract semantics 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 associated with images. This could help in obtaining 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 fall 2004, in the frame of S. Huet Ph.D. thesis, we have also begun the study of the coupling between speech and text together with the METISS Team. This work aims at adapting and inserting methods existing in the text analysis domain into speech recognition models to improve their performances in order to give indexing methods a better access to information such speeches may contain.

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 [111] and several solutions have been proposed to remedy this problem: correlograms [86], weighted histograms [58]...

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 researches have been carried out in the case of specific databases. Face detection and recognition is the most classical and important case, but other studies 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 [110].

The main stages of the method are the following. First, simple features are extracted from each image (interest points in our case, but edges and regions can be used too.) The most widely used extractor is the Harris [83] point detector which provides not very precise but "repeatable" points. Other detectors exist, even for points [95].

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 [57] or by properties established only experimentally [73], [72].

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 [74], Dufournaud [66] and Mikolajczyk [101], photometric invariance was demonstrated for grey-levels by Schmid [110] and for color by Gros [80]. 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 [69], or still image-based descriptors, but which can use the image temporal redundancy [82].

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 [59], [75]. 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 [109] 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 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 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 use domain-specific linguistic resources, both to disambiguate words and to propose equivalent formulations. These domain-specific resources are however not pre-existing and must be automatically extracted from corpora (collections of texts). Moreover, if one wants to use resources really adapted to one's text collection, prior to acquire them, one has to adopt a linguistic framework defining the semantic elements that are to be collected from corpora. In this respect, work in lexical semantics provides different theoretical models; let us cite three of them that are used in the TEXMEX project.

F. Rastier's differential semantic theory [105] 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. Within a given semantic class –group of words that can be exchanged in some contexts–, words share *generic semes* (*i.e.*, generic semantic features) 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).

In J. Pustejovsky's Generative lexicon theory [104], a so-called *qualia structure* is defined. In this 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.

The Meaning-Text Theory is a broad linguistic framework [99], whose lexicology part defines *Lexical Functions* [100] (LF). The LF are designed to encode every semantic relation of a word, such as syntagmatic relations (*e.g. mouse–to click*, *shower–to have*...) or paradigmatic ones (*e.g. professor–student*, *bee–honey*...).

Concerning the corpus-based acquisition of lexical resources, many researches have been undergone in the last decade. While most of them are essentially based on statistical methods, symbolic approaches also present a growing interest [48]. Relying on both methods, machine learning solutions are being developed in TEXMEX; they aim at being automatic and generic enough to give the possibility to extract from a corpus the kind of lexical elements required by a given task (for example, query expansion in an information retrieval system).

3.2.3.1. 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 evaluation campaigns are organized (NIST for sound and speech recognition, TREC for text in English, CLEF for text in various European languages, SENSEVAL or ROMANSEVAL for text disambiguation...).

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 [108].

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 well enough 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 [88] 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) [51], [61], [63], [78], [87]. Many standardized metadata such as in MPEG-7, TV-Anytime, *etc.* and also *ad hoc* content-descriptive metadata can be included in this classification [96], [98].

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...

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 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 (NN)*.

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 [53], [114] that divide the data space according to the distribution of data and *space-partitioning* [84], [113] 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 [81], 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 [114] is an extension that relies on spheres instead of rectangles. The SR-Tree [89] specifies its cells as being the intersection of a bounding sphere and a bounding rectangle.

Space-partitioning techniques like grid-file [102], K-D-B-Tree [107], LSD^{h}-Tree [84] 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 [60]. 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:

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

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

Unfortunately, the “curse of dimensionality” phenomenon makes these filtering rules ineffective in high-dimensional spaces [113], [56], [103], [60], [91].

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 [76] exploit the underlying correlation of vectors and/or their self similarity [91], 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 [112] and Li *et al.* with Clindex [94] 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.* [71] combine this approach 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. Yet, 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.

The VA-BND scheme [112] 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. It does not account for the very different data distributions possible in cells.

The AC-NN scheme for M-Trees [64] 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 [77] 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, it was observed [77] 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 [77] might, in certain cases, return less than k vectors in the result.

Probabilistic Approaches. DBIN [52] 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 [79] 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.

Rank Aggregation-based Approaches. Recently, Fagin *et al.* [70] proposed a framework for very efficiently evaluating single descriptor nearest-neighbor queries over high-dimensional collections. This framework is based on projecting the descriptors onto a limited set of random lines. Each random line is used to give a ranking of the database descriptors with respect to the query descriptor. These rankings are then efficiently aggregated to produce a fairly good approximation of the actual Euclidean k-nearest neighbors. The fastest algorithm to aggregate the rankings was called OMEDRANK.

The OMEDRANK algorithm has several nice properties: it is based on a cheap aggregation of rankings instead of a complex distance function; it uses standard B^+-trees to index the data, therefore handling updates gracefully; and it allows for a clever dimensionality reduction, by varying the number of random lines that are indexed.

4. Application Domains

4.1. Still Image Database Management

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

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's 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.

Traditional image indexing consists in automatically extracting from an image numerical descriptors representing the color, texture, interest points or other similar information. However, such descriptors are not relevant to tackle the problem of a “semantic” querying of an image database: how a customer can find the pictures of a sunset or the pictures of his daughter learning to swim? How an archivist in a news agency can find a relevant picture to illustrate an article dealing with poverty in India? One way to address this problem is to make the most of existing documents in which images and texts appear together, and then use relevant parts of the texts to index the images; nonetheless, the definition of such text-image indexing schemes are up to now under study.

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 switched 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.

Then, one can distinguish two types of digital files: private and professional files. On one hand, the files of private individuals include recordings of broadcasted programs and films recorded 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.

A crucial problem for many professionals is the need to produce documents in many formats for various terminals from the same raw material without multipliying the editing cost. The aim of such a *repurposing* is for example to produce a DVD, a web site or an alert service by mobile phone from a TV program at the minimum cost. The basic idea is to describe the documents in such a way that they can be easily manipulated and reconfigured easily.

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.

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 [68], [62], [97]. 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. Softwares

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 with the “Agence pour la Protection des Programmes” (APP) under the number
IDDN.FR.001.270047.000.S.P.2003.000.21000. (Correspondant: Patrick Gros.)

5.1.2. Asares

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, hyperonymy, qualia, lexical function...), 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. ASARES was deposited with the "Agence pour la Protection des Programmes". (Correspondant: Vincent Claveau.)

5.1.3. *Faestos*

FAESTOS (Fully Automatic Extraction of Sets of keywords for TOpic characterization and Spotting) 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 corpus deals with. 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 of keywords are employed in order to detect the presence of a topic in a paragraph, revealed by a keyword cooccurrence. Morevover, the system extracts from each keyword class a triple of words that permits an intuitive designation of the underlying topic. FAESTOS was deposited with the "Agence pour la Protection des Programmes". (Correspondant: Mathias Rossignol.)

5.1.4. *2PAC*

2PAC (2-Pass Acquisition of semantic Classes) brings together words used in a similar way in a topical sub-corpus such as that extracted by FAESTOS so as to build classes of words of similar meanings (" semantic classes ") specific to the use that is made of them in that given topic. It works by first computing general semantic proximities from the whole corpus, then using that information to perform a more in-depth analysis of one of the topical sub-corpora. The result is a classification tree from which classes must be extracted manually; a graphical user interface has been developed to ease that task of manual exploitation. 2PAC was deposited with the "Agence pour la Protection des Programmes". (Correspondant: Mathias Rossignol.)

5.1.5. *M-Tool*

An authoring tool (M-Tool) to enrich multimedia and audiovisual content with metadata has been developed for the European project ENTHRONE. M-Tool uses a unified metadata model based on MPEG-21 and TV-Anytime to federate a diversity of multimedia resources. For more information, visit the web site: http://www.irisa.fr/enthrone. M-Tool is going through the process of GPL licensing and is being deposited with the "Agence pour la Protection des Programmes". (Correspondant: Laure Berti-Équille.)

5.2. Experimental Platforms

Participants: Patrick Gros, Arnaud Dupuis, Cédric Dufouil.

Following the work done in the FERIA project, the team has developed a new activity around the problem of experimental platforms for automatic video analysis. Processing numerous data becomes quickly unmanageable when using traditional work stations. As an example, computing the local descriptors of 200 000 images takes 3 weeks when the load is distributed on 20 processors!

A first platform has been developed which includes a PC under windows with a TV acquisition device, and 2 RAID disks. Two bi-processor Linux based PCs are devoted to image and video processing. The acquisition and storage of new videos can be controled through a web site which makes it available for all the members of METISS, VISTA and TEXMEX concerned by video analysis, independently of the operating system used. This interface makes it possible to program new recordings, to visualize videos, and to navigate inside the ones associated with a TV-anytime metadata document.

