Acronym of the proposal	**VideoSense**		
Title of the proposal in French	Reconnaissance multimodale de concepts enrichis (statiques, dynamiques, émotionnels) dans des vidéos multilingues au travers de langages pivots.		
Title of the proposal in English	Rich concepts recognition in multilingual videos, throw pivot languages.		
Theme	☐ 1 ■ 2 ☐ 3 ☐ 4		
Type of research	☐ Basic Research ■ Industrial Research ☐ Experimental Development		
Type of scientific project	☐ Platform		
Total requested funding	1 111 090 €	**Project Duration**	36 months

VideoSense project

Partners :

- Eurecom
- Ecole Centrale de Lyon LIRIS
- Laboratoire d'Informatique de Grenoble CNRS
- Laboratoire d'Informatique Fondamentale de Marseille
- Ghanni

1. CONTEXT AND RELEVANCE TO THE CALL

The democratization of digital cameras has led to a proliferation of video data all around us. According to a recent study by the market researcher IDC, digital camera sales rose 15% in 2006 to 105.7 million units worldwide[1]. On the other hand, the progress in information technologies and appearance of more and more powerful front-end devices leads to a constantly growing amount of digitized video in various fields, including video archives, distance learning, communication, entertainment, etc. A content-based access could greatly facilitate navigation in huge video storages. Indexing and retrieving these large quantities of video data is an extremely challenging and increasingly topic problem for both industry and academia.

The present project brings together research teams and companies having strong expertise in their respective field and aims at developing cutting edge techniques for automatic concept-based video sequence tagging. The issues of interest to users would cover not only the "objective" content of video data, but also their "emotional" content, which determines their potential impact. In this context, the VideoSense project aims to explore, develop and experiment new techniques and tools to index and classify video data including aspects related to the independent analysis of its composite modalities, covering spatial-temporal visual, audio and emotional content and the closed captions for the textual modality, and their combination as regard to the multimodal nature of a video data.

Such automatic video content tagging systems can be the basis to support end-user interactions such as online video searching, filtering, mining, content-based routing, personalization, summarization, etc. In developing machine-based video tagging techniques within the VideoSense project, we are particularly targeting two applications, namely video content recommendation and content monetization by targeted ads, in partnership with the French technology innovation labeled Ghanni company (www.ghanni.com) which is specialized in media asset management and recommendation solutions. Once video sequences are tagged with high-level concepts, video to video and ads to video similarity can be applied from a semantic viewpoint. Therefore videos related to a particular video can be recommended based on the content similarity. Moreover, ads can be targeted toward a specific video based on the similarity between its content and the ads' target. Both applications rely on the same underlying video tagging system that extracts semantic concepts or tags from video sequences.

1.1. CONTEXT, ECONOMIC AND SOCIETAL ISSUES

Video tagging systems are major component of the digital asset management (DAM) solutions which target the management of rich digital content, which includes various file types including audio, video, graphics, text, and more, and enable as rich functionalities as the archiving, indexing, search retrieval, ingest, browsing, annotation, repurposing, collaboration, display and transport of digital media. According to a market study by Frost &

[1] « Shipments of Digital Cameras Rose 15% last year, IDC says », The wall street J., 3 Apr. 2007

Sullivan in 2000, Digital Asset Management revenues nearly tripled between 1996 and 2000 and this market has tremendous potential. As digital media applications are spreading across verticals the demand for DAM solutions is also increasing. What was once a niche product aimed at specialty vertical markets such as broadcasting, entertainment and advertising is rapidly moving into the mainstream as organizations realize the need to organize, and manage their growing portfolio of digital assets. Global 2000 companies are moving towards widespread deployment of DAM solutions in order to minimize cost of lost assets and lost time searching for those assets.

With regard to online video resources targeted by the two applications within the VideoSense project, their consumption has become in the last years a major element in the overall video consumption habits. According to a study by Comscore [Source: *Comscore, More than 10 Billion Videos Viewed Online in the U.S. in February, April 16, 2008*] 10 Billion videos viewed online in the US alone in February 2008, a 66% jump versus February 2007. Nearly 135 million U.S. Internet users spent an average of 204 minutes per person viewing online video in February. 72.8 percent of the total U.S. Internet audience viewed online video. The average online video duration was 2.7 minutes. The average online video viewer consumed 75 videos in February 2008.

We can distinguish two main types of videos existing online, User Generated Videos and Professional Videos. User Generated Videos (UGV) are videos generated by amateur users. UGV constitute a major part of videos available on popular video sharing destinations such as YouTube. Professional Videos are videos created by professionals or semi-professionals. Professional videos made up 19.5% of total views on YouTube in 2006 – 2007, and forecast at 26.5% in 2008 [Source: *Professional and UGV Market Size 2005 - 2008: Views, Category and Brand Share Analysis*, Accustream imedia research, March 2008].

One efficient business model for online video market is an ad-supported model. After being efficiently applied to other different types of professional videos, this model has been recently experimented on premium videos, such as fiction series and movies, on sites like Hulu in the US or M6Replay in France. Video ads accounted for 371 Million $ in the US in 2007, and forecast for 1,226 Billion $ in 2010 [Source: eMarketer, November 2008]. Professional videos account for the majority of video ads since they are more attractive to advertisers than UGV. Interesting information is that the CPM (Cost Per Thousand impression) of the video ads is noticeably higher than the CPM of banners or search ads. Video ads CPM is higher than 15$ while the search ads CPM is lower than 10$ [Source: *State of The Industry Q4*, Liverail, 2008].

One important consideration to take into account in the video ads market is ads targeting. Targeting means showing the right ads to the right person at the right time. Targeting is normally based on the content of the video, since this information gives an indication on the users' interests. Targeted ads are more acceptable to the user, generate more revenues to the publisher/producer, and are attractive to the advertiser than random ads.

Presently, targeting is based on a manual indexing and selection of videos. This greatly limits the growth of the market since it's restricted to big name advertisers for high volume ads programs. Automatically indexing videos to permit an efficient targeting of ads is clearly needed to attract small advertisers and small publishers alike.

Automatic video ads targeting will have two main effects. The first one is growing the video ads market by making video ads available to small advertisers and publishers/producers. The second important effect is on the creativity and content creation development. In fact, small publishers and producers that efficiently monetize their videos will be more motivated to improve the quality and availability of their content. Therefore, the creativity process can be obviously enhanced. Combined to intelligent search and recommendation solutions, this can greatly boost the democratization of quality content creation and consumption.

Systems that automatically tag the content of a video with semantic classes (e.g. boat, sea, desert road, beach, Bill Clinton, car pursuit, explosion, people walking, professional video, and amateur video) can be used as the basis for ads targeting and video recommendation. The present VideoSense project aims at developing such systems. Therefore, it can be considered as an effective response to a market need. Additionally, this project can be considered as a catalyst for cultural and informational content creation and dissemination.

1.2. RELEVANCE OF THE PROPOSAL

The development of the web and the huge amount of information that is now accessible has created a tremendous importance for techniques that are able to associate relevant information. A typical example is the Google Adsense technology which can select relevant advertising based on the textual content of the web page being visited. This is currently the basis for the financial success of Google. Nowadays, more and more information comes in the form of video data, as shown by the success of web sites such as YouTube and DailyMotion. The VideoSense project aims at extending similar association mechanisms to video data, by detecting the appearance of certain concepts and categorizing certain segments of video sequences.

The automatic analysis of video sequences has been a very active research area for the past decade. While earlier works focused on certain genres (TV News and sports videos are typical examples), the current trend is build more and more generic systems, that are capable to adapt efficiently to new types of video data or new types of search elements. A major effort in this domain is conducted within the TRECVID international evaluation campaign, which gathers every year the most active research teams on a set of common tasks. One of these tasks is called "High Level Feature Extraction" (HLFE) and is based on the detection of a set of "concepts". Examples of those concepts are: "airplane flying", "cityscape", "telephone", "protest", "singing", etc…

The importance of video tagging appears also in the activity of several projects:

- The MUMIS EU project studied multimedia and multilingual indexing for video archives,
- The PENG EU project has studied the personalization of News articles to specific user interests,
- The RUSHES project focused on the analysis and reuse of video rushes,
- The VICTORY and SAPIR projects studied indexing in a P2P environment,
- The MUSCLE Network of Excellence has explored a number of areas in video and multimodal indexing
- The Quaero AII-OSEO programme aims to facilitate the extraction of information in unlimited quantities of multimedia and multilingual documents, including written texts, speech and music audio files, and images and videos.

The current project clearly aims to answer the challenges described in the 2nd theme "assemblage, édition et exploitation de contenus et connaissances", in particular innovative solutions for the extraction of static, dynamic and emotional concepts from the multimodal nature of video data composed of synchronized visual and audio channels possibly with closed captions. Within this theme, the project will contribute to:

- Subtheme 2 (metadata, ontologies), by the choice of concepts sets that will be accomplished in the project,
- Subtheme 3 (indexing, fusion, extraction), by the various analysis procedures that will be developed in the project,
- Subtheme 4 (ontologies, adaptation, profiling), by the association between video data and other informational material, such as automatic advertising.

In order to deal with the closed captions in several languages, the current project also proposes to make use of a pivot language and to extract disambiguated textual concepts. We will experiment with three types of pivots, including UNL.

The proposed project builds on the expertise of the academic partners, to support the development of innovative applications for the industrial partner of the project. The teams involved in the VideoSense project are active at the European and international levels on the topics addressed by the project. Eurecom and LIG have been regularly participating to the international TRECVID evaluation campaigns since 2001; LIG has recently organized the active learning annotation collaboration. ECL LIRIS has a strong experience in video and audio analysis, and a participation in several national and European research projects. LIF has strong skills in machine learning domain, especially statistical and kernels methods, applied to multimedia analysis. Ghanni brings to the project the knowledge of the market in terms of application requirement, its media asset management platform and a large corpus of 50 000 videos in different genres from professionals or amateurs.

2. SCIENTIFIC AND TECHNICAL DESCRIPTION

2.1. STATE OF THE ART

High level concept recognition from video data is a challenging task because of complex motion, cluttered backgrounds, occlusions, and geometric and photometric variances of objects. Despite these difficulties, increasingly powerful techniques have emerged for machine tagging video content. However, fundamental questions remain. In the following subsequent sections, we propose to overview related works on some of these fundamental questions as compared to our application scenario targeted within the project.

2.1.1 SEMANTIC CONCEPT LEXICON

An overview of publications on automated tagging techniques over the last 10 years reveals some few recurrent semantic concepts, such as indoors versus outdoors, cityscape versus landscape, people, faces, and so forth. Recently, a collaborative effort gathering multimedia researchers, library scientists, and end-users was carried out in USA to develop a large standardized taxonomy for describing broadcast news video, leading to the Large-Scale Concept Ontology for Multimedia (LSCOM). This LSCOM covers 834 semantic concepts containing several broad categories, such as objects, activities/events, scenes/location, people and graphics. VideoSense is targeting a subset of the LSCOM, in particular the ones used in TRECVID [1] (international evaluation campaign of systems for content based video indexing and retrieval, organized by NIST), for which there exist a significant amount of annotated video data for learning and testing. While LSCOM target broadcast news video, VideoSense aims at a big range of video genres, including news, comedies, educational programs, etc., but also professional videos as well as amateur ones.
Therefore, a first filtering of LSCOM's concepts may be necessary. Some specific concepts, for instance professional videos versus amateur videos, may be added. In addition, we also want to take into account the perceptual feelings of the end-users while experiencing video data, namely emotional effect provoked by viewing video document, such as excitement, calmness, enchantment, stress, anxiety, violence, etc.

2.1.2 VIDEO CONTENT DESCRIPTION

A video document is a multimodal data composed of visual channel, auditory channel and possibly textual resources, such as closed captions, describing the content of the video document. Video content modeling aims at representing the semantic content, i.e. consistent properties, of a video from the visual content, the auditory content, the temporal properties and possibly the textual content of a video data.

- *Visual descriptors*

The visual content can be described by well-known low level features at various granularities such as color moment, Gabor texture and edge direction histogram or shape information [2]. However, there can be large variations in lighting and viewing conditions for images representing real world scenes, making difficult stable description of visual content. Salient point detection methods and the associated regions descriptors, e.g. SIFT, can robustly detect regions which are translation, rotation and scale invariant, addressing the problem of viewpoint changes [3,4]. To increase illumination invariance and discriminative power, color descriptors, e.g. color SIFT [5], Hue SIFT [6], have been proposed [7,8]. The visual content of an image can also be modeled by some midlevel descriptors such as line-based

descriptors [9] or midlevel semantic concept scores [10]. All these features were frequently used in Pascal VOC challenge [11,12] or TRECVID [13].
The VideoSense project will adopt a mix representation of visual content of video frames, thus making combined use of low level features (e.g. color SIFT, edge direction histogram) with mid-level concept descriptions as latent semantic directions (e.g. segment, indoor/outdoor [14], city/landscape [15]). Moreover, we also would like to include some basic human visual perception laws and consider Gestalt inspired region segmentation scheme [16].

- *Spatial-temporal descriptors*

The temporal property of the visual content is key information, especially for video activity or event recognition, e.g. car crashing, people walking, etc. The activity or event recognition can be model-based (semi-supervised adapted Hidden Markov Model framework [17,18]) or appearance-based. Appearance-based techniques extract spatiotemporal features in the volumetric regions which can be densely sampled [19] or detected by salient region detection algorithms [20]. The performance of appearance-based techniques usually depends on reliable extraction of spatial-temporal and/or salient regions. This makes the approach sensitive to motion and video quality. Alternatively, a holistic representation for image frames can also be adopted without object tracking or spatiotemporal interest region detection [21]. In this case, a video clip is modeled as a bag of unordered descriptors extracted from all the constituent frames and earth mover's distance is applied to describe temporal similarities among frames from two clips.

Within the VideoSense project, we will adopt a mix description of the video spatiotemporal properties, thus combining some global representation such as color histogram with local representation using trajectories of some feature points.

- *Modeling of low level descriptors*

The visual descriptors can be sparsely extracted from salient image regions, i.e. interest "points" [22,23] or more densely from points extracted using several grids at multiple scales [24]. Once extracted from an image or a video clip, these visual or spatial-temporal descriptors need to be further modeled by a fixed size feature vector for machine-based learning and classification as regard to the target semantic concept or event. Recently, the "bag of features" kind of approach [25,26], which tries to adapt the "bag-of-words" representation for text categorization to "Visual Object Categorization" (VOC) problem, has been widely used and has shown its effectiveness in Pascal VOC and TRECVID contests. The bag of features approach describes an image or a video clip as a bag of discrete "visual words", where the histogram containing the number of occurrences of each visual word is used for further object or concept categorization.
Although this approach has achieved the best performance in Pascal or TRECVID contests, the overall performance, with an average precision less than 60% over 20 classes achieved by the best classifier, is still far from real application-oriented requirements. In particular, the size of visual vocabulary which is the basis of this approach is hard to be fixed as there are no evident similar concepts in images as compared to a textual document. The basic problem

is that bag of features does not necessarily correspond to a human visual perception process which seems to be ruled by some Gestalt principles according to several studies on visual perception [27,28] and supposed to perform a holistic analysis combined with a local one through a fusion process. Moreover, the schemes so far proposed in the literature for automatic generic visual object classification also suffer the problem of a small and biased training dataset, in particular with an imbalanced ratio of positives versus negative samples.
In our own works in multimedia analysis, we have proposed several other modeling schemes, including polynomial modeling [29] and previously Zipf modeling [30,31]. Within the current project, we propose to drive some comparative studies on the various modeling schemes and to deepen these works in order to better take into account contextual spatial and spatiotemporal information of video data.

2.1.3 SEMANTIC AUDIO CLASSIFICATION

The audio channel conveys rich semantic clues for content-based video analysis. Besides widely known speech extraction and speaker identification problems, interesting audio semantic analysis also includes speech/music segmentation [32], speaker gender detection [33,34], special effect recognition such as gun shots or car pursuit, etc. All these problems can be considered as an audio semantic classification problem which needs to generate a semantic label from low audio signal analysis. An overview of the related works can be found in [35].

Some popular features in the literature include bandwidth, power, band power, Zero Crossing Rate (ZCR), Mel Frequency Cepstral Coefficients (MFCC), pitch, etc. An interesting dataset used by several researchers is the MuscleFish database [36] which includes a set of 400 sounds containing classes such as water, bells, telephone, male, female etc.

While the existing audio analysis techniques in the literature are mostly problem specific, we aim to develop within the VideoSense project a general framework for audio semantic classification [37] which can be easily adapted to specific audio events needed for a multimodal recognition of video concepts targeted by the project.

2.1.4 EMOTIONAL CONTENT RECOGNITION

Automated analysis of human affective behaviour aims at capturing changes in the user's affective states and has attracted increasing attention from research community of several disciplines [38]. Automated emotional content recognition is of the deepest interest in video high-level concept recognition as emotional feelings while experiencing a video data directly measure the impact of the video and they are closely related to the cultural background and aesthetic standards of a user. However, as one of the most abstract or subjective semantic concept, its automatic recognition is difficult. The first research topic in the field is the description of affect. There exist in the literature two major theories on emotion description: the discrete and the dimensional emotion theories. For the discrete emotion theory, different numbers and different types of emotions are proposed by researchers. The term "big six" gained attention implying the existence of a fundamental set of six basic emotions while there does not seem to be any agreement on which six these should be [39]. In the

dimensional emotion approach, different emotional states are mapped into a two or three-dimensional space [40].

