Semantic Concept Detection from Visual Content with Statistical Learning

Dehong Wang

B. Eng. (Hons.), HUST, P.R.China

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL OF COMPUTING

NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgements

I am deeply indebted to my supervisors, Dr Sheng Gao and Professor Qi Tian, for their guidance and inspiration throughout my graduate career at National University of Singapore. I am truly grateful to Dr. Sheng Gao, for helping me to identify a number of key issues in my research work, and always leaving his door for open discussion. Also I am very grateful to Professor Qi Tian, for his insightful prospective on numerous technical issues. Besides, I would like to thank my co-supervisor Professor Wing-Kin Sung, Computer science, school of computing of National University of Singapore, for his helpful comments during my research work.

Thanks are also due to the I2R media Group, Dr. Qibin Sun, Dr. Changsheng Xu, Dr. Yongwei Zhu, Dr. Lingyu Duan, Mr. Junsong Yuan, Mr. Zixiang Yang, Mr. Shuiming, Ye, Mr. Xianfeng Yang to name a few, for their help and encouragement.

Finally, I would like to express my deepest gratitude to my wife and her parents, my mother, brother and sister, for the continuous love, support and patience given to me.

Contents

Summary

Visual information plays an important role in collections of digital video world. Since humans tend to use high-level semantic concepts while querying and browsing video/image databases, it is critical to develop techniques for semantic concept detection (SCD) from visual content. Generally, there are three level semantic concepts in videos, namely genre, event and object. Genre is the highest level semantic concept to characterize video segments. Object is the lowest level semantic concept to represent a meaningful concept. A good solution to semantic concept detection will facilitate video/image searching, surveillance and authentication, human computer interaction, video skimming and summarization etc.

However recent research works indicate that SCD is difficult and challenging. In this thesis, we pay attention to two main challenges i. e. tremendous variability and uncertainty of the concepts and multi-modality information fusion. Our work consists of two parts, sports news genre identification and SCD in images i.e. automatic image annotation (AIA).

For the former, the challenges are: first, the length of the video shots change greatly and some of them are very short due to the characteristics of news; second, apart from field shots of different games, there are also some non-field shots such as close-up to people which is confusing with field shots. Previous method attempted to catch motion pattern failed in this scenario because the pattern becomes unstable since shot length is too small. Other work attempted to classify shots by features extracted from key frame also didn’t work since temporal pattern are ignored. In this thesis, we proposed a novel feature extraction method to overcome the above shortcomings.

First, two novel features are extracted from frames, the features are sports field color ratio based on pre-determined field colors for specific types of sports, and background motion, ratio consistent with the background in motion; then compact features are calculated to characterize temporal patterns represented by aforementioned features (it is a few sequences for a shot). The advantage of our method is they are extracted from sample frames rather than key frames of a shot, more over they are compact and have some semantic meaning. The effectiveness of the method is demonstrated by our experiments conducted with challenging dataset from TRECVID 2003.

Some challenges of AIA lie in image representation and multi-modality information fusion. Related work either ignores the contextual information i.e. information from neighbor regions or represents the contextual information by complicated models; moreover most work combine different features in a naive way. To meet the above challenges, we proposed a novel automatic image annotation framework and achieved promising results which outperform the state of the art works in two frequently used dataset: Corel CD images and TREC2003 videos. Our contributions can be summarized from two aspects: first, proposed a novel image representation scheme with which an image can be treated as a text document, while the term of the document catch the contextual information effectively, so many text document techniques can be employed; second, proposed two flexible information fusion methods for fusing diverse visual features and multiple modalities.

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays, vast volumes of digital video data are generated in our daily life. How to effectively classify and retrieve the desired information from huge collections of digital video world is being one of the most crucial and challenging problems. In past years, researches on content-based image and video retrieval have been actively deployed in many research communities. There have emerged a lot of successful paradigms for video parsing, indexing, summarization, classification and retrieval [17,28,16,14,9]. Although fruitful results have been achieved in last decade, more challenging problems need to be addressed and overcome in the future.

