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 - Bretagne Atlantique

THEME SYM

Table of contents

1. Team

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

Head of project-team

Patrick Gros [Senior Research Scientist INRIA, HdR]

Administrative Assistant

Loïc Lesage [Secretary INRIA, partial position in the project-team]

CNRS Staff Members

Laurent Amsaleg [Research Scientist]
Vincent Claveau [Research Scientist]

Faculty Members (National Institute of Applied Sciences of Rennes)

Nicolas Bonnel [Assistant Professor, until August 31st]
Pascale Sébillot [Professor, HdR]

Faculty Members (University of Rennes 1)

Stéphane Huet [Assistant Professor, since October 1st]
Ewa Kijak [Associate Professor]
Fabienne Moreau [Assistant Professor, until August 31st]
Annie Morin [Associate Professor, HdR]
François Poulet [Associate Professor, since September 1st, HdR]

Faculty Members (University of Rennes 2)

Emmanuelle Martienne [Associate Professor]
Fabienne Moreau [Associate Professor, since September 1st]
Laurent Ughetto [Associate Professor]

Ph.D. Students

Siwar Baghdadi [CIFRE grant with Thomson]
Stéphane Huet [MESR grant, also with METISS, until September 30st]
Zied Jemai [INRIA grant, until January 31st]
Gwénolé Lecorvé [MESR grant, since September 1st]
Gaël Manson [CIFRE grant with Orange]
Xavier Naturel [INRIA grant, until June 30th]
Nguyen Khang Pham [also with Institut francophone pour l'informatique, Hanoï, Vietnam]
Romain Tavenard [ENS Cachan grant, since September 1st]
Pierre Tirilly [CNRS - Brittany Council grant]
Joaquin Zepeda [ICOS-HD project, also with TEMICS, since September 1st]

Post-Doctoral Fellows

Antoine Doucet [INRIA grant, until July 31st]
Ariane Herbulot [INRIA grant, also with TEMICS, since November 1st]

Technical Staff

Sebastien Campion [INRIA Research Engineer, since September 17th]
Cédric Dufouil [INRIA Junior Technical Staff, until August 31st]
Arnaud Dupuis [INRIA Technical Staff]
Pierre-Hugues Joalland [AceMedia project, until June 30th]

Visiting Scientists

Kári Harðarson [Lecturer at Reykjavík University, from September 10th to December 15th]
Ichiro Ide [Associate Professor at Nagoya University, from August 20th to October 29th]
Selja Seppälä [PhD student at University of Geneva, from September 19th to December 21st]

2. Overall Objectives

2.1. Overall Objectives

Keywords: *data analysis*, *databases*, *documents*, *image recognition*, *indexing*, *machine learning*, *multimedia*, *natural language processing*, *search*, *television*, *video*.

With the success of sites like Youtube or DailyMotion, with the development of the Digital Terrestrial TV, it is now obvious that the digital videos have invaded our usual information channels like the web. If such new documents are now available in huge quantities, using them remains difficult. Beyond the storage problem, they are not easy to manipulate, browse, describe, search, summarize, visualize as soon as the simple scenario "1. search the title by keywords 2. watch the complete document" does not fulfil the user's needs. That is, in most cases.

Most usages are linked with the key concept of repurposing. Videos are a raw material that each user recombines in a new way, to offer new views of the content, to adapt it to new devices (ranging from HD TV sets to mobile phones), to mix it with other videos, to answer information queries... Somehow, each use of a video gives raise to a new ephemere document that exists only while it is viewed. Achieving such a repurposing process implies to be able to manipulate videos extracts as easily as words in a text.

Many applications exist in both professional and domestic areas. On the professional side, such applications include transforming a TV broacasted program into a web site, a DVD or a mobile phone service, switching from a traditional TV program to an interactive one, better exploiting TV and video archives, constructing new video services (video on demand, video edition...). On the domestic side, video summarization can be of great help, as can a better management of the videos locally recorded, or simple tools to face the exponential number of TV channels available that increase the quantity of interesting documents available but make them really hard to find.

In order to face such new application needs, we propose a multi-field work, gathering in a same team specialists able to deal with the various media and aspects of large video collections: image, video, text, sound and speech, but also data analysis, indexing, machine learning... The main goal of this work is to segment, structure, describe, delinearize the multimedia content in order to be able to recombine or re-use it in new conditions. The focus on the document analysis aspect of the problem is an explicit choice since it is the first mandatory step of any further application, but using the descriptions obtained by the processing tools we develop is also an important goal of our activity.

To summarize our research project in one short sentence, let us say that we would like our PCs to be able to watch TV and use what they have seen and understood in new innovative services. The main challenges to be faced to reach such a goal are: the size of the documents and of the document collections to be processsed, the necessity to process several media in a joint manner and to obtain a high level of semantics, the variety of contents, of contexts, of needs and usages, linked to the difficulty to manage such documents on a traditional interface. Our own research is organized in three directions: 1- developing advanced algorithms of data analysis, description and indexing, 2- searching new techniques for linguistic information acquisition and use, 3- building new processing tools for audiovisual documents.

2.1.1. Advanced Algorithms of Data Analysis, Description and Indexing

Processing multimedia documents produces most of the time lots of descriptive metadata. These metadata can take many different aspects ranging from a simple label issued from a limited list, to high dimensional vectors or matrices of any kind; they can be numeric or symbolic, exact, approximate or noisy... As examples, image descriptors are usually vectors whose dimension can vary between 2 and 900, while text descriptors are vectors of much higher dimension, up to 100,000 but that are very sparse. Real size collections of documents can produce sets of billions of such vectors.

Most of the operations to be achieved on the documents are in fact translated in terms of operations of their metadata and the former appears as key objects to be manipulated. Although they are of much simpler nature that the data used to compute them, these metadata require specific tools and algorithms to face their particular structure and volume. Our work concerns mainly three domains:

- data analysis techniques, eventually coupled to data visualization techniques, to study the structure of large sets of metadata, with applications to classical problems like data classification, clustering, sampling, or modeling,
- advanced data indexing techniques in order to speed-up the retrieval of these metadata in retrieval or query answering problems,
- description of compressed, watermarked or attacked data.

2.1.2. New Techniques for Linguistic Information Acquisition and Use

Natural languages are a privileged way to carry high level semantic information. Used in speech on an audio track, in textual format or overlaid in images or videos, alone or associated with images, graphics or tables, organized linearly or with hyperlink, expressed in English, French, or Chinese, this linguistic information may take many different forms, but always exhibits a common basic structure: it is composed of sequences of words. Building techniques that respect the subtle link existing between these words, their representations with letters or other symbols and the semantics they carry is a difficult challenge.

As an example, actual search engines work at the representation level (they search sequences of letters), and do not consider the meaning of the searched words. As a consequence, they do not use the fact that "bike" and "bicycle" represent a single concept while "bank" has at least two different meanings (a river bank and a financial institution).

Extracting a high level of information is the goal of our work. First, acquisition techniques that allow us to associate pieces of semantics with words, to create links between words are still an active field of research. Once this linguistic information is available, its use opens new problems. For example, in search engines, new pieces of information can be stored and the representation of the data can be improved in order to increase the quality of the results.

2.1.3. New Processing Tools for Audiovisual Documents

Audiovisual documents have the main characteristic to be temporal documents. One of the consequences is that cannot be watched or listened to globally, but only by a linear process that takes some time. On the processing side, these documents often mix several media (image track, sound track, some text) that should be all taken into account to understand the meaning and the structure of the document. They can also have an endless stream structure with no clear temporal boundaries, like on most TV or radio channels. As a consequence there is an important need to segment and structure them, at various scales, before describing the pieces that were obtained.

Our work is organized in three directions. Segmenting and structuring long TV streams (up to several weeks, 24 hours a day) is a first goal that allows to extract program and non program segments in these streams. These programs can be structured themselves at a finer level; we study the use of stochastic models to achieve such structuring. Finally, once the structure is extracted, we use the linguistic information to describe and characterize the various segments. In all this work, the interaction between the various media is a constant source of difficulty, but also of inspiration.

2.2. Highlights of the Year

In 2007, three events in TEXMEX are worth noting:

1. The team has edited and co-authored a book describing our approach on multimodal analysis of multimedia documents. This book, entitled "L'indexation multimédia – Description et recherche automatiques" and published by Hermes - Lavoisier, provides the reader with a wide introduction to

all the fields studied in the team: image, video, sound and text description, multimodal description, indexing techniques, but also evaluation campaigns, use studies...

2. Xavier Naturel has received a special award at the *Prix de la recherche de l'Inathèque de France 2007* for his Ph.D. thesis entitled "Structuration automatique de flux vidéo de télévision".
3. Three new permanent members have joined the team. Fabienne Moreau and François Poulet as associate professors, Sébastien Campion as research engineer. This will allow us to develop new approaches using scientific visualization (F. Poulet's specialty) and to strengthen the development of our experimental platform and of new software.

3. Scientific Foundations

3.1. Background

This section presents some of the basic techniques used in the team. They are basic blocks used in many research projects as well as research topics by themselves, since we usually use them in new contexts.

3.2. Local Image Description

Keywords: *Image Description*, *Local Descriptors*, *Metadata*.

In most contexts where images are to be compared, a direct comparison is impossible. Images are compressed in different formats, most formats are error-prone, images can be re-sized, cropped... The solution consists in computing descriptors from the images and to translate the image comparison problem into a descriptor comparison one. This can be done if, on the one hand, the descriptors contain some information on the image content while, on the other hand, they do not depend on the image format, size or on transformations the image can undergo.

The most classical method associates a unique global descriptor to each image, e.g. a color histogram or correlogram, a texture descriptor. Such descriptors are easy to compute and use, but they usually fail to resist to crops and cannot be used for object recognition. A second method consists in extracting regions in the image and to associate a descriptor to each of these regions. Most of the time, this is done by extracting points (called interest points) with a detector like Harris' [66], and by considering a circular or elliptic region around each of these points.

Among the first local descriptors used were the differential invariants. Established by Florack [63], their use for image comparison was proposed by Schmid [74]. Each descriptor is a combination of the first derivatives of the signal at the interest point. These descriptors appeared experimentally very robust to geometric and photometric transforms. An even more powerful descriptor was then proposed by Lowe: the SIFT descriptor [68]. It is composed of 16 local histograms of gradient directions around the interest point.

Such descriptors can be used in many applications: image comparison for object recognition, image copy detection, detection of repeats in television streams... While they are very reliable, such descriptors are not without problems. As many descriptors can be computed for a single image, a collection of one million of images can generate a database of one billion of descriptors. That is why specific indexing techniques are required to handle them. Up to now, most of them are computed from decompressed images, while most formats are based on image compression. Thus it would be interesting to directly compute the descriptors in the compressed domain. Finally, their evaluation for very large image collection (several millions of images) is still an open and interesting problem.