Research on vocal emotion recognition so far is largely driven by a basic emotion theory, making use of acted speech datasets for training and testing. Most of the existing approaches rely on acoustic features correlated to emotion expressions [41,42]. With the research shift toward the analysis of spontaneous human emotion, some tentative also made use of linguistic-paralinguistic features [43,44]. However, reliable extraction of linguistic-paralinguistic features, being language dependent, is also a difficult problem. In our own work, we have developed two new features, namely harmonic features and Zipf features, to better capture the prosodic properties and their structural patterns of speech signals [45]. As emotional states are application dependent, we also have developed a dimensional model driven hierarchical emotion recognition scheme based on evidence theory [46,47]. Moreover, this work has been extended to perform music mood recognition [48].

Most of the vision-based affect recognition studies focus on facial expression analysis. Although a lot of progress on automatic analysis of facial expressions has been made so far [49,50], the problem of the automatic analysis of facial affect in unconstrained environments is still far from being solved. Indeed, existing methods typically assume that the input data are near frontal or profile-view face image sequences, showing non occluded facial displays captured under constant lighting conditions against a static background.

The VideoSense project will consider rather video data, being professional or amateur, in unconstrained conditions. Therefore, we are not dealing with facial expression recognition but rather the global perceived emotional or aesthetic feelings from users when experiencing a video document. Few works have studied such global emotion recognition from the perception of visual or video content. A key problem is extraction of emotion sensitive features but few researchers extract image features in an emotional perspective [51,52]. Based on colour theory of Itten [53], Colombo et al. [54] retrieved art paintings by mapping expressive and perceptual features to four broad emotional categories. In our preliminary work, based on studies psychological effect of colours [55], we developed some basic color descriptors to categorize still images into four broad emotional classes [56]. Although some novel features have been proposed, research in this field is still at the primary stage.

Within the present project, we are focusing on automatic analysis of emotional content from video data. For this purpose, we are deepening our work on emotional speech recognition and extending it to emotional analysis on audio channel associated with video data. Visual emotion analysis will also be carried out, using some aesthetic features, such as balance and luminance of global and local color or density and thickness of segments, and spatiotemporal features including for instance object speed and/or trajectories and visual rhythm changes. Moreover, considering the composite nature of video data, we will also investigate fusion strategies combining audio and visual modalities for a joint emotion analysis. Such a fusion strategy can be feature-level fusion, decision-level fusion or model-level fusion (e.g. HMM) taking into account temporal structures of the modalities and their temporal correlations.

2.1.5 MULTILINGUAL TEXTUAL DESCRIPTOR

Textual data is often available in video collections. Either as an automatic transcription of the sound track, as closed captions, or as textual metadata associated to the video file. When such data is available in different languages, most projects use automatic translation systems as black boxes to translate the textual data in a common (pivot) language (usually in English, as in TRECVID). Such an approach suffers from two main problems. Firstly, the translated text suffers from all translated languages problems (ambiguity, synonymy, etc.). It has been shown that taking the source language into account (as a disambiguation data) raises the results drastically [57]. Secondly, translating does not scale well when new languages are involved. Developing translators requires great amounts of parallel corpus and/or dictionaries/grammars that do not exist in many language pairs.

In order to correctly handle multilingual textual content, we want to evaluate the use of an artificial pivot language to define textual descriptors. Such an artificial language uses disambiguated concepts. Some tentative has been conducted using UNL [58,59] as a pivot representation for Information Retrieval. UNL is an artificial language that allows the encoding of a textual utterance as a graph that uses "concepts" (called Universal Words or UW) as nodes related by deep syntactic relations (agent, patient, object, time, etc.). UWs are represented as English headwords along with annotation to disambiguate them. Others use WordNet synsets (e.g. [60]) as a pivot language, along with EuroWordNet to provide a translation from text to synsets. Such an approach shows other problems, mainly due to the fact that synsets are sometimes too refined for classification or information retrieval purposes. Moreover, the quality of language data available in EuroWordNet is varying.

For all these approaches, the pivot is a set of symbols (used as indexing terms) defined by a convention, norm or resource (UNL UWs, WordNet synsets) and interpreted as a word sense. [61,62] use conceptual vectors to represent a word sense. In this approach, all word senses in all languages are represented as vectors in the same vector space.

Nobody has ever experimented using conceptual vectors to define an ad-hoc pivot for classification. This will be our approach in VideoSense. Such a pivot will avoid the over-refinement problem found using UNL UWs or WordNet synsets. Moreover, conceptual vector can be associated to WordNet synsets or (to some extent) to UNL UWs. Hence they are a promising tool to determine what kind of pivot language is the best fitted for classification purposes.

2.1.6 ACTIVE LEARNING

Supervised learning consists in training a system from sets of positive and negative examples. The learning systems' performance depends a lot on the implementation choices and details but it also strongly depends upon the size and quality of the training examples. While it is quite easy and cheap to get large amounts of raw data, it is very costly to have them annotated because it involves human intervention for the judging of the "ground truth". While the volume of data that can be manually annotated is limited due to the cost of manual intervention, there remains the possibility to select the data sample that will be annotated so that their annotation is as useful as possible. Deciding which samples will be the more useful is not trivial. Active learning approach uses an existing system to predict the

usefulness of new samples. This approach is a particular case of incremental learning in which a system is trained several times with a growing set of samples.
Several strategies or heuristics can be considered to predict the samples' usefulness. The most popular ones include "query by committee" [63] where the system chose the samples which maximize the disagreement amongst several classifiers, "uncertainty sampling" [64] which tries to increase the sample density in the neighborhood of the frontier between positives and negatives, and "Relevance sampling" [65] which tries to maximize the size of the set of positive. More complex strategies can be used including combinations of these. For instance, the system may choose the samples for annotation amongst the most probable ones and amongst the farthest from the already evaluated ones. Another possibility is to select samples by groups in which maximize the expected global knowledge gain [66].

In [67], an annotation tool taking advantage of active learning process has been design as a web application and used for the annotation of the TRECVID'07 and TRECVID'08 corpora. The web-based approach allowed several users to annotate at same time and from various places. Based on a previous study in [68], the active learning process has been implemented with, first, the "relevance sampling", and then switched to the "most uncertain" strategy. An additional strategy was used in combination with others in order to boost the annotation of positive samples: "neighborhood sampling" has been proposed especially for the annotation of video corpus and consists in looking for new positive samples in the temporal neighborhood of already found positive samples.
Active learning has been shown to be very powerful to select the most informative samples for the annotation of training set. It has been shown that the most useful samples are quickly selected: classification based on the 15% first annotated samples gives satisfying performance, while the classification based on the 35% first annotated samples gives the best performance.

2.1.7 CONCEPTS CLASSIFICATION AND FUSION

Indexing video documents by concepts is a key issue for an efficient management of multimedia archives. Depending on the targeted application, the amount of concepts needed may be large in order to ensure a sufficient level of performance [69]. Due to the high quantity of concepts handled, it is intractable to perfectly optimize the model for each concept. Thus, a number of researchers focus on generic system for semantic video indexing. This problem is challenging because, unlike text documents, there is no simple correspondence between basic elements (numerical values of image pixels and/or of audio samples) and the semantic information (typically concepts). This is usually referred to as the "semantic gap" (between signal and semantics) problem [70]. In the context of generalist video archives, concepts are related to *objects* (car, person, telephone, boat …), *scenes* (cityscape, office, harbor …) or *events* (person entering in a car …).

Hence, the first thing that is commonly done for bridging the semantic gap is to model the documents by extracting low-level descriptors which must be robust to changes such as illumination, noise or rotation in order to allow discriminating between documents. Another critical step for filling the semantic gap consists in mapping the low-level features with

semantic descriptions (the concepts). This step is usually done by means of supervised classification, which aims at finding regularities from a set of annotated examples. Many algorithms have been employed for concepts detection, such as Support Vector Machines (SVM), Hidden Markov Model (HMM) and K-Nearest Neighbors (KNN) [71]. However, even if the low-level features and the classifier are carefully chosen, the link between the input (low-level features) and the output (concepts) is still too weak for efficiently bridging the semantic gap [72].

Nevertheless, in the context of multimedia indexing, some solutions arise when considering multimodality and the use of several classification frameworks. This is usually known as the fusion process, it aims at enhancing concept detection by merging the various modalities (image, sound, and text), features (from those modalities) and/or classification evidences from other concepts. Classical approaches propose either to merge data at the feature level before the classification step or at the classifier level by combining decisions [73]. Respectively called "early" and "late" fusion [74], those approaches reach state of the art performance. As an advanced early fusion, a recent approach consists in taking advantage of kernel classifiers (eg. SVM). Early fusion is able to catch correlation between components of the various features while late fusion can take advantage of the correlation between several concepts. This last notion, called context fusion, is able to enhance the detection of concepts, which are semantically linked to others [75,76,77,78,79]. Kernel fusion aims at merging features at kernel level by learning a linear combination of unimodal kernels. This fusion scheme has been successively applied in multimedia semantic indexing [73,80,81] thanks to the flexibility provided by kernel designing and combination. Despite its good performance, this early fusion approach is not suitable for taking advantage of the relationship between the concepts.

2.2. S & T OBJECTIVES, PROGRESS BEYOND THE STATE OF THE ART

The present project aims to achieve breakthroughs in the following areas:

- Categorization of video sequences based on their semantic content. The semantic concepts considered within the project are a subset of the LSCOM covering broad categories such as objects (e.g. boat, aircraft, road), activities/event (e.g. people walking, explosion, airplane takeoff), scenes/location (mountain, building exterior, beach), people (e.g. Bill Clinton), graphics. Most recent works in the literature make use of a ``bag of features" kind of approach [82,83] which tries to adapt the "bag-of-words" representation used for text categorization to "Visual Object Categorization" (VOC) problem and to video semantic concept recognition. The "bag-of-features" kind of approach is an implementation of the distributional semantic approach which describes the semantic content of a video sequence according to some statistics of "visual words" from a visual vocabulary. While such an approach has shown its effectiveness, obtaining the best performance in Pascal VOC contest [84] and TRECVID contest [85], the average precision achieved so far is still far from real application requirements. In this project, we want to complement the 'bag-of-features" kind of

approach by a componential semantic like approach trying to describe a high level semantic concept by some related visual components based on for instance visual conceptual vectors.

- Innovations on low-level features aiming to characterize the audio-visual content of video sequences. Besides some popular low-level features, namely color moment, Gabor texture, Edge Direction Histogram, we also want to adapt some segment-based and interest point trajectory-based features already respectively developed by LIRIS at ECL LIRIS and Eurecom to better capture some geometric properties of the visual content and spatial-temporal properties of a video sequence. To characterize the audio content of video sequences, we also want to make use of Variable Resolution Transform (VRT) and some advanced audio features such as harmonic and Zipf features developed at ECL LIRIS in complement to popular MFCC like features.
- Categorization of video data based on their emotional content. Recent advances in the automatic emotion recognition from speech, music and text suggest a scenario in which the same kind of associative analysis can be applied to video sequences. While the existing techniques for speech emotion detection or music mood recognition can be generalized to the audio channel of a video sequence, keeping in mind however that the audio signal from the audio channel is quite different from the ones so far dealt with by the research community, we also want to develop new visual "emotional" descriptor. Image processing techniques will be applied at several levels of abstraction. In particular, at single still image level, some aesthetic features such as balance and luminance of global and local color or density and thickness of segments will be exploited. At the video sequence level, the speed or movement intensity and trajectories of feature points or objects will be analyzed to capture emotional properties such as calmness or violence. In terms of objects (forms recognized in the video sequence), face detection and identification techniques can also be used to characterize human expression in the image. Emotional analysis of the video data will result from an optimized fusion strategy, thus combining audio and visual modalities, which can be at feature-level fusion, decision-level fusion or a mix-level fusion such as model-level fusion (e.g. HMM). The emotional content can be described and labeled, at several granularity level (e.g. shot, scene, sequence), with one or more emotional states, according to a basic discrete emotional model, thus enabling distinctions between fundamental emotions including sadness, fear, anger, surprise, etc., , or by combinations of subtle emotions according to a dimensional emotion model, respectively according to arousal and appraisal axis.
- In terms of languages, we want to evaluate the use a pivot language, to deal with closed captions in several languages (English, French, and German) and to annotate these texts with some Interlingua concepts which are to be linked with the hierarchies of semantic concepts generated from analysis on visual, audio and emotional content. 3 different pivot languages will be evaluated.
- Cross-concept detection and automatic hierarchical organization of semantic concepts. Semantic concepts often appear paired or associated within video sequences. For instance, beach often appears paired with sea, driver with car. There exists thus a possibility to leverage the context between multiple concepts. We propose to explore

the pair-wise correlation between all the semantic concepts in the lexicon. Moreover, as the number of semantic concepts is rather important and we need to consider dynamic addition of new semantic concepts for application requirement, we also propose to investigate automatic generation of hierarchical organization of semantic concepts based on audio-visual content of video data and to link it with other hierarchies on associated content such as closed captions in the textual modality. This should improve the performance of the classifier and the addition of new semantic concept can be an adaptation of the existing models, thus achieving an efficient learning with less data.

- Multimodal Analysis for semantic concept recognition in order to leverage the correlations between different synchronized content channels: video, audio, emotion and closed captions. In particular, we will investigate the best fusion strategies for semantic concept recognition in a fusion strategy space ranging from early fusion to late fusion, including a combinatorial number of other intermediate fusion strategies.
- In order to improve the generalization performance of the concept detectors, an active learning approach will be used for dynamically building an appropriate training set for each concept using a large and very varied set of unlabeled videos. The principle of active learning is to use a system which is already existing or under development for selecting for being annotated the samples that are predicted to be the most informative for the system being trained. The most informative samples may be for instance those that are the most likely to be positive for the concept (in the case the concept is infrequent, which is often the case), those that are the most uncertain and/or those that are the most different from the already judged ones. This approach has proven to be very efficient [68] and it permits to obtain the best detection performance while only a small fraction of the training set is annotated. It is also very efficient for system retraining when new types of video contents are encountered.
- The technologies developed in this project will be evaluated against a state-of-the-art benchmark used in the TRECVID campaign. Several partners are active participants in this campaign. They will also be evaluated in a real industrial application provided by Ghanni, so as to tune their parameters and validate their impact on the regular operation of Ghanni's services.
- The technologies developed in this project will be implemented in the Ghanni production environment, this will allow the industrial partner of the project to benefit from the latest technologies and provide innovative services to their customers.

All the concept classifiers integrating the above innovations will be integrated into Ghanni's platform for media asset management. Moreover, they will be experimented and validated, in two ways: through the targeted two applications provided by Ghanni, namely video recommendation and ads monetization, but also through the active participation of the project partners to TRECVID contest. The current Ghanni corpus contains about 50,000 video whose average length is of about 120 seconds. About 7% of the video are longer than 400s. The developed indexing techniques should be scalable enough to deal with video collection of this size and higher. This implies developing descriptors that are of small size and fast to compute. Classification and fusion techniques should also be fast enough, at least for the prediction part.

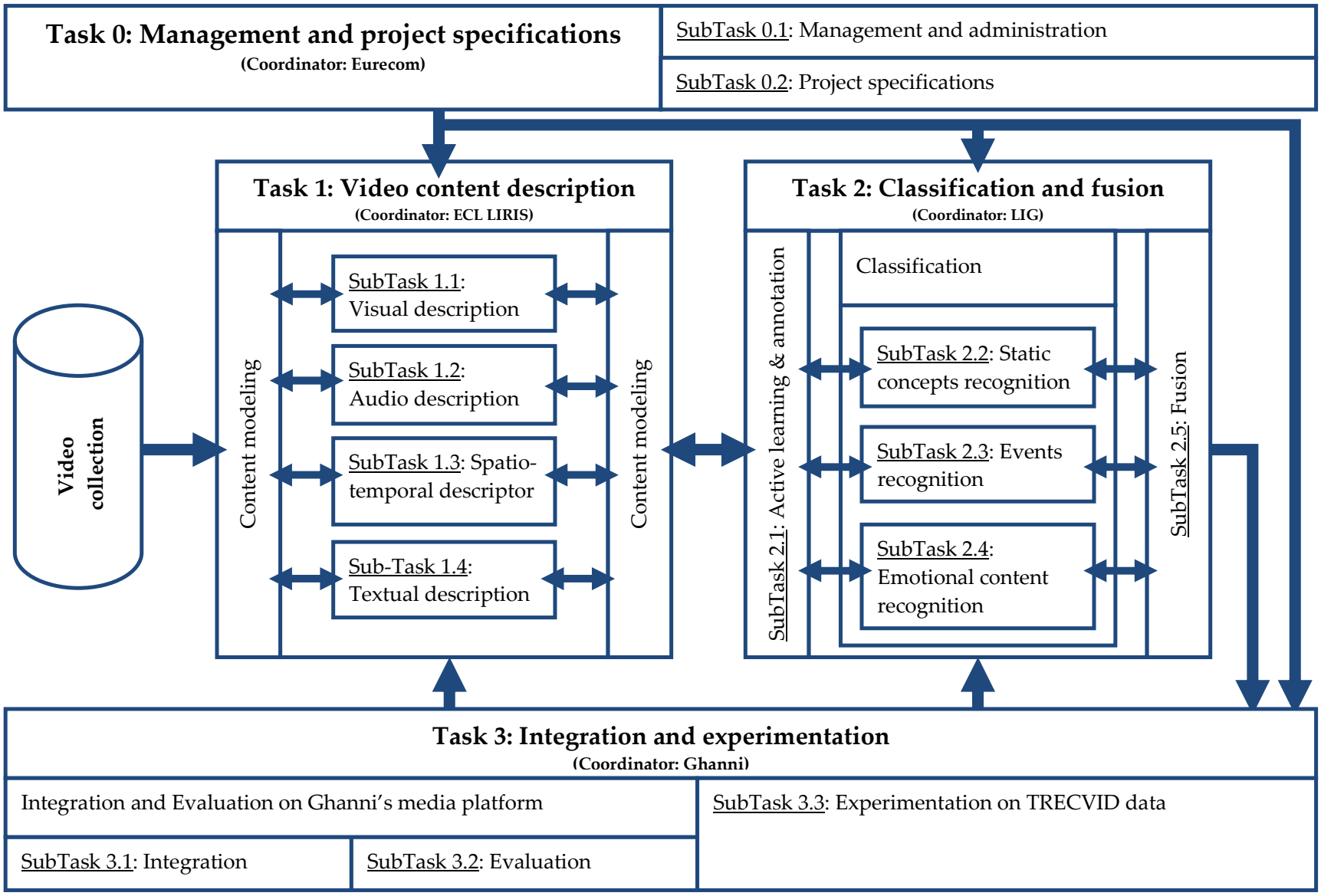
3. SCIENTIFIC AND TECHNICAL OBJECTIVES / PROJECT DESCRIPTION