A second platform uses a cluster of Apple servers available in the PARIS team. This cluster will be completed by new storage capacities. The goal of this platform is to test how a cluster architecture can solve our processing problems and to help the specification of the next generation platform which will be a cluster dedicated to video analysis only. A technical staff should be hired in December 2005 to specify this new platform, thanks to the support of Préfecture de Région Bretagne.

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 proved that content-based image description techniques are suitable for robust detection of illegal copies. As big databases are made of compressed images, a large amount of time could be saved by calculating signatures directly from the compressed images. Thus we designed compression schemes relevant to robust extraction of feature points and local signatures.

A crucial and preliminary point is the choice of the image representation which should be jointly suitable for compression and description. Critically sampled wavelets, such as those used in JPEG'2000 format, 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. We investigated several redundant transforms such as the Laplacian pyramid, the contourlet transform, the complex wavelet transform and the steerable transform. The latter one offers appealing characteristics for local description. Its energy stability allows us to design a point extractor whose robustness is just 10 % less than the multiresolution Harris point extractor. Its orientation selectivity is sufficient to build robust and discriminative descriptors based on D. Lowe's ideas.

The final problem is rarely addressed in the literature and consists in the proper quantization of transformed coefficients for both good image reconstruction and preservation of description quality. As the transform is redundant, one image has many possible representations. We first use POCS (Projection onto Convex Sets) to find a sparser representation and we adapt the classical technique in order to preserve the content of the neighborhoods of extracted points. Then, we design a compression scheme allowing the reconstruction of steerable coefficients from the information required by description, which is reduced to an energy and an orientation coefficient for each spatial point. The final step is the quantization of these coefficients. This compression scheme allows to detect illegal copies in image bases compressed at one bit per pixel. The two main perspectives are the reduction of compression time and the identification of the worst attacks for detection.

6.1.2. PvS: Projections versus Segmentations

Keywords: *Approximate Search Schemes, Curse of Dimensionality, Databases, Local Descriptors, Multidimensional Indexing Techniques, OMEDRANK, Random Projections*.

Participants: Laurent Amsaleg, Patrick Gros, Zied Jemai, Annie Morin.

This is a joint work with researchers from Reykjavík University. This work is done in the context of the INRIA Associate teams program. This program links two research teams (one INRIA, one foreign) willing to cross-leverage their respective excellence and their complementarity. The researchers directly involved in Iceland are Björn Þór-Jónsson (Associate Professor), Herwig Lejsek (MS) and Friðrik-Heiðar Ásmundsson (MS).

While OMEDRANK (briefly described in Section 3.3.2.2) performed very well in the experiments of R. Fagin *et al.* [70], which used rather small collections of global descriptors, it was not clear that it would perform well for large collections of local descriptors. The goal of our work was therefore to study the performance of OMEDRANK with local descriptors, and to find ways to make the performance acceptable for large image collections.

By analyzing the behavior of OMEDRANK with our settings (i.e. when dealing with large collections of local descriptors), we discovered that it performs poorly despite its nice properties. In a nutshell, while OMEDRANK reads only a small percentage of the whole database for every query descriptor, this turns out to be a significant amount of data when the database is very large and when there are many query descriptors. While the processing time of OMEDRANK can be improved through dimensionality reduction, the quality of the results suffers instead.

The key problem of the OMEDRANK indexing approach is that when projecting multiple dimensions onto a single line, many unrelated descriptors may be projected in between near neighbors, or “into-the-way” of the search. The indices used for OMEDRANK are formed by a single projection of each descriptor onto a random line. It can be shown that descriptors that are far apart on the projected line, are also far apart in the original space, while descriptors that are close on the projected line, may or may not be close in reality. With a large collection, it is likely that many descriptors will be projected into-the-way of the search in this manner.

The PvS-index [35] addresses this problem by 1) segmenting the projected line to separate descriptors that are known to be far apart, and 2) re-projecting the data of each segment onto a *new random line* to put distance between descriptors that appeared to be close on the first projected line, but were far apart in reality.

Therefore, our indexing strategy is based on repeatedly segmenting the descriptor collection into overlapping segments and projecting the data in the segments onto new random lines. While the index size grows due to the overlapping segments, our indexing strategy significantly reduces the amount of data that needs to be read for each query descriptor; in fact, we have designed the index to require only 3 disk reads per query descriptor, which results in very efficient query evaluation.

Finally, we ran a detailed performance study using a large collection of real descriptor data. Our results show that our proposed indexing scheme results in query execution that is both efficient and effective. This study shows, for example, that we can search if an image belongs to database in typically less than 16 seconds while an optimized sequential scan takes roughly 10mn. In this case, the database holds more than 30,000 images, or about 20,000,000 descriptors and the query images has about 500 descriptors. It is key to note here that the overall quality of the search is very good, with a recognition rate quite similar to the sequential scan.

Because searching with PvS only requires one disk read per query descriptor per index, we guarantee an upper bound on the query processing time, regardless of the size of the image collection. We therefore strongly believe that our scheme is able to scale to very large collections of images.

Guaranteeing effectiveness when searching a very large collection is clearly crucial. Our experiments gave evidences that a collection of descriptors has to be segmented enough to provide high-quality results. This translates into creating deep PvS indices consuming disk space but saving some processing cost at search time. If disk space is an issue, the quality of the search might be improved by increasing the number of PvS-indices, which in turn increase the CPU consumption counter-balancing disk-space consumption. Overall, we believe trading disk-space for CPU is a very nice tradeoff.

Clustering for Indexing. In parallel to the study mentioned 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 analyzed existing clustering algorithms and investigated their adequacy to be used in the specific context of CBIR systems [67]. 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 tradeoff between the size of such clusters and the cost to process them [42]. 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 [42]. We are currently designing a clustering scheme mixing both approaches.

In addition, we are investigating local clustering issues that seem better suited for high dimensional datasets. Recent work suggests the use of dimension reduction techniques to efficiently index lower dimensional spaces. Local clustering (or subspace clustering) deals with the problem of finding clusters in different subspaces all embedded into the same global dataset. Some experiments conducted on 30 157 images representing about 20 millions descriptors of dimension 24, showed that working on subspaces (with lower dimensions) can drastically reduce response time of on-line searching with no significant loss in effectiveness. We try to combine local clustering techniques to random projections followed by rank aggregations as mentioned above.

User Interaction. Traditionally, CBIRs take a long time to construct very detailed exact answers that are then presented to the user only once the calculation is over. Our studies showed that, in general, a fairly good intermediate result is available soon after the start of the search. Converging to the exact and final answer, however, requires a long time. We believe that users have visual perception skills that are unmatched by any software developed thus far. Therefore, it is beneficial to allow users to get access to the evolving result as it is incrementally built by the system. In this way, they can examine potential answers and intervene in the process, either accepting the answer or aborting the search. Building on this idea, we came up with specifications that resulted in a first implementation of a user interface, which allows to on-the-fly spy what the image search engine is doing and provides users with means to intervene in the process. That interface evolved a lot and its third version is currently in the final stages of construction. An early description of it was published recently [29].

6.1.3. Management of Visual Features for Content-Based Image Retrieval

Keywords: *Content-Based Image Retrieval*, *Progressive Query*, *Query-by-Example Execution Plans*, *Visual Features*.

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

The current state of the literature presents various approaches of visual content description for still images. Among these approaches, while some are specific to a field or to a particular use, others are general and it is not always easy for users to choose those which would satisfy them in a context of content retrieval. We focus our research on general global visual features such as color, texture, layout or shape at the pixel level for still images retrieval. In that context, the typical mode for querying in an image content-based information system is query-by-example, which allows the user to provide an image as a query and to search for similar images (*i.e.*, the nearest neighbors) based on one or a combination of low-level multidimensional features of the query example. Off-line, this requires the time-consuming pre-computing of the whole set of visual descriptors over the image database. On the other hand, the administration of very large collections of images accentuates the classical problems of indexing and efficiently querying information.

On-line, one major drawback of image content-based retrieval systems is that multidimensional sequential NN-search is usually exhaustive over the whole image set face to the user who has a very limited patience. Systems 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 and they return the best few matches to the user in a ranked result list, the most similar images first followed by the less similar ones. But, a subset of features could be sufficient enough to answer very quickly while offering an acceptable quality of results. Our research focuses on the elaboration of fully automatic and generic strategies of visual features usage for content-based retrieval on very large still image databases.

This year, we have proposed a technique for improving the performance of image query-by-example execution strategies over multiple visual features[33]. The work falls on the continuity of those proposed last year [92], [93] and includes first, the pre-clustering of the large image database and then, the scheduling of the processing of the feature clusters before providing progressively the query results (*i.e.*, intermediate results are sent continuously before the end of the exhaustive scan over the whole database). A cluster eligibility criterion and two filtering rules are proposed to select the most relevant clusters to a query-by-example. Experiments over more than 110 000 images and five MPEG-7 global features show that our approach significantly reduces the query time in two experimental cases: the query time is divided by 4.8 for 100 clusters per descriptor type and by 7 for 200 clusters per descriptor type compared to a "blind" sequential NN-search with keeping the same final query result. This constitutes a promising perspective for optimizing image query-by-example execution.