3.3. Corpus-based Text Description and Machine Learning

Keywords: *Machine-learning*, *Text Description*.

Our work on textual material (textual documents, transcriptions of speech documents, captions in images or videos, etc.) is characterized by a chiefly corpus-based approach, as opposed to an introspective one. A corpus is for us a huge collection of textual documents, gathered or used for a precise objective. We thus exploit specialized (abstract of biomedical articles, computer science texts, etc.) or non specialized (newspapers, broadcast news, etc.) collections for our various studies. In TEXMEX, according to our applications, different kinds of knowledge can be extracted from the textual material. For example, we automatically extract terms characteristic of each successive topic in a corpus with no a priori knowledge; we produce representations for documents in an indexing perspective [73]; we acquire lexical resources from the collections (morphological families, semantic relations, translation equivalences, etc.) in order to better grasp relations between segments of texts in which a same idea is expressed with different terms or in different languages...

In the domain of the corpus-based text processing, 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 [58]. For our various problems involving language processing, we use both approaches, making the most of existing machine learning techniques or proposing new ones. Relying on advantages of both methods, we aim at developing machine learning solutions that are automatic and generic enough to make it possible to extract, from a corpus, the kind of elements required by a given task.

3.4. Stochastic Models for Multimodal Analysis

Keywords: *Image stochastic models*, *multimedia*.

Describing multimedia documents, i.e. documents that contain several modalities (e.g. text, images, sound) requires to take all these modalities into account since these modalities can contain complementary pieces of information. The problem is that the various modalities are only weakly synchronized, they do not have the same rate and combining the information that can be extracted from them is not obvious. Of course, we would like to find generic ways to combine these pieces of information. Stochastic models appear as a well dedicated tool for such combinations, especially for image and sound information.

Markov models are composed of a set of states, of transition probabilities between these states and of emission probabilities that provide the probability to emit a given symbol in a given state. Such models allow to generate sequences. Starting from an initial state, they iteratively emit a symbol and then switch in a subsequent state according to the respective probability distributions. These models can be used in an indirect way. Given a sequence of symbols (called observations), hidden Markov models [71] aims at finding the best sequence of states that can explain this sequence. The Viterbi algorithm provides an optimal solution to this problem.

For such HMM, the structure and probability distributions need to be a priori determined. They can be fixed manually (this is the case for the structure : number of states and their topology), or estimated from example data (this is often the case for the probability distributions). Given a document, such an HMM can be used to retrieve its structure from the features that can be extracted. As a matter of fact, these models allow an audiovisual analysis of the videos, the symbols being composed of a video and an audio component.

Two of the main drawbacks of the HMM is that they can only emit a unique symbol per state, and that they imply that the duration in a given state follows an exponential distribution. Such drawbacks can be circumvented by segment models [70]. These models are an extension of HMM were each state can emit several symbols and contains a duration model that governs the number of symbols emitted (or observed) for this state. Such a scheme allows us to process features at different rates.

Bayesian networks are an even more general model family. Static Bayesian networks [61] are composed of a set of random variables linked by edges indicating their conditional dependency. Such models allow us to learn from example data the distributions and links between the variables. A key point is that both the network structure and the distributions of the variables can be learned. As such, these networks are difficult to use in the case of temporal phenomenon.

Dynamic Bayesian [69] networks are a generalization of the previous models. Such networks are composed of an elementary network that is replicated at each time stamp. Duration variable can be added in order to provide some flexibility on the time processing, like it was the case with segment models.

While HMM and segment models are well suited for dense segmentation of video streams, Bayesian networks offer better capabilities for sparse event detection. Defining a trash state that correspond to non event segment is a well known problem is speech recognition: computing the observation probabilities in such a state is very difficult.

3.5. Multidimensional Indexing Techniques

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

Techniques for indexing multimedia data are needed to preserve the efficiency of search processes, as soon as the data to search in becomes large in volume and/or in dimension. These technique aims at reducing the number of I/Os and CPU cycles needed to perform a search. Two classes of multi-dimensional indexing methods can be distinguished: exact nearest neighbor searches and approximate NN-search schemes.

Traditional multidimensional indexing techniques typically divide the data space into cells containing vectors [59]. Cell construction strategies can be classified in two broad categories: *data-partitioning* indexing methods that divide the data space according to the distribution of data, and *space-partitioning* indexing methods that divide the data space along predefined lines and store each descriptor in the appropriate cell. NN-algorithms typically use the geometrical properties of (minimum bounding) cells to eliminate cells that cannot have any impact on the result of the current query [60].

Many data-partitioning index methods derive from the seminal R-Tree [65], and their differences lie in the properties of the shapes used to build cells and/or in the degree of overlapping between cells. Well known space-partitioning techniques are somehow related to the K-D-B-Tree [72], and differ on the way space is split and cells encoded.

Unfortunately, the “curse of dimensionality” phenomenon makes these traditional approaches ineffective in high-dimensional spaces [15]. 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 – their description can be found in [15].

Some approaches simply rely on dimensionality reduction techniques, such as PCA, but their use remains problematic when facing very high-dimensional datasets. Other approaches abort the search process early, after having accessed an arbitrary and predetermined number of cells. While this is highly effective, it does not give any clue on the quality of the result returned to the user. Some other approaches consider an approximation of the sizes of cells instead of considering their exact sizes, making somehow cells “smaller”. Shrunk cells increase efficiency of retrievals as they reduce overlap in space, but interesting vectors might be missed, however.

Recently, several approaches have transformed costly nearest neighbor searches in multidimensional space into efficient uni-dimensional accesses. One approach using locality sensitive hashing (LSH) techniques [64] use several hash functions such that co-located vectors are likely to collide in buckets. Fagin *et al.* [62] proposed a framework based on projecting the descriptors onto a limited set of random lines, each line giving a ranking of the database descriptors with respect to the query descriptor.

3.6. Data Mining methods

Data Mining (DM) is the core of KDD (Knowledge Discovery in Databases) whatever the content of the databases are. We focus here on some aspects of DM we use to describe documents and to retrieve information. There are two major goals to DM : description and prediction. The descriptive part includes unsupervised and visualization aspects while prediction is often referred to as supervised mining.

The description step very often includes feature extraction and dimensional reduction. As we deal mainly with contingency tables crossing "documents and words", we intensively use factorial correspondence analysis. "Documents" in this context can be a text document as well as an image.

Correspondence analysis is a descriptive/exploratory technique designed to analyze simple two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information which is similar in nature to those produced by factor analysis techniques, and they allow one to explore the structure of categorical variables included in the table. The most common kind of table of this type is the two-way frequency cross-tabulation table. There are several parallels in interpretation between correspondence analysis and factor analysis: suppose one could find a lower-dimensional space, in which to position the row points in a manner that retains all, or almost all, of the information about the differences between the rows. One could then present all information about the similarities between the rows in a simple 1, 2, or 3- dimensional graph. The presentation and interpretation of very large tables could greatly benefit from the simplification that can be achieved via correspondence analysis (CA).

One of the most important concept in CA is inertia *i.e.* the dispersion of either row points or column points around their gravity center. The inertia is linked to the total Pearson χ^2 for the two-way table. Some rows and/or some columns will be more important due to their quality in a reduced dimensional space and their relative inertia. The quality of a point represents the proportion of the contribution of that point to the overall inertia that can be accounted for by the chosen number of dimensions. However, it does not indicate whether or not, and to what extent, the respective point does in fact contribute to the overall inertia (χ^2 value). The relative inertia represents the proportion of the total inertia accounted for by the respective point, and it is independent of the number of dimensions chosen by the user. We use the relative inertia and quality of points to characterize clusters of documents. The outputs of CA are generally very large. At this step, we use different visualization methods to focus on the most important results of the analysis.

In the supervised classification task, a lot of algorithms can be used, the most popular ones are the decision trees and more recently the Support Vector Machines (SVM). SVMs provide very good results in supervised classification but they are used as "black boxes" (their results are difficult to explain). We use graphical methods to help the user understanding the SVM results, they are based on the data distribution according to the distance to the separating boundary computed by the SVM and another visualisation method (like scatter matrices or parallel coordinates) to try to explain this boundary.

4. Application Domains

4.1. Copyright Protection of Images

Keywords: *Copyright*, *Digital Pictures*, *IPR*, *Image Databases*, *Movie Makers*, *Photo Agencies*.

With the proliferation of high-speed Internet access, piracy of multimedia data has developed into a major problem and media distributors, such as photo agencies, are making strong efforts to protect their digital property. Today, many photo agencies expose their collections on the web with a view to selling access to the images. They typically create web pages of thumbnails, from which it is possible to purchase high-resolution images that can be used for professional publications. Enforcing intellectual property rights and fighting against copyright violations is particularly important for these agencies as these images are a key source of revenue. The most problematic cases, and the ones that induce the largest losses, occur when “pirates” steal the images that are available on the Web and then make money by illegally reselling those images.

This applies to photo agencies, and also to producers of videos and movies. Despite the poor image quality, thousands of (low-resolution) videos are uploaded every day to video-sharing sites such as YouTube, eDonkey or BitTorrent. In 2005, a study conducted by the Motion Picture Association of America was published, which estimated that their members lost 2,3 billion US$ in sales due to video piracy over the Internet. Due to the high risk of piracy, movie producers have tried many means to restrict illegal distribution of their material, albeit with very limited success.

Photo and video pirates have found many ways to circumvent even the most clever protection mechanisms. In order to cover up their tracks, stolen photos are typically cropped, scaled, their colors are slightly modified; videos, once ripped, are typically compressed, modified and re-encoded, making them more suitable for easy downloading. Another very popular method for stealing videos is camcording, where pirates smuggle digital camcorders into a movie theater and record what is projected on the screen. Once back home, that goes to the web.

Clearly, this environment calls for an automatic content-based copyright enforcement system, for images, videos, and also audio as music gets heavily pirated. Such a system needs to be effective as it must cope with often severe attacks against the contents to protect, and efficient as it must rapidly spot the original contents from a huge reference collection.

4.2. Video Database Management

Keywords: *Video Bases*, *Video Structuring*.

The existing video databases are generally little digitized. The progressive migration to digital television should quickly change this point. As a matter of fact, the French TV channel 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.

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. It is unlikely that users will rigorously manage such collections; thus, there is a great need for tools to help the user: automatic creation of summaries and synopses to allow finding information easily, or to have within few minutes a general idea of a program. Even if the service is rustic, it is initially evaluated according to the added-value brought to a system (video tape recorder, decoder), must remain not very expensive, but will benefit from a large diffusion.