3.1. SCIENTIFIC PROGRAMME, PROJECT STRUCTURE

Recall that the VideoSense project aims at automatic video tagging by high level concepts, including static concepts (e.g. object, scene, people, etc.), events, and emotions, while targeting two applications, namely video recommendation and ads monetization, on the Ghanni's media assess management platform. The innovations targeted by the project include video content description by low-level features, emotional video content recognition, cross-concept detection and, multimodal fusion, and the use of a pivot language for dealing with multilingual textual resources associated with video data.
The figure below illustrates the overall architecture of the VideoSense project which is split into four tasks:

- The first task (Task 0), led by Eurecom, is dedicated to project management and specification. While the compulsory task of management aims at the project monitoring throughout its entire life, the project specification will define a set of high level concepts as regard to the two applications targeted by the project.
- From raw video data, the second task (Task 1), led by ECL LIRIS, focuses on the video content description through its different modalities: visual description, audio description, spatiotemporal description and textual description through a pivot language. The innovations on low level features targeted by the project will be made here.
- The third task (Task 2), led by LIG, takes as input the video content description from Task 1 and deals with classification and fusion for concept categorization. It covers a task for active learning and annotation in order to create a significant training and test data sets, three tasks for the recognition of static concepts, events and emotions. All these recognition tasks will leverage the results from the task for active learning and annotation and make use of some optimized fusion strategies in order to best take advantage of the composite modalities, such as visual modality, audio one and possibly the textual one as well. This task will be the place for innovations on emotional video content recognition, cross-concept detection and automatic hierarchical organization of semantic concepts, and multimodal analysis for semantic concept.
- The last task (Task 3), led by Ghanni, aims at integration of the concept classifiers developed within the Task 2 into the current Ghanni's media assess management platform, "youCircle" and experiment the two applications targeted by the VideoSense project. Moreover, all the academic partners will test their concept classifiers within the TRECVID framework for the purpose of dissemination and worldwide comparison.

As we can see from the figure below, all these tasks have their own finalities while keeping strong interaction between each other.


Task 0: Management and project specifications
(Coordinator: Eurecom)
SubTask 0.1: Management and administration
SubTask 0.2: Project specifications
Task 1: Video content description
(Coordinator: ECL LIRIS)
Content modeling
SubTask 1.1: Visual description
SubTask 1.2: Audio description
SubTask 1.3: Spatio-temporal descriptor
Sub-Task 1.4: Textual description
Content modeling
Video collection
Task 2: Classification and fusion
(Coordinator: LIG)
SubTask 2.1: Active learning & annotation
Classification
SubTask 2.2: Static concepts recognition
SubTask 2.3: Events recognition
SubTask 2.4: Emotional content recognition
SubTask 2.5: Fusion
Task 3: Integration and experimentation
(Coordinator: Ghanni)
Integration and Evaluation on Ghanni's media platform
SubTask 3.1: Integration
SubTask 3.2: Evaluation
SubTask 3.3: Experimentation on TRECVID data

3.2. PROJECT MANAGEMENT

3.2.1 STRUCTURE AND CO-ORDINATION,

A specific task (Task 0) is dedicated to the coordination and management of the project. It includes the management procedures and instances.

The project is lead by Eurecom, the Project Coordinator. The Project Coordinator is responsible of:

- The overall supervision of the project: monitoring the overall progress, the deliverables and milestones, the interrelations between tasks, etc;
- The communication between the project and ANR (representation in audits, meetings, annual reports);
- The internal communication between the Partners;
- The representation of the Project in external Organizations.

The Project Coordinator organizes reviews and progress meetings. He deals with the administrative issues with the responsible contact person nominated by each partner.

To coordinate the project, a Project Management Board (PMB) will be established. Each partner will designate one representative person to participate in the PMB. The PMB will be chaired by the Project Coordinator. The PMB will monitor all issues related to the internal progress of the project (project work plan, objectives, manpower allocation, changes) as well as external policy issues (status of deliverables, external communication).

The PMB will also carry out the technical management and ensures through Task leaders that its directives are implemented in the tasks. Each Task leader is responsible for the work to be done in the Task and the preparation of the corresponding deliverables.

The PMB will also be responsible for resolving conflicts when they appear, based on the technical inputs from Task leaders. The PMB will always try to achieve consensus. In case of disagreement, a seventy five percent (3/4) of the votes of the PMB representatives present or represented by proxies is required to pass a resolution. The procedure will be detailed in the Consortium Agreement to be signed at the beginning of the project.

3.2.2 COMMUNICATION MANAGEMENT

The Project Coordinator will setup several mailing lists to allow easy communication between the partners, one for the PMB, one for all the participants in the project, and other if needed by specific actions during the project. A web site with collaborative tools (shared space, wiki, etc…) will be setup to exchange information and data among the partners. A central repository will be made available for common data. A directory will list all the

persons involved in the project and their respective role, so that each partner may have a complete view of the project activity, and is able to direct questions to the right person.

The Project Coordinator will organize the project general meetings, approximately every three to four months. The location of the meeting will alternate between the partners' facilities. For each meeting, an agenda will be prepared, and minutes will be produced and distributed to the project partners.

When needed, audio conferences will be organized to discuss the progress of the project, or specific questions that may arise. If needed, specific meetings will be organized among the relevant participants.

The Project Coordinator is responsible for forwarding deliverables to the ANR, when necessary.

3.2.3 DELIVERY MANAGEMENT

- *Deliverables*

The deliverables and milestone documents are produced by the Tasks. They are then made available to the PMB and to the other Tasks, as well as additional information related to their preparation. The deliverables and milestone documents are then revised by other Task and possibly by experts specially appointed by the PMB, in order to ensure adherence to the initial declared goals. Finally, the deliverables are formally approved by the PMB, and then sent to the ANR. The deliveries to the ANR will be made on electronic form (ex: e-mail or publishing on the Web) or in hard electronic support (ex: CD-ROM).

- *Planning and Reporting*

The planning of the task internal and external deliveries will be made by the Task leader in coordination with other Task leaders, and the Project Coordinator so as to meet the contractual commitments of the project. The project operational planning is derived from the contractual planning by the PMB taking into account the Tasks operational planning.

Reporting on the progress of the project will be done during the PMB meetings (progress report vs. resources spent, issues and problems …). The PMB will assess these reports and will discuss the technical issues highlighting possible conflicts and proposing solutions. PMB meeting minutes will be distributed to all experts involved in the project and published on the project Web Server (private part). Administrative reports will be circulated among the partner contact persons responsible for administrative matters.

3.2.4 MANAGEMENT OF OTHER ISSUES

- *Meeting Legal and Ethical Obligations*

The project management will guarantee that legal obligations are duly met and will manage the project so that all ethical obligations are met.

3.3. DESCRIPTION OF THE TASKS

3.3.1 TASK 0. MANAGEMENT

• *Task 0*	Management and Project Specifications
Coordinator:	EURECOM
• ***Partners:***	• ECL LIRIS • LIG • LIF • Ghanni
Starting date: T0	Duration: 36
Task description and objectives: This Task will manage the progress of the project, through a reporting mechanism, will ensure the timely preparation of the deliverables, and will organize the reviews of the project.. It will act accordingly in case of unexpected events during the project. It will also organize the collaborative definition of the project specifications.	

• ***Task T0.1***	Project Management
Partners:	**EURECOM**, ECL LIRIS, LIG, LIF, Ghanni
Duration :	T0 to T0+36
SubTask description: This task will organize the management of the project. Regular meetings among the participants will be organized, alternatively at each participant location, with a frequency of approximately three or four per year. An agenda will be prepared for each meeting, and minutes will be taken. The project manager will maintain a list of action points that will be updated at each meeting. At each semester, a progress report will be elaborated by all the participants. Annual reviews will be organized in the format requested by the ANR.	

• ***Task T0.2***	Concepts definition and specifications
Partners:	**Ghanni,** EURECOM, ECL LIRIS, LIG, LIF

Duration :	T0 to T0+6 and T0+21 to T0+24
SubTask description: The first task to be done before developing and applying supervised and semi-supervised classifiers is the definition of classes. The aim of this subtask is to identify a set of classes or concepts that are of use in the context of the project's real world applications, ads targeting and video recommendation. We first propose to study existing market segmentation and consumer behaviour research work. In parallel we will interview different kind of advertisers to analyse their vision of ads targeting. Data collected from existing research work and interviews will be then analysed and a set of concepts will be generated. Furthermore, this set of concepts will be discussed with the different research teams partners of the project in order to define the final set of concepts that will be used. Once significant technical advancements are made, we propose to reassess the adaptability of the originally specified concepts. Hence we can update the specified concepts to take into account the technical progress and the latest market needs.	

Task delivrables	**Date**	*Deliverable title*	**Nature**
D0.1.n	T0	Progress report (every 6 month)	Report
D0.2.1	T0+6	Market-Oriented concepts	Report
D0.2.2	T0+24	Concepts specifications Update	Report

3.3.2 TASK 1. VIDEO CONTENT DESCRIPTION:

• *Task 1*	Video Content Description	
Coordinator:	ECL LIRIS	
• ***Partners:***	• LIG • EURECOM • LIF • Ghanni	
Starting date: T0		Duration: 18
Task description and objectives: A video document is a multimodal data composed of visual channel, auditory channel and possibly textual resources, such as closed captions, describing the content of the video document. Video content description aims at representing the semantic content of a video data, thus trying to capture consistent properties respectively from the visual content, the auditory content, the temporal properties and possibly the textual content of a video data.		

• Task T1.1	Visual description
Partners:	**ECL LIRIS**, LIG, EURECOM, LIF
Duration :	T0+0 to T0+18

Subtask description:

The visual content of an image can be described by some well-known low level features at various granularities such as color moment, Gabor texture and edge direction histogram or shape information. However, there can be large variations in lighting and viewing conditions for images representing real world scenes, making difficult stable description of visual content. The visual content of an image can also be modeled by some midlevel descriptors such as line-based descriptors or midlevel semantic concept scores. In the later case, the midlevel concepts need first to be carefully selected and the associated classifiers trained on a large dataset.

Within the VideoSense project, we will adopt a mix representation of visual content of video frames, thus making combined use of some popular low level features (e.g. SIFT, color SIFT, edge direction histogram) with some mid-level concept descriptions as latent semantic directions (e.g. segment, indoor/outdoor [86], city/landscape [87], etc.). Moreover, we also would like to deepen our work of visual content description based on some basic human visual perception laws and consider Gestalt inspired region segmentation scheme for the derivation of visual features on segmented partial Gestalts [88].

The visual description of video frames will also be considered from the aesthetic perspective for perceptual feelings analysis. In particular, at single video frame level, some aesthetic features such as balance and luminance of global and local color or density and thickness of segments will be exploited.

• Task T1.2	Audio description
Partners:	**ECL LIRIS**, Ghanni
Duration :	T0 to T0+18

Subtask description:

The audio channel conveys rich semantic clues for content-based video analysis. Besides widely known speech extraction and speaker identification problems, interesting audio semantic analysis also includes speech/music segmentation, speaker gender detection, special effect recognition such as gun shots or car pursuit, etc. All these problems can be considered as an audio semantic classification problem which needs to generate a semantic

label from low audio signal analysis.
In addition to some popular features (e.g. MFCC, pitch, energy, Zero Crossing Rate, etc.), we also propose to deepen and extend our existing works, including emotion sensitive features, e.g. harmonic and Zipf features, Harmonic features, Variable Resolution Transform (VRT) and our human perception motivated model PGM, for better description of the audio channel from a video data.

• ***Task T1.3***	Spatiotemporal description
Partners:	**EURECOM**, LIG, ECL LIRIS
Duration :	T0 to T0+18

Subtask description:

This task will extract visual descriptors on consecutive frames of the video sequences. These descriptors will be both global (such as colour histograms on various grid regions, texture analysis using DCT, Gabor and other measures, regions segmentation, motion vector fields, …) and local (such as SIFT or Harris key points, associated to their local descriptors). Some of these elements will be compared on consecutive frames to compose trajectories. The dynamical characteristics of these elements will serve as a description of the content of the video sequences.

• ***Task T1.4***	Textual description
Partners:	**LIG**, EURECOM, ECL LIRIS
Duration :	T0 to T0+18

Subtask description:

This task will extract textual descriptors from texts present in video metadata and captions. Considered texts are available in French, English or German. These textual descriptors will be based on a pivot language; hence textual descriptors will be handled by categorizers regardless of the source language of the text they come from. Several pivot language will be considered: 1) a pivot language consisting in a reduced set of basic concepts as defined in existing thesauri (~1000 index terms); 2) an ad hoc pivot language tailored for the task (~3000-4000 index terms); 3) a general pivot language (WordNet or UNL, depending on available lexical data). Conceptual vectors will be used in all experiment to extract textual descriptors.

Task deliverables	**Date**	*Deliverable title*	**Nature**
D1.1	T0+18	set of visual descriptors	Report
D1.2	T0+18	set of audio descriptors	Report
D1.3	T0+18	Set of spatiotemporal descriptors	Report
D1.4	T0+18	Set of textual descriptors	Report

3.3.3 TASK 2 CLASSIFICATION AND FUSION

• *Task 2*	Classification and Fusion
Coordinator:	LIG
• ***Partners:***	• LIF • EURECOM • ECL LIRIS • Ghanni
Starting date: T0+6	Duration: 30

Task description and objectives:

Video is a multimodal document, which embeds semantic concepts such as objects, scenes, events and emotions. Classification and Fusion aims at extracting the semantic content of video shots by training classifiers from sets of descriptors and merging data in order to take benefit of multimodality. This task also aims at building the required annotations to train and test the classifiers.

This task is divided into five subtasks. Subtask 2.1 concerns the building of annotated corpora for the training of the concept detectors developed in the other subtasks. Subtasks 2.2, 2.3 and 2.4 concerns the development of concept classifiers. Subtask 2.5 concerns the development of fusion methods for improved concept recognition. Subtasks 2.2, 2.3 and 2.4 focuses respectively on static concept classification, on video event recognition and on video emotion recognition. The corresponding concept detectors will use the descriptors developed in task 1. The final product of task 2 is a set of concept detectors. The actual set of concepts to be detected will be defined in subtask 0.2.

• ***Task T2.1***	Active Learning and annotation
Partners:	**LIF**, LIG, Ghanni
Duration :	T0+6 to T0+18

Subtask description:

This task will take advantage of an active learning process to reduce the effort of annotation. We will extend the system described in [67] and manage the annotation process. We propose to study new sample selection functions, as well as the possibility of taking advantage of relationships between concepts. As the most obvious, we will study the genericity/specificity relationships and integrate such knowledge in the selection function of active learning process. Furthermore, in order to annotate aesthetics, statics and dynamics concepts as targeted in VideoSense, we will have to propose new modes of visualization.

Ghanni will be responsible for the annotating effort. Based on our expertise, we estimate the

time required for the annotation of one subshot and one single concept as 3 to 5 key frames per 10 seconds, depending of the kind of concept. As we plan to focus on about 50 concepts in VideoSense, we therefore estimate about 2 x 6 months of full work to annotate around 2000000 subshots, corresponding to 50 concepts x 40000 subshots. However, emotional patterns embedded within a video data are subjective and typically vary in time. The collection and manual labeling will be specifically dealt within task 2.4.

• ***Task T2.2***	Static-concept Classification
Partners:	**LIG**, LIF
Duration :	T0+12 to T0+36
Subtask description: The main objective of this subtask is the design concept detectors with good generalization capabilities. All types of descriptors developed in subtasks 1.1 (visual), 1.2 (audio) and 1.4 (visual) will be used. The work will focus on selecting the best combination of descriptors and the best set of parameter for each descriptor for each concept to be detected. The best combination concerns both the selection of descriptors and the way they are combined. The quality of the combination will be evaluated both on the absolute detection performance and on the generalization capability (the ability of the detector to detect the concept in data from different sources and types than those used for training the system). The last point should be achieved by using the annotations obtained in subtask 2.1 that will be as varied as possible thanks to an appropriate active learning strategy and by looking for a tuning of classifiers' parameters that maximize their generalization capabilities. Specific studies will be done on this last point. Finally, the textual descriptors based on a pivot language will allow working with documents in different languages both for system training and for classification.	

• ***Task T2.3***	Video Event Recognition
Partners:	**EURECOM**, LIG, ECL LIRIS
Duration :	T0+12 to T0+36
Subtask description: In this subtask, we will use the multimedia descriptors that have been extracted in Task 1 and the annotation that will be provided by Task 2.1 to construct detectors for the video events that have been defined in the specifications (Task 0.2). In order to better capture the dynamical aspects of the events, we will explore two complementary approaches: • The classification of dynamical features, for example the aspect of the motion vector fields, or the shape of a key point trajectory, • The use of dynamical models on static features, for example HMMs can be used to model the evolution of certain parameters like camera movement or object position.	

The result of this task will be a set of events, including their category, their time limits. These results will be evaluated by comparing the results with a manually annotated ground truth. An analysis of the results will be performed.