Most traditional efforts focused on retrieving video content by text annotation and low level features of images. However, they are questioned and challenged as following reasons. Firstly, the cost of manual text annotation is unreasonable expensive when the collections of videos are huge. Secondly, it is difficult to express semantic concept using low level features. Therefore, in order to effectively access the content of image and video data, many problems still need to be addressed and tackled. One of important issue is how to segment, classify and index the image and video data automatically or semi-automatically. Another crucial and challenging issue is how to bridge the gap between the low level features and high level semantic concepts.

Video databases serve as a perfect example of how the acute need for tools has severely constrained the use of multimedia content. A study by Smith and Chang reveals that 95% of the queries to their search engine [42] that supports low-level matching were semantic and key-word-based. Therefore, it is a urgent need to conduct research on semantic concept detection in image/video.

Research in understanding the semantics of image/video will open up several new applications. Multimedia databases can be better accessed if the index generated contains semantic concepts. Surveillance and authentication can definitely benefit from semantic analysis. Filtering of multimedia content can enable automatic rating of Internet sites and restrict access to violent content. Semantic understanding could mean better and natural interfaces in human computer interaction. Very low bit-rate video coding, video skimming, summarization, and transcoding are among the several applications that could benefit from semantic multimedia analysis.

1.1 Problem definition and challenges

As Figure 1-1 shows, the content of video can be semantically divided into three levels [43], the highest level is genre, which can themselves in turn be made up of genre. For example a sports program includes several kinds of sports games, basketball, ice hockey, baseball etc. The genre of a given video can be, and is often contested by reviewers or journalists. The determination of a genre is made by viewing the video content and often comes down to subjective views and semantic subtleties.

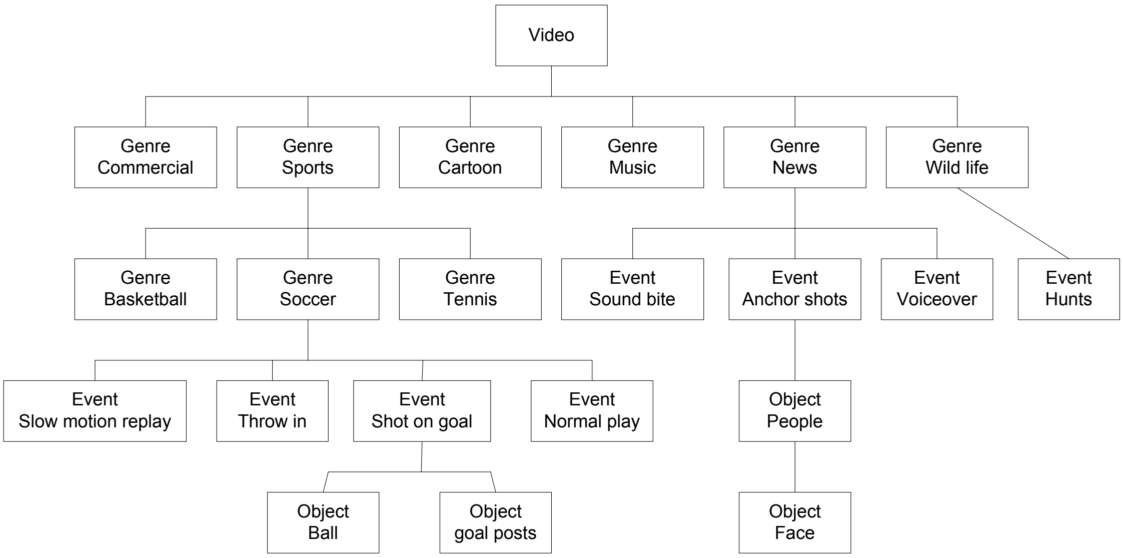


Figure 1-1 Semantic concepts of video

The second level is events. Events are made up of objects and are defined by the objects interactions and interrelations over a finite period of time. Event detection approaches in general add complexity to the feature extraction process to determine the more specific nature of events when compared with genre classification.

Finally, each event is shown to be made up of a number of objects. Objects are conceptually the lowest level of classification that can affect the semantic meaning of the video content.