Besides, Annie Morin is currently developing local principal component analysis on the descriptor features, starting by a clustering of the most correlated dimensiosn for a given descriptor and then performing a PCA in each cluster of variables

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).

This work considers the control of robot motions, using visual information provided by an on-board camera. The navigation problem comes from the fact that the initial position of the robotic system and its desired position are described by two totally different images. This make impossible to use classical vision-based control laws, like visual servoing.

The approach we are working on belongs to the topological family, in which the environment is described by an image data-base acquired off-line. The localization of the robotic system, within this context, is nothing but the search in the data-base of images that are similar to the one describing the current position of the system. This is performed by using image retrieval techniques. The definition of the path to realize is then obtained by performing a search path into a graph associated to the image data base. This image path describes visually the environment the robot has to go through.

Specific visual measures are defined to control the conditions of observation in the current view of the points that have been matched between the consecutive images from the path. An original control law derived from the classical visual servoing, called a qualitative visual servoing, has been proposed to make these measures reach a confident interval. The robot progresses towards its desired position without being imposed to reach each position associated to the images from the path.

Results obtained on the articulated arm for planar motion, and in simulation for $3D$ motions and $3D$ environments confirm the validity of our approach [40]. We are now considering the application of this formalism onto our real car-like robot, the Cycab. Therefore, the different steps involved within the control loop has to be reconsidered in order to be robust, and well-adapted to this kind of robot. This work is realized in collaboration with Sinisa Segvic, Post-Doc in the Lagadic team.

This year, The first step considered has been the point tracking. This step is a critical issue in vision-based robot navigation, since the robot motions are defined with respect to these landmarks. One needs then features that are robustly located, but also that have a sufficiently long life-time in order to define the robot motion during all the navigation.

A classical procedure for monitoring the tracking quality consists in requiring that the current features nicely warp towards their appearance in a reference frame. This procedure is valid when considering features that are projection of planar $3D$ patches. However, in real application, a non negligible amount of points contain occluding boundaries. This fact becomes even more problematic when considering a camera on-boarded in car moving along straight sections of a road, since the motion of the camera induces a permanent increase of the apparent feature size.

To deal with such points, a multi-scale monitoring is implemented. It maximizes the point lifetime, while also detecting the tracking failures. A technique is proposed for inferring the parts of the reference which are not projected from the same $3D$ surface as the patch that has been consistently tracked until the present moment. Experiments realized on real sequences taken from cars driving trough urban environments have shown that the proposed method is quite effective in increasing the average feature lifetime, especially in sequences with occlusions and large photometric variations.

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, Information Retrieval, Lexical Semantics, Machine Learning, Natural Language Processing, Semi-Supervised Learning.*

Participants: Laurent Amsaleg, Vincent Claveau, Son Doan, 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 three 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 and other relations defined in the Meaning-Text Theory framework, 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 2005, our work has especially concerned the 2 following points:

1. Acquisition of semantic lexicons based on Rastier's differential semantics: automatic detection of semes within semantic classes
   We have completed a first round of research aiming at the automatic acquisition from a morpho-syntactically tagged corpus of lexical semantic information based on F. Rastier's Interpretative Semantics. Following the principles of meaning representation defended by Rastier, we had developed a methodology in 3 stages: the vocabulary used in the corpus is first split into *domains* gathering words used to address a certain topic; then, the words of each domain are arranged into *taxemes*, semantic classes of words that have a similar meaning *in that domain*; finally, we put into light *specific semes*, that is, fine-grained semantic information distinguishing the meanings of words within a taxeme.
   The two first tasks are performed by two systems, FAESTOS and 2PAC, respectively presented in the 2003 and 2004 activity reports, which have since then been finalized (and incidentally named). This year has been especially dedicated to the third and last stage of semantic information extraction, the

detection of specific semes, that corresponds to a level of lexical semantic subtlety quite unexplored by automatic means so far. By studying the common points and differences between the words used in the sentences where the compared words occur, we are able to suggest parallels between the “ distinction axes ” drawn by pairs of words: for example, the system can point out the similarities between the oppositions *data/information*, *hardware/software* and *device/system*, that can be interpreted as a *material/immaterial* distinction. The developed system makes a series of propositions of such “ similarities of distinction ” between pairs of words, which must still be weeded out manually[11].

2. 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 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. After a state-of-the-art of the wide domain of NLP+IR[46], this year has been dedicated to the two following points:

   – determining a way to model and integrate inside a unique IRS all the various kinds of linguistic knowledge (morphological, syntactic, semantic...) that NLP techniques can acquire, contrary to the option generally followed by systems that only focus on one kind of resources;

   – finding accurate scenarii to actually use this knowledge during a real human-IRS interaction.

   A first prototype of linguistically informed-IRS, based on VSM, is currently under development.
   Concurrently to this aim of efficiency, we also study a second aspect of this quest of performances: IRS have to be relevant and to provide better answers, but they also have to be rapid and to answer quickly to users, even when they are questioning huge textual databases. Son Doan begins in December 2005 a post-doctoral research which aims at finding a way to merge the models that are the most suitable to integrate linguistic information and the best search algorithms developed within the scientific field of text databases (TDB). In this domain, very efficient algorithms for IR on huge textual databases have been developed, that however use over-simplified models of linguistic knowledge.

6.2.2. Intensive Use of Factorial Analysis for Text Mining: Indicators and Displays

Keywords: *Correspondance Analysis*, *Vizualisation*.

Participant: Annie Morin.

Textual data can be easily transformed in frequency tables and any method working on contingency tables can be used to process them. Besides, with the important amount of available textual data, we need to find convenient ways to process the data and to get invaluable information. It appears that the use of factorial correspondence analysis allows to get most of the information included in the data. But even after the data processing, we still have a big amount of material and we need visualization tools to display it. We study the relevance of different indicators used to cluster the words on one side and the documents on the other side and we are concerned by the vizualisation of the outputs of factorial analysis: we need to help the user to go through the huge amount of information we get and to select the most relevant points. Most of the time, we do not pre-process the texts: that means that there is no lemmatization.

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.

We are currently working on the dynamic generation of 2D and 3D multimedia interactive presentations. The context of this work is the visualization of web search results. So the issue is how to effectively represent the results coming from a textual search engine. This work is done in cooperation with France Telecom R&D (thanks to a CIFRE contract).

In the topic of this thesis, we can distinguish two main steps. The first one consists of organizing efficiently the results of a web query. In other words, we investigate on-the-fly clustering methods and we only consider statistical approaches which have to be deterministic too. More precisely we are interested in a particular unsupervised clustering method: the self-organizing maps. Indeed, this method enables to cluster the results **and** to organize the clusters thanks to a non-linear projection on a predefined map. To correctly adapt this method to our context, we discuss on many points such as the distance to use (comparisons between various distances or dissimilarities) [14] or the weighting schemes to apply. Another important point is the quality evaluation of the proposed classification.

The second step concerns the visualization of the organized results. The goal is to define visualization metaphors to improve the graphical representation of search results. Various interactive and adaptive metaphors are then proposed. The main one is the city metaphor [24] which enables the user to interact with web search results which are represented in a 3D virtual city. Some key points for designing successful 3D metaphors are also discussed [25]. All the proposed interfaces are hybrid *i.e.* composed of a 2D part (Java applet) and a 3D scene (defined in VRML). This visualization step has to be evaluated too. However a reflexion is needed before to be able to successfully evaluate this kind of interface.

These researches are integrated in a prototype developed by France Télécom R&D [23] and a first user study has already been carried out. Finally we can consider our approach as a post-processing step for a search engine. Obviously the necessity to have an effective indexing and an effective retrieval process is a complementary work to this one [26].

6.3.2. Quality-Aware Data Management and Mining

Keywords: *Data Cleaning Techniques*, *Data Quality Metrics*, *Quality-Adaptive Query Processing*, *Quality-Aware Knowlegde Discovery in Databases*.

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 [65], [85], [106]. But data quality problems can be adressed with several methods from many disciplines such as statistics, database techniques and metadata [54]. Our approach is threefold:

- the preventive approach: we propose system-oriented techniques for continuous controlling and measuring various dimensions of the quality of data flows (*e.g.*, freshness, completeness, consistency) entering into the database or data warehouse systems [22];
- the diagnostic approach: we adapt and optimize data cleaning and data mining techniques for detecting data anomalies (*e.g.*, duplicate, inconsistency, outlier) on structured and semi-structured data (*i.e.*, relational records or XML data) [20] and also for ensuring the optimal cost/quality trade-off of knowledge discovery and data mining (KDD) processes[19];
- the adaptive approach: we propose et optimize the processing of quality-extended queries of both structured and semi-structured data query languages (SQL and XQuery) and we propose quality-adaptive recommendation of data [21].

