Multimedia classification of movie shots using low-level and semantic features

1st Author
1st author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
1st author's email address

2nd Author
2nd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
2nd E-mail

3rd Author
3rd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
3rd E-mail

ABSTRACT

Movie shots categorization may be approached by using audio and visual features for inferring high-level information about a movie shot. Low-level audio and visual features such as color and MFCC and mid-level features such as sky and speech detection have been used in multimedia understanding research. However, integrating all this features in a classifier remains a subject of study. In this paper, we propose a multimedia SVM fusion model, presented in Figure 1, for integrating knowledge from low-level and semantic features extracted from auditory and visual signal for scene classification of movie shots. We also compare our method with common approaches for feature integration based on Bayesian Network. Our computational results show that our model can achieve significantly better and more stable performance than other strategies.

Categories and Subject Descriptors

D.3.3 [**Artificial Intelligence**]: Vision and scene understanding – *video analysis*.

General Terms

Theory, Design, Performance.

Keywords

Multimodal information fusion, statistical modeling, video indexing, SVM, Bayesian network, early fusion, late fusion.

1. INTRODUCTION

Large digital video libraries require tools for representing, searching, retrieving content. One possibility is the query-by-example (QBE) approach, in which users provide (usually visual) examples of the content they seek. However, such schemes have some obvious limitations, and since most users wish to search in terms of semantic-concepts rather than by visual content [1], work in the video retrieval area has begun to shift from QBE to query-by-keyword (QBK) approaches, which allow the users to search by specifying their query in terms of a limited vocabulary of semantic concepts.

Semantic concepts are estimated by classification algorithm. Shot classification commonly means grouping shots into semantically meaningful categories based on the available training data. However, although some work [2,3,4] has been done in this field, little results have been obtained to satisfy user's expectations.

Shot classification is still a challenging and important problem in compute vision recently.

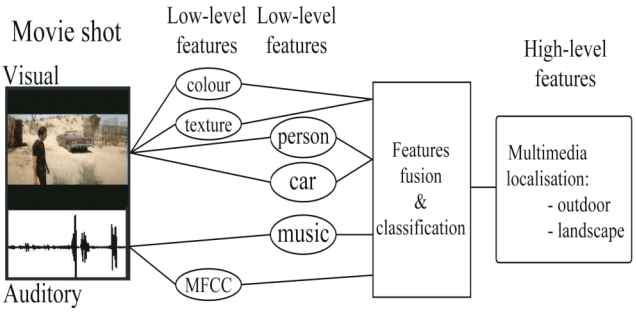

Figure 1. Multimedia shot localization scheme.

The scene categorization, which classifies movie shots into meaningful semantic scenes, is a major issue in movie analysis. Knowledge of the scene type of a movie shot is useful in shot event classification that constitutes a fundamental component of automatic albuming systems [5]. Scene categorization is also valuable in shot retrieval from databases because it provides understanding of scene content that can be used along with color, texture, and shape for database browsing. This semantic representation is also used for movie scene segmentation, which constitutes fundamental components of movie management systems. These systems performs structuring and categorization of the image and audio signal from movies.

The general problem of automatic scene categorization is difficult to solve and is best approached by a divide-and-conquer strategy. A good first step is to consider only two classes such as indoor vs outdoor [6,7], which may be further subdivided into city vs landscape [8,9], etc.

Scene categorization is often approached by computing low-level features from the visual signal, which are processed with a classifier engine for inferring high-level information about the shot [6,8]. One problem with the methods using low-level features in scene categorization is that it is often difficult to

Conference'04, Month 1–2, 2004, City, State, Country.

generalize these methods to diverse shot data beyond the training set. More importantly, they lack semantic interpretation that is extremely valuable in determining the scene type. Scene content such as the presence of people, sky, grass, etc., may be used as cues for improving the classification performance obtained by low-level features alone [7].

In general, two modalities exist in video, namely the auditory and the visual modality. Most of the actual systems only use visual signal for scene classification. But at present, there is enough experimental evidence to state that semantic video analysis yields the most effective index when a multimodal approach is adhered [10,11]. Environmental cues related to place and type of activity, can be utilized to improve classification performance. Environmental sounds and background sounds in places like the office, classrooms, streets, train stations and cafes can be a rich source of information for inferring scene types.

Using low-level and semantic features from auditory and visual signal for inferring high-level information may thus approach movie shots categorization. One of the issues when dealing with a diverse set of features is how to integrate them into a classification engine. Pioneering approaches for features fusion focused on indexing of specific concepts only, e.g. [12]. In these cases a rule-based combination method yields adequate results. Drawbacks of such approaches, however, are the lack of scalability and robustness. To cope with both issues, a recent trend in semantic video analysis is generic indexing approaches using machine learning [10,11].

In this article we propose a SVM approach for the integration of low-level features and semantic features from auditory and visual signal of movie shots. This approach improves the classification performance over using visual low-level features alone.

The rest of the paper is organized as follows. In the next section, related work on feature fusion is reviewed. The multimedia feature set used for inferring shot categorization is presented in Sect. 3. In Sect. 4, we present a comparison of the two principal classifiers used for feature fusion, namely Bayesian network and SVM. In Sect. 5, we introduce and compare two general schemes for feature fusion, early and late fusion. In Sect. 6, we determine the best fusion model by experiments over real-world movie shots, and some potential application domains of the proposed strategies are outlined. Concluding remarks are given in Sect. 7.

2. RELATED WORK

Query using keywords representing semantic-concepts has motivated recent research in semantic media indexing [13,14, 7, 8]. Recent attempts to introduce semantics in the structuring and classification of videos include [16,17].

Naphade et al. [13] present a novel probabilistic framework for semantic video indexing by learning probabilistic multimedia representations of semantic events to represent keywords and key concepts. They define probabilistic multimedia objects (multijects) to map low-level media features to semantic labels. A graphical network of such multijects (multinet) captures scene context by discovering intra-frame as well as inter-frame dependency relations between the semantic concepts. The authors place all the concepts on the same semantic level. They can belong to different basic categories: objects (car, man, helicopter), place (external, beach), or events (explosion, man which goes), and they are connected by a statistical relation to the low-level features they are associated. Intuitively it is clear that the presence of certain multijects suggests a high possibility of detecting certain other multijects. Similarly some multijects are less likely to occur in the presence of others. The detection of *sky* and *water* boosts the chances of detecting a *beach*, and reduces the chances of detecting *Indoor*. It might also be possible to detect some concepts and infer more complex concepts based on their relation with the detected ones. Detection of human speech in the audio stream and a face in the video stream may lead to the inference of *human talking*.