On the other hand, these are professional files: TV channel archives, 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 archivists, 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 multiplying 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 these documents. The index validation is carried out by searching information in a database more general than the one used to build the index and by studying the retrieved 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.

5. Software

5.1. Software

5.1.1. I-Description

Participant: Patrick Gros [correspondant].

I-DESCRIPTION allows a user 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.

5.1.2. Asares

Participant: Vincent Claveau [correspondant].

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. These semi-supervised versions rival the supervised one of ASARES. ASARES was deposited with the "Agence pour la Protection des Programmes" (APP) under the number IDDN.FR.001.0032.000.S.C.2005.000.20900.

5.1.3. AnaMorpho

Participant: Vincent Claveau [correspondant].

ANAMORPHO is a tool that makes it possible to automatically detect morphological relations between words in many languages (like installer-désinstallation in French). It relies on analogy processing that are automatically bootstrapped, resulting in a wholly unsupervised technique. Such morphological relations are proven useful in many natural language processing tasks, particularly in the information retrieval domain. ANAMORPHO is currently in a process of being deposited with the "Agence pour la Protection des Programmes" (APP).

5.1.4. Faestos

Participant: Pascale Sébillot [correspondant].

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. Moreover, 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" under the number IDDN.FR.001.470029.000.S.P.2006.000.40000.

5.1.5. 2PAC

Participant: Pascale Sébillot [correspondant].

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” under the number IDDN.FR.001.470028.000.S.P.2006.000.40000.

5.1.6. DiVATex

Participants: Cédric Dufouil, Arnaud Dupuis [correspondant], Patrick Gros.

The library DIVATEX (*Di*stant *V*ideo *A*ccess *Tex*mex) is the evolution of a previous library called TMX-DIVA-SOLUTION. This software enables an easy access to a collection of audio, video and picture contents stored on a remote machine. It contains a server on the remote machine, in charge of extracting from content files (video, audio or picture), the image or sound data needed by the user, and of sending it to the client. The client part of the software decodes what it receives from the server and provides the user with the exact data needed (the exact image or sound information). The first version of this software (TMX-DIVA-SOLUTION) was deposited with the "Agence pour la Protection des Programmes” (APP) under the number IDDN.FR.001.320006.000.S.P.2006.000.40000. The new version (DIVATEX) is currently in a process of being deposited with the APP.

5.1.7. NaviTex

Participants: Cédric Dufouil, Arnaud Dupuis, Patrick Gros [correspondant].

The NAVITEX software was developed with the help of DIVATEX. It allows a user to index and annotate videos using TV-Anytime (XML) metadata format. Those metadata can be exported to the server TELEMEX, in order to easily access some parts of a video directly from the web page. NAVITEX was deposited with the "Agence pour la Protection des Programmes” (APP) under the number IDDN.FR.001.190034.000.S.P.2007.000.40000.

5.1.8. Telemex

Participants: Cédric Dufouil, Arnaud Dupuis [correspondant], Sébastien Campion, Patrick Gros.

The TELEMEX web interface was made to manage huge audiovisuals corpora whatever the user’s operating system. Using this interface, each authorized user can record TV programs, visualize it and navigate inside video files. Each video can be associated with a TV-anytime file, generated by the software NAVITEX. For each video sequence defined by TV-anytime files, TELEMEX generates a unique identifier. Then, this video can be used by DIVATEX.

5.2. Experimental Platform

Participants: Laurent Amsaleg, Sébastien Campion, Patrick Gros, Arnaud Dupuis, Cédric Dufouil, Pascale Sébillot.

Until 2005, we used various computers to store our data and to make our experiments. In 2005, we began some work to specify and set-up dedicated equipment to experiment on very large collections of data. During 2006 and 2007, we specified, bought and installed our first complete platform. It is organized around a very large storage capacity (24TB in 2006 and 70TB in 2007), and contains 4 acquisition devices (for Digital Terrestrial TV), 2 video servers, and some computing servers partially included in the IRISA cluster architecture.

Three softwares were developed: one to manage the acquisition process and all the data already stored (TELEMEX), the second one to annotate audiovisual corpora (NAVITEX) stored on the server, and the third one to ensure an easy access to the data from processing programs DIVATEX (video, audio and pictures). The platform will be completed with dedicated software to manage all the metadata associated with the data. This platform is funded by a joint effort of INRIA, INSA Rennes and University of Rennes 1.

6. New Results

6.1. Advanced Algorithms of Data Analysis, Description and Indexing

6.1.1. Advanced Description Techniques

6.1.1.1. Image Joint Description and Compression

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

Participants: Ewa Kijak, Joaquin Zepeda, Ariane Herbulot.

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

The objective of the study initiated, in collaboration with Christine Guillemot from TEMICS is to design scalable signal representation and approximation methods amenable to both compression (that is with sparseness properties) and description. During the last two decades, image representations obtained with various transforms, e.g., Laplacian pyramid, separable wavelet transforms, curvelets and bandlets have been considered for compression and de-noising applications. Yet, these critically-sampled transforms do not allow the extraction of low level signal features (points, edges, ridges, blobs) or of local descriptors. Feature extraction requires the image representation to be covariant under a set of admissible transformations, which ideally is the set of perspective transformations. Reducing this set of transformations to the group of isometries, and adding the constraint of causality, the image representation is uniquely characterized by the Gaussian scale space. The Gaussian scale space is however not amenable to compression. Through the Ph.D thesis of J. Zepeda, we have thus started investigating subspace-based approaches and sparse representations for local image texture description. These aspects form the core of the ICOS-HD project started in 2007.

Another aspect of this project is to study descriptions able to deal with the various resolution of images provided by scalable coding. Indeed, scalable coding allows an easy distribution of the content on many platforms, and some functionalities are required by the professionals of the domain in order to develop their services, as copy detection as a complementary tool to DRM. Ariane Herbulot has joined TEXMEX in November 2007 for a 16-months Inria post-doctoral position; the research aims at finding a "scalable" descriptor that is the most suitable for video copy detection in the context of scalable coding.

6.1.1.2. Image and Text Joint Description

Keywords: *Image-Text Interaction*.

Participants: Vincent Claveau, Patrick Gros, Pierre Tirilly.

In current image retrieval engines, similarity between images is computed as a similarity between image descriptors. This implies, for the user, to give an image or an image descriptor as a query, whereas using natural language would be a much more intuitive way to express his information need.

A way to authorize such textual queries is to find relations between keywords extracted from the texts and image descriptors in corpora mixing images and texts. Our first approach was to verify that such relations exist in a real huge corpus. The corpus we used contains generalist news articles including a text, pictures illustrating the text and captions describing precisely each picture content. On one side, we showed that, on such data, there is no link between the caption text and low-level descriptors like color histograms [55] or Gabor filter coefficients (characterizing texture). On the other side, we showed that high-level clues extracted from both text and images, such as named entities and the presence of faces, can be used together to give relevant annotations of the pictures [56].

On this basis, we studied another representation of images based on the so-called *video-google* approach [76]. This representation uses regions of interest (called *visual words*) detected in the images, which seem more expressive semantically than low-level features like color. Moreover, it gives a *bag-of-visual words* representation of the images, comparable to the classic *bag-of-words* representation used in text retrieval. It allows us to use techniques primarily developed in the scope of text retrieval or natural language processing on images. For example, machine translation techniques may thus allow us to "translate" a document from visual words to textual words.

However, one difficulty, before using such techniques, is to extract the most semantically significant visual words, since the visual words building process is very coarse. Many visual words are noisy or, by analogy with language processing, polysemous or synonyms. We studied a few ways to isolate significant words from noisy words : stop-lists, which remove words according to their frequency within the corpus, advanced weighting schemes such as Okapi BM25, which are supposed to differentiate significant words from non-significant words, and latent semantic analysis (LSA), which deals with synonymy and polysemy by grouping words in topics. At this time, each of these approaches only gave a very small improvement in image classification results.

6.1.2. Advanced Data Analysis Techniques

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

Keywords: *Correspondence Analysis*, *Visualization*.

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 us to get most of the information included in the data [48]. 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 visualization 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.

6.1.2.2. Intensive Use of SVM for Text Mining

Keywords: *Support Vector Machine*, *Visualization*.

Participants: Nguyen Khang Pham, François Poulet.

SVM and kernel methods are known to provide accurate models but the learning task usually needs a quadratic program, so this task for very large datasets requires a large memory capacity and a long time. We have developed a new algorithm: a boosting of least squares Support Vector Machine to classify very large datasets on standard personal computers. We have extended the least squares SVM algorithm (LS-SVM). The first step was to adapt the algorithm to deal with datasets having a very large number of dimensions (like in text or image mining). Then we have applied boosting to LS-SVM for mining huge datasets having simultaneously a very large number of datapoints and of dimensions on standard computers. The performance of the new algorithm has been evaluated on large datasets from Machine Learning repository [19] like Reuters-21578 or Forest Cover Type. The accuracy is increased in almost all datasets compared to LibSVM.

6.1.2.3. Intensive Use of Factorial Analysis for Image Mining

Keywords: *Correspondence Analysis*, *Visualization*.

Participants: Patrick Gros, Annie Morin, Nguyen Khang Pham, François Poulet.

This work is done with Institut francophone pour l'informatique, Hanoï, Vietnam.

To analyze and retrieve information in image databases, we use the same method as in textual data analysis. This work is apart of the Ph.D. thesis of Pham Nguyen Khang.That means that in order to apply CA to images, we must define "words" in images. This is usually achieved by 2 stages: (1) vector quantizing automatically extracted local image descriptors (*i.e.* SIFT) and (2) applying a clustering algorithm (*i.e.* k-mean clustering) on the set of descriptors to form "visual words". Once the visual words are defined, we then construct a contingency table by crossing visual words and images.

We began experiments by applying CA on the "Alogic" database (contains 961 images). First, a vocabulary of 1000 visual words was computed from 60 first images using SIFT descriptors (code of D. Lowe) and a k-mean algorithm. CA was then applied on a contingency table of 961 x 1000. We have kept only 30 first axes and used those axes for computing image similarity (Euclidean distance was used). Surprisingly, we found some groups of images which described same category (toys, houses, Eiffel tower).

Motivated by this promising result we continued our approach on "caltech4" database (contains 4090 images of 5 categories: faces, airplane, motorbikes, cars (rear) and backgrounds) [75]. About 3000 descriptors sampled from all of descriptors (a third for every category) was clustered to form 2224 visual words. We explored this database on 2 tasks: image categorization and image retrieval. For the first task, we applied CA on the contingency table representing the database and kept only 7 axes (for comparison to PLSA trained with 7 hidden topics). A k-mean algorithm was then invoked to form clusters (categories). The result showed that CA performed slightly better than PLSA.