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 previous work [55], we proposed 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*.

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: Siwar Baghdadi, 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 Thomson as external partner.

We address the problem of an efficient multimodal fusion for video structure analysis. Hidden Markov Models provide a powerful framework for multimodal video modeling and structure analysis. The drawback of HMMs is the requirement all the modalities of a video document to share the same frame rate and the same topology, as well. In this year, we have studied video indexing with Segment Models, a generalization of HMMs, where the fusion of different modalities can be performed in a more flexible way by relaxing the synchrony constraint. We extend previous work carried out by E. Kijak [90] on tennis broadcasts analysis based on HMMs. In this type of video, game rules as well as production rules result in a structured document. This prior structural knowledge we have on the video is easily incorporated in Segment Models, as in Hidden Markov Models. Our aim is to recover this structure in order to construct the table of contents of the video by segmenting it in human meaningful scenes. The semantic indexes thus obtained can meet the high-level user needs or can be used for managing large video databases.

Before proceeding to the modeling stage, we need to identify the large homogeneous segments that exist in both the video and audio tracks. The video is segmented into shots by hard cut and dissolve transitions detection. We then extract from each shot a feature vector whose features are the presence/absence of dissolve, the shot length and a visual similarity to a key frame representing the prototype court view of the video. This key frame is extracted by a fully automated procedure that results into a key frame unique for every tennis match. Similarly, the soundtrack is segmented into homogeneous segments in which we detect the presence/absence of the following audio events: ‘ball hits’, ‘applause’ and ‘music’, based on GMMs and training data annotated by human listeners.

Our aim is to decode the tennis match according to some pre-identified scenes, namely first missed serve and exchange, exchange, replay and break, defined on top of the video shot segmentation. The succession of the scenes is modeled by an ergodic HMM. In the HMM framework, the content of each scene is modeled also by an HMM, resulting into a large ergodic HMM of 12 hidden states that models the video as a succession of shots. The video is decoded via the Viterbi algorithm and finally we segment the video into scenes having the semantic label of each shot. The multimodal integration with HMMs is achieved by simply concatenating to the visual feature vectors binary descriptors of the audio events occurring in the corresponding shot.

In the Segment Models framework, each scene corresponds to a segment. The Viterbi decoding for the Segment Models involves the detection of the scene boundaries and the semantic labeling of each scene as well. Using some straightforward optimization techniques, the computational cost of this enhanced decoding can scale linearly with time, as with HMMs. The video content of each scene is modeled by HMMs, which now serve as observation scorers. For the audio content, we can use shot-based audio descriptors as in HMMs. But in Segment Models we can build independent models for the audio modality that operate in their own sample rate. In this way we can fuse audio and video features that can be asynchronous inside the scene boundaries. We have tried to model the audio content firstly by bigram models of succession of audio events. We then attempted to model it directly on top of the audio cepstral coefficients instead of using manually extracted mid-level features. With this approach we avoid the pre-segmentation and pre-classification stages for the detection of sound classes which are always difficult to perform in a satisfactory level. We used again HMMs as observation scorers, trained with the HTK toolkit.

When adding no audio information, Segment Models outperform HMMs. The addition of shot-based audio descriptors or asynchronous scene-based audio bigrams give a performance at the level and better compared to HMMs. The integration of audio models of low-level features can marginally improve the performance of the unimodal SMs. However, the results so far are encouraging as these models are built from-the-scratch and operate in a high-dimensional space. Further experimentation may give better performance.

Future work involves the integration of text in the models. This could be at first the speech transcription of the speakers and the referee. Another source of textual information are the score indications appearing on the screen, a modality that is even more asynchronous and harder to integrate. A second direction of future work includes the incorporation of high-level game semantics into the Segment Model.

6.4.2. *Automatic Structuring and Labeling of Television Streams*

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

Participants: Patrick Gros, Xavier Naturel.

Our work on this topic is done in collaboration with INA.

Television archives are very difficult to handle. The size of these archives is in itself challenging but retrieving meaningful information from a TV archive is also painstakely hard. Video is not a structured document and for now, only manual indexing can deal with the queries that a video database system might face.

The first step of a television structuring process is to isolate regular programs from commercials and other non-program segments. Various methods have been proposed for detecting commercials in a television stream. However, these previous researches are not satisfactory since they are not applied on real-world, large television corpus, and are evasive about whether they detect only commercials or the whole non-program segment. This might be questionned since these methods have not been seen as a way of structuring television streams but just as a tool for ad-skipping.

Detecting program segments is only the first step. Our goal is to index a television stream as thoroughly as possible. Each identified program segment should have a precise label, and the various elements in the non-program segments should also be identified and labeled. This is of course very interesting for a retrieval application, but it may also be used for computing statistics on channel habits (percentage of auto-promotion, average timelapse between supposed diffusion time and actual broadcast). This is also useful for enforcing legal regulations such as monitoring the time slots devoted to commercials.

In our work, a standard method has first been used to detect non-program segments. It consists in detecting simultaneous occurrence of monochrome frames and silence. This proves very effective for detecting commercials although this is not precise enough for detecting the precise boundaries of non-program segments.

Our contribution lies in the way we use repetitions to help both the macrosegmentation and the labeling. Non-program segments are composed of elements that are repeated, with various frequencies and lifetime. Thus, the idea is to detect non-program segments as a concatenation of previously seen elements. We therefore developped a method for detecting duplicate or near-duplicate video elements.

This method differs strongly from other video search methods since it is not based on similarity search but on exact matching. This allows much more efficient retrieval but also need great care in the construction of the feature on which the matching is based. We have chosen to build a signature for each image of a video from the low frequency DCT coefficients. These coefficients are binarised using their median value as a threshold. Using this signature, we make the assumption that two video duplicates share at least one common signature, and that they can be retrived using exact matching, and then further analysed by a similarity distance. This assumption holds while the set of transformations used for transforming a video sequence into its near-duplicate is very reduced. This set nonetheless includes gaussian noise, temporal editing, inserted text, color change... This method is however not suited for detecting tampering or similarity search, but is effective to detect near-duplicates in a video structuring purpose. Results on television sequences are satisfactory, with precision and recall close to 98 %.

We have used this method coupled with a standard monochrome frames/silence detector to perform an enhanced macrosegmentation. This segmentation is labeled with the information coming from the TV program guide using a dynamic time warping algorithm (DTW). The DTW allows us to deal with inacurracies in both the macro-segmentation and in the TV guide. Results show up to 98 % of the shots correctly labeled on a 24-hour video.

6.4.3. Image and Text Joint Description

Keywords: *Image-Text Interaction*.

Participants: Vincent Claveau, Patrick Gros, 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 texts to disambiguate or improve the understanding of the text.

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: Vincent Claveau, Stéphane Huet, Pascale Sébillot.

A lot of sound documents contain speech. Speech recognition is used to automatically index and exploit them, particularly when the database to analyze is voluminous. Speech recognition systems (SRS) produce a textual transcription from sound documents, thanks to acoustic and linguistic clues. However, most of current SRS are more based on purely statistical methods (study of n-grams in corpora) than on linguistics. Stéphane Huet's Ph.D. thesis (under the double supervision of Guillaume Gravier, from the METISS Project-Team, and Pascale Sébillot) aims at adapting Natural Language Processing (NLP) techniques to help the transcription of sound documents.

During 2005, a state-of-the art of the domain has been realized. Both the way linguistics that can be introduced during the transcription process, and the attempts previously made to integrate some morphological, syntactic, or semantic information, have thus been studied. This work will be published in 2006 as an internal report. In the same time, we realized first experiments on the behavior of NLP tools on spoken documents, including automatic transcriptions. The main difficulties we were faced with were the lack of structure of these documents, where there aren't paragraphs or even sentences, and the noise of automatic transcriptions, as the hypotheses made by the SRS may include many erroneous words. We studied several Part of Speech (PoS) taggers (tools at the base of any NLP process that identify if a word is a noun, a verb, an adjectif, *etc.*, and if it is at the singular or plural form, conjugated, *etc.*): rule-based systems like Brill's tagger, as well as HMM-based taggers. Without an adjustment of theses taggers to the specificities of spoken documents, more than 90 % of words of automatic transcription were already correctly tagged, while the usual results for written documents are of about 95 %. We decided to use information brought by PoS tagging to improve transcriptions produced by METISS Project-Team's SRS. Indeed, for example, French language has numerous homophones that only differ according to the number and gender agreement, which can introduce mistakes in the word error rate of SRS. We tried to improve transcription by rescoring the 10-best hypotheses. Some mistakes were corrected, like number and gender information, but this tagging introduces in turn new mistakes as it does not use any acoustic cues brought by the SRS.