Usually, semantic concepts in video include genre, event and object. Many works have been conducted in detecting semantic concepts from video. The most famous activity is high-level feature extraction in TRECVID [56]. 17 semantic concepts are defined in this task in 2003. They are: Outdoors, News subject face, People, Building, Road, Vegetation, Animal, Female speech, Car/truck/bus, Aircraft, News subject monologue, Non-studio setting, Sporting event, Weather news, Zoom in, Physical violence and Person x. 10 concepts are defined in 2004, they are boat/ship, train, beach, road, Bill Clinton, Madeleine Albright, Basketball scored, airplane takeoff, people walking/running, physical violence. The task is as follows: given the feature test collection, the common shot boundary reference for the feature extraction test collection, and the list of feature definitions, participants will return for each feature the list of at most 2000 shots from the test collection, ranked according to the highest possibility of detecting the presence of the feature. Each feature is assumed to be binary, i.e., it is either present or absent in the given reference shot.

Semantic concept detection in TRECVID and automatic image annotation or automatic image captioning [21] is very similar tasks if we ignore the temporal-related features in video (e.g. temporal structure and motion), both of which are to decide which concepts are related to a given shot (the former) or an image. Automatic image annotation (AIA) is a task to automatically assign some keywords from a predefined set to an image based on its content. Generally these keywords are the semantic descriptors for the image content; it is obviously that each keyword represents a semantic concept. Therefore, assign keyword to the image is equivalent to detect a concept in the image.

Recognizing class of objects is one of the fundamental challenges in computer vision. Some people address general object recognition [8] problem by defining the object as a semantic concept. From this point of view, object recognition is a semantic concept detection problem in computer vision. However, many researchers conduct their research work in a constrained data set [6,11,13,27,46,48,52,53]. Usually the object to be recognized is the main part of the image, the background is either very simple or with a little cluttered. Although research work on object recognition also considers object occlusion and unknown location in training data, in general, object class recognition can be looked as a special case of semantic concept detection in constrained dataset and more importantly is a binary classification problem, say, for an image only object and background are considered.

Research work conducted on above area indicate that there are some challenges to be met, which can be summarized from three main issues that need to be tackled in design a semantic concept detection system, namely representation, detection and learning. The first challenge is coming up with models that can capture the 'essence' of a concept, i.e. what is common to the concepts that belong to it, and yet are flexible enough to accommodate concept variability (e.g. presence/absence of distinctive parts such as mustache and glasses, variability in overall shape, changing appearance due to lighting conditions, viewpoint etc). Because there is a tremendous variability and uncertainty of the concepts, and also in most of the training data the location of the concept is not given, people represent the concept with features extracted from a whole image. However, to make certain the features carry sufficient information and the learning method tractable, capturing the contextual information from different part of the image is critical.

The challenge of detection is defining metrics and inventing algorithms that are suitable for matching models to images efficiently. Usually we extracted high dimensional feature from images, the curse of dimensionality make feature selection indispensable. And also various feature will be extracted, how to fuse them is also a very challenging problem.

Learning is the ultimate challenge. If we wish to be able to design visual systems that can detect, say, 1000 concepts, then effortless learning is a crucial step. This means that the training sets should be small and that the operator assisted steps that are required (e.g. label the image regions etc) should be reduced to a minimum or eliminated. So we have to face the challenge of insufficient training data and the inconsistency between training data and test data. We also need to consider use part of incomplete labeled data or even unlabeled data for training, which is hot issue in machine learning and have no good solution so far.

1.2 Objectives

The problem of detecting semantic concept from visual content is high challenging; it involved video/image representation, feature extraction and fusion and machine learning. To make the problem more focused, we conduct our research on two parts, namely sports news genre identification and automatic image annotation.

For the former, the challenges come from the length of the video shot change greatly and some of them are very short. Previous method attempted to catch motion pattern failed in this scenario because the pattern becomes unstable since shot length is too small. We need to explore a new feature extraction method to overcome the above shortcomings. For the latter, challenges of AIA lie in image representation and multi-modality information fusion. Related works either ignore the contextual information i.e. information from neighbor regions or represent the contextual information by complicated models; A new method to represent image will be explored which can effectively capture and process the contextual information.

To sum up, efficiently capturing temporal pattern for video or contextual information for image, compactly representing the extracted information and effectively learning the concept model are key points to the solution. In addition, more and more digital images/videos are becoming available over the World Wide Web, but it seems that search engines are still in their infancy. While existing search engines normally retrieve images/videos based on low-level features, users often have a more abstract notion of what will satisfy them. In fact, there is still a big gap between user-based semantic concepts and system-based low-level features. Thus, high-level semantic concepts should make contribution to image/video retrieval in the internet. Hence in this research, we hope to extend our techniques to improve web-based image/video retrieval system to some extent.