The disadvantage of the multinet is that it must consider the relations between all the concepts of the ontology. In our case the significant number of concepts and features makes computing time prohibitory. It is thus necessary to eliminate certain irrelevant relations between concepts as between *man* and *sky* for example. A solution is to create a hierarchy to explicitly represent semantic-concepts using a basis of other semantic-concepts.

[18] begins by assuming the a priori definition of a set of *atomic* semantic-concepts or mid-level features (objects, scenes, and events) which is assumed to be broad enough to cover the semantic query space of interest. By atomic mid-level features, they mean concepts such as sky, music, water, speech, and so forth, which cannot be decomposed or represented straightforwardly in terms of other concepts. Concepts that can be described in terms of other concepts are then defined as high-level concepts. Clearly, the definition of high-level concepts depends, to some extent, on the variety of mid-level concepts defined. Note that these concepts are defined independently of the modality in which they are naturally expressed (i.e., an atomic concept can be multimodal and a high-level concept can be unimodal etc.).

The challenges are then: Firstly, high-level concepts must be linked to the presence (or absence) of other concepts (either within a modality or across) and statistical models for combining these concept models into a high-level model must be chosen. Secondly, cutting across these levels, information from multiple modalities must be integrated or *fused*. Fusion could occur at various levels: low-level features, within atomic concept models, or by combining several atomic-concept models within a multimodal high-level concept models.

The first challenge is considered by assuming a priori definition of the set of low-level and mid-level features relevant to the higher-level concept. Then retrieval of the high-level concepts is a multiclass classification problem. It is amenable by the modeling of class conditional densities with Bayesian network [17,19,20] or more discriminative techniques such as SVMs [18].

For the second challenge, we identify two general fusion strategies within the machine-learning trend to semantic video analysis, namely: early fusion [21] and late fusion [10,11,22,41]. In the early fusion model, after analysis of the various unimodal streams, the extracted features are combined into a single representation before classification. The late fusion learns semantic high-level concepts directly from unimodal features. Then scores from auditory and visual classification are combined to yield a final multimedia classification. The question arises whether early fusion or late fusion is the preferred method for semantic multimedia video analysis.

In this paper we present a comparison of multiple fusion models based on these techniques. The analysis of the particular dependence relations between low-level and mid-level semantic features, on one hand, and between auditory and visual features, on the other hand, will permit to select the best fusion/classification model for the multimedia localization of movie shots. In the next section we present the set of features relevant to the high-level localization concepts

3. DATA EXTRACTION

One of the keys of any content analysis algorithm is the type of features employed for the analysis. These features must be able to discriminate among different target scene classes. Some of them may be designed for specific tasks, while others are more general and may be useful for a variety of applications. In this section we remind the definition of a movie shot, then we review the features used for our classification task.

3.1 Shot segmentation

The apparition of shots is leaned to the invention of camera. A shot is a video sequence that consists of continuous movie images for one camera action. We extend this definition to include auditory signal by supposing that a multimedia shot is the images sequence between two visual transitions (cut, fade…) associated with the synchronous auditory signal.

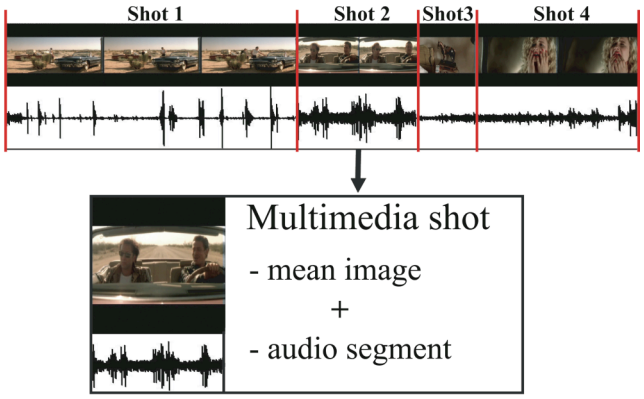

Figure 2. Movie shots segmentation.

Although we consider that shots are multimedia as they convey auditory and visual signal, movie shots are segmented using image alone. Indeed, the auditory signal is often continuous along two consecutive shots. After the seminal work of Nagasaka and Tanaka in 1990 [23], work has been done for detecting shot boundaries in a video flow. Many researchers [24,25] have focused on trying to detect the different kinds of shot transitions that occur in video. The TREC video track 2001 [26] compared different temporal segmentation approached, and if the detection of cuts between shots is usually successful, the detection of fades does not achieve very high success rates.

For our use, a standard algorithm was developed by [27]; it realizes shots segmentation by local maxima detection of an observation function. This function is based on a wavelet transform of the colour and luminosity of images within the film. The algorithm is optimized to over-segment the signal, nearly every transition is detected as recall=98% and some false alarm are spotted as precision=86%.

The audio features describing a shot are extracted from the integrality of the auditory signal from a shot.

As far as image is concerned, we remarked that extracting the set of visual features from each image or frame of a shot would make the processing time prohibitory. In addition, we suppose that content within a shot is homogeneous. This observation results in supporting a treatment of segments of images rather than a treatment image by image. Many approaches thus reduce the problem of the content extraction of a segment to the features processing of only one image per considered segment. Several algorithms were developed in order to determine the average image or summary of an image sequence [28]. It is also possible to extract statistics (e.g.: mean, standard deviation) from the low-level features of the images of the sequence [29]. But this technique supposes features extraction from every image, which is time consuming. Thus we choused to extract from each shot a "mean" image. This image is the local minima of the observation function within a shot: it is the closest image (visually speaking) from its neighbours images in the shot. Content features are extracted from this image, which, we suppose, contains enough information about the shot it summarizes. The Figure 2 presents the multimedia shots segmentation.

3.2 Low-level features

3.2.1 Auditory low-level features

Auditory low-level features should be selected so that they are relevant for the specific scene classification task. Environmental sounds provide many contextual cues that enable us to recognize important aspects of our surroundings. We selected cepstral coefficients features, as they are known to be a good signature of environmental sounds [28]. They express the energy distribution of sounds in a mel-scaled frequency space.