We plan to go on this study of PoS tagging and find a way to select from their results only correct and relevant information. Examining other areas of linguistics like thematic detection is also foreseen to improve SRS results. The work previously done by TEXMEX Project-Team in this domain[11] can indeed be adapted to bring useful information during the transcription process.

7. Contracts and Grants with Industry

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

7.1.1. Pôle de Compétitivité

Participant: Patrick Gros.

The French government organized in 2005 competitiveness poles (*pôles de competitivité*) in France to strengthen ties in given region between industries (big and small companies), research labs (both public and private ones) and teaching institutions (universities and schools of engineering). We play an active role in the pole called "Images and networks" whose main actors are Thomson, France Télécom and Thales and which is located in Brittany and Pays de la Loire. In this frame, we are part of a common project called "Semantic images" with France Télécom, Thomson, TDF, and the VISTA, METISS and R2D2 teams of IRISA. This project should be 2 year long and should start in 2006.

7.1.2. Quaero

Participant: Patrick Gros.

The French and German governments have launched a new initiative to build big common projects. We are part of the project named Quaero and which aims at developing technologies to create and exploit multimedia documents. The main partners are Thomson, France Télécom, DGA, CNRS. Many SME participates to this project like LTO, Exalead, Versys, Jouve, Bertin... The project is organized in four parts: content creation, content finding, core technologies and corpora. We are involved in the third of these parts. Our participation will mainly concern multimedia analysis and fast search algorithms.

7.1.3. GIS Æternum Multimedia

Participant: Patrick Gros.

During this year, a long negociation was conducted between INRIA, CNRS, INA and Thomson to define the terms of a long term collaboration contract (*GIS : groupement d'intérêt scientifique*). An agreement was found, especially on intellectual property exploitation and the contract should be signed in December.

7.1.4. Contract with Thomson

Participants: Patrick Gros, Siwar Baghdadi.

Duration: 36 months, starting in December 2005.

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

7.1.5. Contract with France Télécom

Participants: Annie Morin, Nicolas Bonnel.

Duration: 36 months, starting in December 2002.

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

7.1.6. RIAM FERIA Project

Keywords: *TV Archives, Video Databases*.

Participants: Patrick Gros, Philippe Daubias, Brigitte Fauvet.

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.

The basic tools of still image analysis we were in charge of were almost finished. We focused our work on the development of complete processing sequences. This implied to refine the input and output of many modules and to develop 5 new modules (in conjunction with VISTA). At the end, two sequences have been completed: one for the FIDELIO demonstrator which allow an easy navigation in an opera broadcast, the second for the PACE application which allow to publish a collection of programs on the web and to navigate this collection easily.

7.2. European Initiatives

7.2.1. 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, Manolis Delakis.

Duration: 4 years, starting in April 2004. 42 partners. Prime: ERCIM, scientific coordinator: Eric Pauwels, 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).

TEXMEX is mainly active in the WP6 (multimodal analysis) through the work of M. Delakis (see Section 6.4.1).

7.2.2. 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, Pierre-Hugues Joalland.

Duration: 48 months, started in January 2004. 15 partners. Prime: Motorola Ltd.

The goal of this project is to encode multimedia document for their diffusion on networks like Internet, telecommunication networks or broadcasting systems. This new encoding scheme is based on autonomous entities called ACEs (standing for Autonomous Content Entity). Each entity is made of data, related metadata and an intelligence layer.

ACEs are dedicated at storing, retrieving and communicating documents in an efficient and autonomous way. It supports self-organisation, self-annotation and self-adaptation according to the current user's preferences and devices. Additional embedded mechanisms are semantic detection, fast retrieval and relevance feedback.

TEXMEX team is responsible for the methods related to indexing, intelligent search and fast retrieval of ACE documents. The ACE documents will be described by both conceptual and content-based descriptors. Our algorithms compute the list of the most similar documents according to a query, by matching their numerical descriptors. Currently, the provided methods are the optimised exhaustive search and a based-on kd-trees method. Other methods will be evaluated.

7.2.3. European Integrated Project ENTHRONE Phase 1: 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: 2 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 harmonize 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 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 [49]. Our work under the work package 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. We developed the metadata model based on those studies and implemented the metadata authoring tool that serves as a proof of concept for this work[41].

8. Other Grants and Activities

8.1. National Initiatives

8.1.1. ACI masses de données Remix: Indexing Large Amounts of Data with Reconfigurable Chips

Participant: Laurent Amsaleg.

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

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.

8.1.2. ACI masses de données MDP2P: Managing Large Amounts of Data in P2P Systems

Participant: Laurent Amsaleg.

Duration: 3 years, starting in September 2003. Joint work with PARIS and ATLAS.

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).

8.1.3. ACI masses de données DEMI-TON: Multimodal Description for Automatic Structuring of TV Streams

Participants: Manolis Delakis, Stéphane Huet, Patrick Gros, Xavier Naturel, Pascale Sébillot, Arnaud Dupuis, Cédric Dufouil.

Duration: 3 years, starting in April 2005. Partners: INA, METISS project-team.

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 have 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.

The first year of this project was dedicated to define places of collaborations between the three involved groups and to start the work in these three directions:

- A first common work has been carried out to coordinate the effort done by the three partners to develop experimental platforms. This work will continue what has been achieved within the FERIA project, with new developments to take the hardware available at each partner's site;
- Macrosegmentation of TV streams has been studied: the learning based approach of INA will be coupled with the signal based approach of TEXMEX;
- The interaction between speech transcription and natural language processing tools has been studied in a collaboration between TEXMEX and METISS.

At IRISA, a first platform to record and analyze TV streams has been developed (see Section 5.2). It is based on a Windows server in charge of recording and distributing the content and two Linux PCs in charge of processing tasks. It is now available to all the members of METISS, TEXMEX and VISTA through a simple web site.

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

Keywords: *Architecture Reconfigurable, FPGA, Genomics*.

Participant: Laurent Amsaleg.

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

This work 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.

8.1.5. INRIA R&D 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.

8.1.6. Participation to National Working Groups

- 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 RTP 33 (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 6^{th} Framework Program. Since then, it has no independent activitee anymore.

8.2.2. Collaboration with Reykjavík University, Iceland

Keywords: *Approximate Search Schemes, Content-Based Image Retrieval Systems, Curse of Dimensionality, Local Descriptors, Random Projections, User Interface.*

Participant: Laurent Amsaleg.

This collaboration is done in the context of the INRIA Associate Teams program. This program links two research teams (one INRIA, one foreign) willing to cross-leverage their respective excellence and their complementarity. The researchers directly involved in Iceland are Björn Þòr-Jònsson (Associate Professor), Herwig Lejsek (MS) and Friðrik Heiðar Ásmundsson (MS).

Image databases, and content-based image retrieval systems in particular, have become increasingly important in many applications areas. Moreover, new applications exploiting fine detail of images are now fast emerging thanks to recent and modern image processing techniques. While extremely effective (they return high quality results), these image processing techniques are very inefficient (they answer very slowly) due to their complexity and because of the inadequacy of traditional lower layers of software. The goal of our collaboration is to work on a project called Eff^2 in order to research and develop techniques that integrate efficiency and effectiveness in content-based image retrieval systems.

The Eff^2 project has a history of collaboration: many visits have taken place, and many exchanges of code and data have fostered our work. B. Þór-Jónsson visited IRISA four times (1 week in March 2002, 4 weeks in June 2003, 3 days in September 2004, 1 week in October 2004, with H. Lejsek). L. Amsaleg visited Reykjavík University twice (1 week in December 2003, 2 months in summer 2004). Further visits by L. Amsaleg, B. Þór-Jónsson and 3 of its students are planned for fall 2005. Overall, B. Þór-Jónsson and L. Amsaleg have jointly worked with 14 students at the undergraduate and MS level. Extensive work with four of them was achieved during the summer of 2004.

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, department of Electronics, Microelectronics, Computer and Intelligent systems, 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.

Last year, we started a cooperation on medical data and studies (a report is available on http://lis.irb.hr/IDADM/). For the second year of our Egide program, we focused first on the problems of overfitting during the learning step with the members of the Institute Rudjer Boskovic implied in our project and the laboratory ERIC of the university of Lyon. With members of the faculty of Electrical engineering and computing of the university of Zagreb, we mined textual data in Croatian using on one side some specific linguistic tools for Croatian language and.on the other hand, factorial correspondence analysis. The last part of our collaboration is concerned by Orange, a data mining library for Python developed by Blaz Zupan at the university of Ljubljana.

8.2.4. Collaboration with Japan

Keywords: *Computational Linguistics, Natural Language Processing Interactive Environment for Human Learning, Video Retrieval.*

Participants: Vincent Claveau, Annie Morin, Pascale Sébillot.