• ***Task T2.4***	Video emotion recognition
Partners:	**ECL LIRIS**, LIG, EURECOM
Duration :	T0+12 to T0+36

Subtask description:

In this subtask, we are interested by the "emotional content" of a video data as it can be perceived by users and its automatic recognition is of the deepest interest, especially within the VideoSense project as it directly concerns the potential impact of a video document. The aim here is to take into account the perceptual feelings of the end-users while experiencing video data, namely emotional effect provoked by viewing video document, such as excitement, calmness, enchantment, stress, anxiety, violence, etc.

On the basis of our previous work and a state of the art on the topic, the following issues will be investigated:

- The description of emotional content conveyed by video data. The emotional content of a video data can be described and labeled, at several granularity levels (e.g. shot, scene or sequence), by one or more discrete emotional states, according to a discrete emotion model, thus enabling distinctions between fundamental emotions including sadness, fear, anger, surprise, etc., or combination of subtle emotions according to a dimensional emotion model. This work will be made in connection with Task 0.2.
- Collection, by active learning and annotation in connection with Task 2.1, of a representative dataset from the emotion perspective for learning and testing.
- Study of emotion sensitive features both in audio and visual channels, in connection with Tasks 1.1, 1.2, 1.3, for the purpose of the further classification and fusion step.
- Emotional analysis of the video data from an optimized fusion strategy, in connection with Task 2.3, thus combining audio and visual modalities, which can be at feature-level fusion, decision-level fusion or a mix-level fusion such as model-level fusion (e.g. HMM).

Moreover, the video emotion recognizer will be trained and tested on significant manually labeled emotional video resources. However, as emotion patterns typically vary in time within a video sequence and their perception is subjective and hence possibly multiple by different people , the collection of such a significant and representative corpus typically is human resource consuming task and such manual annotation will also require a specific interface. This task of the emotional video collection and its manual annotation will be subcontracted.

• *Task T2.5*	Fusion
Partners:	**LIF**, LIG, EURECOM, ECL
Duration :	T0+12 to T0+36

Subtask description:

This task aims at exploiting knowledge coming from the available modalities to enhance the detection of the various targeted concepts (static, dynamic, emotional).
In the multimedia setting, most fusion schemes are empirically designed and evaluated, while only few studies endeavour a theoretical study of machine learning-based fusion algorithms. Indeed, machine learning algorithms and theoretical results would be of benefits to study the properties of various fusion schemes that include learning steps. Our proposal is, first, to design a theoretical framework for further promoting guidelines for fusion experiments of real multimedia data, namely video, whether the fusion is early, late, or incremental. Second, based on previous theoretical issues, we promote the implementation and evaluation of the following approaches:

- **Kernel fusion** offers the benefits to adapt the kernel function according to the modalities. We want to consider such fusion framework to evaluate performance and complementarities of specific kernel functions for concept detection in video documents.
- **"Bag of features"** is a recent and efficient approach for image and video classification, which performs classification based on a visual vocabulary. In this task, we will investigate deeply the use of bag of features for multimedia document indexing. We will construct vocabularies from the available low-level features extracted at several granularity levels. We will study how to combine vocabularies extracted from visual, sound and textual modalities.
- We plan to study the feasibility to infer some concepts from others, by exploiting relationship among them. This **context fusion** involves being able to measure the relationship between concepts. When annotations are available, we will study the effect of computing correlation based on low-level features, or from the amounts of annotations shared between concepts. While the former should contain more information, it may be too dependent of the employed feature extractor. From such correlation, we will consider automatic structuring of the concepts to build a specific ontology of our targeted concepts. When annotations are not available for a given concept, we plan to exploit knowledge from external database, such as the WordNet ontology.
- The **co-training algorithm** [89] takes advantage of the specificities of several views on data, within a cooperative framework for learning concepts despite few labelled examples. Considered as a semi-supervised algorithm, co-training benefits from a solid theoretical ground that ensures improved performances under strong assumptions. We plan to investigate both theoretically and experimentally the co-training algorithm for iteratively merging modalities and classifiers decision in the context of video classification.

Task deliverables	Date	*Deliverable title*	Nature
D2.1	T0+18	Active learning and annotations	Report
D2.2a	T0+24	Static concepts classification	Report
D2.2b	T0+36	Static concepts classification	Final Report
D2.3a	T0+24	Video events detection	Report
D2.3b	T0+36	Video events detection	Final Report
D2.4a	T0+24	Video emotion recognition	Report
D2.4b	T0+36	Video emotion recognition	Final Report
D2.5a	T0+24	Fusion	Report
D2.5b	T0+36	Fusion	Final Report

3.3.4 TASK 3. INTEGRATION AND EVALUATION

• *Task 3*	Integration and evaluation	
Coordinator:	Ghanni	
• ***Partners:***	• ECL LIRIS • LIG • EURECOM • LIF	
Starting date: T0+18		Duration: 12
Task description and objectives: The developed techniques need to be evaluated on different kinds of data. This Task aims at integrating the developed modules into the Ghanni's youCircle media management platform. After the module integration, experimentations will be carried out to evaluate the developed techniques on videos from TRECVID and from the youCircle platform.		

• ***Task T3.1***	Integration
Partners:	**Ghanni**,
Duration :	T0+18 to T0+36
SubTask description: The aim of this subtask is to integrate the developed modules into the Ghanni's Media Management platform, youCircle. The integration permits automatic video tags extraction for all of the platforms' videos. The integration will be made in two steps: - A first step where an integration of the basic technical bricks,	

- A second step where the final versions of technical modules will be integrated. A basic video matching system will be developed to enable video to video and ad to video similarity estimation based on simple tags matching. This system is crucial in order to be able to assess the effectiveness of the developed techniques.

• ***Task T3.2***	Subjective and real world usage evaluation
Partners:	**Ghanni**,
Duration :	T0+24 to T0+36
SubTask description: The aim of this subtask is to experiment the developed modules into the Ghanni's Media Management platform, youCircle. We propose to evaluate the different modules in three experiments. The first one is a dedicated experiment which is supervised and voluntary for a selection of youCircle's users. The second one is comparative and based on usage statistics for the entire user base. The third one is an evaluation of the classification accuracy on manually labelled data. In the first experimentation, a selection of current youCircle's users will be asked to participate in a research study concerning video recommendation and ads targeting. A special User Interface will be developed to be used by the participants. The User Interface will let the user search for a video of interest from youCircle and show recommendations of other videos using automatic tags matching between the current video and youCircle's videos. The user will be asked to evaluate the recommended videos. Furthermore, each time the user is selecting a video of interest, the most similar ad and the less similar ad will be shown to the user. The user will be asked to evaluate both ads according to his/her interests. In the second experimentation, we propose to introduce, in the same recommendations for users, videos based on the actually existing Ghanni's recommendation and ones based on the modules developed in the project. The same will be done for ads targeting. Usage statistics, such as click rates on recommended videos and ads, will be analyzed to constitute a measure of performance. Therefore, the system can be evaluated by comparing the performance with or without the use of the automatic tags extraction modules. Furthermore, the performance measurement enables us to fine-tune the system's parameters. In the third experimentation, we propose to evaluate the developed techniques on a set of videos that are manually tagged.	

• ***Task T3.3***	Evaluation on TRECVID data
Partners:	**LIG**, LIF, ECL LIRIS, Eurecom

Duration :	T0+18 to T0+36
SubTask description: The methods used for the design of the detectors will also be tested on TRECVID data in the context of the High Level Feature (HLF) detection task. We will initially use the TRECVID 2007-2009 data and concepts will be considered. This includes a total of about 360 hours of MPEG-1 video. Three sets of 20 concepts (partly overlapping) are also defined. They are included or closely related to the LSCOM ontology. The developed detectors will be evaluated according to the TRECVID methodology and metrics and compared to other state of the art approaches. We will follow the evolution of the TRECVID campaign to include data from subsequent years in our experiments.	

Task deliverables	**Date**	*Deliverable title*	**Nature**
D3.1	T0+36	Modules integration	Report, Prototype
D3.2	T0+36	Subjective and real world usage evaluation	Report
D3.3	T0+36	Evaluation on TRECVID data	Report

3.4. TASKS SCHEDULE, DELIVERABLES AND MILESTONES

3.4.1 TASKS SCHEDULE

Task	Description Months	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
0	**Management**																																				
0.1	Project Management																																				
0.2	Concepts definition and specifications																																				
1	**Specification of Integrated System**																																				
1.1	Visual Description																																				
1.2	Audio Description																																				
1.3	Spatiotemporal Description																																				
1.4	Textual Description																																				
2	**Classification and Fusion**																																				
2.1	Active Learning and annotation																																				
2.2	Static concepts recognition																																				
2.3	Video Event Recognition																																				
2.4	Video emotion recognition																																				
2.5	Fusion																																				
3	**Integration and experimentation**																																				
3.1	Integration																																				
3.2	Subjective and real world usage evaluation																																				
3.3	Evaluation on TRECVID data																																				

3.4.2 LIST OF DELIVERABLES

Task delivrables	Date	*Deliverable title*	Nature
D0.1.n	T0	Progress report (every 6 month)	Report
D0.2.1	T0+6	Market-Oriented concepts	Report
D0.2.2	T0+24	Concepts specifications Update	Report
D1.1	T0+18	set of visual descriptors	Report
D1.2	T0+18	set of audio descriptors	Report
D1.3	T0+18	Set of spatiotemporal descriptors	Report
D1.4	T0+18	Set of textual descriptors	Report
D2.1	T0+18	Active learning and annotations	Report
D2.2a	T0+24	Static concepts detection	Report
D2.2b	T0+36	Static concepts detection	Final Report
D2.3a	T0+24	Video events detection	Report
D2.3b	T0+36	Video events detection	Final Report
D2.4a	T0+24	Video emotion recognition	Report
D2.4b	T0+36	Video emotion recognition	Final Report
D2.5a	T0+24	Fusion	Report
D2.5b	T0+36	Fusion	Final Report
D3.1	T0+36	Modules integration	Report, Prototype
D3.2	T0+36	Subjective and real world usage evaluation	Report
D3.3	T0+36	Evaluation on TRECVID data	Report

4. DISSEMINATION AND EXPLOITATION OF RESULTS. MANAGEMENT OF INTELLECTUAL PROPERTY

From a scientific perspective, the academic partners are active in the scientific community and plan to disseminate the results of the project through the usual academic channels. We will submit the research work accomplished in the projects in international workshops, conferences, and journals. Furthermore, we plan to participate to national and/or international evaluation campaigns on concept detection in video documents, including TRECVID.

Ghanni will experiment the developed tools in its media platform for two applications: Ads targeting and video recommendation. In fact, currently ads targeting and video recommendation are both based on the manual indexing of video content. If the developed techniques can be proved efficient, Ghanni will integrate them in its platform and make them available to its clients.
The use case is as follows:

- The video content available to the users via the Ghanni's platform is analyzed and automatically tagged
- Advertisers will select one or more categories for their target when uploading their ads
- When a specific video is watched by a user, the more convenient ads based on the video tags and other criteria will be shown.
- The user can ask for similar videos that are generated based on the video tags

Furthermore, Ghanni is developing an open Video Ads service, SemanticAd, which can be used by any publisher having video content available online and wishing to monetize it. Ghanni will integrate the developed tools for automatic video tagging to enable ads targeting. This will clearly bring a key competitive advantage in comparison to competitors such as LiveRail, Kyte, Brightroll. By using automatic ads targeting system Ghanni can motivate publishers and producers with higher prices ads while helping advertisers to save money and target potentially interested persons.
The use case is as follows:

- The publisher accesses a Webservice provided by Ghanni that analyzes and tags its video content
- The publisher gets a specific code to integrate in its media player's code
- Advertisers will select one or more categories for their target when uploading their ads
- When a specific video is watched by a user on a publisher's site, the more convenient ad based on the video tags and other criteria will be shown.
- The user can ask for similar videos from the publisher's site that are generated based on the video tags

A consortium agreement will be signed at the beginning of the project to handle the relationship between the various partners about the intellectual property of the innovations created during the project. The consortium agreement will define how the responsibility of each partner to facilitate the dissemination and impact of the project outcomes in practical applications.

5. CONSORTIUM DESCRIPTION

5.1. PARTNERS DESCRIPTION & RELEVANCE, COMPLEMENTARITY

5.1.1 EURECOM

EURECOM is a teaching and research center supported by the Institut Télécom. Its activities are in the domains of Mobile communications, Network and Security, and Multimedia communications. The Multimedia department is focused on Multimedia indexing, audio and video analysis, biometry, watermarking. EURECOM has a strong industrial partnership, and collaborate in numerous collaborative projects, at the European as well as at the national level. It also has a strong implication in the scientific animation: for example, we are part of the steering committee of the CBMI workshop series, and we organized the 2009 edition of the MultiMedia Modeling international conference (MMM 2009).

The EURECOM team has a strong experience in multimedia indexing. Eurecom has participated to the TRECVid High Level Concept Detection task since 2003. Within this task we have developed original approaches for video indexing based on feature extraction, classification, and automatic learning. We have also extensively studied the use of fusion mechanisms at all levels, including the conceptual level. EURECOM will be the coordinator of the project, and will be mainly involved on the video analysis, especially for the spatiotemporal event detection and identification. The main persons participating in the project will be Bernard Merialdo (Professor) and Benoit Huet (Lecturer).

5.1.2 ECL LIRIS

With more than 80 faculties and 120 PhD students, LIRIS is large CNRS supported laboratory in information systems and image federating research forces from the University of Claude Bernard, the University of Lumière, INSA de Lyon and Ecole Centrale de Lyon. The imagine team of LIRIS at ECL LIRIS has been working on multimedia analysis and recognition since 1995 and developed strong expertise in the field. More than 15 PhD theses have been defended, including video segmentation & structuring, semantic image classification, face detection and recognition, audio event classification, emotional speech and music mood recognition since then. Imagine team at ECL LIRIS has solid experience in national and European project. It has taken part not only to nationwide research projects, including RNRT Cyrano for personalized video distribution, RNTL Muse for multimedia Search engine, ANR Omnia for categorization and filtering of still images, ANR MusicDiscover for music retrieval, but also European research project Phoenix for multimedia services on mobile.
Within the VideoSense project, the Imagine team at ECL LIRIS will mainly be involved on video content modeling, audio event recognition and emotional video recognition thanks to their previous expertise. The key persons involved within the current project are Prof. Liming Chen, Dr. Charles-Edmond Bichot and Dr. Emmanuel Dellandréa (cf. their CVs in the annex). While Prof. L. Chen and Dr. E. Dellandréa have been working on multimedia analysis and recognition for several years, Dr. Charles-Edmond Bichot will bring to us his confirmed expertise in graph processing techniques for the purpose of video object and event recognition as it has become a major trend in the field and it was featured out by several papers in the PAMI's November 2008 special section on real-world image annotation and retrieval.

5.1.3 LIG-CNRS

The Laboratory of Informatics of Grenoble (LIG) is a research unit whose funding partners are CNRS, INRIA, Grenoble INP, UJF and UPMF. It was created on January 1, 2007. It gathers 500 researchers, lecturers-researchers, students and postdocs, technical and administrative staff members. Research activities are structured around 24 autonomous research groups. Two research groups from LIG will be involved in the VideoSense project: MRIM and GETALP.

The Multimedia information indexing and retrieval (MRIM) group is specialized (as it is shown in its name) with multimedia. This expertise will be crucial for the project. The group actually investigates the following axes:
- Logical models for information retrieval,
- Multimedia document indexing: Still images, Structured documents, Videos,
- Textual and multilingual information retrieval,
- Filtering and collaborative information retrieval,

The MRIM group also organized the 2003 edition of the European Summer School on Information Retrieval and the 2005 French Conference on Information Retrieval (CORIA). Within the VideoSense project, the MRIM group will be involved in video shot indexing, fusion for concept classification and active learning for annotation production.

The GETALP group is organized around five main research topics:
- Machine Translation (MT) and Computer-Aided Translation (CAT)
- Automatic Recognition of Speech, Speakers and Sounds
- Lexical resources and Corpora (software and content)
- Dialogue, communication and emotions
- Specialized Programming Languages and Environments for NLP

The GETALP group is a founder member of the U++ consortium, whose aim is to develop applications and promote the use of the UNL language in multilingual applications.
Within the VideoSense project, the GETALP group expertise involves language resource development, management and usage (acquired through several European projects since 1990) and multilingual language processing (acquired through a unique activity in machine translation of speech and language since 1964). The GETALP group also participated in several Cross Language Information Retrieval experiments in collaboration with the MRIM group (through several participations to CLEF).
The key persons are Dr. Georges Quénot for the MRIM group and Dr. Giles Sérasset for the GETALP Group.

5.1.4 GHANNI

Ghanni is a start-up company specialized in media management and recommendation solutions. Ghanni has developed a unique technology for media recommendation. This technology is currently used in the Ghanni's Media platform, youCircle. When applied to music, it was evaluated as in the Top3 music recommender systems with Pandora and lastfm (Alexandra E. Fox. *Battle of the Music Recommender Systems: User-Centered Evaluation of Collaborative Filtering, Content-Based Analysis and Hybrid Systems.* A Master's paper for the M.S. in I.S. degree. November, University of North Carolina and Chapel Hill, 2007).

Ghanni is a spin-off of the Ecole Centrale de Lyon, incubated by the CREALYS Rhône-Alpes incubator, labelled "Young Innovative Company", Ghanni was also recipient of several awards, including « le prix stratégique de création d'entreprise de la fondation Rhône-Alpes Futur 2005 », « La griffe Lyonnaise 2006 » et « le prix de l'innovation Total-CPE 2006 ».

Companies such as Galleries Lafayette, Thomson, Radio RTL Luxembourg, GrandLink Media use Ghanni's technology.