The feature vector **Laud**, containing the cepstral coefficient, is the final result of the low-level auditory features extraction stage.

3.2.2 Visual low-level features

Visual low-level features are selected so that they describe the physical characteristics of places. Multiple low-level features may describe material and structural content of places. Identifying chromaticity and lighting conditions of a place is a good first approach. These characteristics are represented by a basic color-histogram [30]. The texture of the particular elements present in a place may also be a good signature. We use a texture histogram founded on local-edge pattern (LEP). Contours of the image are initially calculated with a 3x3 Sobel filter. After thresholding, a binary image of contours is obtained. Then, for each pixel of this image, the 3x3 window around this pixel is considered. There is 2^9 = 512 possible configurations. To each central pixel the number of the associated configuration is listed. It is then possible to build a 512 components histogram. It was shown that this feature, called TextureLEP, provides good performances in image retrieval by similarity [30]. Thus we expect it to well reflect the signature of places.

The feature vector **Lvis**, containing the color and TextureLEP histograms, is the final result of the low-level visual features extraction stage.

3.3 Semantic features

3.3.1 Auditory semantic features

We believe that the presence of certain types of audio signal may give some indication on the localization of the action in films. Thus, classification of audio in speech, music, ambient-sounds, silence may present an interest for classifying shots.

These semantic features are extracted by classifying a set of selected low-level features containing zero crossing rate, 4 Hz modulation and others, noted **LMaud**. In the simplest form, we model a mid-level semantic concept Maud (audio type here) as a set of class conditional probability density function over a feature space. For the semantic-concept *audio type* and a feature observation, we choose the label as the set of classes conditional density knowing the observed features, stored in the vector **Maud**. The concept is then represented in a semantic space by the relevance rates of its classification estimated by the classifier.

Each dimension of the vector expresses the confidence in the automatic classification of a shot in the selected class *Ci*.

$$\boldsymbol{M_{aud}} = \{P(LM_{aud}|M_{aud}=C_i)\}_{i=1..n} \quad (1)$$

A large variety of supervised machine learning approaches exists to learn the relation between a concept M and pattern L. For our purpose, the method of choice should handle typical problems related to semantic video analysis. Namely, it must learn from few examples, handle unbalanced data, and account for unknown or erroneous detected data. In such heavy demands, the Support Vector Machine (SVM) framework [31] is a solid choice. We convert the SVM output using Genoud's method [32] to acquire a measure in the form of conditional probabilities used here.

3.3.2 Visual semantic features

Semantic visual features that describe the scene set of an action may also indicate the localization of a shot. Scene content such as the presence of people, sky, grass, etc., may be used as cues for improving the classification performance obtained by low-level features alone [19].

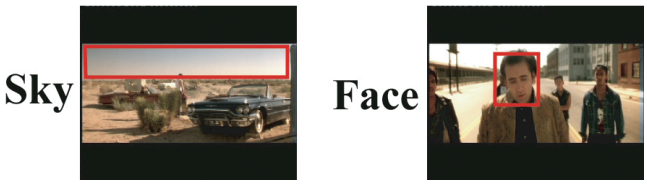

Figure 3. Texture and object detections.

For object detection in images, we use the Adaboost classification scheme developed by [33]. The author describes a visual object detection framework that is capable of processing images extremely rapidly while achieving high detection rates. The system yields faces and objects (cars, artificial lighting) detection performance comparable to the best previous systems.

The identification of particular texture in the background (sky, grass, building, snow, sand, dirt, buildings) in shots is also used [34]. First homogeneous background regions are segmented by a image segmentation algorithm based on hierarchical LPE defined in [35], Then regions are classified using SVM classifier learned on a texture database.

Both identification algorithm shows performance of 80% of detection rate, which is largely sufficient for our application. Figure 3 presents an example of texture and object detections.

These semantic features are extracted by classifying a set of selected low-level features. We model a semantic-concept (car, sky) as binary variable equal to 1 when concept is present and 0 when absent. Each concept is represented by the conditional probability density function that it is present over a feature space. For each semantic-concept Mvis,i and a feature observation LMvis,i, we choose the label as the class equal present conditional density knowing the observed feature. Each concept is then represented by the relevance rate of its classification estimated by the classifier. Then the scores are stored in the vector **Mvis**. Each dimension of the vector expresses the confidence in the associated concept being present.

$$\boldsymbol{M_{vis}} = \{P(LM_{vis,i}|M_{vis,i}=1)\}_{i=1..n} \quad (2)$$

For both classifier the Adaboost and SVM we convert the output using Genoud's method [32] to acquire a measure in the form of conditional probabilities used here.

3.4 Fusion of low-level and semantic features

For both auditory and visual signal, low-level and semantic features are concatenated, so that the segmentation algorithm may analyze the set of features. We obtain one feature vector per shot. Features used here come from different extraction models. They describe multiple physical characteristics and media. Therefore, it is necessary to apply standardization: for each dimension of the vector the mean is set to 0 and the deviation is set to 1. This normalization gives the same weight to every physical characteristic in the classification process.

4. CLASSIFICATION ALGORITHM

Semantic indexing in movie is perceived as a pattern recognition problem. Given a shot S the aim is to obtain a measure that indicates the value of a high-level semantic concept H (e.g. H=indoor or outdoor). To obtain a pattern representation from multimodal movie we rely on feature extraction and classification. Bayesian network and support vector machine differ in the way they classify the results of the features extraction.

4.1 Classification scheme

Our assumption is that high-level semantic concept may be inferred by both low-level and mid-level features. We start our discussion by the simplest case shown in Figure 4.

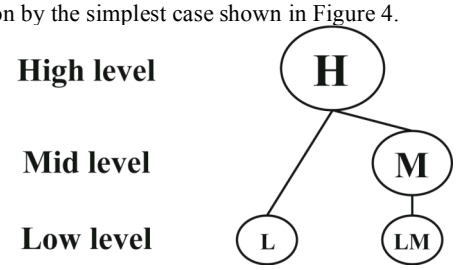

Figure 4. Features relation in the classification scheme.