So we can summarize the three main objectives of our research as follows:

- *To tackle the problem of genre identification in sports news video, due to the variation of sports games and shot length, this problem is high challenging.*
- *To tackle the problem of semantic concept detection in image with a novel framework, with which can partly meet the aforementioned challenges. Concretely, explore a method to represent the image content so that the context information can be caught and processed effectively;*
- *To explore effective multi-modality information fusion methods so that different type of visual features can be combined flexibly with better performance than state of the art works.*

1.3 Outline of the thesis

This thesis is organized as follows. In next chapter, we survey some work related to semantic concept detection from three aspects namely visual information representation, statistical techniques to concept detection and multi-modality information fusion, the challenges are identified.

In chapter 3, we address sports news genre identification problem by propose a novel feature extraction method which can effectively capture and characterize the temporal pattern and classify sports news video shots into predefined classes.

After that we proposed a novel AIA framework to detect semantic concept in images in chapter 4. We firstly proposed a method to transform an image into text-alike document so various text categorization techniques can be employed to address the AIA problem; then introduced a discriminative multi-class classifier to tackle the problem; after that, two novel information fusion schemes are illustrated followed by all experiments analysis. Finally we concluded the thesis and discussed some future work.

Chapter 2

Related Work

2.1 Visual information representation

The visual features extracted include features for representing color, texture, structure, shape, motion etc. This processing is done in different color spaces such as HSV, RGB [2], YIQ, YUV [19], Lab [54] etc. Color has been represented most frequently with histograms, correlograms and moments with varying bin sizes. Texture has been represented by Gabor texture [44], Tamura, Wavelets [2], etc. Structure has been represented by edge direction histograms [2,18], and edge maps. Shape has been represented by moment invariants, templates etc. Motion has been represented by motion direction and magnitude histograms, optical flow and motion patterns in fixed directions [19]. Visual features have been processed from key frames [2,18,54] only or from all I-frames [19,18] within a shot. Temporally extracted features also include temporal color correlogram and temporal gradient correlogram [34]. Features have been extracted from compressed domain [45] as well as decompressed frames. Visual features have also been extracted at global level, and regional level (segmented automatically or use of regular grids, layouts etc to achieve regional localization). Although various visual features could be extracted from videos and images, in this thesis, we will pay our attention to the static visual features, i.e. visual features extracted from key frames or images.

For semantic concept detection, each image can be represented by a set of continuous visual features [25,12] or discrete symbols [10,20,30,8,22]. Duygulu et al [10] described images using a vocabulary of blobs. First, regions are created using a segmentation algorithm like normalized cuts. For each region, features are computed and then blobs are generated by clustering the image features for these regions across images. Each image is generated by using a certain number of these blobs. In [22], Jeon et al proposed the use of the Maximum Entropy approach for the task of automatic image annotation. Given labeled training data, Maximum Entropy is a statistical technique which allows one to predict the probability of a label given test data. The techniques allow for relationships between features to be effectively captured. In the paper, the authors created a discrete image vocabulary similar to that used in Duygulu et al [10] and [20]. The main difference is that the initial regions they used are rectangular and generated by partitioning the image into grids with fixed size rather than using a segmentation algorithm. Features are computed over these rectangular regions and then the regions are clustered across images. These clusters are called visual symbols to acknowledge that they are similar to terms in language. Using these visual symbols, Maximum Entropy can compute the probability and in addition allows for the relationships between visual symbols to be incorporated. The above models use a discrete image vocabulary. In this vocabulary, an image is tokenized by a set of symbols or a sequence of symbol. Because each region may just be represented by a few symbols, its number of dimensions is far lower than represented by continuous feature, it is easy to process those symbols feature. However, symbols are drawn out by unsupervised clustering, some information may be missed.