Pr. Ichiro Ide from the university of Nagoya and also researcher at the NII of Tokyo was a visiting professor at University of Rennes 1 and at IRISA in the TEXMEX project-team during one month in April 2005. His research interests are in integrated media processing, including natural language processing, image processing and audio processing.

Pr. Kumiko Tanaka-Ishii from the university of Tokyo gave a seminar on Predictive Text Entry System for Small Mobile Devices using Adaptive Language Models. We plan to settle in the future some cooperation projects on Multimedia tools for Foreign Multi-Language Learning.

8.2.5. Collaboration with University of Geneva, Switzerland

Participants: Laurent Amsaleg, Patrick Gros.

Following the CVBD'04 and CVDB'05 workshops we have organized, first contacts have been established with the VIPER group of University of Geneva headed by Stéphane Marchand-Maillet. We have submited a common PAI project whose acceptance should be known in December 2005. The main focus of this project is on video processing of very large collections.

8.2.6. Collaboration with Dublin City University, Ireland

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. A corpus of several hundreds of thousands of images has been shared between DCU and TEXMEX to make common experiments.

8.2.7. Collaboration with University of Montreal, Quebec, Canada

Participants: Vincent Claveau, Stéphane Huet, Fabienne Moreau, Pascale Sébillot.

A one-year collaboration agreement has been signed between Montreal (OLST research project) and Rennes 1 (TEXMEX research project) universities in 2005. The cooperation aims at automatically detecting in textual corpora paradigmatic relations between lexical units or terms.

9. Dissemination

9.1. Conference, Workshop and Seminar Organization

- A. Morin organized a seminar at the Institut Henri Poincaré on free software and open source for data mining with Blaz Zupan and Janez Demsar from the university of Ljubljana, Yves Le Chevallier INRIA Rocquencourt and Ricco Rakotomalala from the ERIC laboratory, Lyon.
- L. Amsaleg, B. Þór-Jónsson and Vincent Oria (New Jersey Institute of Technology, USA) were able to convince ACM to co-locate with SIGMOD a workshop they created in 2004 http://cvdb04.irisa.fr. This workshop was entitled "First Computer Vision Meets Databases (CVDB) Workshop". This workshop was held in Paris, France, on June 13, 2004, and was co-located with the 2004 ACM SIGMOD/PODS conferences (the most prestigious international forum on databases) and was attended by forty-two participants from all over the world. We created this workshop because we wanted to understand why only few researches in the computer vision community have adopted any of the indexing schemes that have been designed by database researchers. We discovered many valid scientific reasons but also that there was a great gap between the computer vision and the database communities. Our 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. 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 [47]. We wrote a workshop report in ACM Sigmod Record [50]. 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 held a second edition of CVDB in co-location with SIGMOD/PODS in Baltimore, Maryland in June 2005 http://cvdb05.irisa.fr. 9 papers were presented, 2 keynotes were given and a panel focused on "Multimedia applications: Beyond similarity searches" [9]. We are happy to witness that other scientists share our visions by, among other things, participating to CVDB. We will of course apply again for co-location with SIGMOD/PODS in Chicago, Illinois in June 2006.

- L. Amsaleg was head of the organizing committee of BDA'05, the main French conference in the field of databases. This conference was held in St Malo from October the 17th to Oct. the 20th. It gathered more than one hundred persons.
- L. Berti-Équille, Carlo Batini (University di Milano) and Divesh Srivastava (AT&T Research Lab) organized the "Second International Workshop on Information Quality in Information Systems (IQIS)". This workshop was held in Baltimore, MD, USA, on June 17, 2005. The workshop was co-located with the 2005 ACM SIGMOD/PODS conference and was attended by forty-six participants from all over the world. Two notorious key note speakers were invited to give a talk: Hector Garcia-Molina (Univ. Sanford) and Willam E. Winkler (Census Bureau, USA).
- L. Berti-Équille, and Fabrice Guillet (University of Nantes) organized the "First National Workshop on Data and Knowledge Quality (DKQ)" in conjunction with the French conference “Extraction et Gestion des Connaissances (EGC)” in Paris, on January 18, 2005. The workshop was attended by thirty-two French participants (both academic and practionners).

9.2. Involvement with the Scientific Community

- L. Amsaleg:
  - was a program commitee member of the 2nd Conférence en Recherche d'Infomations et Applications, CORIA'05, Grenoble, March 2005.
  - was a program commitee member of the 1st IEEE International Workshop on Managing Data for Emerging Multimedia Applications, EMMA'05, Tokyo, April 2005.
  - was a program commitee member of the 11th International Workshop on Multimedia Information Systems, MIS 2005, Sorento, Italy, September 2005.
  - was a program commitee member of the 3rd Manifestation des jeunes chercheurs francophones dans les domaines des STIC, MajecStic 2005, Rennes, France, November 2005.
  - is the 2005 ACM SIGMOD Workshops Coordinator.
  - was a workshop co-chair, Second Computer vision meets Databases, CVDB05, Baltimore, USA, June 2005.
- L. Berti-Équille:
  - was a member of the program committee of CORIA'05 (National Conference on Information Retrieval and Applications), March 2005.

- was a member of the program committee of IQIS'05 (International Workshop on Information Quality in Information System) co-located with the ACM SIGMOD/PODS conference, Baltimore, MD, USA, June 2005.
- was prime chair of the organizing committee of IQIS'05 (International Workshop on Information Quality in Information
System) sponsored by the ACM. The workshop was held in conjunction with 2005 ACM SIGMOD/PODS conference, Baltimore, MD, USA, June 2005.
- was a member of the program committee of IQ'05 (International Conference on Information Quality) at the Massachussets Institute of Technology, Cambridge, USA, November 2005.
- was a member of the program committee of QoIS'05 (International Workshop on Quality of Information Systems) in conjunction with the ER International Conference, Klagenfurt, Austria, October 2005.
- was prime chair of the organizing committee of DKQ'05 (National Workshop on Data and Knowledge Quality) co-located sponsoredwith the French conference "Extraction et Gestion des Connaissances (EGC'05)", Paris, January 2005.
- was a member of the organizing committee of BDA'05 (French Conference on Advanced Databases), Saint-Malo, October 2005.
- was prime chair of the organizing committee of a plenary meeting of the European project ENTHRONE, Rennes, January 2005.
- is a permanent member of the editorial board of the International Journal intitled "Journal of Digital Information Management".
- is a permanent member of the editorial board of the French Journal intitled "Ingénierie des Systèmes d'Information".

- P. Gros:

  - was a member of the program committee of the Second COnference on Information Retrieval and Application CORIA'05, Grenoble, March 2005.
  - was a member of the program committee of the CORESA workshop (Compression et repsentation des signaux audiovisuels), Rennes, November 2005.
  - was a member of the program committee of the Conference on Image and Video Retrieval CIVR'05, Singapore, July 2005.
  - was a member of the program committee of the Conference on Content-Based Multimedia Indexing (CBMI'05), Tampere, Finland, June 2005.
  - was a member of the program committee of the workshop TAIMA'05 (Atelier sur le traitement et analyse de l'information : méthodes et applications), Hammamet, Tunisia, September 2005.
  - was a member of the program committee of the Second International Workshop on Computer Vision meets Databases, Baltimore, Maryland, USA, June 2005.
  - was a member of the program committee of the Pacific Rim Conference on Multimedia, Jeju, Korea, November 2005.
  - is a member of the editorial board of the French journal intitled "Traitement du Signal".

- A. Morin:

- is a member of the program committee of ITI 2005 (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 - Cahiers Romans de Sciences Cognitives.
  - is an associate editor of Jedai (Journal Électronique d'Intelligence Artificielle).
  - was a member of the program committee of GL'05 (3^{rd} International Workshop on Generative Approaches to the Lexicon), Geneva, Switzerland, May 2005.
  - was a member of the program committee of LLL'05 (4^{th} International Workshop on Learning Language in Logic), Bonn, Germany, August 2005.
  - was a member of the program committee of CORIA'05 (2^{nd} National Conference on Information Retrieval and Applications), Grenoble, France, March 2005.
  - was a member of the program committee of TIA'05 (6^{th} National Conference on Terminologie et Intelligence Artificielle) Rouen, France, April 2005.
  - was a member of the program committee of TALN'05 (12^{th} National Conference on Traitement Automatique des Langues Naturelles), Dourdan, France, June 2005.
  - was a member of the program committee of MajecStic'05 (3^{rd} Manifestation des jeunes chercheurs francophones dans les domaines des STIC), Rennes, France, November 2005.
  - 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, P. Gros and P. Sébillot: Multimedia Indexing: Techniques and Applications, Research Master in Computer Science, 2^{nd} year, IFSIC, University of Rennes 1.
- L. Amsaleg, P. Gros and P. Sébillot: Digital Documents Indexing and Retrieval, Professional Master in Computer Science, 2^{nd} year,IFSIC, University of Rennes 1.
- L. Amsaleg: Advanced Databases, Professional Master in Computer Science, 2^{nd} year, IFSIC, University of Rennes 1.
- L. Amsaleg: Management of Multimedia Documents, Professional Master in Computer Science, 2^{nd} year, IFSIC, University of Rennes 1.
- L. Amsaleg: Advanced Databases & Modern Information Systems, Diic3, IFSIC, University of Rennes 1.
- P. Sébillot: Advanced Databases and Modern Information Systems, 5^{th} year, Computer Science, INSA Rennes.