Suppose a high-level semantic concept H (e.g. H=indoor/ outdoor). Suppose that this concept may be inferred by one low-level feature L (e.g. L=color histogram) and one medium-level concept M (e.g. M=car=1/0). The value of M is obtained by the classification of a set of low-level features LM. Then the hierarchical schema shown in Figure 4 represents the dependence relationships between all this features. The aim of the classification system is to estimate the value of semantic concepts, the mid-level (car=1/0) and the high-level (H=indoor/outdoor) from the set of observed low-level features.

The challenge of the classification scheme is then to take advantage of the dependence relation between the content features: (1) Mid-level and high-level concept are connected by a particular statistical relation. Intuitively it is clear that the presence of a car suggests a high possibility of detecting outdoor (down-up relation). Similarly the detection of a car is less likely to occur in an indoor scene (up-down relation). The relation between M and H is bidirectional. (2) As both low-level and mid-level feature influence the classification of the high-level concept it is important to use the correlation between low and mid level features in the mixed feature space

Next we will discuss the ability of both Bayesian network and SVM to consider these relations in their classification.

4.2 Bayesian network

Bayesian or belief networks provide an effective knowledge representation and inference engine in artificial intelligence and can be used in a variety of media understanding applications [17,19,20,36].

4.2.1 Generalities on Bayesian network

Bayes networks (BN) are directed, acyclic graphs that encode the cause-effect and conditional independence relationships among variables in the probabilistic reasoning system [37]. The network structure can easily incorporate domain-specific knowledge and a complicated joint probability distribution can be reduced to a set of conditionally independent relationships that are easier to characterize. Thus, a Bayes network can be used to represent the dependence relationships between various features that are represented by random variables at the nodes of the network. The directions of links represent causality and the links between the nodes, or variables, represent the conditional probabilities of inferring the existence of one variable (destination) given the existence of the other variable (source). Probabilistic reasoning uses the joint probability distribution of a given domain to answer a question about this domain. According to Bayes' rule, the posterior probability can be expressed by the joint probability, which can be further expressed by conditional probability and prior probability. With Bayes networks, the computation of the joint probability distribution over the entire system given partial evidences about the state of the system is greatly simplified by using Bayes' rule to exploit the conditional independence relationships among variables.

Bayesian networks offer several advantages: explicit uncertainty characterization, fast and efficient computation, and quick training. They are highly adaptive and easy to build, and provide explicit representation of domain-specific knowledge in human reasoning framework. We found that for our applications, Bayes networks offer good generalization with limited training data, easy maintenance when adding new features or new training data, and convenience in building performance-scalable versions by pruning features.

4.2.2 Bayesian network integration

The hierarchical relation between features in our classification scheme may be represented by the Bayesian network structure shown in Figure 4. The network integrates low-level, mid-level and high-level features. The joint probability function encoded by this Bayesian network is:

$$P(H,L,M,LM)=P(H)P(L|H)P(M|H)P(LM|M) \quad (2)$$

The general classification of the shot S consists in allotting the values of the concepts M and H which maximize the joint probability of observing these concepts and the low-level features. This probability may be expressed by the product of three terms. (1) $P(H)$ is the marginal probability of observing one of the class. This term may be estimated by counting the ground truth examples, here we prefer to set it equal for each class from the considered high-level concept (e.g. $P(H)=0.5$ for *indoor-outdoor*). (2) The second term is the conditional probability of observing the low-level features attached to H, knowing the value of the concept H. This term represents the weight of the low-level features in the global classification. (3) The third term is the produce of the joint probability of observing the mid-level feature M knowing H and the joint probability to observe the low-level feature attached to M knowing M. This term represents the weight of the mid-level feature M in the classification of H.

We will know discuss the ability of the Bayesian network to consider the dependence relation between the content features: correlation and hierarchical relation.

The correlation between the low-level and the mid-level features is expressed by the product of the second and third term. In this product each feature type gives its weight for the classification. The observations of features are considered independent knowing H. In the facts the independence assumption of low and mid-level features does not seem valid. The BN overestimates the importance of the dependent features.

A far as hierarchical relation is concerned; the maximization of the global joint probability authorizes the bidirectional relation. The down-up relation is considered as low and mid-level features will influence the estimated value of the high-level concept H, this is our main goal. Moreover, up-down relation is considered, the estimated value of H may influence the value of M. For example, In a indoor scene, a car is detected with a low confidence ($P(LM|M=car\ present)$=0.6). The probability of M=*car present* knowing H=*indoor* ($P(M=car\ present|H=indoor)$=0.02) is low. So, if the probability of detecting H= *indoor* from the low-level features ($P(L|H=indoor)$) is high, the global maximization will estimate the value of M as M=*car absent* even if the low-level features attached to M gives the opposite estimation. This characteristic may be interesting when mid-level classification may be wrong, but it causes errors for some particular scenes as shots from an underground car park.

4.3 Support vector machine

From Bayes classifier to neural networks, there are many possible choices for an appropriate classifier. Among these, support vector machines (SVMs) would appear to be a good candidate because of their ability to generalize in high-dimensional spaces without the need to add a prior knowledge. The appeal of SVMs is based on their strong connection to the underlying statistical learning theory. That is, an SVM is an appropriate implementation of the structural risk minimization method [31]. For several pattern classification applications [18,38,39], SVMs have been shown to provide better generalization performance than traditional techniques such as neural networks and BN [40].

4.3.1 Support vector machine generalities

The classification task involves training and testing data, which consist of some data instances. Each instance in the training set contains one "target value" (class labels) and several "attributes" (features). The goal of SVM is to produce a model, which predicts target value of data instances in the testing set that are given only the attributes.

Training vectors are mapped into a higher (maybe infinite) dimensional space by a learned function. Then SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. The mapping function is estimated through the determination of a kernel function. Though new kernels are being proposed by researchers, beginners may find in SVM books the following four basic kernels: linear, polynomial, radial basis function, sigmoid. For our use, a 20-polynomial kernel will be largely sufficient.

4.3.2 Support vector machine integration

We will know discuss the ability of the SVM to consider the dependence relation between the content features: correlation and hierarchical relation.

The low-level and the mid-level features are the two initial dimensions of the feature space. They are mapped into a higher dimensional space by a learned function. This function is estimated so that classes from H are separated by a hyperplane in the new features space. This particularity permits to consider the correlation between the features, as through the kernel estimation their inter-relation is learned from the examples mapped in the initial feature space. The hyperplane projected in the initial space gives a non-linear separating curve between the classes.