A couple of other models use the actual (continuous) features computed over each image region. This tends to give improved results. Correlation LDA proposed by Blei and Jordan [7] extends the Latent Dirichlet Allocation (LDA) Model to words and images. Lavrenko et al. proposed the Continuous Relevance Model (CRM) to extend the Cross Media Relevance Model (CMRM) [20] to directly use continuous valued image features. This approach avoids the clustering stage in CMRM. They showed that the performance of the model on the same dataset was a lot better than other models proposed. Feng et al also claimed that continuous feature works better than discrete features [12]. However, because the dimension of continuous feature is very high, it is very hard to consider the relation between different regions represented by those high dimensional features. In other words, most of work assumed the independence between regions in a same image. This is a weak assumption because the contextual information is very important in representing a concept for a image; moreover, it is very hard to fusion different features because they are all high dimensional.

In these proposed models, some assume the set of features extracted from a set of grid or regions for an image representation is independent [10,20,24,12]. It is a well-known weak assumption, especially for the image. As we know, each grid or region has not sufficient discriminative power, while the contextual relation among the grids or regions can improve its expressive capacity. This contextual relation is embedded in the set of cliques for MRF [8] and the set of hidden states and their transitions for 2D-HMM [25].

2.2 Statistical techniques to concept detection

2.1.1 TRECVID

Most participating groups have approached the concept detection problem as a supervised pattern classification problem and have used different pattern classification and machine learning algorithms. One fundamental distinction that can be made for different groups is whether they approach this as a generic classification problem for classes of concepts or whether they approach this as a problem requiring a special algorithm for every concept. For example some groups use a specialized Face detector while others treat all concepts the same and pass the data through the identical processing pipeline for all concepts with the only difference in training being the ground truth used for each concept. Figure 2-1 tries to capture the common elements of the processing pipeline used by most groups. Figure 2-1 shows a processing pipeline that starts from extracted media features. The next block is the feature-based models, classifiers such as Gaussian mixture models [18,44], support vector machines [1,2,18,54], hidden Markov models [19], fuzzy KNN [54]. The next block is the feature specific aggregation. This is achieved by approaches such as weighted averaging [41,1,2], boosting [54] etc. This step involves combining results over models that are built using the same features but with different parameter configurations or assumptions such as scale,

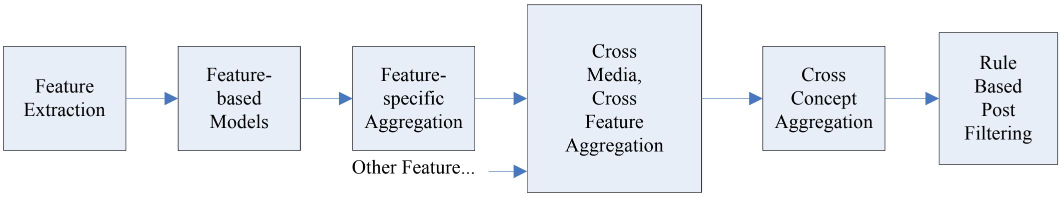

Feature Extraction
Feature-based Models
Feature-specific Aggregation
Other Feature...
Cross Media, Cross Feature Aggregation
Cross Concept Aggregation
Rule Based Post Filtering

Figure 2-1 An abstraction of the common processing pipeline for concept detection [29]

frame etc. The next block shows the aggregation across features and modalities. Apart from approaches identical to the Feature-specific aggregation, However the aggregation here is complicated by the need for synchronization, and for cross feature normalization. The next block is the cross concept aggregation. This is typically the processing module that accounts for inter-conceptual context [29,2], composition of more complex concepts from other primitive models. Then comes the rule based module that is used by several groups for combination of models [33], filtering [2] etc. Different groups may have one or more of these modules in different order. For example the rule based filtering can be the first module [33].

2.1.2 Image annotation modeling

Recently, a number of models have been proposed for image annotation [22,4,7,10,20,24,30,25,8,12]. Up to now, most of works for AIA is to develop a mapping function, which is a joint distribution of the observed visual features and the keywords. Many statistical models have been proposed to learn this joint distribution from the training images. For the AIA task, some proposed models try to directly estimate the joint distribution between the continuous feature or symbols and the keywords. They are the translation model [10], cross-media relevance model (CMRM) [20], multiple Bernoulli relevance model (MBRM) [12], maximum entropy (ME) [22], and Markov random field (MRF) [8]. Other models, such as 2D HMM [25], factor this joint distribution into a conditional distribution on the keywords for the visual feature and the keywords distribution. The first term is easily learned from the training data for a keyword when these keywords are assumed independent. The second term is often assumed uniform in this case. In contrary, the keyword distribution in CMRM, MBRM, and MRF is estimated from the manual labels in the training set. Although the keyword distribution carries some semantic information about the image content, its estimation from the co-occurrence of image and keywords often faces severe data sparsity. So how much benefits can be gotten should be further studied.