9.4. Participation to Seminars, Workshops, Invited Conferences

- L. Amsaleg gave an invited talk at the School of Information Science and Technology, University of Tokyo.
- L. Amsaleg was a member of the board for Herwig Lejsek's Master's thesis in June 2005 and entitled "The PvS-Index an Effcient and Effective Indexing Method for Content Based Image Retrieval using Local Descriptors", University of Reykjavík, Iceland.
- L. Berti-Équille gave an invited talk at CNAM to the "CRM and Data Quality Conference" in June 2005.
- P. Gros gave an invited talk at CVDB'05, Baltimore, Maryland, USA, in June 2005.
- P. Gros gave an invited talk at CMU at the "Vision and Autonomous System Center Seminar" in June 2005.
- S. Huet and P. Sébillot gave an invited talk at OLST-RALI seminar, Montreal, about the introduction of linguistic knowledge in speech recognition systems, in July 2005.
- A. Morin coorganized with J.-H. Chauchat a workshop on data mining and text mining in Ottawa which took place during the Symposium 2005 of Statistics Canada, in November 2005.
- P. Sébillot gave an invited talk at LIR-LIMSI seminar, Orsay, about numerical and symbolic approaches of semantic knowledge acquisition from corpora, in April 2005.

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, March 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 International Conference on Computational Linguistics, Speech and Document Processing, ICCLSDP'98, Calcutta, India", February 1998.

[3] S.-A. BERRANI, L. AMSALEG, P. GROS. *Approximate Searches: k-Neighbors + Precision*, in "Proceedings of the 12th ACM International Conference on Information and Knowledge Management", 2003, p. 24-31.

[4] L. BERTI-ÉQUILLE. *Quality-Extended Query Processing for Mediation Systems*, in "Information Quality Management: Theory and Practice", L. AL-HAKIM (editor). , to be published, IDEA Group Inc., 2006.

[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 "IEEE International Conference 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'02)", vol. 1, n° 4, 2002, p. 407-412.

[8] M. ROSSIGNOL, P. SÉBILLOT. *Combining Statistical Data Analysis Techniques to Extract Topical Keyword Classes from Corpora*, in "IDA (Intelligent Data Analysis)", vol. 9, n° 1, 2005, p. 105-127.

Books and Monographs

[9] L. AMSALEG, B. ÞÓR-JÓNSSON, V. ORIA (editors). *Proceedings of the 2nd International Workshop on Computer Vision meets Databases, CVDB'05*, ACM, Baltimore, MD, USA, June 2005.

[10] L. BERTI-ÉQUILLE, F. GUILLET (editors). *Actes du 1er Atelier Qualité des Données et des Connaissances, DKQ'05*, 2005.

Doctoral dissertations and Habilitation theses

[11] M. ROSSIGNOL. *Acquisition automatique de lexiques sémantiques pour la recherche d'information*, Thèse de doctorat, Université de Rennes 1, France, October 2005.

Articles in refereed journals and book chapters

[12] T. AHMED, A. H. ASGARI, A. MEHAOUA, E. BORCOCI, L. BERTI-ÉQUILLE, G. KORMENTZAS. *End-to-End Quality of Service Provisioning Through an Integrated Management System for Multimedia Content Delivery*, in "Special Issue of Computer Communications on Emerging Middleware for Next Generation Networks", to appear, 2005.

[13] L. AMSALEG, B. ÞÓR-JÓNSSON, V. ORIA. *Report from the First International Workshop on Computer Vision meets Databases, CVDB'04*, in "Sigmod Record", n° 1, March 2005, http://www.sigmod.org/sigmod/record/issues/0503/index.html.

[14] N. BONNEL, A. COTARMANAC'H, A. MORIN. *Classification non supervisée et visualisation 3D de documents*, in "Revue des Nouvelles Technologies de l'Information (RNTI-E-3), numéro spécial EGC'05", vol. 2, January 2005, p. 557-562.

[15] 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, 2005.

[16] A. KOKARAM, N. REA, R. DAHYOT, M. TEKALP, P. BOUTHEMY, P. GROS. *Browsing sports video - the challenge of choice*, in "IEEE Signal Processing Magazine", to appear, 2005.

[17] M. ROSSIGNOL, P. SÉBILLOT. *Combining Statistical Data Analysis Techniques to Extract Topical Keyword Classes from Corpora*, in "IDA (Intelligent Data Analysis)", vol. 9, n° 1, 2005, p. 105-127, http://www.irisa.fr/texmex/publications/versionElect/2005/rossignolsebillotida05.pdf.

Publications in Conferences and Workshops

[18] J.-A. BENVENUTI, L. BERTI-ÉQUILLE, E. JACOPIN. *When Protégé and Rules becomes Parsing for Learning*, in "Proceedings of the 1st International Workshop Protégé with Rules in conjunction with the 8th Intl. Protégé Conference, Madrid, Spain", July 2005.

[19] L. BERTI-ÉQUILLE. *Cost of Low-Quality Data over Association Rules Discovery*, in "Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis, AMSDA'05, Brest, France", May 2005.

[20] L. BERTI-ÉQUILLE. *Nettoyage des données XML : combien ça coûte ?*, in "Actes du 1er Atelier Qualités des Données et des Connaissances, DKQ'05, en conjunction avec la conférence Extraction et Gestion des Connaissances, EGC'05, Paris, France", January 2005.

[21] L. BERTI-ÉQUILLE. *Recommendation of XML Documents exploiting Quality Metadata and Views*, in "Proceedings of the 2nd International Workshop on Data and Information Quality, DIQ'05, in conjunction with the 17th Conference on Advanced Information Systems Engineering, CAiSE'05, Porto, Portugal", June 2005.

[22] L. BERTI-ÉQUILLE, F. MOUSSOUNI. *Quality-Aware Intergation and Warehousing of Genomic Data*, in "Proceedings of the 10th Intl.Conference on Information Quality, IQ'05, Massachusetts Institute of Technology, Cambridge, MA, USA", November 2005.

[23] N. BONNEL, A. COTARMANAC'H, A. MORIN. *Gestion et visualisation des résultats d'une requête*, in "Actes de l'atelier Visualisation et Extraction de Connaissances - EGC'05", January 2005, p. 37-47.

[24] N. BONNEL, A. COTARMANAC'H, A. MORIN. *Meaning Metaphor for Visualizing Search Results*, in "Proceedings of the 9th International Conference on Information Visualisation, IV'05, London, United Kingdom", IEEE Computer Society, July 2005, p. 467-472.

[25] N. BONNEL, A. COTARMANAC'H, A. MORIN. *Visualisation 3D des résultats de recherche : quel avenir ?*, in "Proceedings of the Hypermedias hypertexts, products, tools and methods, H2PTM'05, Paris, France", November 2005.

[26] N. BONNEL, F. MOREAU. *Quel avenir pour les moteurs de recherche ?*, in "Actes de la Manifestation des Jeunes Chercheurs francophones dans les domaines des STIC, MAJECSTIC'05, Rennes, France", October 2005.

[27] M. DELAKIS, G. GRAVIER, P. GROS. *Audiovisual Fusion with Segment Models for Video Structure Analysis*, in "Proceedings of the 2nd European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies, EWIMT'05, London, United Kingdom", to appeear, November 2005.

[28] M. DELAKIS, G. GRAVIER, P. GROS. *Multimodal Segmental-Based Modeling of Tennis Video Broadcasts*, in "Proceedings of IEEE International Conference on Multimedia and Expo, Amsterdam, Netherlands", July 2005, http://www.irisa.fr/texmex/publications/versionElect/2005/icme2005_manolis.pdf.

[29] R. Ý. GRÉTARSDÓTTIR, S. H. EINARSSON, B. ÞÓR-JÓNSSON, L. AMSALEG. *The Eff^2 Image Retrieval System Prototype*, in "Proceedings of the IASTED International Conference on Databases and Applications,

DBA, Innsbruck, Austria", February 2005.

[30] E. GUÉRIN, G. MARQUET, A. BURGUN, O. LORÉAL, L. BERTI-ÉQUILLE, U. LESER, F. MOUSSOUNI. *Integrating and Warehousing Liver Gene Expression Data and Related Biomedical Resources in GEDAW*, in "Proceedings of the 2nd International Workshop on Data Integration in the Life Sciences, DILS, San Diego, CA, USA", July 2005.