For image retrieval task, we compared our approach to PLSA and TF*IDF using L1 distance, L2 distance and cosine similarity. In the case of PLSA and CA, the retrieval was performed very fast because the problem dimension was reduced from 2224 to 7. For PLSA and CA, cosine similarity gave better result than L1 and L2 distance. The performances of CA and PLSA was equivalent and much better than that of TF*IDF.

We have also proposed a method for scaling up the problem using inverted files based on image representation quality on axes. Every inverted file was associated with the well represented images (on the axis to which the file belong). Given an image query, the search began by choosing the appropriate inverted files and intersecting those files. The similarity computation is done only on a subset of images resulting from the intersection of inverted files (about 1/5 to 1/8 of entire database). The performance was degraded only about 0.3% with respect to the exhaustive method.

6.1.2.4. Visualization and Web Mining

Keywords: *3D Metaphors*, *Human-Computer Interfaces*, *Information Retrieval*, *Interface Evaluation*, *Search Result Visualization*, *Self-Organizing Maps*, *Visual Categorization of Web Pages*.

Participants: Nicolas Bonnel, Annie Morin.

This work deals with the dynamic generation of interactive 3D presentations of web search results. Here the issue is how to effectively represent the results matching a query on textual search engine. Our approach focuses on the necessity to dynamically create interactive 3D presentations based on visualization metaphors adapted not only to the end-user, but also to the task to complete as well as to the data. Two main steps can be distinguished in our work [30].

The first one consists of organizing effectively the results of a web query. For this purpose, on-the-fly clustering methods are investigated and only statistical and deterministic approaches are considered. More precisely we focus on a particular unsupervised clustering method: the self-organizing maps.

The second step concerns the visualization of the organized search results. The goal is to define cognitive 3D metaphors of visualization allowing for a richer space representation which efficiently and effectively helps users in their tasks. Various interactive and adaptive metaphors are then proposed but the main one is based on the city concept.

6.1.3. Advanced Indexing Algorithms

6.1.3.1. Indexing for Very Large High Dimensional Spaces

Participants: Laurent Amsaleg, Zied Jemai.

This is a joint work with researchers from Reykjavík University.

We have highlighted in Section 4.1 the need for having automatic content-based video retrieval system for enforcing the protection of copyright. We have developed such a system that was presented at the latest edition of ACM Multimedia [35].

Our system, called Eff^2 Videntifier, builds a reference database of low-level features extracted from the videos to protect. Then, videos streams to be checked are fed into the system to detect near-duplicate content. As pointed out earlier, effectiveness and efficiency are crucial in order for a system to be usable in practice at a realistic scale.

The Eff^2 Videntifier is built on top of a novel high dimensional database technology, the NV-Tree [6], [67], which is able to perform accurate approximate high-dimensional nearest neighbor queries in constant time, independent of the size of the video descriptor collection. The NV-Tree also copes with updates and its design allows for easy distribution of queries across several computers, which increases system throughput linearly.

Video frames are described using the Eff^2 image descriptors [6]. The Eff^2 descriptors are of the SIFT family, but have shown much better recognition power for very large collections, making the Eff^2 Videntifier very effective at detecting pirated movies.

We made initial tests with a database containing about 100 millions of video descriptors from more than 100 full-time videos (about 150 hours of video). We have created "pirated" videos, stored them on the network, and the system continuously downloads videos from the net and check them against our collection. We have shown that the system is able to detect copies of video in real time with many kinds of modifications, classified from medium-hard to hard. Among them are cropping, blurring, brightness and contrast changes, reencoding, camcorder attacks, compression to 56Kbps quality, lens distortion, noise and saturation changes to the recording, affine distortions, small rotations and bordering effects.

6.1.3.2. Describing Sequences for Audio/Video Retrieval

Participants: Laurent Amsaleg, Romain Tavenard.

Our work on this topic is done in close collaboration with Guillaume Gravier from the METISS *project-team of IRISA.*

Today, we can quite well exploit rather large databases of still images and we know how to efficiently query them by content. The next step asks to turn our focus on more complex documents, typically video and audio. There are today several description techniques for audio and video but only very few techniques efficiently perform query-by-content on video or audio databases at large scale. Being able to use such techniques is particularly crucial for professional multimedia archivists.

The state of the art makes such searches possible, but only at a very small scale, i.e., on a very small amount of data. Today, no search technique is efficient enough to allow any practical usage of real-scale audio or video archives. In addition, simply extending existing multidimensional indexing techniques is not possible since they were designed for description schemes in which the concept of sequence is lacking.

We have started investigating this issue [53], [54]. Overall, deciding whether two sequences of descriptors are similar requires to clarify what elements should be compared, and how the comparison should be enforced. We have tried two very different approaches where elements to compare were either the descriptors themselves, or sequence models learned from the values of the descriptors. Directly comparing sequences of descriptors is done using the traditional Dynamic Time Warping approach. It is in fact an *a posteriori* alignment of the sequences to compare. Here, the similarity of sequences is directly related to the similarity of the descriptions. We also compared sequence models, where each sequence is modeled using a Support Vector Machine approach used in regression (and not in classification, as usually done). Each model is somehow a translation of the temporal behavior of its corresponding sequence. Here, comparing models instead of low-level descriptors asks to define a new metrics as direct comparison between models is impossible. Overall, we have shown that relying on models (instead of relying on descriptors) provides a better robustness to severe modifications of sequences, like temporal distortions for example. These results were obtained using a sequence collection made of real audio data broadcast on radio.

These initial results suggest to push forward the investigations. We will look on ways to connect the models to multidimensional indexing techniques, as to handle scale.

6.1.3.3. Browsing Personal Image Collections

Participants: Laurent Amsaleg, Kári Harðarson.

In recent years, the world has seen a tremendous increase in the capability to create, share and store digital images. As a result, personal image collections are growing at an astounding rate and it is clear that in the future individuals will need to access tens of thousands, or even hundreds of thousands, of digital images.

Despite numerous features (effective packing on thumbnails on screen, identifying representative images, zoomable user interfaces...), all current photo browsers share limitations such as using a time-line view or a folder view at each time, failing to use the two dimensions of the screen.

In contrast, each image should be described by a number of attributes, based on image contents and image meta-data (such as camera and time information, stored in so-called EXIF headers). Some of these attributes may be linear or spatial, such as time and location of taking the image, while others may be textual, hierarchical or categorical. These attributes may be considered dimensions in an image hyper-space, which we must be able to traverse dynamically to fully enjoy our digital images.

In on-line analytical processing (OLAP), multi-dimensional data is dealt with by considering a few dimensions at a time and pivoting between dimensions when necessary. In advanced computer games such as EVE online, large three-dimensional worlds are explored by simulating space-travels. Both approaches have been very successful in keeping their users occupied and focused on their task for a long time. We propose that a browsing interface for images should merge these features into a multi-dimensional interface that allows flexible space-travel like exploration of the image hyperspace.

In order to begin exploring the possibilities of such a browsing interface we have implemented a prototype, based on the PartiView browser, which allows us to browse images in a three-dimensional space. The dimensions may be based on image contents and image meta-data and different dimensions may be combined in an arbitrary manner. What is novel in this work is that we want to integrate to an image browser OLAP browsing concepts, such as pivoting and filtering that have typically been designed to facilitate the browsing of huge financial data collections.

We have described in a technical report [57] an image browsing prototype and our experience from using it with two different real-life data sets. Our conclusion is that while the prototype has shortcomings, this is a very promising research, in particular, it suggests to better study the seamless integration of the searching and the browsing tasks.

6.2. New techniques for language processing and applications

6.2.1. NLP for Document Description

6.2.1.1. Analysing Terminological Data

Participants: Vincent Claveau, Pascale Sébillot, Selja Seppälä.

During the 3-month long visit of S. Seppälä, we carried out some experiments aiming at extracting categorized knowledge from existing terminological documents. These documents contain definitions of terms as used in different specialized domains (juridical, mechanical, ecological...). Using a Language Modeling approach, we try to split these definition in relevant portions and to discover the conceptual link these portions share with the term defined. It may allow us to organize knowledge in a specialized field as a network of concepts, which may latter be used in any relevant applications, like, for instance, information retrieval.

6.2.1.2. Analysis of Bibliographical Databases

Participants: Laurent Amsaleg, Patrick Gros, Ichiro Ide, Annie Morin, Pascale Sébillot.

This year, Ichiro Ide's visit was the opportunity to study the use of the topic threading system he built for TV news analysis in another context, the analysis of bibliographical databases. The idea is to check whether it is possible to track the development of scientific topics. It appears that bibliography has a very different structure than news, most authors publishing on only one topic. We plan to enlarge this research to radio news using speech recognition now.

6.2.2. Textual Information Retrieval

Keywords: *Corpus-Based Acquisition of Lexical Relations*, *Information Retrieval*, *Machine Learning*, *Natural Language Processing*.

Our research in this field focuses on two points. First, we aim at extracting linguistic resources needed by a given application from textual corpora; this step is usually carried out by elaborating fully automatic and generic machine learning solutions using both symbolic and statistical approaches. Secondly, we exploit these linguistic resources in information retrieval systems, trying to gracefully and efficiently integrate them into the systems. During 2007, our work has especially concerned the 3 following aspects.

6.2.2.1. Machine-Learning based Acquisition of Morphological Variants of Words

Participants: Vincent Claveau, Fabienne Moreau, Pascale Sébillot.

Information retrieval systems (IRSs) usually suffer from a low ability to recognize a same idea expressed in different forms. A way of improving them is to take into account morphological variants (such as compile/compilation/recompiling...). Following our work initiated last year, we finalized a prototype, called AnaMorpho, that makes it possible to recognize morphological variants. This tool is fully automatic (relying on an unsupervised machine-learning approach). To consolidate the initial results, we have realized several experiments against several IR collections and several languages; AnaMorpho is proven useful in every case. These results have been presented in ECIR, a major conference on Information Retrieval [45].

6.2.2.2. Cross-Language Information Retrieval in the Biomedical Domain

Participant: Vincent Claveau.

In the biomedical domain, the international research framework makes knowledge resources such as multilingual terminologies and thesauri essential to carry out many researches. Within this context, we developed an original method to translate biomedical terms from one language to another, relying on the proximity of such terms in many languages (for instance, consider the French-English examples: ophtalmorragie/ophthalmorrhagia, ophtalmoplastie/ophthalmoplasty, leucorragie/leukorrhagia). The main idea of our work is to automatically learn these regularities with well suited machine-learning techniques, and then to use them to translate new or unknown biomedical terms. Our approach is based on rewriting rules learning and on language modeling. It allows us to generate translations of terms between various languages (English, French, Italian, Portuguese, Russian...) with surprisingly good results [33]. This technique is also proven useful in a cross-language information retrieval task, in which specialized queries are translated into the language of the document collection [34].