Although the proposed models achieve some success, there are still some challenging problems. Here we discuss two problems faced in semantic concept detection. The first is adaptively selecting discriminative features from the various visual features and exploiting their correlations. In the current framework, it is not easy. For example, in [20,24,22,25,12], adding one type of features only means increasing the dimension of the vector or adding another feature specific classifier [2]. They cannot pick out discriminative features for classification and cannot exploit the correlation information among these distinctive features. The second is that the model is often learned without discriminative training except for ME model trained by maximizing conditional entropy [22]. This causes the model much sensitive for mismatch between the training data and testing data. When they matched, good performance can be gotten. Otherwise, the performance is deteriorated. In many real-world applications, mismatch often occurs because it is expensive, even impossible in some cases, to collect large labeled data covering all possible conditions. This is the case we are always facing for concept detection.

2.3 Multi-modality information fusion

In general, multimedia data such as images and videos are represented by features from multiple media sources. Traditionally, images are represented by keywords and perceptual features such as color, texture, and shape. Videos are represented by features embedded in the visual, audio and caption tracks. For example, when detection concept from video, non-visual features were extracted, such as audio features [1,2,18], ASR Transcript Based Features and Video Optical Character Recognition and Metadata [2,18]. These features are extracted and then fused in a complementary way for detecting semantic concept.

Unfortunately, traditional work on multimodal integration has largely been heuristic-based. It lacks theories to answer two fundamental questions:

1) What are the best modalities?

2) How can we optimally fuse information from multiple modalities?

Suppose we extract *l*, *m*, *n* features from the visual, audio, and caption tracks of videos. At one extreme, we could treat all these features as one modality and form a feature vector of l+m+n dimensions, as method 1. At the other extreme, we could treat each of the $l + m + n$ features as one modality, as method 2. We could also regard the extracted features from each media-source as one modality, formulating a visual, audio, and caption modality with *l*, *m*, and *n* features, respectively as method 3. Almost all prior multimodal-fusion work in the multimedia community employs one of these three approaches. But, can any of these feature compositions yield the optimal result?

There are extensive works to study above problems, especially in the task of TRECVID [2,47,50,7,30,22,12,24,55]. Almost all the successful systems in TRECVID apply various fusion methods to exploit the power of every available visual feature (e.g. color, texture, shape) and modality (visual, textual and audio and speech) for improving the system performance [47,2,55].

In method 1, various features are concatenated to construct a high dimensional vector. Then the classifier is trained on it. This method is natural and simple; however, it suffers many weaknesses. The first is the curse of dimensionality. It deteriorates robustness of the classifier, especially in the case of the sparse training samples. The second is that it is difficult to scale the different features for avoiding one of them dominating in the classification. It is not trivial to fuse features from the various scales often occurred. Since all features are intertwined, it is not easy to analyze the contribution of each type of feature. Some people employ PCA or ICA to select optimal features, however, PCA and ICA cannot perfectly identify independent components for at least two reasons. First, like the way the k-means algorithm works, all well-known ICA algorithms need a good estimate of the number of independent components k to find them effectively. Second, ICA only performs the best attempt under some error-minimization criteria to find k independent components. But the resulting components, may still exhibit interdependencies.

In method 3, two–step learning is often used. The first step is to learn a set of classifiers, each corresponding to one type of visual feature or one modality. Since each classifier only handles a few types of features, the affect of the curse of dimensionality is reduced to some degree. Analyzing the contribution of each feature is also easy. Adding another type of feature can be done by training a new classifier and will not affect others, which is very flexible. Each classifier makes its decision and a confidence value will be output. The second step is to fuse multiple decisions or confidence scores for the final decision. The popular fusion scheme is to learn a second classifier on the confidence scores as a feature, e.g.

product combination and weighted sum [15,55]. Other methods are the maximization / minimization product or average [2].