[31] W. JOUVE, C. DUFOUIL. *Stratégies de gestion et d'utilisation des métadonnées dans les réseaux de diffusion de contenus multimédia*, in "Actes de la Manifestation des Jeunes Chercheurs francophones dans les domaines des STIC, MAJECSTIC'05, Rennes, France", October 2005.

[32] A. KAROUI, N. BENNACER, M.-A. AUFAURE, A. MORIN. *Découverte de connaissances à partir du Web*, in "Actes du 2ième atelier sur la "Fouille de données complexes dans un processus d'extraction des connaissances, Conférence EGC'05, Paris, France", January 2005, p. 37-42.

[33] A. KOUOMOU-CHOUPO, L. BERTI-ÉQUILLE, A. MORIN. *Optimizing Progressive Query-By-Example over Pre-Clustered Large Image Databases*, in "Proceedings of the 2nd ACM SIGMOD International Workshop on Computer Vision meets DataBases, CVDB'05, Baltimore, MD, USA", June 2005.

[34] A. KOUOMOU-CHOUPO, L. BERTI-ÉQUILLE, A. MORIN. *Optimizing Progressive Query-By-Example over Pre-Clustered Large Image Databases*, in "Actes des Journées Bases de Données Avancées, BDA'05, Saint-Malo, France", October 2005.

[35] H. LEJSEK, F.-H. ÁSMUNDSSON, B. ÞÓR-JÓNSSON, L. AMSALEG. *Efficient and Effective Image Copyright Enforcement*, in "Actes des Journées Bases de Données Avancées, BDA'05, Saint Malo, France", October 2005.

[36] A. MORIN, A. KOUOMOU-CHOUPO, J.-H. CHAUCHAT. *Dimension reduction and clustering for query-by-example in huge image databases*, in "Proceedings of the 3rd International Conference on Computational Statistics and Data Analysis, Limassol, Cyprus", October 2005.

[37] A. MORIN. *Intensive use of correspondence analysis for text mining*, in "Proceedings of the ISI Meeting, Sydney, Australia", April 2005.

[38] X. NATUREL, G. GRAVIER, P. GROS. *Étiquetage Automatique de Programmes de Télévision*, in "Actes de Compression et représentation des signaux audiovisuels, CORESA'05, Rennes, France", November 2005.

[39] X. NATUREL, P. GROS. *A Fast Shot Matching Strategy for detecting duplicate sequences in a television stream*, in "Proceedings of the 2nd ACM SIGMOD International Workshop on Computer Vision meets DataBases, CVDB'05, Baltimore, MD, USA", June 2005, http://www.irisa.fr/texmex/publications/versionElect/2005/CVDB2005_Naturel.pdf.

[40] A. REMAZEILLES, F. CHAUMETTE, P. GROS. *Image based robot navigation in 3D environments*, in "Proceedings of the International Symposium on Optomechatronic Technologies, ISOT'05, Sapporo, Japan", December 2005.

[41] B. ROUSSEAU, W. JOUVE, L. BERTI-ÉQUILLE. *Enriching Multimedia Content Description for Broadcast*

Environments: From A Unified Metadata Model to A New Generation of Authoring Tool, in "Proceedings of the IEEE International Symposium on Multimedia, ISM'05, Irvine, CA, USA", December 2005.

[42] R. SIGURÐARDOTTIR, H. HAUKSSON, B. ÞÓR-JÓNSSON, L. AMSALEG. *A Case Study of the Quality vs. Time Trade-off for Approximate Image Descriptor Search*, in "Proceedings of the 1st IEEE International Workshop on Managing Data for Emerging Multimedia Applications, EMMA, in cooperation with 21th IEEE Conference on Data Engineering, ICDE, Tokyo, Japan", April 2005.

[43] P. SÉBILLOT. *Symbolic Machine Learning: A Different Answer to the Problem of the Acquisition of Lexical Knowledge from Corpora*, in "Proceedings of the European conference on Computing and Philosophy, E-CAP'05, Västerås, Sweden", June 2005, http://www.irisa.fr/texmex/publications/versionElect/2005/sebillot05.pdf.

[44] F. TONNIN, P. GROS, C. GUILLEMOT. *Analyse de représentations multirésolution pour la compression et la description locale d'images*, in "Actes de Compression et représentation des signaux audiovisuels, CORESA'05, Rennes, France", November 2005.

Internal Reports

[45] H. LEJSEK, F.-H. ÁSMUNDSSON, B. ÞÓR-JÓNSSON, L. AMSALEG. *Efficient and Effective Image Copyright Enforcement*, Technical report, n° 1699, IRISA, March 2005, http://www.irisa.fr/bibli/publi/pi/2005/1699/1699.html.

[46] F. MOREAU, P. SÉBILLOT. *Contributions des techniques du traitement automatique des langues à la recherche d'information*, Technical report, n° 1690, IRISA, February 2005, http://www.irisa.fr/bibli/publi/pi/2005/1690/1690.html.

Bibliography in notes

[47] L. AMSALEG, B. ÞÓR-JÓNSSON, V. ORIA (editors). *Proceedings of the 1st International Workshop on Computer Vision meets Databases (CVDB'04)*, June 2004.

[48] 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.

[49] T. AHMED, A. H. ASGARI, A. MEHAOUA, E. BORCOCI, L. BERTI-ÉQUILLE, G. KORMENTZAS. *End-to-End Quality of Service Provisioning Through an Integrated Management System for Multimedia Content Delivery*, in "Special Issue of Computer Communications on Emerging Middleware for Next Generation Networks", 2005.

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

[51] J. ANDERSON, M. STONEBRAKER. *Sequoia 2000 Metadata Schema for Satellite Images*, in "acm sigmod Record, Special issue on Metadata for Digital Media", vol. 23, n° 4, 1994.

[52] K. BENNETT, U. FAYYAD, D. GEIGER. *Density-Based Indexing for Approximate Nearest-Neighbor Queries*,

in "Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA", August 1999.

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

[54] 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.

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

[56] K. BEYER, J. GOLDSTEIN, R. RAMAKRISHNAN, U. SHAFT. *When Is "Nearest Neighbor" Meaningful?*, in "Proceedings of the 8th International Conference on Database Theory, London, United Kingdom", January 1999.

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

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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] P. CIACCIA, M. PATELLA. *PAC Nearest Neighbor Queries: Approximate and Controlled Search in High-Dimensional and Metric Spaces*, in "Proceedings of the 16th International Conference on Data Engineering, San Diego, CA, USA", February 2000.

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

[66] 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.

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

[68] 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.

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

[70] R. FAGIN, R. KUMAR, D. SIVAKUMAR. *Efficient similarity search and classification via rank aggregation*, in "Proceedings of the ACM SIGMOD Conference, San Diego, CA", 2003.

[71] H. FERHATOSMANOGLU, E. TUNCEL, D. AGRAWAL, A. E. ABBADI. *Approximate Nearest Neighbor Searching in Multimedia Databases*, in "Proceedings of the 17th International Conference on Data Engineering, Heidelberg, Germany", April 2001, p. 503–511.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

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

[77] A. GIONIS, P. INDYK, R. MOTWANI. *Similarity Search in High Dimensions via Hashing*, in "Proceedings of the 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, United Kingdom", September 1999, p. 518–529.

[78] 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.

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

[80] 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.

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

[82] 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.

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

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

[85] 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^o 1, 1998, p. 9-37.

[86] 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.

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

[88] 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.

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

[90] E. KIJAK. *Structuration multimodale des vidéos de sports par modèles stochastiques*, Thèse de doctorat, Université de Rennes 1, France, December 2003.

[91] 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^o 1, January 2001, p. 96–111.

[92] 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 2004.

[93] 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.

[94] 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^o 4, July 2002, p. 792–808.

[95] 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.

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

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

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

[99] I. MEL'ČUK. *Vers une linguistique Sens-Texte*, January 1997, Leçon inaugurale, Collège de France, Chaire internationale.

[100] I. MEL'ČUK. *Collocations and Lexical Functions*, in "Phraseology. Theory, Analysis, and Applications, Oxford", A. P. COWIE (editor). , chap. 2, Clarendon Press, 1998.

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

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

[103] B. PAGEL, F. KORN, C. FALOUTSOS. *Deflating the Dimensionality Curse Using Multiple Fractal Dimensions*, in "Proceedings of the 16th International Conference on Data Engineering, San Diego, California, USA", March 2000.

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

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

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

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

[108] 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, Toulouse, France", October 1999.

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

[110] 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.

[111] 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.

[112] R. WEBER, K. BÖHM. *Trading Quality for Time with Nearest Neighbor Search*, in "Proceedings of the 7th Conference on Extending Database Technology, Konstanz, Germany", March 2000.

[113] 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.

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