6.2.2.3. Linguistic Knowledge and Information Retrieval (IR)

Participants: Laurent Amsaleg, Vincent Claveau, Antoine Doucet, Fabienne Moreau, Pascale Sébillot, Laurent Ughetto.

Our aim in this domain is to explore methods that enable information retrieval systems 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 [22]. The main research conducted in this domain has focused on two axes: for the first one, our goal was to find another way to couple NLP and IR in order to achieve more categorical results about the contribution of the linguistic knowledge in IR; the second one aims at exploring new IR models more suited to integrate the richness of these kinds of linguistic information.

For the first axis, the work done this year aimed at consolidating the results obtained during F. Moreau's Ph.D. thesis in 2006 that proposed to revisit the coupling between NLP and IR. In particular, we studied the exploitation of various linguistic information in IR:

- Contrary to many studies already performed, which usually exploit a single kind of linguistic information within the IRSs, we have studied the impact of integrating various types of linguistic information (*i.e.* information that belongs to morphological, syntactic and semantic levels of language) on the IRS performances. In particular, an original and thorough study was conducted to determine whether those various kinds of knowledge, when integrated within a single information

retrieval system, are either complementary or redundant for finding relevant documents. To perform these experiments, we have used our prototype of linguistically informed-IRS, based on VSM (Vector Space Model), that enables us to integrate in parallel multiple kinds of linguistic knowledge. The results obtained from several IR collections prove the effective interest of integrating several sorts of knowledge (especially morphological and semantic information) within the IRSs [47].

- In order to automatically detect the best way to combine these kinds of information within an IRS, we proposed a supervised machine-learning approach (based on a neural network) that merges the lists of documents produced by each linguistic index, and automatically adapts its behavior to the characteristics of the queries. Our different experiments prove the interest of our fusion method which offers more balanced precision-recall compromises and consequently obtains more stable results than those got by the better individual index [46].

For the second axis, the goal was to exploit other IR models most likely to integrate the richness of the information gleaned from NLP techniques. Indeed, the traditional systems based for example on Salton's VSM represent document and queries as sets of words, without regard for the relationship or even linear order between them. Consequently, although the linguistic knowledge extracted from the texts are relevant to represent the textual content of documents and queries, systems are ill-suited to receive richer information than simple keywords. Antoine Doucet has joined TEXMEX in October 2006 for a one-year INRIA post-doctoral position; the research aims at studying more efficient models. After considering various alternatives, we decided to focus on one recent type of IR model: language modeling, which offers promising solutions for the exploitation and integration of different kinds of linguistic information. The other avenue of research for this problem relies on the use of fuzzy logic, through the work of L. Ughetto, in order to nicely integrate linguistic information in IRS. This on-going work may also be useful to model the user's preferences and profile.

6.3. New processing tools for audiovisual documents

6.3.1. TV Stream Structuring

6.3.1.1. TV Structuring based on a Reference Video Set

Participants: Xavier Naturel, Patrick Gros.

The first technique we developed was based on the use of a Referenced Video Set (RVS), i.e. a small portion of the stream manually labeled. In our experiments, we used a day of video. In this excerpt, all non-program segments (commercials, self-promotion, trailers...) are segmented and labeled. This RVS is used to detect new non-program sequences in the rest of the stream. A few rules allow to extend these detections (a small unlabeled sequence surrounded by commercials is a commercial...) Programs are *a contrario* defined as all sequences that are not labeled as non program sequences.

An EPG is then used to label the segments. The DTW algorithm is used to align the stream with the EPG. A post processing is used to insure the best coherency as possible. This process yields good results in terms of correctly labeled images, since the longest programs are the easiest to recognize. The results are lower in terms of correctly segmented and recognized programs; all the errors come from very small programs that are sometimes not present in the EPG and that are difficult to distinguish from non-program sequences.

6.3.1.2. Repetition Detection based TV Structuring

Participants: Emmanuelle Martienne, Patrick Gros.

One of the drawback of the previous method is to require the manual annotation of the excerpt of video, a very tedious task. Our goal is to suppress this stage. In order to do that, we propose to start by detecting all image repetitions. This is very computer intensive, but the comparison was not the more complex part in our previous work. Computing appropriate descriptors is the longest part.

This work is done in the frame of the Semimages project in collaboration with the CAIRN team of IRISA. The idea is to use the Remix chip developed by this team to speed up the computation. Once these repetitions have been detected, they must be post-processed in order to switch from image repetitions to sequence repetitions.

6.3.1.3. Clustering-based TV Structuring

Participants: Gael Manson, Patrick Gros.

This work is done in the frame of the CIFRE thesis of Gael Manson at Orange Labs (former France-Telecom R&D).

In a second attempt to avoid any manual annotation, we work on a novel clustering-based method. It clusters key-frame descriptors and then analyzes them in order to select clusters that would generate repeated sequences. A final refinement processing step extends detected repeated sequences and exactly delimits them. This approach does not make any hypothesis on the frequency or the size of repeated sequences.

The following step we are working on is how to make use of the detected repeated sequences. We are in particular investigating their use within a global classification-based process in which metadata (like EPG and EIT) would be also integrated in order to recover the high-level structure of the stream.

On the other hand, in order to have a clear idea on the imprecision of metadata, we have performed a deeper study on a manually and precisely annotated 24 hour TV stream. This study shows that metadata are generally imprecise (more than 40% of the programs start more than 5 minutes earlier or later than that expected in the metadata.), incomplete (short programs are in general not mentioned) and not coherent (EIT and EPG may be very different). This study will be very helpful for further developments.

6.3.2. Program Structuring

6.3.2.1. Stochastic Models for Video Description

Keywords: *Dynamic Bayesian Networks*, *Hidden Markov Models*, *Image-Sound Interaction*, *Segment models*, *Video Structuring*.

Participants: Siwar Baghdadi, Patrick Gros.

Our work on this topic is done in close collaboration with Guillaume Gravier from the METISS *project-team of IRISA and Thomson as external partner.*

Bayesian Networks are an elegant and powerful semantic analysis tool. They combine an intuitive graphical representation with efficient algorithms for inference and learning. They also allow the representation in a comprehensive manner of the interaction between a system variable and the integrating of external knowledge. Unlike HMM and segment models, structures of Bayesian Networks are very flexible and can be learned from data. We explored the idea of using Bayesian Networks and their temporal extension Dynamic Bayesians Networks to do event detection in video streams. As a first application we have chosen commercial detection. We modeled the video stream by a Dynamic Bayesian Network. According to this model, the video stream is a sequence of observations (a set of multimodal features). Each observation is generated according to the state of the system (program or commercial). The model is fed with knowledge about the commercial segment duration. Detecting commercial segment is then a problem of inferring the optimal sequence of hidden nodes with the convenient duration. Structure learning allowed us to learn the optimal interaction between variables. Future work involves the extension of our model to do event detection in soccer games. The challenge of this part is to take into consideration all kinds of feature interactions (spatial, temporal at short or long term).

6.3.2.2. Automatic Character Indexing in Audiovisual Document

Keywords: *Image-Sound Interaction*, *Video Structuring*.

Participants: Ewa Kijak, Arnaud Dupuis.

This work is done in collaboration with J. Pinquier from IRIT, in the context of the European Network of Excellence MUSCLE.

We propose a system which permits to describe and structure audiovisual documents without training, nor corpus knowledge, and to visualize the principal interventions with an interface. It displays the most significant person list of the processed documents (news, TV games, variety programs, film, etc.). A person is considered as significant if she/he speaks or appears on the screen during a minimum time lapse. We have specified this threshold in order to keep only the principal and significant characters. This list is presented with representative labels of the character (face or/and sound extract for example). Thanks to this person list, it is possible to listen and/or to view all interventions of each character by clicking on the representation of the selected one. The system is based on the INRIA/TEXMEX face detection tool and the IRIT speaker and costume segmentation tools. The interface allows to visualize (and/or to listen) the only segments where the character of interest appears, without a priori knowledge. We also have the statistics over the speaking time and appearance time of each character. This leads to a demonstrator shown in the framework of the Network of Excellence (NoE) MUSCLE showcases.

6.3.3. Using Speech to Describe and Structure Video

Keywords: *Text-Speech Interaction.*

Participants: Stéphane Huet, Gwénolé Lecorvé, Pascale Sébillot.

Our work on this topic is done in close collaboration with Guillaume Gravier from the METISS *project-team of IRISA.*

Automatic speech recognition (ASR) can be used to structure and index large collections of spoken documents (videos, audio streams...), based on a linguistic analysis of the ASR transcript. However, most current ASR systems rely on purely statistical methods (study of n-grams in corpora) and make little use of linguistic knowledge therefore limiting the quality of their outputs. Our research aims at combining knowledge coming from natural language processing (NLP) and ASR to increase the quality of the transcriptions but also to better combine the information extraction task with the transcription one.

S. Huet's Ph. D. thesis [13] (under the joint supervision of P. Sébillot and G. Gravier) focuses on those points by incorporating morpho-syntactic information and thematic adaptation into the transcription process. To experiment our methods, we use the ASR system of the METISS project-team dedicated to the transcription of French broadcast news data.

Three aspects have been particularly considered this year:

- studying morpho-syntactic information to compute confidence measures that indicate how reliable is a transcribed word. Confidence measure computation is a key point for the interface between ASR and NLP. We show that the processing of N-best lists generated by the ASR system with morpho-syntactic knowledge improves confidence measures [40];
- segmenting broadcast news into thematically coherent parts. We extended a statistical method based on lexical cohesion [77] for topic shift detection to take into account additional knowledge such as semantic relation between words, discourse markers (like "`and now, thank you`"), and acoustic cues. Our technique enables us to improve segmentation, although a few parts —particularly those corresponding to the headlines of the program— have still to be refined.
- adapting the language model based on corpora retrieved from the Internet. We propose a method to adapt the language model of an ASR system for each thematically coherent segment encountered, using neither *a priori* knowledge about topics nor a static collection of texts. The idea is to gather textual adaptation data for each segment, based on information retrieval (IR) methods to extract keywords which are used to retrieve documents from the Web. IR techniques, used both for keyword extraction and for document selection, have been adapted to take into account the particularity of ASR transcripts (*e.g.* misrecognized words, named entities). Results indicate a large improvement of the language model, which finally yields a small improvement of the word error rate [21].

Gwénolé Lecorvé's Ph. D. thesis (sup.: P. Sébillot, G. Gravier), which started in October 2007, aims at combining thematic segmentation with vocabulary and language model adaptation to improve the transcription of large audio streams.

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 compétitivité*) 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 are part, through our participation to the two projects Semim@ges and ICOS-HD, to the pole called "Images and networks" whose main actors are Thomson and Orange Labs and which is located in Brittany and Pays de la Loire.