Ghanni will bring to the project its knowledge of the digital audio/video market needs. Additionally, Ghanni will provide real-world experimentations and evaluations of the developed techniques as regard to the two targeted applications on its media platform that manages more than 50,000 videos. The key persons are Dr. Hadi Harb (CEO) and Dr. Aliaksandr Paradzinets (CTO).

5.1.5 LIF

LIF is the laboratory of fundamental computer science (UMR 6166), supported by CNRS, University of Méditerranée and University of Provence. The LIF contains about 70 researchers and 25 PhD students working within 5 teams, which mainly focus on fundamental aspect of computer science. The team Databases and Automatic Learning (BDAA) will be involved in the VideoSense project. This team is working on theoretical aspects of statistical machine learning and knowledge representation, with application to web mining, multimedia indexing, natural language processing and bioinformatics.

Recently, BDAA has organized the 36th edition of the French spring school on theoretical computer science (EPIT'08), dedicated to machine learning.

Within the VideoSense project, the BDAA team at LIF will mainly be involved on the classification and fusion tasks. The key members involved within VideoSense are Prof. François Denis, Dr. Stéphane Ayache, Dr. Cécile Capponi and Dr. Amaury Habrard (cf. their CVs in the annex). While Stéphane Ayache have been working on multimedia indexing for several years and has specifically been working on the fusion issues, Prof. François Denis, Dr. Cécile Capponi and Dr. Amaury Habrard will bring their confirmed expertise in machine learning techniques for the purpose of multimodal fusion. Others members of BDAA will contribute to the project, bringing specific skills in Machine Learning. Most of the BDAA researchers are members of the PASCAL European excellence network.

The VideoSense partners gather all the necessary skills to complete the project:

- Eurecom has experience on feature extraction, classification, and automatic learning and fusion;
- LIF has expertise in machine learning techniques for the purpose of multimodal fusion;
- LIG has experience on general concept classification, on multimodal fusion and on active learning for corpus annotation;
- LIRIS has experience in video content modeling, audio event recognition and emotional concept recognition;
- Ghanni has a real industrial application that will be a target; it also has a large database of multimedia contents that will be available for the development and the

evaluation of the developed techniques; and it has the industrial experience necessary for an efficient integration of these techniques.

5.2. RELEVANT EXPERIENCE OF THE PROJECT COORDINATOR

Eurecom has a long standing experience in participating and managing collaborative projects, at the national and European levels. For example, we are currently involved in about 20 national projects, and are already a coordinator for 2 ANR projects. Our administrative structure includes specific personal in charge of contract management. The responsibility of these persons is to comply with the contract reporting procedure to ensure that the correct data for the activity of the project in collected and transmitted. We also have a lawyer to handle the aspects related to the project contract, the consortium agreement, and potential negotiations for intellectual property and valorization activities.

Prof. Merialdo has been participating in several European projects, the last one being PorTiVity (2006-2008) and K-Space (2006-2008), and he is currently involved in one ANR project RPM2 (2008-2010).

6. SCIENTIFIC JUSTIFICATION FOR THE MOBILISATION OF THE RESOURCES

6.1. PARTNER 1 : EURECOM

- *Equipment*

The equipment consists of one workstation for the PhD student, and a storage server for storing the video data and processing results. The project will also be using the existing computing facilities at Eurecom (two clusters are available).

Personnel costs

This project will fund one PhD student for 36 months. The student will mainly focus on the spatiotemporal analysis of video sequences. His work will also include collaboration with the industrial partner for the improvements and optimization of the algorithms to better match the specific requirements of the industrial application.

- *Subcontrating*

None.

- *Travel*

Travels to the project meetings (three or four per year) and to the annual review. Travels to relevant national and international conferences in the field to present the project results.

- *Expenses for inward billing (Costs justified by internal procedures of invoicing)*

N/A

- *Other working costs*

N/A

6.2. PARTNER 2 : ECL LIRIS

- *Equipment*

- Some computers of development for the PhD student, the postdoc, engineer trainees and faculties.
- A server for the storage of video data and computation. The project will also be using the existing computing facilities at ECL LIRIS.

Personnel costs

ECL LIRIS requires support of:

- - a PhD student (36 months) working on video emotion recognition (Task 1.1, 1.2, Task 2.4). also see the phd proposal by ECL Liris 7.1.2..
- - A 12 months Post-Doctoral fellow working on emotion sensitive features (task 1.2) and emotive patterns (Task 2.4) in the audio channel within a video data. All these audio emotion sensitive features or audio emotion patterns will be further fused according to an optimized fusion strategy with visual sensitive features or visual emotion patterns to deliver an emotion label of the video data under study.

- *Subcontrating*

Video emotion recognizer needs to be trained and tested on significant manually labeled emotional video resources. However, as emotion patterns typically vary in time within a video sequence and their perception is subjective and hence possibly multiple by different people, the collection of such a significant and representative corpus typically is human resource consuming task and such manual annotation will also require a specific interface. This task of the emotional video collection and its manual annotation will be subcontracted.

- *Travel*

Travels to the project meetings (three or four per year) and to the annual review.
Travels to relevant conferences in the field to present the project results.

- *Other working costs*

Expenses for small equipment or software, etc.

6.3. PARTNER 3 : LIG-CNRS

- *Equipement*

For the computation of visual descriptors and for experiments on classification and fusion, the LIG partner will need a high performance computing server (around 10,000 €). This server will come as a complement of the computing and storage server servers that the MRIM team already has. These servers will be shared between the MRIM and GETALP groups of the LIG partner.

- *Personnel costs*

We ask funds to support workforces, participating into VideoSense at 100%:

- one 12-month postdoc working on the development of textual descriptors based on pivot languages (Subtask 1.4). This will include development of a set of pivot elements suited for video content indexing and the development of tools for extracting them from text input associated to video segments (metadata, speech transcription, closed captions …);
- one 12-month postdoc working on the development of visual descriptors (Subtask 1.1), on active learning for corpus annotation (Subtask 2.1) and on static-concept classification (Subtask 2.2) and multimodal fusion (Subtask 2.5).

- *Travel*

We shall have to travel to three or four project meetings per year and to relevant conferences in the field for result dissemination. We estimate the global cost both for permanent and non-permanent to about 16000 € for the whole project.

- *Other working costs*

We will need to provide PC (laptop and/or office station) for each non-permanent participant (except internships). We also ask funds to provide one machine per key permanent involved with VideoSense. We estimate the global cost to about 8,000 €.

- *Subcontrating*

We will sub-contract the implementation of a database for the storage and management of the descriptors and classification results. We estimate the global cost to about 10,000 €.

6.4. PARTNER 4 : LIF

- *Equipement*

The BDAA team of the LIF partner is not enough equipped for intensive data analysis. Thus, for the purpose of this project, we need one storage server (around 5000 €) and one computation server (around 7000€). This request will be completed with others supports from *Région PACA*, University of Méditerranée and University of Provence.

- *Personnel costs*

We ask funds to support workforces, participating into VideoSense at 100%:

- One PhD student working on classification and fusion task. See the PhD proposal by LIF;
- One 6 month engineer working on the co-training algorithm (Task 2.3). The engineer will be responsible of the design, the development, and the test of the co-training plat-form. Ideally such a platform will be developed over Weka using Java.

- *Travel*

Travels to three or four project meetings per year and to relevant conferences in the field to present the project results. We estimate this outlay as globally for permanent and non-permanent to about 23000€ for the full project.

- *Other working costs*

We will need to provide PC (laptop + office station) for each non-permanent participant (except internships). We also ask funds to provide one machine per key permanent involved with VideoSense. We estimate this outlay to about 10000€ (5 x 2000€).

- *Subcontrating*

One usefull task in VideoSense will aim at exchanging data (descriptors, classifier decisions, fusion decisions) within team members and with others partners. We will sub-contract the implementation of a web platform to that purpose. We estimate this cost to about 5000€.

6.5. PARTNER 5 : GHANNI

- *Equipment*

The PCs and Software used for the development in the context of this project have been already amortized.

- *Personnel costs*

Ghanni's participation in this project is concentrated on the subworkpackages T0.2 (specifications), T1.2 (audio descritpion), T2.1(corpus annotation), T3.1 (integration) and T3.2 (experimentation).

Aliaksandr Paradzinets will participate as a full-time senior research engineer in tasks T0.2, T2.1 and T3.1. This corresponds to 36 months full-time in senior research engineering resources.
Hadi Harb will participate as a part-time research manager in tasks T0.1, T0.2, T1.2, T2.1, T3.1 and T3.2. This corresponds to 22 months full-time in project management resources.
Ghanni will hire a junior engineer to work as a full-time non-permanent junior research engineer on tasks T1.2, T2.1, T3.1 and T3.2. This corresponds to 33 months full-time non-permanent junior development engineering resources.

Permanent Project Manager: PPM [4000 €/month]
Permanent Senior Engineer: PSE [3500 €/month]
Non-permanent Junior Engineer: NJE [2500 €/month]

Total ressources : 22 PPM, 36 PSE, 33 NJE

- *Subcontrating*

N/A

- *Travel*

Hadi Harb and Aliaksandr Paradzinets will participate in Advertizing-Oriented and Media recommendation-Oriented conferences.
Hadi Harb and Aliaksandr Paradzinets will participate in 4 Videosense project meetings/year.
Moreover, during the T0.2 subwokrpackage, Aliaksandr Paradzinets will go on missions to interview different advertisers.

- *Expenses for inward billing (Costs justified by internal procedures of invoicing)*

N/A

- *Other working costs*

N/A

7. ANNEXES

7.1. PHD TOPICS

7.1.1 EURECOM

Title: Event detection and recognition in video sequences
Advisors: Bernard Merialdo (Professor, Eurecom), Benoit Huet (Lecturer, Eurecom)

SUBJECT

1. Context and objectives

Video analysis is gaining increasing importance for advanced multimedia applications. Most research has been built upon Image Processing techniques, which allow to static situations in a video sequence. Following the current trend, we will focus on extending these techniques for dynamic events involving a time dimension. Most of the current approaches have introduced dynamic parameters such as motion vectors as extra features in static classifiers. Some work has been done with dynamical models such as HMM, but they are generally focused on very specific situations, such as the analysis of sport games. Our objective is to extend such dynamical models to more general cases, as addressed for example in the TRECVID evaluation campaign, or in the Ghanni database.

2. Scientific approach

The goal of this thesis is to develop analysis techniques to allow the detection and recognition of a set of predefined events. The research will in particular focus on dynamical events, which include an important time dimension. Those events will be identified through the specification of the industrial application by Ghanni, and from the TRECVID evaluation campaign. The scientific approach to solve this task will be first to study the various possible spatiotemporal descriptors, in order to identify those which are the most relevant to describe the concepts selected. Then, we will focus on how to best use these descriptors, in combination with the standard ones, in particular, we will investigate how dynamic models can be used for the detection. The proposed algorithms will be implemented and tested on the database that will be constructed for the Ghanni application, as well as data available from the TRECVID campaign.

7.1.2 ECL LIRIS

Title: Emotional concept recognition from video data
Advisors: Liming Chen (Professor, ECL LIRIS), Dr. Charles-Edmond Bichot (Lecturer, ECL LIRIS), Dr. Emmanuel Dellandréa (Lecturer, ECL LIRIS)

SUBJECT

1. Context

Automatic concept recognition from video data is an extremely challenging and increasingly topic problem for both industry and academia within the context of proliferation of video data all around us as a result of the democratization of digital cameras and more and more powerful front-end devices. In this thesis, we are interested by the "emotional content" of a video data as it can be perceived by users and its automatic recognition is of the deepest interest [90], especially within the VideoSense project as it directly concerns the potential impact of a video document.

2. Hypothesis under investigation and main aims

The video data considered within this thesis includes different genres, including news, comedies, educational programs, etc. They can be professional videos or amateur ones. While professional videos are in general of very good quality, making use of some well-known editing rules in the post-production, this is not the case for amateur videos. The aim of this thesis here is to take into account the perceptual feelings of the end-users while experiencing video data, namely emotional effect provoked by viewing video document, such as excitement, calmness, enchantment, stress, anxiety, violence, etc.

3. Research strategy and main milestones

On the basis of our previous work [91,92,93,94,95] and a state of the art on the topic, the following issues will be investigated:

- The description of emotional content conveyed by video data. The emotional content of a video data can be described and labeled, at several granularity level (e.g. shot, scene or sequence), by one or more discrete emotional states, according to a discrete emotion model [96], thus enabling distinctions between fundamental emotions including sadness, fear, anger, surprise, etc., or combination of subtle emotions according to a dimensional emotion model [97]. *See Task 0.2.*
- Collection, by active learning and annotation, of a representative dataset from the emotion perspective for learning and testing. *See Task 2.1.*
- Study of emotion sensitive features both in audio and visual channels for the purpose of the further classification and fusion step. *See Task 1.1, Task 1.2, Task 1.3.*

4. Methods of research

While the existing techniques for speech emotion detection or music mood recognition can be generalized to the audio channel of a video sequence, keeping in mind however that the audio signal is quite different from the ones considered by the most works in the literature, namely acted emotional speech, we also want to develop new visual "emotional" descriptor. Image processing techniques will be applied at several levels of abstraction. In particular, at single still image level, some aesthetic features such as balance and luminance of global and local color [98] or density and thickness of segments

will be exploited [99]. At the video sequence level, the speed or movement intensity and trajectories of feature points or objects will be analyzed to capture emotional properties such as calmness or violence. In terms of objects (forms recognized in the video sequence), face detection and identification techniques can also be used to characterize human expression in the image. However, they are not reliable in unconstrained environment. Emotional analysis of the video data will result from an optimized fusion strategy, thus combining audio and visual modalities, which can be at feature-level fusion, decision-level fusion or a mix-level fusion such as model-level fusion (e.g. HMM). *See Task 2.5.*

7.1.3 LIF

Title: Optimal Fusion for semantic multimedia indexing.
Advisors: François Denis (Professor, LIF), Stéphane Ayache (MCF, LIF), Georges Quénot (CR1, LIG).

SUBJECT

1. Context

Automatic multimedia indexing is an extremely challenging and increasingly topic to ensure the future of Search Engine. In order to satisfy the user needs, one should achieve indexing at semantic level by automatically representing multimedia documents with set of concepts. Commonly, indexing systems implements feature extractors, classifiers and fusion schemes, and are generically designed in order to handle various kinds of concepts similarly. In this thesis, we are interested by the design and the choice of the Fusion process to globally improve concepts detection. This work will be directly involved within the VideoSense project, where concepts and modalities will be especially heterogeneous.

2. Hypothesis under investigation and main aims

While generic indexing systems allow to index multimedia documents with a lot of concepts, the performance in terms of quality is still not satisfying (Mean Average Precision in the 0.1-0.2 range for the world best systems on last TRECVID evaluation campaigns. In other hand, specific approaches are designed for a particular concept and cannot be deployed to detect many concepts. This thesis aims at study specific fusion approaches in order to optimally improve the detection of large amount of concepts. Expected results would be an indexing system with "good" performance able to detect "a lot" of concepts.

3. Research strategy and main milestones

On the basis of our previous work and a state of the art on the topic, the following issues will be investigated:

- Video shots content description and its relation to concepts modeling. This part aims at studying the discriminating power of low-level features with respects to concepts or class of concepts. *See Tasks 1.1 to 1.4.*
- Building of Multimodal vocabulary for Bag of Words. *See Tasks 1.2 and 2.5.*
- Kernel design and combination for multimedia indexing. *See Task 2.5.*
- Context based concept learning. Study the feasibility to infer concepts from others based on their relationships and/or statistical knowledge. *See Task 2.5.*

4. Methods of research

First, an inventory of visual, sound and textual descriptors as well as machine-learning algorithms will be studied for the context of video content analysis. Especially, we will consider the “Bag of Words” feature representations, and the Kernel based machine-learning approaches. This work will be based upon existing toolbox and tools previously developed within the team. We want to study the correlation between modalities and concepts in terms of annotations, descriptors, or decision scores. This study will be the

basis to further investigate the role and the theoretical issues of the fusion process. Then, based on recent kernel advances in machine learning, we will investigate the design and the combination of specifics kernels to optimally model the video data. For instance, we may investigate the use of specifics distances (like Edit distance) to compare video shots with "bag of words" representation. Evaluation of the proposed methods will be conducted in the context of international campaign (like TREC VIDEO), as well as real case data (from Ghanni) and will strongly contribute to the VideoSense project. *See Task 3.*

5. References

Cees G.M. Snoek and Marcel Worring. *Multimodal video indexing : A review of the state-of-the-art*. In Multimedia Tools and Applicationsi, 2005.

Weng, M. and Chuang, Y. *Multi-cue fusion for semantic video indexing*. In Proceeding of the 16th ACM international Conference on Multimedia (ACM MM), 2008.

C.G.M. Snoek, M. Worring, and A.W.M. Smeulders. *Early versus late fusion in semantic video analysis*. In Proceedings of ACM Multimedia, 2005.

Stéphane Ayache, Georges Quénot and Jérôme Gensel. *Classifier Fusion for SVM-Based Multimedia Semantic Indexing*. In European Conference of Information Retrieval (ECIR), 2007.

J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid. *Local features and kernels for classification of texture and object categories: a comprehensive study*. In International Journal of Computer Vision, 2007.

Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, *Evaluation of Color Descriptors for Object and Scene Recognition*. Proceedings of CVPR, 2008.