A far as hierarchical relation is concerned, the down-up relation is considered as low and mid-level features will influence the estimated value of the high-level concept H. But SVM prohibits up-down relation: it is a one-way classification.

5. FUSION SCHEME

We perceive semantic indexing in video as a pattern recognition problem. Given pattern (**Maud**, **Mvis**, **Laud, Lvis**) describing the shot S, the aim is to obtain a measure that indicates the value of the high-level semantic concept H. To obtain a pattern representation from multimodal video we rely on feature extraction. Early fusion and late fusion differ in the way they integrate the results from feature extraction on the various features types (low/mid level, auditory/visual). The general schemes for early and late fusion are illustrated in Figure 5.

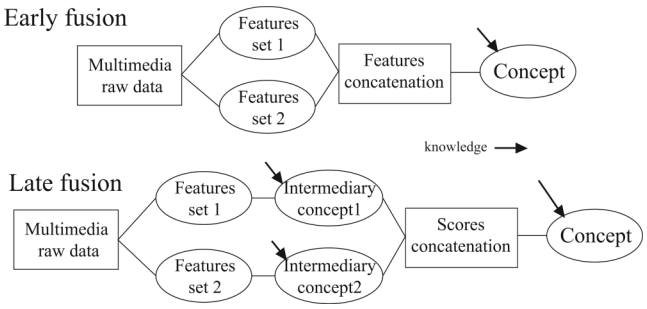

Figure 5. Early and late fusion scheme.

5.1 Early fusion

Indexing approaches that rely on early fusion first extract features. After analysis of the various streams, the extracted features are combined into a single global representation. In [21] author used concatenation of unimodal low-level feature vectors to obtain a fused multimedia representation. The authors of [20] introduce a probabilistic layered framework based on Bayesian network and early fusion that combines low-level and mid-level features for multimedia classification.

After combination of features in a multimodal representation, early fusion methods rely on supervised learning to classify semantic concepts.

5.2 Late fusion

Indexing approaches that rely on late fusion also start with extraction of the whole features. In contrast to early fusion, where features are then combined into a global representation, approaches for late fusion learn intermediary concepts directly from each type of features (low-level/auditory, low-level/visual, mid-level/auditory, mid-level/visual). These scores are combined afterwards to yield a final detection score.

In [41] separate generative probabilistic models are learned for low-level features extracted from the visual and textual modality. In [19] low-level and mid-level visual features are fused by a late fusion scheme to classify image localization. In general, late fusion schemes combine learned intermediary scores into a multimodal score representation. Then late fusion methods rely on supervised learning to classify semantic concepts.

5.3 Early vs late fusion

In [42] the authors compare early and late fusion by experiment on 184 hours of broadcast video data and for 20 semantic concepts. They show that late fusion tends to give slightly better performance for most concepts. However, for those concepts where early fusion performs better the difference is more significant.

An advantage of the early fusion is the requirement of one learning phase only. Disadvantage of the approach is the difficulty to combine features into a common representation.

Late fusion focuses on the individual strength of features sets. Intermediary concept detection scores are fused into a multimodal semantic representation rather than a feature representation. A big disadvantage of late fusion schemes is its expensiveness in terms of the learning effort, as every modality requires a separate supervised learning stage. Moreover, the combined representation requires an additional learning stage. Another disadvantage of the late fusion approach is the potential loss of correlation in mixed feature space, which could be penalizing, as we believe that correlation between features plays an important role in classification tasks.

6. EXPERIMENT

In this section we evaluate the classification and fusion models on a hierarchical localization of a shots in indoor-outdoor, then outdoor shots are classified in city-landscape, and indoor shots in store-office-dwelling-underground. The algorithm is founded on the fusion of low-level (MFCC, color) and semantic (face, car) features from audio and image signal to determine the localization of a shot.

In order to emphasize the performances of each model, we study their performances on all the fusion levels. We start by testing the fusion of the visual low-level features on the classification of the basic scene concept, *indoor/outdoor*. Then we study the fusion of visual low-level and mid-level features on the global hierarchical localization of visual shots. We finish our experiment with the multimodal fusion of low-level and mid-level features extracted from auditory and visual signal of movie shots.

6.1 Evaluation metrics and database

The shots are extracted from 8 commercial movies. The examples of localization concepts have been manually annotated in the corpus of 10.000 shots. The high-level classification algorithms are trained from 7.000 shots. The mid-level audio and visual concepts were already trained from another experience on a large example databases. Therefore, we use the mid-level classification results of training examples for training the high-level classification and not ground-truth. This way, the performance of the algorithm may be shown in a real situation.

We measure classification performance using the confidence in the classification. The confidence is defined as the percentage of objects annotated with the i^{st} class and classified by the model in

i[st] class. We use this strategy as our baseline because of its popularity in the literature.

6.2 Low-level unimodal feature fusion

We first experiment the classification of the high-level semantic concept, indoor/outdoor, using visual low-level features alone. We compare the performance classification and fusion model. In Figure 6, we present the hierarchical dependence relation between features from both low-level early and late fusion scheme.

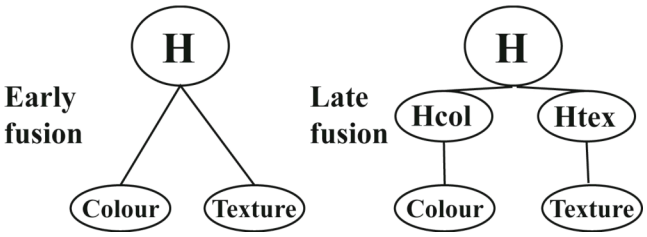

Figure 6. Low-level features fusion scheme.

6.2.1 Results

The confidence in the indoor/outdoor classification of visual low-level features is presented in the table 1.

Table 1. Confidence in the classification of visual low-level features

Classes	BN early	SVM early	SVM late
Indoor	79	**82**	81
Outdoor	83	85	**86**

This table shows that the performances obtained vary for the two classes. The classification of the outdoor shots provides better results than that of the indoor shots for the tests. The badly classified indoor shots correspond to shots whose lights and colors seem to reveal an outdoor shot. They are mainly: shots comprising a window; shots where the top is clearer than the bottom, indicating the potential presence of a "sky"; or shots of large enlightened indoor spaces (as a cathedral or a mall). The badly classified outdoor shots represent, in general, close-ups on objects or landscapes where the sight is blocked by a dark element of the scene set: a wall, a forest. It seems that this kind of pathological cases are more current for the indoor shots than for outdoor shots, which would explain such a difference in the performances observed.