7.1.2. Contract with Thomson

Participants: Patrick Gros, Siwar Baghdadi.

Duration: 36 months, starting in December 2005.

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

7.1.3. Contract with Orange Labs

Participants: Patrick Gros, Gaël Manson.

Duration: 36 months, starting in November 2006.

G. Manson's Ph.D. thesis is supported by a CIFRE grant in the framework of a contract between Orange Labs and TEXMEX.

7.2. European Initiatives

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

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

Participants: Patrick Gros, Laurent Amsaleg, Pascale Sébillot, Gwénolé Lecorvé, Xavier Naturel, Ewa Kijak.

Duration: 4 years, starting in April 2004. 42 partners. Prime: ERCIM, scientific coordinator: Nozha Boujemaa, INRIA.

This project aims at developing the collaboration in the domain of automatic multimedia document analysis, in particular to be able to handle and exploit 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 us to combine the information extracted from each media, and by the common techniques needed to handle such data (optimization, classification, intensive computation).

In 2007, we were active in three showcase projects. The first one on character recognition in video structuring through a common use of face and costume recognition (E. Kijak), the second one on audiovisual speech recognition (WP 6, through the work of S. Huet, G. Lecorvé, P. Sébillot and G. Gravier), the third one on film summarizing (X. Naturel).

P. Gros is co-editing a book on multimodal processing and interaction with P. Maragos from the National Technical University of Athens, and A. Potamianos from Technical University of Chania. The book will be published by Springer in 2008.

7.2.2. European Integrated Project aceMedia: Integrating Knowledge, 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 to storing, retrieving and communicating documents in an efficient and autonomous way. It supports self-organization, self-annotation and self-adaptation according to current user's preferences and devices. Additional embedded mechanisms are semantic detection, fast retrieval and relevance feedback.

TEXMEX team provided indexing algorithms and 2007 was devoted to evaluation and dissemination of these results.

8. Other Grants and Activities

8.1. National Initiatives

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

Participants: Stéphane Huet, Patrick Gros, Xavier Naturel, Gwénolé Lecorvé, Pascale Sébillot, Arnaud Dupuis, Cédric Dufouil, Sébastien Campion.

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 records and index more than 50 channels 24 hours a day. As the number of available archivists did not increase as fast as the number of channels to index, they have to rely on more automatic processes. The first need is to verify that the programs in 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.

In 2007, Xavier Naturel ended his thesis on TV streams structuring. Our work was then mainly dedicated to finding ways to use speech in this structuring process. This was the research topic of Stéphane Huet and Gwénolé Lecorvé.

After two years mainly dedicated to software development, our video analysis platform was mainly upgraded on the hardware side this year. A NAS server of 24 terabytes of memory was installed with several side servers to record and distribute data. These machines are connected to the grid of IRISA (more than 20 CPUs) to obtain the necessary computing power.

8.1.2. ANR project Semimages

Participants: Patrick Gros, Laurent Amsaleg, Emmanuelle Martienne, Sébastien Campion.

Duration : 27 months, starting in January 2007. Partners: Orange Labs, TDF, Kersonic, Telisma, CAIRN team.

The project is devoted to TV data exploitation and repurposing. Two main applications will be considered: TV news analysis, and TV streams structuring. TEXMEX project-team will mainly be involved in the second one. The aim of our work is to structure automatically long TV streams in more usable units like programs or non-program sequences, exactly like it was done in Xavier Naturel's thesis. But we would like to achieve this same goal completely automatically this time, by removing most of the manual annotation required and by relying on non-supervised classification techniques.

The first step of this work consists in detecting all replicated images in the stream. This will be done by a massively parallel approach in collaboration with the CAIRN team of IRISA.

8.1.3. Participation to National Working Groups

- P. Sébillot is a member of the thematic network "Information and knowledge: discovering and abstracting" of the ST2I department of CNRS.
- P. Sébillot is a member of the working group PRC I3-AFIA TIA (terminologie et intelligence artificielle)

8.2. International Collaborations

8.2.1. 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. Björn Þór Jónsson (Associate Professor) leads the team of researchers involved in Iceland.

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. This is particularly prevalent at large scale when dealing with image collections of realistic sizes. The goal of this project is to research and develop new database support that integrates efficiency and effectiveness for modern large-scale computer-vision related applications and problems.

Together, we came up with the NV-tree framework that provides an efficient and scalable support for local description based recognition applications. While this work is still very active, we have initiated another thread of research by investigating the browsing of personal image collections. Today, everyone can witness the tremendous increase in the capability to create, share and store digital images. As a result, personal image collections are growing at an astounding rate and it is clear that in the future individuals will need to access tens of thousands, or even hundreds of thousands, of digital images. It is therefore imperative to start studying ways to access these images in a useful and interesting manner. Addressing this topic is a new development in our cooperation.

8.2.2. Collaboration with Croatia and Slovenia

Participant: Annie Morin.

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 Lyon2

A. Morin got two Egide contracts with Slovenia (Proteus) and Croatia (Cogito) for 2007 and 2008 on knowledge discovery and visualization for textual data. In Slovenia, we work with Blaz Zupan and Janez Demsar from faculty of Computer and Information Science, University of Ljubljana and in Croatia with Bojana Dalbelo Basic from faculty of Electrical Engineering and Computing, university of Zagreb. The French laboratory ERIC (university of Lyon 2) is the other French partner.

The concerned research teams have different expertise on the same subject: machine learning for the Slovenian and Croationa teams, statistics for the French teams and common abilities such as development of open source data mining software and visualization tools. They have been in touch since a first meeting in 2004 on intelligent data mining. We plan to implement a new visualization system for textual data. Proposed collaboration includes sharing of a number of Ph.D. students.

8.2.3. Collaboration with Nagoya University, Japan

Participants: Laurent Amsaleg, Patrick Gros, Ichiro Ide, Annie Morin, Pascale Sébillot.

Ichiro Ide is associate professor at Nagoya University and visiting associate Professor at the National Institute of Informatics in Tokyo.

Asian research teams are among the most dynamic ones in multimedia research. For several years, we have had a collaboration with Ichiro Ide from Nagoya University and some contacts with Shin'ichi Satoh from NII in Tokyo. In 2007, taking advantage of the visit of Ichiro Ide (see Section 6.2.1.2), we proposed to build a common associate team.

8.3. Visits of foreign researchers, Invitations to foreign labs

8.3.1. Visit to the University of Montreal

Participants: Vincent Claveau, Pascale Sébillot.

Partners: Alain Polguère, Professor at the University of Montreal, Canada, and Marie-Claude L'Homme, head of OLST lab and Professor at the University of Montreal.

In the framework of an informal collaboration between TEXMEX and OLST, Vincent Claveau and Pascale Sébillot spent 10 days at OLST in January 2007. During this period, they worked on the lexical acquisition of semantic relations between terms and had discussions with Alain Auger, group leader of the Information and knowledge management section at Defense R&D Canada - Valcartier.

8.3.2. Visit from Members of the University of Ljubljana

Participant: Annie Morin.

Blaz Zupan and Janez Demsar came in IRISA in november 2007. We finalized a document on text mining.

8.3.3. Visit from Members of the University of Zagreb

Participant: Annie Morin.

Bojana Dalbelo Basic and Jan Snalders spent one week in IRISA in April 2007. Sasa Petrovic worked on his Master Degree on "Collocation extraction measures for text mining applications" during a 3-month INRIA internship. The topic of the thesis is the use of collocations and factorial correspondence analysis in text mining and its implementation in the free software Orange, a data mining tool developed by Blaz Zupan at the university of Ljubljana. In December, Josip Krapac will spend 3 days in our laboratory.

8.3.4. Visit to the University of Zagreb

Participant: Annie Morin.

A Morin spent one week in April 2007 at the university of Zagreb. We started writing a paper which has been presented at EPIA 2008 and which is published in a LNCS document [52].

9. Dissemination

9.1. Conference, Workshop and Seminar Organization

- L. Amsaleg organized the third edition of the Computer Vision meets Databases Workshop (CVDB), held in co-location with the International Conference ACM-SIGMOD/PODS, June 2007, Beijing, P. R. China. The CVDB workshop series tries to bridge the gap between the computer vision and the database communities 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.

9.2. Involvement with the Scientific Community

- L. Amsaleg:
  - was a program committee member of the 23rd Bases de données avancées, BDA 2007, Marseille, France;
  - was a program committee member of the 4th conférence en Recherche d'Information et Applications, CORIA'07, Saint-Étienne, March 2007;
  - was a program committee member of the Multimedia Data Mining (MDM07) workshop, associated with the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD07), San Jose, CA, USA, August 2007;
  - was a program committee member of the International Workshop on Ambient Intelligence, Media, and Sensing, associated with the International Conference on Data Engineering (ICDE 2007), Istanbul, Turkey, April 2007;
  - was a program committee member of the second edition of the EXPDB workshop on the performance and evaluation of data management systems, co-located with ACM-SIGMOD 2007, Beijing, P. R. China;
  - was a program committee member and the local chairman of the Topic 11 for the Europar 2007 International Conference, Rennes, France, August 2007;
  - is the 2007 ACM SIGMOD Workshops Coordinator.
- N. Bonnel:
  - was a reviewing committee member of the 5th Manifestation des jeunes chercheurs francophones dans les domaines des STIC, MajecStic'07, Caen, France, November 2007.
- V. Claveau:
  - was a program committee member of the 4rd COnference on Information Retrieval and Application CORIA'07, Saint-Étienne, March 2007;
  - was a program committee member of TALN'07 (14^{e} conférence nationale Traitement automatique des langues naturelles), Toulouse, France, June 2007;
  - was a program committee member of RECITAL'07 (Rencontres Jeunes Chercheurs en Traitement automatique des langues), Toulouse, France, June 2007;
  - was a program committee member of the special issue on Pattern-based approaches to semantic relation extraction of the Terminology journal (vol. 14(1), to appear in 2008).
- P. Gros:
  - is a member of the editorial board of the French journal entitled "Traitement du signal";

  - was program committee member of the fourth COnference on Information Retrieval and Application CORIA'07, Saint-Étienne, March 2007;
  - was a program committee member of the conference on Content-Based Multimedia indexing CBMI'07, Bordeaux, June 2007.
  - was a program committee member of the Conference "COmpression et représentation des signaux audiovisuels" CORESA'07, Montpellier, November 2007.
  - was a program committee member of the Workshop "Computer Vision meets DataBases", Beijing, June 2007.
  - was a program committee member of the GRETSI conference, Troyes, June 2007.
  - was a program committee member of the Conference on Multimedia Signal Processing MMSP'07, Chania, October 2007.
  - was a program committee member of the International conference on signal processing and multimedia applications SIGMAP'07, Barcelona, July 2007.
  - was a program committee member of the Conference on Visual Information Engineering VIE'07, London, July 2007.