7.2. RESUME OF THE MAIN PROJECT PARTICIPANTS

7.2.1 BERNARD MERIALDO (EURECOM)

Surname: MERIALDO **First Name:** Bernard
Employment: Professor and department head, EURECOM
Professional address:
EURECOM, BP 193, 06904 Sophia Antipolis
Direct tel.: +33 (0)4 93 00 81 29 **Fax:** +33 (0)4 93 00 82 00
Mail: merialdo@eurecom.fr **Webpage:** http://www.eurecom.fr/

1) BIOGRAPHY

Education:

- 1992: Habilitation à diriger des recherches, Université de Paris 7

Research Interests:

- Multimedia indexing
- Video analysis

Student-researcher advising:

- Supervision of 10 PhD students

Recent participations in research projects:

- EU projets PorTiVity, EU NoE K-Space
- ANR RPM2

Recent professional activities:

- Steering committee CBMI Workshop

2) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

- Dumont, Emilie;Merialdo, Bernard;Essid, Slim;Bailer, Werner;Byrne, Daragh;Bredin, Hervé;O'Connor, Noel E.; Jones,Gareth J.F.;Haller, Martin; Krutz,Andreas; Sikora, Thomas;Piatrik, Tomas **A collaborative approach to video summarization** SAMT 2008, 3rd International Conference on Semantic and Digital Media Technologies, December 3-5, 2008, Koblenz, Germany
- Bailer, Werner;Dumont, Emilie;Essid, Slim;Mérialdo, Bernard **A collaborative approach to automatic rushes video summarization** 1st IEEE ICIP Workshop on Multimedia Information Retrieval: New Trends and Challenges, October 12, 2008, San Diego, USA
- Trichet, Rémi;Mérialdo, Bernard **Keypoints labeling for backgroung substraction in tracking applications** ICME 2008, IEEE International Conference on Multimedia & Expo, June 23-26, 2008, Hannover, Germany
- Dumont, Emilie;Mérialdo, Bernard **Redundancy removing and event clustering for video summarization** WIAMIS 2008, 9th International Workshop on Image Analysis for Multimedia Interative Services, May 7-9, 2008, Klagenfurt, Austria
- Neuschmied, Helmut;Trichet, Rémi;Mérialdo, Bernard **Fast annotation of video objects for interactive TV** MM 2007, 15th international ACM conference on multimedia, September 24-29, 2007, Augsburg, Germany
- Dumont, Emilie;Mérialdo, Bernard **Video search using a visual dictionary** CBMI 2007, 5th International Workshop on Content-Based Multimedia Indexing, June 25-27, 2007, Bordeaux, France

7.2.2 BENOIT HUET (EURECOM)

Surname: Huet **First Name:** Benoit
Employment: Maître de Conférences, EURECOM
Professional address:
EURECOM, BP 193, 06904 Sophia Antipolis
Direct tel.: +33 (0)4 93 00 81 79 **Fax:** +33 (0)4 93 00 82 00
Mail: Benoit.Huet@eurecom.fr **Webpage:** http://www.eurecom.fr/

1) BIOGRAPHY

Education:

- 1999: PhD, York University (GB)

Research Interests:

- Multimedia indexing
- Emotion recognition

Student-researcher advising:

- Supervision of 3 PhD students

Recent participations in research projects:

- EU projets PorTiVity, EU NoE K-Space

Recent professional activities:

- General Chair, MultiMedia Modeling Conference MMM 2009

2) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

- Huet, Benoit; Smeaton, Alan F.; Mayer-Patel, Ketan; Avrithis, Yannis **Advances in Multimedia Modeling** Springer : Lecture Notes in Computer Science, Subseries: Information Systems and Applications, incl. Internet/Web, and HCI , Vol. 5371, ISBN: 978-3-540-92891-1
- Benmokhtar, Rachid;Huet, Benoit **Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features** MIR 2008, ACM International Conference on Multimedia Information Retrieval 2008, October 27- November 01, 2008, Vancouver, BC, Canada
- Galmar, Eric;Huet, Benoit **Spatiotemporal modeling and matching of video shots** 1st ICIP Workshop on Multimedia Information Retrieval : New Trends and Challenges, October 12-15, 2008, San Diego, California, USA , pp *5-8*
- Benmokhtar, Rachid;Huet, Benoit, Berrani, Sid-Ahmed **Low-level feature fusion models for soccer scene classification** 2008 IEEE International Conference on Multimedia & Expo, June 23-26, 2008, Hannover, Germany
- Paleari, Marco;Huet, Benoit **Toward emotion indexing of multimedia excerpts** CBMI 2008, 6th International Workshop on Content Based Multimedia Indexing, June, 18-20th 2008, London, UK

7.2.3 LIMING CHEN (ECL LIRIS)

Surname: CHEN **First Name:** Liming Born Sept. 30, 1963. Married. 3 children.
Employment: Professor at Ecole Centrale de Lyon, Head of the department mathematics-Informatics
Professional address:
Ecole Centrale de Lyon, dept. of Mathematics and Informatics, laboratory LIRIS CNRS UMR 5205
36 avenue Guy de Collongue F-69134 Ecully Cedex
Direct tel.: +33 (0)4 72 18 65 76 **Fax:** +33 (0)4 72 18 64 43
Mail: liming.chen@ec-lyon.fr **Webpage:** http://www.ec-lyon.fr/ , http://liris.cnrs.fr

1) BIOGRAPHY

Education:

- 1989: Ph.D. in Computer Science, University of Paris VI

Research Interests:

- multimedia analysis and indexing
- Face modeling, detection, analysis and recognition

Student-researcher advising:

- Since 1992, 14 former Ph.D. students and currently 12 PhD under supervision or co-supervision.

Recent participations in research projects:

- European project PHENIX on multimedia value added services in mobile environment
- National French Projects: RNRT Cyrano, RNTL Muse, ACI MusicDiscover, Technovision IV2, ANR Omnia, ANR FAR3D
- Industrial Collaborations: France Telecom research, Anaveo, Telem Sécurité Electronique, Xerox.

Recent professional activities:

- Chairman, PC member or reviewers of many international conferences and journals including for instance Medianet, IEEE SITIS, IEEE Signal Processing Letters, Computer vision and Image Understanding, IEEE Transactions on Systems, Man, and Cybernetics, etc.
- member of the IEEE Society, one of six guest editors for the special issue on Automatic Audio Classification of EURASIP Journal on Audio, Speech, and Music Processing.
- Former Chief scientific officer (2001/2003) at Avivias specialized in media asset management and Scientific expert multimedia at France Telecom R&D China in 2005

2) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

Liming Chen is an author or coauthor of more than 100 publications that have appeared as journal papers or proceeding articles, 6 book chapters, 1 book, 3 international patents and 1 protected software on face detection.

- Huanzhang Fu, Alain Pujol, Emmanuel Dellandréa, Liming Chen, "Region based Visual Object Categorization using Segment Features and Polynomial Modeling", Joint IAPR International Workshops on Structural and Syntactic Pattern Recognition (SSPR 2008) and Statistical Techniques in Pattern Recognition (SPR 2008), Orlando, Florida, USA, December 4-6, 2008
- Hadi Harb, Liming Chen, "A General Audio Semantic Classifier based on human perception motivated model", Multimedia Tools and Applications, Eds. Springer Netherlands, ISSN 1380-7501 (Print) 1573-7721 (Online), http://dx.doi.org/10.1007/s11042-007-0108-9, March 05, 2007

- Dr. Parshin and Dr. Chen, "Statistical Audio-Visual Data Fusion for Video Scene Segmentation", Idea Group Inc. book, Semantic-Based Visual Information Retrieval, edited by Dr. Yu-Jin Zhang, Sept. 2006, ISBN 1-59904-370-X, pp.68-89
- Hadi Harb, Liming Chen, "Audio-based visualizing and structuring of videos", International Journal on Digital Libraries, Special issue on Multimedia Contents and Management in Digital Libraries, Vol 6(1), Springer-Verlag, pp.70-81, 2006, published online : 22 October 2005, http://dx.doi.org/10.1007/s00799-005-0120-5
- Hadi Harb, Liming Chen, "voice-based Gender Identification in multimedia applications", Journal of intelligent information systems, J. Intell. Inf. Syst. 24, vol 24(2), 2005, pp.179-198

7.2.4 CHARLES-EDMOND BICHOT (ECL LIRIS)

Surname: BICHOT **First Name:** Charles-Edmond Born Nov. 20, 1981. Married.
Employment: Assistant professor at Ecole Centrale de Lyon (Université de Lyon)
Professional address:
Ecole Centrale de Lyon, dept. of Mathematics and Informatics, laboratory LIRIS CNRS UMR 5205
36 avenue Guy de Collongue F-69134 Ecully Cedex
Direct tel.: +33 (0)4 72 18 65 83 **Fax:** +33 (0)4 72 18 64 43
Mail: charles-edmond.bichot@ec-lyon.fr **Webpage:** http://www.ec-lyon.fr/ , http://liris.cnrs.fr

1) BIOGRAPHY

Education:

- 2007: Ph.D. in Computer Science, Institut National Polytechnique de Toulouse, France
- 2006: Master of Social Science at the Institut d'Etudes Politiques de Toulouse, France
- 2004: Engineering degree at the Ecole Nationale de l'Aviation Civile, Toulouse, France
- 2004: M.Sc. in Computer Science, Université Paul Sabatier de Toulouse, France

Research Interests:

- Graph partitioning, data clustering
- Image segmentation and video emotional recognition
- Optimization and operation research

Student-researcher advising:

- 2 Ph.D. students under co-supervision.

Recent participations in research projects:

- National French Project: ANR Solstice (Solveurs et simulation en calcul extrême)
- Collaborations: Direction des services de la navigation aérienne (DSNA).

Recent professional activities:

- Chairman, PC member or reviewers of IEEE CEC (Congress on Evolutionary Computation), INCOM, META, special issue of EJOR
- Member of the IEEE Society, of the Roadef and of the AFPC.

2) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

Charles-Edmond Bichot is an author or co-author of 2 journal papers and 8 proceeding articles.

- Charles-Edmond Bichot, "Application of the fusion-fission metaheuristic to document clustering", In proceedings of the International Conference on Metaheuristics and Nature Inspired Computing, 2008
- Charles-Edmond Bichot, "**A New Meta-method for Graph Partitioning**", In proceedings of IEEE Congress on Evolutionary Computation, 2008
- Charles-Edmond Bichot, "**A new Method, the Fusion Fission, for the relaxed k-way graph partitioning problem, and comparisons with some multilevel algorithms".** Journal of Mathematical Modeling and Algorithms, 2007, 6(3): 319-344
- Charles-Edmond Bichot,, "**A metaheuristic based on fusion and fission for partitioning problems**", In proceedings of the 20th IEEE International Parallel and Distributed Processing Symposium, 2006

7.2.5 EMMANUEL DELLANDREA (ECL LIRIS)

Surname: DELLANDREA **First Name:** Emmanuel Born March, 10, 1977. Married. 1 child.
Employment: Associate Professor at Ecole Centrale de Lyon
Professional address:
Ecole Centrale de Lyon, dept. of Mathematics and Informatics, laboratory LIRIS CNRS UMR 5205
36 avenue Guy de Collongue F-69134 Ecully Cedex
Direct tel.: +33 (0)4 72 18 65 26 **Fax:** +33 (0)4 72 18 64 43
Mail: emmanuel.dellandrea@ec-lyon.fr **Webpage:** http://perso.ec-lyon.fr/emmanuel.dellandrea

1) BIOGRAPHY

Education:

- 2003: Ph.D. in Computer Science, Université de Tours, France.
- 2000: M. Sc in Computer Science, Université de Tours, France.
- 2000: Engineer in Computer Science, Ecole Polytechnique de l'Université de Tours, France.

Research Interests:

- image, audio and video analysis, pattern recognition, multimedia indexing.

Student-researcher advising:

- 1 former Ph.D. student and currently 3 PhD under co-supervision.

Recent participations in research projects:

- ACI MusicDiscover
- ANR OMNIA (LIRIS scientific responsible)

2) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

- Z Xiao, E Dellandréa, W. Dou, L. Chen. Ambiguous classification of emotional speech. Dans International Workshop on EMOTION - satellite of INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 2008.
- Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "What is the Best Segment Duration for Music Mood Analysis ?", International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 17-24, 2008.
- Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Automatic Hierarchical Classification of Emotional Speech", Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007), Taichung, Taiwan, pp. 291-296, 2007.
- Z. Xiao, E. Dellandréa, W. Dou, L. Chen, "Two-stage Classification of Emotional Speech", International Conference on Digital Telecommunications (ICDT'06). Cap Esterel (France), pp. 32-37, 2006.
- E. Dellandréa, H. Harb, L. Chen, "Zipf, Neural Networks and SVM for Musical Genre Classification", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2005), Athènes (Grèce), pp. 57-62, 2005.
- E Dellandréa, P. Makris, N. Vincent, "Zipf Analysis of Audio Signals", Fractals, World Scientific Publishing Company, vol. 12(1), pp. 73-85, 2004.

7.2.6 GEORGES QUENOT (LIG-CNRS)

Last name: QUÉNOT **First Name:** Georges **Born:** May 14, 1960. Married. 2 children.
Employment: Researcher (CNRS) at Laboratoire d'Informatique de Grenoble.
Professional address:
Laboratoire d'Informatique de Grenoble – CNRS UMR 5217
Bâtiment B, 385, rue de la Bibliothèque, B.P. 53, 38041 Grenoble Cedex 9
Direct tel.: +33 (0)4 76 63 58 55 **Fax:** +33 (0)4 76 63 56 86
Mail: Georges.Quenot@imag.fr **Webpage:** http://clips.imag.fr/mrim/georges.quenot/

1) BIOGRAPHY

Education:

- 1988: Ph.D. in Computer Science, University of Orsay – Paris XI.

Research Interests:

- Multimedia information indexing and retrieval;
- Concept indexing in image and video documents;
- Machine learning.

Student-researcher advising:

- 6 former Ph.D. students and currently 2 PhD students.

Recent participations in research projects:

- European project (STREP) PENG: PErsonalised News content programming;
- National French projects: ANR AVEIR : Annotation automatique et extraction de concepts visuels pour la recherche d'images ; TechnoVision ARGOS : Campagne d'évaluation d'outils de surveillance de contenus vidéos ; OSEO-AII Quaero : La recherche et la reconnaissance de contenus numériques ;
- ICT ASIA project: MoSAIC: Mobile Search and Annotation using Images in Context.

Recent professional activities:

- PC member or reviewers of many international conferences and journals including for instance: Proceedings of the IEEE, ACM Transactions on Multimedia Computing Communications and Applications, IEEE Transactions on Multimedia, Information Processing and Management, IEEE Transactions on Pattern Analysis and Machine Intelligence, Multimedia Tools and Applications, and Signal Processing: Image Communication.
- Organization of the first École d'Automne en Recherche d'Information et Application (EARIA'06).
- Member of the TechnoVision steering committee.

2) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

Georges Quénot is an author or coauthor of 13 journal papers, 74 proceeding articles, 5 book chapters, 1 international patent, 3 French patents and 1 protected software on motion analysis.

- Stéphane Ayache and Georges Quénot, "Image and Video Indexing using Networks of Operators", in EURASIP Journal on Image and Video Processing, Vol. 2007, Article ID 56928, 13 pages, 2007.
- Stéphane Ayache and Georges Quénot, "Evaluation of active learning strategies for video indexing", in Signal Processing: Image Communication, Vol. 22/7-8 pp 692-704, August-September 2007.
- Philippe Joly, Jenny Benois-Pineau, Ewa Kijak and Georges Quénot, "The ARGOS campaign: Evaluation of Video Analysis Tools", in Signal Processing: Image Communication, Vol. 22/7-8 pp 705-717, August-September 2007.

- Stéphane Ayache and Georges Quénot, "Video Corpus Annotation using Active Learning", in 30th European Conference on Information Retrieval (ECIR'08), Glasgow, Scotland, 30th March - 3rd April, 2008.
- Stéphane Ayache, Georges Quénot, Jérôme Gensel and Shin'ichi Satoh, "Using Topic Concepts for Semantic Video Shots Classification", in International Conference on Image and Video Retrieval (CIVR'06), Tempe, AZ, USA, July 13-15, 2006.
- Mbarek Charhad, Daniel Moraru, Stéphane Ayache and Georges Quénot, "Speaker Identity Indexing In Audio-Visual Documents", in Content-Based Multimedia Indexing (CBMI'05), Riga, Latvia, June 21-23, 2005.

7.2.7 GILLES SERASSET (LIG-CNRS)

Surname: SÉRASSET **First Name:** Gilles Born Aug., 15, 1968. Married. 1 child.
Employment: Associate Professor at Université Joseph Fourier – Grenoble 1
Professional address:
Laboratoire d'Informatique de Grenoble, Équipe GETALP, BP 53, 38051 Grenoble cedex 9
Direct tel.: +33 (0)4 76 51 43 80 **Fax:** +33 (0)4 76 63 56 86
Mail: Gilles.Serasset@imag.fr **Webpage:** http://getalp.imag.fr

3) BIOGRAPHY

Education:

- 1994: Ph.D. in Computer Science, Université Joseph Fourier – Grenoble 1

Research Interests:

- Multilingual Lexical Data Modeling and Management
- Multilingual Communication

Student-researcher advising:

- Since 1995, 3 former Ph.D. students and currently 1 PhD under co-supervision.

Recent participations in research projects:

- Europe's INTERREG IIIb LexALP project (project team leader);
- International cooperation: Franco-Thai project (Thai Ph.D. supervision); Leader of the Papillon dictionary project;

Recent professional activities:

- Chairman, PC member or reviewers of several international workshops; guest editor of the forthcoming special issue on Multilingual Language Resources and Interoperability of LRE (Language Resources and Evaluation) Journal; member of the ACL and ATALA.

4) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

- Francis Brunet-Manquat and Gilles Sérasset. Création d'une base terminologique juridique multilingue à l'aide de la plateforme générique jibiki : le projet LexALP. In P. Mertens, C. Fairon, A. Dister, and P. Watrin, editors, TALN06: Actes de la 13e conférence sur le traitement automatique du langage, pages 435–444, Louvain-la-Neuve, April 2006. Presses universitaires de Louvain.
- Jean-Pierre Chevallet and Gilles Sérasset. Using Surface-Syntactic Parser and Deviation from Randomness. X-IOTA IR System Used for CLIPS Mono and Bilingual Experiments for CLEF 2004. In Carol Peters, et al. editors, Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers, volume 3491/2005 of LNCS, pages 38–49. Springer Verlag, 2005.
- Verena Lyding, Elena Chiocchetti, Gilles Sérasset, and Francis Brunet-Manquat. The LexALP information system: Term bank and corpus for multilingual legal terminology consolidated. In Proceedings of the Workshop on Multilingual Language Resources and Interoperability, pages 25–31,

Sydney, Australia, July 2006. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W/W06/W06-1004.

- Loïc Maisonnasse, Gilles Sérasset, and Jean-Pierre Chevallet. Using the X-IOTA system in mono- and bilingual experiments at clef 2005. Lecture Notes in Computer Science, 4022 (2006):69–78, 2006.
- Gilles Sérasset and Étienne Blanc. Remaining issues that could prevent UNL to be accepted as a standard. In Jesus Cardeñosa, Alexander Gelbukh, and Edmundo Tovar, editors, Universal Networking Language, Advances in Theory and applications, volume 12 of Research on Computing Science, pages 117–124. 2005. URL http://www-clips.imag.fr/geta/gilles.serasset/convergences03-serasset.pdf.
- Gilles Sérasset, Francis Brunet-Manquat, and Elena Chiocchetti. Multilingual legal terminology on the jibiki platform: The lexalp project. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006), pages 937–944, Sydney, Australia, July 2006. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P06/P06-1118.
- Andreas Witt, Gilles Sérasset, Susan Armstrong, Jim Breen, Ulrich Heid, and Felix Sasaki, editors. Proceedings of the Workshop on Multilingual Language Resources and Interoperability. Association for Computational Linguistics, Sydney, Australia, July 2006. URL http://www.aclweb.org/anthology/W/W06/W06-10.

7.2.8 STÉPHANE AYACHE (LIF)

Last name: AYACHE **First Name:** Stéphane **Born:** February 01, 1979. Married. 1 child.
Employment: Associate Professor at University of Mediterranée - Laboratoire d'Informatique Fondamentale de Marseille
Professional address:
Laboratoire d'Informatique Fondamentale de Marseille – CNRS UMR 6166
Centre de Mathématiques et Informatique - 39 rue Joliot-Curie - F-13453 Marseille Cedex13.
Direct tel.: +33 (0) 4 91 82 86 74 **Fax:** +33 (0) 4 91 82 86 71
Mail: stephane.ayache@univmed.fr **Webpage:** http://stephane.ayache.perso.esil.univmed.fr

3) BIOGRAPHY

Education:

- 2007: Ph.D. in Computer Science, National Institute of Polytechnics, Grenoble.
- 2003: M. Sc. In Computer Science, University Joseph Fourier, Grenoble.

Research Interests:

- Multimedia information indexing and retrieval;
- Concept indexing in image and video documents;
- Machine learning.

Recent participations in research projects:

- Projet Franco-asiatique ISERE (Intermedia Semantic Extraction and REasoning)
- ICT ASIA project: MoSAIC: Mobile Search and Annotation using Images in Context.

Recent professional activities:

- PC member or reviewers of international conferences and journals including: IEEE Transactions on Systems, Man, and Cybernetics, Part B; International Workshop on Content-Based Multimedia Indexing; ACM SIGIR.

4) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

Stéphane Ayache is an author or coauthor of 2 international journal papers and 12 proceeding articles.

- Stéphane Ayache and Georges Quénot, "Image and Video Indexing using Networks of Operators", in EURASIP Journal on Image and Video Processing, Vol. 2007, Article ID 56928, 13 pages, 2007.
- Stéphane Ayache and Georges Quénot, "Evaluation of active learning strategies for video indexing", in Signal Processing: Image Communication, Vol. 22/7-8 pp 692-704, August-September 2007.
- Stéphane Ayache and Georges Quénot, "Video Corpus Annotation using Active Learning", in 30th European Conference on Information Retrieval (ECIR'08), Glasgow, Scotland, 30th March - 3rd April, 2008.
- Stéphane Ayache, Georges Quénot and Jérôme Gensel. "Classifier Fusion for SVM-Based Multimedia Semantic Indexing". In European Conference of Information Retrieval (ECIR), 2007.
- Stéphane Ayache, Georges Quénot, Jérôme Gensel and Shin'ichi Satoh, "Using Topic Concepts for Semantic Video Shots Classification", in International Conference on Image and Video Retrieval (CIVR'06), Tempe, AZ, USA, July 13-15, 2006.
- Laurent Besacier, Georges Quénot, Stéphane Ayache and Daniel Moraru, "Video Story Segmentation with Multi-Modal Features: Experiments on TRECvid 2003". In 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR), 2004.

7.2.9 CÉCILE CAPPONI (LIF)

Last name: Capponi **First Name:** Cécile **Born:** July 29, 1970. Married. 2 children.
Employment: Associate Professor at University of Provence - Laboratoire d'Informatique Fondamentale de Marseille.
Professional address:
Laboratoire d'Informatique Fondamentale de Marseille – CNRS UMR 6166 Centre de Mathématiques et Informatique - 39 rue Joliot-Curie - F-13453 Marseille Cedex13.
Direct tel.: +33 (0) 4 91 11 36 11 **Fax:** +33 (0) 4 91 11 36 02
Mail: Cecile.Capponi@lif.univ-mrs.fr **Webpage:** http://www.lif.univ-mrs.fr/~capponi

BIOGRAPHY

Education:

- 1995: Ph.D. in Computer Science, Joseph Fourier University – INRIA

Research Interests:

- Machine Learning,
- Knowledge Representation,
- Bioinformatics

Student-researcher advising:

- 2 former Ph.D. Students (2004 and 2007)

Recent participations in research projects:

- ACI Genoto3D
- ACI ISYMOD+

Recent professional activities:

- Reviewer of IEEE Transactions on Data and Knowledge Engineering (2008,2009). Reviewer of national conferences JOBIM, LMO since 2004.
- Head of the organization of the 36th French Spring School on Theoretical Computer Sciences (Porquerolles, 2008). Theme: Machine Learning (around 80 attendees, http://epit08.univ-mrs.fr). Member of the organization comitee of JOBIM 2007 (Marseille, around 400 attendees).
- Co-head of the Web Multimedia Mining Group, Marseille.
- Reviewer of scientific ATIP-CNRS projects (2008)
- In charge of industrial property management and technology transfert of the Laboratoire d'Informatique Fondamentale de Marseille.

MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

- C. N. Magnan, C. Capponi, F. Denis (2007). *A Protocol to Detect Local Affinities Involved in Proteins Distant Interactions.* In proc. of the 2007 IEEE International Conference on Bioinformatics and Biomedicine, San Jose (CA, US), IEEE Press, pp. 252-257.
- Capponi, C., Fichant, G., Quentin, Y., & Denis, F. (2005). *Boosting Blast*. In proc. of the 11th Symposium on Applied Stochastic Models and Data Analysis (ASMDA), Brest, France.
- Chabalier, J., Capponi, C., Quentin, Y., & Fichant, G. (2005). ISYMOD: a Knowledge Warehouse for the identification, assembly and analysis of bacterial integrated systems. *Bioinformatics*, 21(7), 1246-56.
- Moisuc, B., Capponi, C., Genoud, P., Gensel, J., & Ziébelin, D. (2007). Modélisation algébrique et représentation des connaissances par objets en AROM. *Revue Sciences et Technologie de l'Informatique, série L'Objet*, 13, 83-98.

7.2.10 FRANÇOIS DENIS (LIF)

Last name: DENIS **First Name:** François **Born:** April 15, 1956. 1 child.
Employment: Professor at University of Provence – Laboratoire d'Informatique Fondamentale de Marseille.
Professional address:
Laboratoire d'Informatique Fondamentale de Marseille – CNRS UMR 6166 Centre de Mathématiques et Informatique - ⊚39 rue Joliot-Curie - F-13453 Marseille Cedex13.
Direct tel.: +33 (0) 4 91 11 36 05 **Fax:** +33 (0) 4 91 11 36 02
Mail: mailto:francois.denis@lif.univ-mrs.fr **Webpage:** http://www.lif.univ-mrs.fr/~fdenis/

5) BIOGRAPHY

Education:

- 2000: HDR in Computer Science, University of Lille.
- 1990: Ph.D. in Computer Science, University of Lille.
- 1979 – 1991: Professor « agrégé » of mathematcs in high school.

Research Interests:

- Machine learning;
- Semi-supervised learning;
- Statistical grammatical inference.

Recent participations in research projects:

- ANR Masse de données Marmota (MAchine learning pRobabilistic MOdels Tree lAnguages), avril 2006 -> avril 2009
- ANR Domaine Emergents SEQUOIA (analySEur syntaxiQUe prObabIliste à large couverture pour le françAis), janvier 2009 -> décembre 2011

Recent professional activities:

- Director of LIF since 2001;
- Member of the PASCAL2 network of excellence.

6) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

François Denis is an author or coauthor of 10 international journal papers and 24 proceeding articles.

- R. Bailly and F. Denis: Absolute Convergence of Rational Series is Semi-decidable. Proceedings of 3rd International Conference on Language and Automata Theory and Applications, 2009.
- Denis, F. and Esposito, Y.: On Rational Stochastic Languages. Journal of Fundamental Informatics. Volume 86, 2008.
- Denis, François and Habrard, Amaury. Learning rational stochastic tree languages. In Proceedings of the 18th international Conference on Algorithmic Learning theory, 2007.
- Denis, F. and Esposito, Y. and Habrard, A. Learning rational stochastic languages. Proceedings of COLT, 2006.
- François Denis, Christophe Nicolas Magnan and Liva Ralaivola. Efficient learning of Naive Bayes classifiers under class-conditional classification noise. Proceedings of ICML, 2006.

7.2.11 AMAURY HABRARD (LIF)

Last name: HABRARD **First Name:** Amaury **Born:** December 09,1978.
Employment: Associate Professor at University of Provence – Laboratoire d'Informatique Fondamentale de Marseille.
Professional address:
Laboratoire d'Informatique Fondamentale de Marseille – CNRS UMR 6166 Centre de Mathématiques et Informatique - 39 rue Joliot-Curie - F-13453 Marseille Cedex13.
Direct tel.: +33 (0) 4 91 11 35 71 **Fax:** +33 (0) 4 91 11 36 02
Mail: mailto:amaury.habrard@lif.univ-mrs.fr **Webpage:** http://www.lif.univ-mrs.fr/~habrard/

7) BIOGRAPHY

Education:

- 2004: Ph.D. in Computer Science, University of Saint-Etienne.

Research Interests:

- Machine learning;
- Learning similarity measures;
- Dealing with structured data;
- Probabilistic models.

Recent participations in research projects:

- ANR Masse de données Marmota (MAchine learning pRobabilistic MOdels Tree lAnguages), avril 2006 -> avril 2009
- ANR Domaine Emergents SEQUOIA (analySEur syntaxiQUe prObabIliste à large couverture pour le françAis), janvier 2009 -> décembre 2011

Recent professional activities:

- PC member or reviewers for international conferences in Machine Learning including ICML, ICDM, ECML, ICGI
- Member of the PASCAL2 network of excellence

8) MOST SIGNIFICANT PUBLICATIONS IN THE FIVE PAST YEARS

Amaury Habrard is an author or coauthor of 2 international journal papers and 14 proceeding articles.

- Marc Bernard, Laurent Boyer, Amaury Habrard, Marc Sebban: Learning probabilistic models of tree edit distance. Pattern Recognition, volume 41, number 8, 2008. Elsevier Science.
- Amaury Habrard, José Manuel Inesta Quereda, David Rizo, Marc Sebban: Melody Recognition with Learned Edit Distances. Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition (SSPR & SPR), volume 5342 of LNCS, 2008. Springer.
- Laurent Boyer, Amaury Habrard, Marc Sebban: Learning Metrics Between Tree Structured Data: Application to Image Recognition. Proceedings of the European Conference on Machine Learning (ECML), volume 4701 of LNCS, 2007. Springer.
- François Denis, Yann Esposito, Amaury Habrard: Learning Rational Stochastic Languages. Proceedings of the International Conference on Computational Learning Theory (COLT), volume 4005 of LNCS, 2006. Springer.
- Amaury Habrard, Marc Bernard, Marc Sebban: Detecting Irrelevant Subtrees to Improve Probabilistic Learning from Tree-structured Data. Fundamenta Informaticae, volume 66, number 1-2, 2005. IOS Press.

7.2.12 HADI HARB (GHANNI)

Surname: HARB **First Name:** Hadi Born Feb., 21, 1978. Single.
Employment: CEO Ghanni
Professional address:
14, Cité Griset, 75011, PARIS
Direct tel.: +33 (0)970.446.733 **Mail:** hadi.harb@ghanni.com **Webpage:**

3) BIOGRAPHY

Education:

- 2000: M.Eng in Electrical Engineering, Lebanese University
- 2003: Ph.D. in Computer Science, Ecole Centrale de Lyon

Research Interests:

- Media Recommender Systems
- Ads Targeting

Recent Projects:

- Ghanni Music Explorer: a music management and automatic playlist generation Software for PCs
- gRadio: a personalized music/video streaming service with advanced personalization, recommendation and licences from 3 major labels (Universal Music, Warner, Sony-BMG)
- MusicInStore: a music management and streaming service for professionals (Bars, Restaurants...)
- Symbio Personalized Radio: An integration of personalized radio for the Thomson's Symbio telephone
- youCircle: a media management community-enabled platform

Recent professional activities:

- Co-Founder and CEO of Ghanni, a start-up specialized in Media Recommendation and Management Solutions

4) MOST SIGNIFICANT SCIENTIFIC PUBLICATIONS IN THE FIVE PAST YEARS

- Hadi Harb and Liming Chen, A general audio classifier based on human perception motivated model, Multimedia Tools and Applications, Volume 34, Number 3 / septembre 2007
- Hadi Harb and Liming Chen, Audio-based description and structuring of videos, International Journal on Digital Libraries, Volume 6, Number 1 / février 2006
- Hadi Harb and Liming Chen, Voice-Based Gender Identification in Multimedia Applications, Journal of Intelligent Information Systems, Volume 24, Numbers 2-3 / mars 2005
- Hadi Harb, Liming Chen, Mixture of experts for audio classification: an application to male female classification and musical genre recognition. In the proceedings of IEEE International Conference on Multimedia and Expo ICME 2004, 2004

7.2.13 ALIAKSANDR PARADZINETS (GHANNI)

Surname: PARADZINETS **First Name:** Aliaksandr Born May, 7, 1980. Married.
Employment: CTO Ghanni
Professional address:
14, Cité Griset, 75011, PARIS
Direct tel.: +33 (0)623.524.940 **Mail:**
aliaksandr.paradzinets@ghannimusic.com

5) BIOGRAPHY

Education:

- 2003: M.Sc. in Radiophysics and electronics, Belarussian State University
- 2007: Ph.D. in Computer Science, Ecole Centrale de Lyon

Research Interests:

- Media Recommender Systems
- Ads Targeting

Recent Projects:

- Ghanni Music Explorer: a music management and automatic playlist generation Software for PCs
- gRadio: a personalized music/video streaming service with advanced personalization, recommendation and licences from 3 major labels (Universal Music, Warner, Sony-BMG)
- MusicInStore: a music management and streaming service for professionals (Bars, Restaurants...)
- Symbio Personalized Radio: An integration of personalized radio for the Thomson's Symbio telephone
- youCircle: a media management community-enabled platform

Recent professional activities:

- Co-Founder and CTO of Ghanni, a start-up specialized in Media Recommendation and Management Solutions

6) MOST SIGNIFICANT SCIENTIFIC PUBLICATIONS IN THE FIVE PAST YEARS

- Paradzinets Aliaksandr., Kotov Oleg., Harb Hadi., Chen Liming. Continuous Wavelet-like Transform Based Music Similarity Features for Intelligent Music Navigtion, Proceedings of the CBMI07, Bordeaux - France, 2007
- Paradzinets Aliaksandr., Harb Hadi., Chen Liming. Use of Continuous Wavelet-like Transform in Automated Music Transcription, Proceedings of the EUSIPCO06, Florence - Italy, 2006
- Kotov Oleg, Paradzinets Aliksandr, Bovbel Eugene., Musical Genre Classification using Modified Wavelet-like Features and Support Vector Machines, Proceedings of the EuroIMSA, Chamonix - France, 2007
- Paradzinets Aliaksandr, Chen Liming, Speaker segments regroupment, Proceedings of the RIAO'2004, Avignon - France, 2004

7.3. REFERENCES

7.3.1 PARTNERS' REFERENCES

The partners' references are available in the resume of the main people involved (section 7.2)

7.3.2 REFERENCES

[1] http://www-nlpir.nist.gov/projects/trecvid/

[2] Dong Xu, Shih-Fu Chang. Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1985-1997, November 2008

[3] D.Lowe, « Object recognition from local scale-invariant features », In Proc. ICCV, pages 1150-1157, 1999

[4] Mikolajczyk, K.; Schmid,C, **A performance evaluation of local descriptors**, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 27, Issue 10, Oct. 2005 Page(s):1615 - 1630 Digital Object Identifier 10.1109/TPAMI.2005.188

[5] A.Bosch, A.Zisserman, X.Muoz, « Scene classification using a hybrid generative/discriminative approach », IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(04):712-727, 2008

[6] J.Van de Weijer, T.Gevers, A.Bagdanov, « Boosting color saliency in image feature detection », IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(01):150-156, 2006

[7] Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, **Evaluation of Color Descriptors for Object and Scene Recognition**. *Proceedings of CVPR*. Anchorage, Alaska, USA, June 2008

[8] Koen E. A. van de Sande, Theo Gevers, Cees G. M. Snoek: A comparison of color features for visual concept classification. CIVR 2008: 141-150

[9] H. Fu, A. Pujol, E. Dellandréa and L. Chen, "Region based visual object categorization using segment features and polynomial image modeling", *7th International Workshop on Statistical Pattern Recognition*, 2008

[10] A. *Yanagawa*, S.-F. *Chang*, L. *Kennedy*, and W. Hsu "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts," Columbia University, technical report, Mar. 2007

[11] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

[12] The PASCAL Visual Object Classes Challenge Workshop 2008, 17th October 2008, ECCV 2008, Marseille, France (http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/index.html)

[13] M. Campbell, A. Haubold, S. Ebadollahi, M.R. Naphade, A. Natsev, J.R. Smith, J. Tesic, L.Xie, "IBM Research TRECVID-2006 Video Retrieval System", *TRECVID 2006*, November 2006

[14] W.Mahdi, M.Ardebilian, L.Chen, "Procédé de classification d'une image couleur selon la prise de vue en extérieur ou intérieur", PCT/FR01/03 869, Décembre 2000 ;

[15] Alain Pujol and Liming Chen, "A Hough Transform Based Cityscape Classifier", 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux (Switzerland), April 2005.