6.2.2 SVM vs Bayesian network

We will now compare both classification algorithm, namely SVM and Bayesian network (BN).

The classification of the low-level features by BN and SVM seems to provide similar results. However, we notice a light advantage for the SVM model of classification (approximately 2%). In the facts the independence assumption of low-level features does not seem valid. The BN overestimates the importance of the dependent low-level features. As SVM better consider the correlation between features, it integrates more information from the features in the classification scheme, and then gives better results.

6.2.3 Early vs late fusion

We will now compare both fusion techniques, namely early and late fusion.

The table reveals that the performances vary little according to the selected model. Within sight of these results and those observed by other studies, we notice that the late fusion does not make significant improvement of the performances of classification within the framework of low-level features fusion. The disadvantage of late fusion, its expensiveness in terms of the learning effort, makes us prefer the early fusion in this framework.

6.2.4 Conclusion

The results from these tests make it possible to conclude on the best model for low-level features classification/fusion: early fusion associated with a SVM classifier. These results will be regarded as allowed for the experiments presented in the following subsections. The examples of this chapter are used as experimental reference marks of the classifications performances used within the framework of shots localization.

6.3 Low-level and mid-level unimodal feature fusion

We then experiment the classification of the whole high-level semantic concepts, using visual low-level and mid-level features. We compare the performances of the classification and fusion models. In Figure 7 we present the hierarchical dependence relation between features from both low and mid-level features early and late fusion scheme.

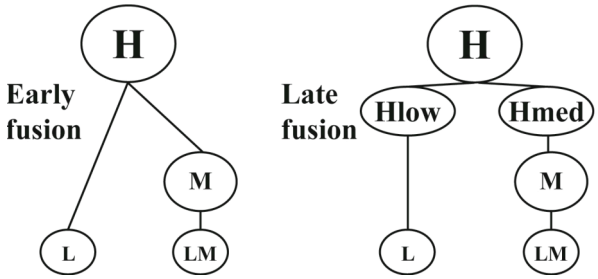

Figure 7. Low-level and mid-level features fusion scheme.

6.3.1 Results

The confidences in the three classifications of visual low-level features alone, mid-level features alone and the fusion of both sets are presented in the table 2.

Several remarks can be made from this table.

Firstly the automatic localization of shots in *indoor/outdoor* by mid-level concepts alone obtains good performances. Compared with those obtained by classification of the low-level features, we note a clear increase of approximately 5%. The mid-level concepts seem to contain more information on classification than low-level features. But the use of these concepts involves an increase in the computing time. However, for certain shots, classification remains complex. We notice for example that an indoor shot of car park containing a car is classified in outdoor, which constitutes an error. That is due to the fact that the majority of the cars are present in outdoor training shots. When the presence of certain concepts strongly indicates the class of the shot, the system makes errors. Also let us notice that the absence of concept in a shot actually gives no indication on the nature of the place. However, in the facts, the two algorithms classify such

an image in indoor. That involves errors for the outdoor shots containing none of the selected mid-level concepts.

Table 2. Confidence in the classification of visual low-level and mid-level features

Classes	SVM low	SVM med	BN early	SVM early	SVM late
Indoor	82	86	87	**89**	87
Outdoor	85	89	90	91	**92**
Indoor Store	85	100	83	**85**	85
Indoor Office	82	0	76	**84**	83
Indoor Dwelling	79	0	60	**78**	76
Indoor Underground	70	22	52	**77**	72
Outdoor Landscape	87	62	86	**86**	81
Outdoor City	92	88	90	**92**	86

Secondly, the classification of indoor shots in *store-office-dwelling-underground* by the concepts obtains bad results. The concepts used describe the presence of outdoor objects (and textures) and a great part of indoor shots does not contain any concept. Thus, they seem inadequate for indoor shots classification. Only some elements of the class *underground* were well classified thanks to the presence of cars in the shot (car park).
Secondly, the classification of outdoor shots in *city-landscape* obtains rather good results; the mid-level concepts employed for classification carry out a good discrimination of the outdoor classes.
Thirdly, the fusion of the mid-level concepts and the low-level features improves the performances of classification of the majority of the concepts of localization (3\% on average), compared to the low-level models. We note that the concepts bring additional and relevant information to the classification models.

6.3.2 SVM vs Bayesian network

We will now compare both classification algorithm, namely SVM and Bayesian network (BN).
Like for low-level features alone, the classification of the set of low and mid-level features by BN and SVM seems to provide similar results. However, we notice a light advantage for the SVM model of classification (approximately 1%). In the facts the independence assumption does not seem valid. The BN overestimates the importance of the dependent features. As SVM better consider the correlation between features, it integrates more information from the features in the classification scheme, and then gives better results. For example, a outdoor shot containing a tree and grass, but whose automatic classification is mistaken and obtains for these mid-level concepts a rate of confidence of 40%, is classified in indoor by the BN and outdoor by the SVM.

6.3.3 Early vs late fusion

We will now compare both fusion techniques of low-level and mid-level features, namely early and late fusion. The classification results bring to several reflections.
For both concepts, *indoor/outdoor* and *city-landscape*, for which mid-level concepts are relevant, the SVM classification models applied to early and late fusion obtain equivalent results. The late fusion shows slightly better performances, compared with the early fusion.
For the localization of indoor shots in *store-office-dwelling-underground*, only the early fusion model does not involve a reduction of the performances of categorization, compared with low-level model. If a dimension of the feature vector is not relevant for classification, it will not influence this one. Thus irrelevant mid-level concepts for the classification are not considered. The concept *car* is the only useful concept for this categorization; its presence indicates that the shot is an underground car park. For the classification in store, office and dwelling, none of the concepts is taken into account and we observe equivalent performances with and without mid-level features. We conclude from it that, when the mid-level concepts provide bad performances of classification, only early fusion improves, to a significant degree, the performances of classification compared to the low–level model. This characteristic brings that, on average, early fusion show better performances of classification than late fusion, in the framework of low and mid-level features fusion.