- F. Moreau:
  - was a reviewing committee member of RECITAL'07 (Rencontres Jeunes Chercheurs en Traitement automatique des langues), Toulouse, France, June 2007.

- A. Morin:
  - is a program committee member of ITI 2007 (Information technology interfaces);
  - is a member of the CNU (National Council of the University) in the computer science section.

- F. Poulet:
  - was a program committee member of AusDM'07, Australasian Data Mining Conference, Gold Coast, Queensland, Australia, Dec.2007.

- P. Sébillot:
  - was a program committee member of GL2007 (4^{th} international workshop on Generative approaches to the Lexicon), Paris, France, May 2007;
  - was a program committee member of TALN'07 (14^{e} conférence nationale Traitement automatique des langues naturelles), Toulouse, France, June 2007;
  - was a program committee member of DEFT'07 (3^{e} Défi fouille de textes), Grenoble, France, July 2007;
  - was a program committee member of TIA'2007 (7^{e} conférence nationale Terminologie et intelligence artificielle), Sophia-Antipolis, France, October 2007;
  - was a program committee member of the special issue on Pattern-based approaches to semantic relation extraction of the Terminology journal (vol. 14(1), to appear in 2008).

- L. Ughetto:
  - was a reviewing committee member for the Fuzzy Sets and Systems journal, Elsevier;
  - was a reviewing committee member for the IEEE Trans on Fuzzy Systems journal;
  - was a program committee member of the Rencontres Francophones sur la Logique Floue et ses Applications (LFA'07);

9.3. Teaching Activities

- L. Amsaleg: Managing Large Collections of Digital Data. Research Master in Computer Science, 2nd year, University of Rennes 1.
- L. Amsaleg: Advanced Databases, Professional Master in Computer Science, 2nd year, IFSIC, University of Rennes 1.
- P. Gros coordinates the track "From data to knowledge: Machine Learning, Modeling and Indexing Multimedia Contents and Symbolic Data" of the Research Master 2 in computer science, at university of Rennes 1.
- E. Kijak: Analysis of audiovisual documents and flows for indexing, Research Master in Computer Science, 2nd year, IFSIC, University of Rennes 1.
- E. Kijak and P. Tirilly: Digital Documents Indexing and Retrieval, Professional Master in Computer Science, 2nd year, IFSIC, University of Rennes 1.
- P. Sébillot, V. Claveau and P. Tirilly: Advanced Databases and Modern Information Systems, 5th year, Computer Science, INSA Rennes.
- V. Claveau: Symbolic Sequential data, Research Master in Computer Science, 2nd year, University of Rennes 1.

9.4. Participation to Seminars, Workshops, Invited Conferences

- L. Amsaleg and his co-authors were asked to present an invited paper [43] at the 2007 CBMI International Workshop on Content-Based Multimedia Indexing.
- V. Claveau gave an invited talk at Orange Labs (France Télécom R&D), Lannion, about Inductive Logic Programming in Natural Language Processing, in October 2007.
- V. Claveau gave an invited talk at the University of Montreal (joint seminar for the RALI and OLST labs) about the use of his tool AnaMorpho in information retrieval, in January 2007.

9.5. Popular work, Press articles

- L. Amsaleg published an article in *Techniques pour l'ingénieur*, entitled "Contrer le piratage d'images : un logiciel précis et rapide", February 2007.
- P. Gros published an article entitled "Description et indexation automatiques des documents multimédias : du fantasme à la réalité" on the web site *Interstices* (http://interstices.info/display.jsp?id=c_19256) which aims at disseminating scientific results and culture for a broad audience.

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] 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.

[4] 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.

[5] E. KIJAK, G. GRAVIER, L. OISEL, P. GROS. *Audiovisual integration for sport broadcast structuring*, in "Multimedia Tools and Applications", vol. 30, 2006, p. 289-312, http://www.springerlink.com/content/24h61433843r474l/.

[6] H. LEJSEK, F. ÁSMUNDSSON, B. Þ. JÓNSSON, L. AMSALEG. *Scalability of Local Image Descriptors: A Comparative Study*, in "Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, CA, USA", October 2006, http://www.irisa.fr/texmex/publications/versionElect/2006/fp33a-lejsek.pdf.

[7] R. PRIAM, A. MORIN. *Visualisation des corpus textuels par treillis de multinomiales auto-organisées - Généralisation de l'analyse factorielle des correspondances*, in "Revue Extraction des Connaissances et Apprentissage (Actes EGC'02)", vol. 1, n^{o} 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^{o} 1, 2005, p. 105-127.

[9] F. TONNIN, P. GROS, C. GUILLEMOT. *Analysis of multiresolution representations for compression and local description of images*, in "Proceedings of the 8th International Conference on Visual Information and Information Systems, VISUAL'05, Amsterdam, The Netherlands", 2005.

Year Publications

Books and Monographs

[10] L. AMSALEG, B. Þ. JÓNSSON, V. ORIA (editors). *Proceedings of the 3rd International Workshop on Computer Vision meets Databases, CVDB'07*, ACM, Beijing, P. R. China, June 2007.

[11] P. GROS (editor). *L'indexation multimédia – Description et recherche automatiques*, Traité IC2, série Traitement du signal et de l'image, Hermes - Lavoisier, May 2007.

[12] F. POULET, B. LE GRAND, L. BOUDJELOUD (editors). *Actes des 5ème Atelier Visualisation et Extraction de Connaissances, Extraction et Gestion de Connaissances (EGC'07)*, January 2007.

Doctoral dissertations and Habilitation theses

[13] S. HUET. *Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole*, Ph. D. Thesis, University of Rennes 1, December 2007.

[14] X. NATUREL. *Structuration automatique de flux vidéos de télévision*, Ph. D. Thesis, University of Rennes 1, April 2007, http://www.irisa.fr/centredoc/publis/theses/irisafolder.2007-03-01.1633154603/irisapublication.2007-04-26.0496201623.

Articles in refereed journals and book chapters

[15] L. AMSALEG. *Indexation multidimensionnelle*, in "L'indexation multimédia. Description et recherche automatiques", P. GROS (editor), traité IC2, chap. 8, Lavoisier - Hermès Science Publications, 2007, p. 215-244.

[16] L. BOUDJELOUD, F. POULET. *Algorithme génétique interactif pour la fouille de données*, in "I3, Information Interaction Intelligence", 2007, p. 31-56.

[17] 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, 2007, http://www.irisa.fr/texmex/publications/versionElect/2007/claveausebillot07.pdf.

[18] M. DELAKIS, G. GRAVIER, P. GROS. *Audiovisual Integration with Segment Models for Tennis Video Parsing*, in "Computer Vision and Image Understanding", to appear, 2007.

[19] T.-N. DO, N.-K. PHAM, F. POULET. *Exploration interactive de résultats d'arbre de décision*, in "RNTI-E-9, Revue des Nouvelles Technologies de l'Information", vol. 1, 2007, p. 157-168.

[20] E. FANGSEU BADJIO, F. POULET. *Algorithme génétique interactif pour la fouille de données*, in "I3, Information Interaction Intelligence", 2007, p. 99-129.

[21] S. HUET, G. LECORVÉ, G. GRAVIER, P. SÉBILLOT. *Toward the Integration of NLP and ASR Techniques: POS Tagging, Topic Adaptation and Transcription*, in "Multimodal Processing and Interaction: Audio, Video, Text", P. MARAGOS, A. POTAMIANOS, P. GROS (editors), to appear, Springer, 2007.

[22] F. MOREAU, V. CLAVEAU, P. SÉBILLOT. *Description des textes*, in "Indexation multimédia. Description et recherche automatiques", P. GROS (editor), traité IC2, chap. 6, Lavoisier - Hermès Science Publications, June 2007, p. 163-190.

[23] X. NATUREL, P. GROS. *Detecting Repeats for Video Structuring*, in "Multimedia Tools and Application", to appear, 2007.

[24] F. POULET, T.-N. DO. *Interactive Decision Tree Construction for Interval and Taxonomical data*, in "Visual Data Mining: Theory, Techniques and Tools for Visual Analytics", A. M. S. SIMOFF (editor), Springer-Verlag, 2007.

[25] F. POULET. *Towards Effective Visual Data Mining with Cooperative Tools*, A. M. S. SIMOFF (editor), Springer-Verlag, 2007.

[26] A. REMAZEILLES, F. CHAUMETTE, P. GROS. *Asservissement visuel qualitatif pour la navigation*, in "Traitement du signal", vol. 24, n^{o} 3, June 2007, p. 191–209.

[27] W. VOGLOZIN, G. RASCHIA, L. UGHETTO, N. MOUADDIB. *From User Requirements to Evaluation Strategies of Flexible Queries in Databases*, J. GALINDO (editor), Information Science Reference, to appear, 2007.

Publications in Conferences and Workshops

[28] J. BENOIS-PINEAU, P. JOLY, E. KIJAK, G. QUÉNOT. *ARGOS: French Evaluation Campaign for benchmarking of video content analysis methods*, in "Proceedings of the SPIE-IS&T Electronic Imaging conference, San Jose, CA, USA", A. HANJALIC, R. SCHETTINI, N. SEBE (editors), Multimedia Content Access: Algorithms and Systems, vol. 6506, 65060O, The International Society for Optical Engineering (SPIE), 2007, p. 1–11.

[29] S.-A. BERRANI, P. LECHAT, G. MANSON. *TV broadcast macro-segmentation: Metadata-based vs content-based approaches*, in "Proceedings of the 6th ACM International Conference on Image and Video Retrieval, CIVR, Amsterdam, Netherlands", July 2007, p. 325-332.

[30] N. BONNEL. *SmartWeb: Display of Web Search Results in a Virtual City*, in "Proceedings of the ACM SIGCHI 2007 Workshop on "Exploratory Search and HCI: Designing and Evaluating Interfaces to Support Exploratory Search Interaction", San Jose, CA, USA", April 2007, p. 27-31.

[31] J.-H. CHAUCHAT, A. MORIN. *Quelques erreurs et abus courants dans l'interprétation des résultats de la fouille des données*, in "Actes du Troisième atelier : qualité des données et des connaissances, 7ème conférence Extraction et Gestion des Connaissances (EGC'07), Namur, Belgium", January 2007, p. 65-70.

[32] J.-H. CHAUCHAT, A. MORIN, R. RAKOTOMALALA. *Correcting the error rate estimation bias in Data Mining when the dataset comes from a two-stage sampling*, in "Proceedings of IASC 07 - Statistics for Data Mining, Learning and Knowledge Extraction, Aveiro, Portugal", August 2007.