[16] Alain Pujol, Liming Chen, "Coarse Adaptive Color Image Segmentation for Visual Object Classification", Proceedings of the 15th International Conference on Systems, Signals and Image Processing (IWSSIP 2008), Bratislava, Slovak Republic, during June 25 - June 28, 2008

[17] D.Zhang, D.Perez, S.Bengio, I.McCowan, "Semi-supervised adapted HMMs for unusual event detection", Proc. IEEE CVPR, pp.611-618, 2005

[18] M.Brand, N.Oliver, A.Pentland, "Coupled Hidden Markov Models for Complex Action Recognition", Proc. IEEE CVPR, pp.994-999, 1997

[19] O.Boiman and M.Irani, "Detecting Irregularities in Images and in Video", Proc. IEEE Int'l Conf. Computer Vision, pp.1166-1173, 2005

[20] Y.Ke, R.Sukthankar, M.Hebert,"Efficient Visual Event Detection Using Volumetric Features", Proc. IEEE Int'l Conf. Computer Vision, pp.166-173, 2005

[21] Dong Xu, Shih-Fu Chang, "Video Event Recognition Using Kernel Methods with Multilevel Temporal Alignment", IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(11):1985-1997, Nov.2008

[22] Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision 30 (1998) 79–116

[23] Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International Journal of Computer Vision 60 (2004) 63–86

[24] Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2005) 524–531

[25] Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision (2004)

[26] Perronnin, F., Dance, C., Csurka, G., Bressan, M.: Adapted vocabularies for generic visual categorization. In: Proceedings of the European Conference on Computer Vision 4 (2006) 464-475

[27] Kaniza, G.: Grammatica del vedere. Il Mulino, Bologna (1997)

[28] Wertheirmer, M.: Untersuchungen zur lehre der gestalt II. Psychologische Forschung (4) (1923) 301-350

[29] Huanzhang Fu, Alain Pujol, Emmanuel Dellandréa, Liming Chen, "Region based Visual Object Categorization using Segment Features and Polynomial Modeling", Joint IAPR International

Workshops on Structural and Syntactic Pattern Recognition (SSPR 2008) and Statistical Techniques in Pattern Recognition (SPR 2008), Orlando, Florida, USA, December 4-6, 2008

[30] E Dellandréa, P. Makris, N. Vincent, "Zipf Analysis of Audio Signals", Fractals, World Scientific Publishing Company, vol. 12(1), pp. 73-85, 2004.

[31] E. Dellandréa, H. Harb, L. Chen, "Zipf, Neural Networks and SVM for Musical Genre Classification", IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2005), Athènes (Grèce), pp. 57-62, 2005.

[32] J. Pinquier, C. Sénac and R. André-Obrecht, "Speech and music classification in audio documents", Proceedings of the IEEE ICASSP'2002, Orlando, May 2002

[33] Neti C., Roukos S., Phone-context specific gender-dependent acoustic-models for continuous speech recognition, Proceedings., 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, 1997., 14-17, Page(s): 192 -198, Dec. 1997

[34] Hadi Harb, Liming Chen, "voice-based Gender Identification in multimedia applications", Journal of intelligent information systems, J. Intell. Inf. Syst. 24, vol 24(2), 2005, pp.179-198

[35] Hadi Harb, Liming Chen, "A General Audio Semantic Classifier based on human perception motivated model", Multimedia Tools and Applications, Eds. Springer Netherlands, ISSN 1380-7501 (Print) 1573-7721 (Online), http://dx.doi.org/10.1007/s11042-007-0108-9, March 05, 2007

[36] Wold E., Blum T., Keislar D., Wheaton J (1996), "Content-based classification search and retrieval of audio", IEEE Multimedia Magazine 3(3):27-36

[37] Hadi Harb, "Classification Sémantique du Signal Sonore en Vue d'une Indexation par le Contenu des Documents Multimédias", Ecole Centrale de Lyon, 11 décembre 2003, devant le jury composé de MM. Frédéric Bimbot (Rapporteur), Claude Montacié (rapporteur), Jean-Paul Haton (Président), Edouard Geoffrois, Liming CHEN (Directeur de thèse), Jean-Yves Auloge

[38] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, Thomas S. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.31, No.1, Jan 2009, pp.39-58

[39] Scherer, K, R., Vocal communication of emotion: A review of research paradigms, Speech Communication 40, pp. 227-256, 2002

[40] Pereira, C., Dimensions of emotional meaning in speech, Proceedings of the ISCA workshop on Speech and Emotion, p 25 – 28, Newcastle, Northern Ireland, 2000

[41] P.-Y.Oudeyer, "The production and recognition of emotions in speech: features and algorithms", Int'l J. Human-Computer Studies, vol.59, pp.157-183, 2003

[42] Zhongzhe Xiao, « Classification of emotion in audio signals », thèse de l'Ecole Centrale de Lyon, soutenue le 25 Janvier 2008, devant le jury composé de Mme.André-Obrecht Régine (Pr Université de Paul Sabatier, Rapporteur), MM.Laurent Besacier (MdC HDR Université Joseph Fourier, Rapporteur), Jean-Paul Haton (Pr Université Henri Poincaré, Président), Liming Chen (Directeur de thèse),

Emmanuel Dellandréa (MdF ECL LIRIS, Co-directeur), Dou Weibei (Pr Tsinghua University, Co-directrice)

[43] L.Devillers and I.Vsilescu, "Real-life emotions with lexical and paralinguistic cues on human-human call center dialogs", Proc.Ninth Int'l Conf. Spoken language processing (ICSLP), 2006

[44] B.Schuller, R.J.Villar, G.Rigoll, M.Lang, "Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition", Proc.IEEE Int'l Conf. Accoustics, speech, and signal processing (ICASSP'05), pp.325-328, 2005

[45] Z Xiao, E Dellandréa, W. Dou, L. Chen.Features Extraction And Selection In Emotional Speech. International Conference on Advanced Video and Signal based Surveillance (AVSS 2005). Como (Italie). p. 411-416. Septembre. 2005.

[46] Z Xiao, E Dellandréa, W. Dou, L. Chen. Automatic Hierarchical Classification of Emotional Speech. Dans Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007), Taichung, Taiwan. pp. 291-296. 2007.

[47] A Dimensional Emotion Model Driven Multi-stage Classification of Emotional Speech. Z Xiao, E Dellandréa, W. Dou, L. Chen. Rapport de recherche RR-LIRIS-2007-033, 2007. Submitted to Multimedia Tools and Applications.

[48] Z Xiao, E Dellandréa, W. Dou, L. Chen. What is the Best Segment Duration for Music Mood Analysis ?. Dans International Workshop on Content-Based Multimedia Indexing (CBMI), . pp. 17-24. 2008.

[49] M. Pantic, L.J.M Rothkrantz –Automatic analysis of facial expressions: the state of the art- IEEE trans. on PAMI, vol.22, N°12, pp. 1424-1445, December 2000.

[50] B. Fasel, J. Luettin – Automatic Facial Expression analysis: a survey- Pattern Recognition, vol.36, pp.259-275, 2003

[51] Hayashi. T., Hagiwara, M., "Image Query by Impression Words-The IQI System", IEEE Trans. Consumer Electronics, vol.44, No.2, pp.347-352, 1998

[52] A. Mojsilovic, J. Gomes and B. Rogowitz, "Semantic-Friendly Indexing and Querying of Images Based on the Extraction of the Objective Semantic Cues", Intl. J. of Computer Vision, vol. 56, no.1/3, pp. 79-107, 2004.

[53] J. Itten, "The Art of Color", Otto Maier Verlag, Ravensburg, Germany, 1961

[54] C. Colombo, A. Del Bimbo, and P. Pala, "Semantics in Visual Information Retrieval", IEEE Multimedia, Vol.6, No.3, pp.38-53, 1999.

[55] A. Colin, "Introduction à la couleur : des discours aux images", 1994

[56] M.Hammami, C.Li, B. Ben Amor, C.Vial, « Classification d'images par concepts », Journées francophones sur l'Accès Intelligent aux Documents Multimédias (MediaNet'2002), 17-21 Juin 2002, Sousse, Tunisie, Hermes, ISBN 2-7462-0500-9, pp.380-385

[57] A.G. Hauptmann, M. Christel, R. Concescu, J. Gao, Q. Jin, W.-H. Lin, J.-Y. Pan, S. M. Stevens, R. Yan, J. Yang, Y. Zhang, *CMU Informedia's TRECVID 2005 Skirmishes*, In TRECVID Workshop, 2005.

[58] UNL Specifications (2005) available at http://www.undl.org/unlsys/unl/unl2005/

[59] Jesús Cardeñosa, Carolina Gallardo and Luis Iraola *UNL as a Text Content Representation Language for Information Extraction* , in Flexible Query Answering Systems, LNCS 4027/2006, pp.507-518.

[60] Gonzalo J.; Verdejo F.; Peters C.; Calzolari N. *Applying EuroWordNet to Cross-Language Text Retrieval*, Computers and the Humanities, Volume 32, Number 2-3, 1998 , pp. 185-207(23).

[61] Schwab Didier, Lim Lian Tze, Lafourcade Mathieu, *Conceptual vectors, a complementary tool to lexical networks*, NLPCS 2007 : The 4th International Workshop on Natural Language Processing and Cognitive Science, Funchal, Madeira - Portugal, 12-13 Juin, 2007.

[62] Lim Lian Tze, Schwab Didier, *Limits of Lexical Semantic Relatedness with Ontology-based Conceptual Vectors*, NLPCS 2008 : The 5th International Workshop on Natural Language Processing and Cognitive Science, Barcelona, Espagne, 12-13 Juin 2008.

[63] H. S. Seung, M. Opper, H. Sompolinsky, *Query by Committee*, in *Proceedings: COLT92*, p. 287-294

[64] D. Lewis and W. Gale. Training text classifiers by uncertainty sampling. In Proceedings of International ACM Conference on Research and Development in Information Retrieval, pages 3-12, 1994.

[65] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. Technical report, Cornell University, 1988.

[66] Fabrice Souvannavong, Bernard Merialdo, Benoit Huet. Partition sampling for active video database annotation. 5th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'04), 2004.

[67] Stéphane Ayache and Georges Quénot. Video Corpus Annotation using Active Learning, in European Conference on Information Retrieval (ECIR), 2008.

[68] Stéphane Ayache and Georges Quénot. Evaluation of active learning strategies for video indexing, Signal Processing : Image Communication. Volume 22, Issues 7-8, August-September 2007. Pages 692-704. "Special Issue on Content-Based Multimedia Indexing and Retrieval".

[69] Hauptmann, A., Yan, R., and Lin, W. How many high-level concepts will fill the semantic gap in news video retrieval?. In Proceedings of the 6th ACM international Conference on Image and Video Retrieval (CIVR), 2007.

[70] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-Based Image Retrieval at the End of the Early Years. IEEE Trans Pattern Anal Mach Intell 2000;22(12):1349-80.

[71] Cees G.M. Snoek and Marcel Worring. Multimodal video indexing : A review of the state-of-the-art. In Multimedia Tools and Applications, 2005.

[72] Stéphane Ayache and Georges Quénot. Image and Video Indexing using Networks of Operators.

EURASIP Journal on Image and Video Processing, 2007.

[73] Stéphane Ayache, Georges Quénot and Jérôme Gensel. Classifier Fusion for SVM-Based Multimedia Semantic Indexing. In European Conference of Information Retrieval (ECIR), 2007.

[74] C.G.M. Snoek, M. Worring, and A.W.M. Smeulders. Early versus late fusion in semantic video analysis. In Proceedings of ACM Multimedia, 2005.

[75] D. H. Wolpert. Stacked generalization. In Journal of Neural Networks, 1990.

[76] M.R.Naphade and T.S.Huang. Semantic Video Indexing using a probabilistic framework. In Proceedings of the ICPR, 2000.

[77] W. Jiang, S.-F. Chang, and A. Loui. Context-based concept fusion with boosted conditional random fields. In Proc. of ICASSP, 2007.

[78] Gao, S., Lim, J., and Sun, Q. An integrated statistical model for multimedia evidence combination. In Proceedings of the 15th international Conference on Multimedia (ACM MM), 2007.

[79] Weng, M. and Chuang, Y. Multi-cue fusion for semantic video indexing. In Proceeding of the 16th ACM international Conference on Multimedia (ACM MM), 2008.

[80] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. In International Journal of Computer Vision, 2007.

[81] Koen E. A. van de Sande, Theo Gevers and Cees G. M. Snoek, Evaluation of Color Descriptors for Object and Scene Recognition. Proceedings of CVPR, 2008.

[82] Dance, C., Willamowski, J., Fan, L., Bray, C., Csurka, G.: Visual categorization with bags of keypoints. In: ECCV International Workshop on Statistical Learning in Computer Vision (2004)

[83] Perronnin, F., Dance, C., Csurka, G., Bressan, M.: Adapted vocabularies for generic visual categorization. In: Proceedings of the European Conference on Computer Vision 4 (2006) 464-475

[84] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/workshop/index.html

[85] http://www-nlpir.nist.gov/projects/tv2008/tv2008.html

[86] W.Mahdi, M.Ardebilian, L.Chen, "Procédé de classification d'une image couleur selon la prise de vue en extérieur ou intérieur", PCT/FR01/03 869, Décembre 2000 ;

[87] Alain Pujol and Liming Chen, "A Hough Transform Based Cityscape Classifier", 6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), Montreux (Switzerland), April 2005.

[88] Alain Pujol, Liming Chen, "Coarse Adaptive Color Image Segmentation for Visual Object Classification", Proceedings of the 15th International Conference on Systems, Signals and Image Processing (IWSSIP 2008), Bratislava, Slovak Republic, during June 25 - June 28, 2008

[89] Blum, A. and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In

Proceedings of the Eleventh Annual Conference on Computational Learning Theory, 1998.

[90] Zhihong Zeng, Maja Pantic, Glenn I. Roisman, Thomas S. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.31, No.1, Jan 2009, pp.39-58

[91] Z Xiao, E Dellandréa, W. Dou, L. Chen. Features Extraction And Selection In Emotional Speech. International Conference on Advanced Video and Signal based Surveillance (AVSS 2005). Como (Italie). p. 411-416. Septembre. 2005.

[92] Z Xiao, E Dellandréa, W. Dou, L. Chen. Automatic Hierarchical Classification of Emotional Speech. Dans Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007), Taichung, Taiwan. pp. 291-296. 2007.

[93] A Dimensional Emotion Model Driven Multi-stage Classification of Emotional Speech. Z Xiao, E Dellandréa, W. Dou, L. Chen. Rapport de recherche RR-LIRIS-2007-033, 2007. Submitted to Multimedia Tools and Applications.

[94] Zhongzhe Xiao, « Classification of emotion in audio signals », thèse de l'Ecole Centrale de Lyon, soutenue le 25 Janvier 2008, devant le jury composé de Mme.André-Obrecht Régine (Pr Université de Paul Sabatier, Rapporteur), MM.Laurent Besacier (MdC HDR Université Joseph Fourier, Rapporteur), Jean-Paul Haton (Pr Université Henri Poincaré, Président), Liming Chen (Directeur de thèse), Emmanuel Dellandréa (MdF ECL LIRIS, Co-directeur), Dou Weibei (Pr Tsinghua University, Co-directrice)

[95] M.Hammami, C.Li, B. Ben Amor, C.Vial, « Classification d'images par concepts », Journées francophones sur l'Accès Intelligent aux Documents Multimédias (MediaNet'2002), 17-21 Juin 2002, Sousse, Tunisie, Hermes, ISBN 2-7462-0500-9, pp.380-385

[96] Scherer, K, R., Vocal communication of emotion: A review of research paradigms, Speech Communication 40, pp. 227-256, 2002

[97] Pereira, C., Dimensions of emotional meaning in speech, Proceedings of the ISCA workshop on Speech and Emotion, p 25 – 28, Newcastle, Northern Ireland, 2000

[98] Alain Pujol, Liming Chen, "Coarse Adaptive Color Image Segmentation for Visual Object Classification", Proceedings of the 15th International Conference on Systems, Signals and Image Processing (IWSSIP 2008), Bratislava, Slovak Republic, during June 25 - June 28, 2008

[99] Alain Pujol and Liming Chen, "Line Segment Based Edge Feature Using Hough transform", The 7th IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP), 2007.