6.3.4 Conclusion

The results from these tests make it possible to conclude on the best model for low-level and mid-level features classification/fusion: early fusion associated with a hierarchical SVM classifier. These results will be regarded as allowed for the experiments presented in the following subsections. The examples of this chapter are used as experimental reference marks of the classifications performances used within the framework of multimedia shots localization.

6.4 Low-level and mid-level multimodal feature fusion

We then experiment the classification of the whole high-level semantic concepts, using visual and auditory low-level and mid-level features. We compare the performance of the classification and fusion models. In Figure 8 we present the hierarchical dependence relations from both early and late fusion scheme of auditory and visual features.

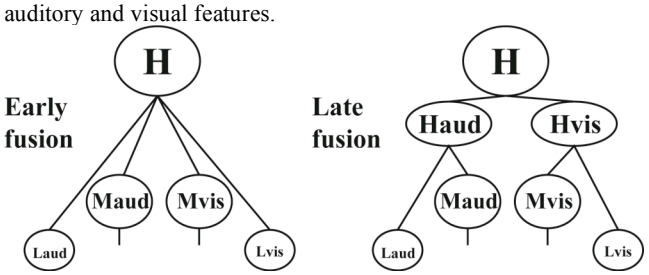

Figure 1. Auditory and visual features fusion scheme.

6.4.1 Results

The confidences in the three classifications of auditory features alone and the fusion of both visual and auditory features are presented in the table 3.
We can deduce several observations from this table.
First of all, the comparison of the unimodal classification results shows that according to the considered high-level concept, the different information brought by each medium involves variable performances. With regard to the concepts *indoor-outdoor* and *city-landscape* the image processing is more effective than that of

the sound. And for the concept *store-office-dwelling-underground* we observe the opposite phenomenon.
Secondly, the collaboration of audio and image for the classification provide better performances compared with the classification of visual features alone (approximately 5%). For the auditory signal the fusion of the low and mid-level features obtains good localization rates, despite some irrelevant mid-level concepts. Thus, the fusion of the sound and the image improves, definitively, the shots localization.

Table 3. Confidence in the classification of visual and auditory low-level and mid-level features features

Classes	SVM audio	SVM visual	SVM early	SVM late
Indoor	88	89	93	**95**
Outdoor	91	91	92	**94**
Indoor Store	**94**	85	88	89
Indoor Office	85	84	87	**90**
Indoor Dwelling	92	78	93	**94**
Indoor Underground	87	77	86	**89**
Outdoor Landscape	89	86	89	**93**
Outdoor City	88	92	93	**95**

6.4.2 Early vs Late fusion

We will now compare both fusion techniques of auditory and visual features, namely early and late fusion. The classification results induce some essential conclusion.
At the opposite of the preceding subsection experiment, early fusion obtains worse performances than the late fusion of the sound and image (1.5\% on average). This is disconcerting because, a priori, early fusion keeps the informative correlations between features and exceeds the late fusion in term of performances. However, we give an explanation to this variation: this model well consider "masking" phenomenon where the image (or the sound) does not carry enough information to identify the place of the action. For example: the image is a close-up on a character face, or the sound contains only music. In general, this phenomenon only involves one media at a shot. The late fusion separately treats the two sets of features, and is thus more effective in these cases. Let us notice that in the fiction movie framework this kind of masking is frequent, due to the tendency to non-realistic editing. This is why this type of fusion is well adapted to films. It would be interesting to test these models for other types of audio-visual documents: documentary, television news, sporting meetings, for which the audio-visual editing is more "realistic" in its representation of the world: the transmitters are often synchronously present in both media. It is possible that in these cases, the late fusion is less effective, because of the loss of correlation between the two media.

6.4.3 Conclusion

The results from these tests make it possible to conclude on the best model for visual and auditory features classification/fusion: late fusion associated with a hierarchical SVM classifier.

7. CONCLUSION

In this paper, we have addressed the problem of multimedia movie shots localization. We proposed the fusion of low-level and semantic features extracted from auditory and visual signal of movie shots for the automatic labeling of high-level semantic-concepts.
Feasibility of the framework was demonstrated for the semantic-concepts of localization as *indoor-outdoor,* first for concept classification using low-level and mid-level information in single modalities and then for concept classification using information from multiple modalities (auditory and visual).
These experimental results, whilst preliminary, suffice to draw two conclusions: first, mid-level concepts information may be added to low-level features to improve the classification performance, second information from multiple modalities (visual and auditory) can be successfully integrated to improve semantic labeling performance over that achieved by any single modality.
There is considerable potential for improving the schemes described for mid-level and high-level concepts classification. Future research directions include the utility of multimodal fusion in mid-level concepts models (e.g. cars, telephone, dogs… detection). Schemes must also be extended to much larger numbers of high-level semantic-concepts. In the framework of movie analysis, we believe that it is possible to label abstract concepts as suspense or love. The model developed here may also be applied to other applications domains as news or sport indexing. Specific events detection as sport highlights (e.g. goals) or news events (e.g. Presidential allocution) should be investigated.

8. REFERENCES

[1] Smith, J.R. and Chang, S.F., VisualSEEk: a fully automated content-based image query system. in *Proc. 4th ACM International Conference on Multimedia*, 1996.

[2] Forsyth, D.A., Malick, J., Fleck, M.M. etc. Finding pictures of objects in large collections of images. In *International Workshop on Object recognition for computer vision*, 1996.

[3] Yu, H.H. and Wolf, W. Scene classification methods for image and video databases. *In Proc. SPIE, Digital Image Storage and archiving Systems*, 1995, 363-371.

[4] Gorlcani, M.M. and Picard, R.W. Texture orientation for sorting photos " at a glance". In *12th Intl conference on Pattern Recognition*, 1994, 459-464.

[5] Loui, A.C. and Savakis, A.E. Automatic Image Event Segmentation and Quality Screening for Albuming Applications. *Proc. Int. Conf. Multimedia and Expo*, 2000.

[6] Szummer, M. and Picard, R.W. Indoor-Outdoor Image Classification, *IEEEInternational Workshop on Content-Based Access of Image and Video Databases, ICCV '98*, 1998.

[7] Paek, S., Sable, C.L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B., Chang, S. and McKeown, K.R. Integration of Visual and Text Based Approaches for the Content Labeling and Classification of Photographs. *ACM SIGIR '99 Workshop on Multimedia Indexing and Retrieval*, 1999.