[33] V. CLAVEAU. *Inférence de règles de réécriture pour la traduction de termes biomédicaux*, in "Actes de la conférence Traitement automatique des langues naturelles, TALN'07, Toulouse, France", June 2007, http://www.irisa.fr/texmex/people/claveau/publis/Claveau-TALN07.pdf.

[34] V. CLAVEAU. *Traduction de termes biomédicaux pour la recherche d'information interlingue*, in "Actes de la 4ème Conférence en Recherche d'Informations et Applications, CORIA'07, Saint-Étienne, France", March 2007, http://www.irisa.fr/texmex/people/claveau/publis/Moreau-Claveau_Sebillot_CORIA07.pdf.

[35] K. DAÐASON, H. LEJSEK, F. H. ÁSMUNDSSON, B. Þ. JÓNSSON, L. AMSALEG. *Eff2 Videntifier: Identifying Pirated Videos in Real-Time*, in "Proceedings of the 15th ACM International Conference on Multimedia, Augsburg, Germany", September 2007.

[36] T.-N. DO, N.-K. PHAM, H.-T. DO, D.-L. NGO, T.-V. NGUYEN. *Classifying large datasets with SVM*, in "Proceedings of the 3rd National Symposium Fundamental and Applied IT Research, FAIR'07, Vietnam", 2007.

[37] T.-N. DO, F. POULET. *Classification de grands ensembles de données avec un nouvel algorithme de SVM*, in "Actes des 7èmes journées Extraction et Gestion de Connaissances (EGC'07), Namur, Belgium", January 2007, p. 739-750.

[38] T.-N. DO, F. POULET, J.-D. FEKETE. *Massive Data Mining via Boosting of Least Squares SVM Algorithm*, in "Proceedings of the 5th IEEE International Conference on Computer Sciences: Research & Innovation and Vision for the Future (RIVF'07), Hanoï, Vietnam", 2007, p. 47-52.

[39] F. GIANNI, J. PINQUIER, E. KIJAK. *ACADI Showcase - Automatic Character Indexing in Audiovisual Document*, in "Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR", ACM, 2007, p. 109–112, http://doi.acm.org/10.1145/1282280.1282298.

[40] S. HUET, G. GRAVIER, P. SÉBILLOT. *Morphosyntactic Processing of N-Best Lists for Improved Recognition and Confidence Measure Computation*, in "Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech), Antwerp, Belgium", August 2007, p. 1741-1744, http://www.irisa.fr/texmex/publications/versionElect/2007/shuet07a.pdf.

[41] P. JOLY, J. BENOIS-PINEAU, E. KIJAK, G. QUÉNOT. *The ARGOS Campaign: Evaluation of Video Analysis Tools*, in "Proceedings of the International Workshop on Content-Based Multimedia Indexing, CBMI, Bordeaux, France", IEEE, 2007, p. 130–137, http://ieeexplore.ieee.org/iel5/4275036/4275037/04275065.pdf.

[42] H. KOSCH, L. AMSALEG, E. J. PAUWELS, B. Þ. JÓNSSON. *Topic 11 Distributed and High-Performance Multimedia*, in "Proceedings of Euro-Par conference", 2007, p. 767-767, http://dx.doi.org/10.1007/978-3-540-74466-5_81.

[43] H. LEJSEK, B. Þ. JÓNSSON, L. AMSALEG. *The Eff2 Project: Towards Efficient and Effective Support for Large-Scale High-Dimensional Indexing*, in "Invited Paper. Proceedings of the 5th International Workshop on Content-Based Multimedia Indexing, CBMI, Bordeaux, France", June 2007.

[44] P. P. LÉVY, B. LE GRAND, F. POULET, M. SOTO, L. DARAGO, L. TOUBIANA, J.-F. VIBERT. *Pixelization Paradigm*, in "Proceedings of the First Visual Information Expert Workshop (VIEW 2006), Paris, France", Lecture Notes in Computer Science, vol. 4370, Springer, 2007.

[45] F. MOREAU, V. CLAVEAU, P. SÉBILLOT. *Automatic Morphological Query Expansion Using Analogy-Based Machine Learning*, in "Proceedings of the 29th European Conference on Information Retrieval, ECIR 2007, Roma, Italy", April 2007, http://www.irisa.fr/texmex/people/claveau/publis/Moreau_Claveau_Sebillot_ECIR07.pdf.

[46] F. MOREAU, V. CLAVEAU, P. SÉBILLOT. *Combining linguisitic indexes to improve information retrieval systems: a machine learning based solution*, in "Proceedings of the RIAO 07 conference "Large-Scale Semantic Access to Content (Text, Image, Video and Sound)", Pittsburgh, USA", May-June 2007, http://www.irisa.fr/texmex/publications/versionElect/2007/Moreau_Claveau_Sebillot_RIAO07-1.pdf.

[47] F. MOREAU, V. CLAVEAU, P. SÉBILLOT. *Intégrer plus de connaissances linguisitiques en recherche d'information peut-il augmenter les performances des systèmes ?*, in "Actes de la 4ème Conférence en Recherche d'Informations et Applications, CORIA'07, Saint-Étienne, France", March 2007, http://www.irisa.fr/texmex/publications/versionElect/2007/Moreau_Claveau_Sebillot_CORIA07.pdf.

[48] A. MORIN. *Recherche d'information et analyse des correspondances : une étude des publications récentes sur la malaria*, in "Actes du colloque international de Statistique Appliquée pour le Développement en Afrique, SADA'07, Cotonou, Bénin", 2007.

[49] X. NATUREL. *Applications de la détection de texte pour l'indexation de vidéos de télévision*, in "Proceedings of the 12èmes journées d'étude et d'échange Compression et Représentation des Signaux Audiovisuels, CORESA'07", 2007.

[50] N.-K. PHAM, T.-N. DO, F. POULET, A. MORIN. *Exploration interactive des arbres de décision*, in "Actes du 5ème Atelier Visualisation et extraction de connaissances, 7èmes Journées d'Extraction et Gestion des Connaissances, EGC'07, Namur, Belgium", 2007, p. 79-91.

[51] N.-K. PHAM, T.-N. DO, F. POULET, A. MORIN. *Interactive Exploration of Decision Tree Results*, in "Proceedings of the International Symposium on Applied Stochastic Models and Data Analysis, ASMDA'07, Chania, Greece", 2007, p. 152-160.

[52] A. SILIC, J.-H. CHAUCHAT, B. DALBELO BASIC, A. MORIN. *N-grams and morphological normalization in text classification : a comparison on a Croatian-English corpus*, in "Proceedings of EPIA 2007", To appear., Springer-Verlag, 2007.

[53] R. TAVENARD, L. AMSALEG, G. GRAVIER. *Estimation de similarité entre séquences de descripteurs à l'aide de machines à vecteurs supports*, in "Actes des 23èmes journées Bases de Données Avancées, BDA'07, Marseille, France", October 2007, http://www.irisa.fr/texmex/publications/versionElect/2007/tavenard07a.pdf.

[54] R. TAVENARD, L. AMSALEG, G. GRAVIER. *Machines à vecteurs supports pour la comparaison de séquences de descripteurs*, in "Actes des 12èmes journées d'étude et d'échange Compression et Représentation des Signaux Audiovisuels, CORESA'07, Montpellier, France", November 2007, http://www.irisa.fr/texmex/publications/versionElect/2007/tavenard07b.pdf.

[55] P. TIRILLY, V. CLAVEAU, P. GROS. *Annotation d'images sur de grands corpus réels de données : limites des histogrammes de couleurs*, in "Actes des 2èmes Rencontres Jeunes Chercheurs en Recherche d'Information, RJCRI'07, Saint-Étienne, France", March 2007, http://www.irisa.fr/texmex/publications/versionElect/2007/tirilly07a.pdf.

[56] P. TIRILLY, E. MARTIENNE, V. CLAVEAU, P. GROS. *Association d'un détecteur de visages et d'un détecteur d'entités nommées pour l'annotation automatique d'images*, in "Actes des 2èmes Rencontres Jeunes Chercheurs en Recherche d'Information, RJCRI'07, Saint-Étienne, France", March 2007, http://www.irisa.fr/texmex/publications/versionElect/2007/tirilly07b.pdf.

Internal Reports

[57] K. HARÐARSON, B. Þ. JÓNSSON, L. AMSALEG. *Breaking out of the Shoebox: Towards Having Fun with Digital Images*, Technical report, n^{o} RUTR-CS07002, Reykjavík University, August 2007.

References in notes

[58] 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.

[59] S.-A. BERRANI, L. AMSALEG, P. GROS. *Recherche par similarités dans les bases de données multidimensionnelles : panorama des techniques d'indexation*, in "Ingénierie des Systèmes d'Information", vol. 7, n^{o} 5/6, 2002.

[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^{o} 3, September 2001.

[61] T. DEAN, K. KANAZAWA. *A model for reasoning about persistence and causation*, in "Articial Intelligence Journal", vol. 93, n^{o} 1, 1989.

[62] 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.

[63] 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^{o} 2, 1994, p. 171-187.

[64] 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.

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

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

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

[68] D. G. LOWE. *Distinctive image features from scale-invariant keypoints*, in "International Journal of Computer Vision", vol. 60, n^{o} 2, 2004, p. 91–110.

[69] K. MURPHY. *Dynamic Bayesian Networks: Representation, Inference and Learning*, Ph. D. Thesis, University of California, Berkeley, 2002.

[70] M. OSTENDORF. *From HMMs to Segment Models*, in "Automatic Speech and Speaker Recognition - Advanced Topics", chap. 8, Kluwer Academic Publishers, 1996.

[71] L. RABINER, B.-H. JUANG. *Fundamentals of speech recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

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

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

[74] C. SCHMID, R. MOHR. *Local Grayvalue Invariants for Image Retrieval*, in "ieee Transactions on Pattern Analysis and Machine Intelligence", vol. 19, n^{o} 5, May 1997, p. 530–534.

[75] J. SIVIC, B. C. RUSSELL, A. A. EFROS, A. ZISSERMAN, W. T. FREEMAN. *Discovering Object Categories in Image Collections*, in "Proceedings of the International Conference on Computer Vision, Beijing, China", 2005.

[76] J. SIVIC, A. ZISSERMAN. *Video Google: A Text Retrieval Approach to Object Matching in Videos*, in "Proceedings of the International Conference on Computer Vision", vol. 2, October 2003, p. 1470–1477.

[77] M. UTIYAMA, H. ISAHARA. *A Statistical Model for Domain-Independent Text Segmentation*, in "Proceedings of the 39th Annual Meeting of Association for Computational Linguistics, ACL'01, Toulouse, France", July 2001.