However, for product combination, supposing that modalities are independent of each other, and we can estimate posterior probability for each modality accurately, the product-combination rule is the optimal fusion model from the Bayesian perspective. However, in addition to the fact that we will not have truly independent modalities, we generally cannot estimate posterior probability with high accuracy, product-combination rule is highly sensitive to noise, this strategy is not appropriate. The weighted-sum strategy is more tolerant to noise because sum does not magnify noise as severely as product. Weighted-sum is a linear model, not equipped to explore the interdependencies between modalities. Recently, Yan and Hauptmann [59] presented a theoretical framework for bounding the average precision of a linear combination function in video retrieval. Concluding that the linear combination functions have limitations, they suggested that non-linearity and cross-media relationships should be introduced to achieve better performance. So, Yi Wu et al [55] proposed the super-kernel fusion scheme finds the best combination of modalities through supervised training. or treating all features as one modality.

The above methods have been proved successful in the tasks of TRECVID. However, they lack capturing the multi-category discriminative power of the training sample and features. It is well known that semantic concept detection are a multi-category, multi-label (i.e. an image may have more than one annotations) classification problem. The traditional methods often treat them as multiple independent binary classification problems. They cannot efficiently exploit the multi-category discriminative power.

Chapter 3

Sports news video genre detection

3.1 Introduction

The extensive amount of multimedia necessitates content-based video indexing and retrieval methods. Sports video, due to rich spatial-temporal patterns and having tremendous commercial potentials, has been widely studied. However, published papers seldom cover sports news video genre identification. Recently researchers seldom considered to detect text keywords through automatic speech recognition to address the problem. Because available speech recognition systems are known to be mature for applications with a single speaker and a limited vocabulary. However, their performance degrades when they are used in real world applications instead of a lab environment. This is especially caused by the sensitivity of the acoustic model to different microphones and different environmental conditions. Since conversion of speech into transcripts still seems problematic, integration with other modalities might prove beneficial [43]. In this research we cast our light into detecting genre by visual features. The challenges come from the content varieties of same sports and short shot length. Many works have been done for one kind of sports analysis, including segmentation and shots/scenes classification [60,32,26,37]; These are quite different from sports news shot classification in which shots may come from various sports and also non-sports such as leadin, text caption and people talking.

Other people studied sports genre classification [35,57,40,3] i.e. classifying sports video file into some predefined classes. Considering that different sports often present different motion patterns, some researchers attempt to catch motion patterns from motion vector [9,42,43]. However their methods may not work when the length of the video clip is short since the motion pattern becomes unstable at that time.

On the other hand, Jurgen Assfalg et al [3] classified sports video clip by color histogram of the shot keyframes. However, keyframes may not contain significant part of the field; and because the number of the keyframes is small, they cannot represent the color distribution in all frames of the shot. Xavier et al [57] use dominant color of each frame to replace the color histogram. However, they only pick one color for each frame, so they may not differentiate sports with the same field color, for instance, golf and baseball. Moreover, the introduction of "do not care" color makes some shots such as pitching in baseball unclassifiable.

In addition, some of the above papers indicate that classification accuracy can be improved by filter non-field shots [35,3]. Assfalg et al [3] proposed to differentiate sports field shot with player shots and audience shots using edge features, Drew D. Saur et al [35] proposed a method to differentiate wide-angle from close-up by computing camera motion parameter and intra-macroblock in a P frame. However their method may not work in classifying close-up shots because the former make too strict background assumption for close-up shots and the latter assumes that camera moves in wide-angle shots.

In this chapter we proposed two kinds of features to address the above problems. The first is three field color ratio namely yellow (for basketball), green (for baseball and golf), and white (for ice hockey). The second is background motion and consistency motion ratio; the latter is the ratio of inner MB (Ref. Figure 3-3) whose motion is consistent with the background. We compute these two features once for every four frames. Based on them, a 11-dimension feature vector is calculated for each shot; using them, a decision tree is employed to classify the sports news shots.

To demonstrate the effectiveness of our methods, we select TRECVID 2003 dataset to conduct our experiment. Basketball, ice hockey, baseball and golf, which compose of 90% of sports field shots in CNN headline news, are four predefined classes of field shots. Other three classes are: leadin, which is the fixed introduction and ending of the sports news; text, which is text caption shots; and non-field sports shots including close-up to people, surroundings of the field and audience.