[8] Vailaya, A., Jain, A. and Zhang, H.J. On Image Classification: City Images vs Landscapes. *Pattern Recognition*, 1998, 1921-1935.

[9] Gorkani, M. and Picard R.W. Texture Orientation for Sorting Photos at a Glance. *IEEE Conference on Pattern Recognition,* 1994.

[10] Amir, A. et al. IBM research TRECVID-2003 video retrieval system. In *Proc. TRECVID Workshop*, 2003.

[11] Iyengar, G., Nock, H. and Neti. C. Discriminative model fusion for semantic concept detection and annotation in video. In *ACM Multimedia*, 2003, 255-258.

[12] Tsekeridou S. and Pitas, I. Content-based video parsing and indexing based on audio-visual interaction. *IEEE Trans. CSVT, 11(4)*, 2001, 522-535.

[13] Naphade, M., Kristjansson, T., Frey, B. and Huang, T.S. Probabilistic multimedia objects (multijects): a novel approach to video indexing and retrieval in multimedia systems," in *Proc. IEEE International Conference on Image Processing, 3*, 1998, 536–540.

[14] Ellis, D. *Prediction-driven computational auditory scene analysis*, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, Mass, USA, 1996.

[15] Barnard, K. and Forsyth, D. Learning the semantics of words and pictures. In *Proc. International Conf. on Computer Vision, 2,* 2001, 408–415.

[16] Wolf, W. Hidden Markov model parsing of video programs. In *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, 4*, 1997, 2609–2611.

[17] N. Vasconcelos and A. Lippman, "Bayesian modeling of video editing and structure: semantic features for video summarization and browsing. In *Proc. IEEE International Conference on Image Processing, 3*, 1998, 153–157.

[18] Adams, W.H., Iyengar, G., Lin, C.Y., Naphade, M.R., Neti, C., Nock, H.J. and Smith, J.R. Semantic Indexing of Multimedia Content Using Visual, Audio, and Text Cues, In *JASP(2003), 2*, 2003, 170.

[19] Savakis, A., Serrano, N. and Luo, J. A computationally efficient approach to indoor/outdoor scene classification. In *International Conference on Pattern Recognition,* 2002.

[20] Jasinschi, R.S., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J., Li, D. and Louie, J.A probabilistic layered framework for integrating multimedia content and context information. In *Porceedings ICASSP*, 2002.

[21] Snoek, C. et al. The MediaMill TRECVID 2004 semantic video search engine. In *Proc. TRECVID Workshop*, 2004.

[22] Wu, Y., Chang, E., Chang, K.C. and Smith, J. Optimal multimodal fusion for multimedia data analysis. In *ACM Multimedia*, 2004

[23] Nagasaka, A. and Tanaka, A. Automatic Scene-Change Detection Method for Video Works. *2nd Working Conference on Visual Database Systems*, 1991, 119-133.

[24] Zabih, R., Miller, J. and Mai, K. Feature-based algorithms for detecting and classifying scene breaks. In *Proceedings of the Third ACM Conference on Multimedia*, 1995, 189-200.

[25] Aigrain, P. and Joly, P. The Automatic Real-Time Analysis of Film Editing and Transition Effects and Its Applications. *Computer and Graphics. 18, 1*, 1994, 93-103.

[26] Smeaton, A.F., Over, P. and Taban, R. The TREC-2001 Video Track Report. *The Tenth Text Retrieval Conference (TREC 2001)*, 2001, 500-250.

[27] Josserand, P. *Detection de transitions à l'intérieur d'une séquence video en vue de son indexation*, Master's thesis, Université du Littoral de Calais, 2000.

[28] Sundaram, H. and Chang, S.F. Determining computable scenes in films and their structures using audio-visual memory models, 2000.

[29] Assfalg, J., Colombo, C., Del Bimbo, A. and Pala, P. Embodying Visual Cues in Video Retrieval. In *IAPR International Workshop on Multimedia Information Analysis and Retrieval,* 1998, 47-59.

[30] Cheng, Y.C. and Chen, S.Y. Image classification using color, texture and regions. *Image and Vision Computing, 21*, 2003, 759–776.

[31] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, USA, 2000.

[32] Genoud, D., Bimbot, F., Gravier, G. and Chollet. G. Combining methods to improve speaker verification decision. In *Proc. International Conference on Spoken Language Processing (ICSLP), 3*, 1996, 1756–1759.

[33] Viola, P. and Jones, M. Robust real-time object detection. *International Journal of Computer Vision*, 2002.

[34] Millet, C. *Génération de sémantiques spatiales de différents niveaux*. Master's thesis, CEA/LIC2M, 2004.

[35] Angulo López, J. *Morphologie mathématique et indexation d'images couleur. Application à la microscopie en biomédecine*. Ph.D. Thesis, École des Mines de Paris, Paris, France, 2003.

[36] Luo, J., Savakis, A., Etz, S. and Singhal, A. On the Application of Bayes Networks to Semantic Understanding of Consumer Photographs," *Proc. ICIP*, 2000.

[37] Pearl, J. *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufmann, San Francisco, 1988.

[38] Wang, Y., Chen, L. and Hu, B. Semantic extraction of the building images using support vector machines. In *Proceedings of First International Conference on Machine Learning and Cybernetics*, 2002, 1608-1613.

[39] Gao D., ,Zhou, J. and Xia, L. Svm-based detection of moving vehicles for automatic traffic monitoring. In *IEEE Intelligent Transportation Systems Conference Proceeding*, 2001, 745-749.

[40] Scholkof, B., Sung, K., Burges, C., Girosi, G., Pogio, T. and Vapnik, V. Comparing support vector machines with Gaussian Kernels to Radial Basis Fuction Classifiers. *IEEE ,Trans. Signal Processing, 45, 11,* 1997, 2758-2765.

[41] Westerveld T. and al. A probabilistic multimedia retrieval model and its evaluation. *EURASIP JASP, (2),* 2003, 186-197.

[42] Snoek, C., Worring, M. and Smeulders, A. Early versus Late Fusion in Semantic Video Analysis. In ACM Multimedia, 2005.

[43] Delezoide, B. *Modèles d'indexation multimedia pour l'analyse automatique de films de cinéma*. Ph.D. Thesis, Université Pierre et Marie Curie, Paris, France, 2006