3.2 Feature extraction

3.2.1 Color features

For each sports type, the color of the playing field is either fixed or it varies in a small set of possibilities [3]. The most common field colors in sports video are green (golf, baseball, etc) , brown/yellow (basketball, volleyball, etc), and white (ice hockey etc). Therefore, we define these three field colors at the current stage. After frame decoding, each pixel in the frame is represented by 24-bit colors. We quantize the 24-bit colors into standard 256 colors. Then, by learning from some example areas, three color sets are learned, namely yellow set (denoted by Y), green set (G) and white set (W). In our experiment, there are 20 elements in Y, 12 in G, and 8 in W.

When a frame is quantized into 256 colors, representing it with *C*(*i*, *j*); let *Y*(*i*, *j*), *G*(*i*, *j*), *W*(*i*, *j*) denote the binary mark of yellow, green and white color respectively, where (*i*, *j*) is the coordinate of the pixels in the frame. As a example, we give yellow mark formula:

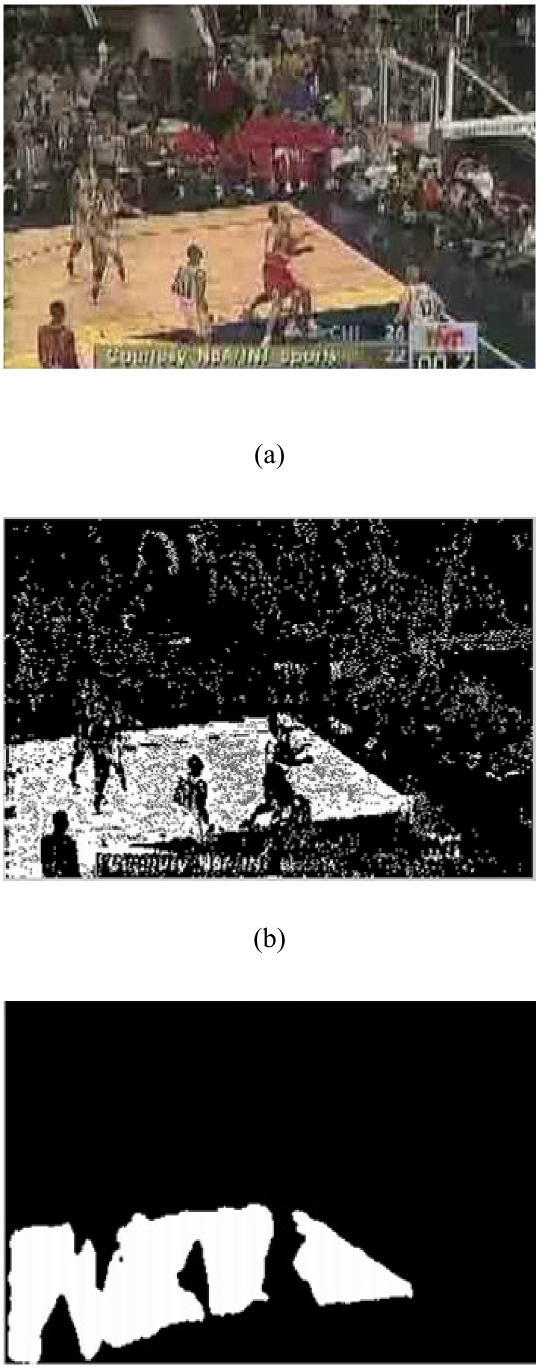
$$Y(i,j)=\begin{cases}1 & C(i,j)\in Y\\ 0 & C(i,j)\notin Y\end{cases}$$

Since most of field shots are wide-angle or middle-angle shots, we can remove other non-field pixel with yellow in field shot by conducting morphological operation. Figure 3-1 shows an example. (a) is an original frame of a basketball game and Figure (b) shows the yellow binary mark before morphological operation. The last result is given in (c).

Then the yellow field color ratio can be calculated as follows,

$$y\_ratio=\sum_{i=1}^{W}\sum_{j=1}^{H}Y(i,j)/(W*H),$$

where *W* is the width of the frame, *H* is the height of the frame. Similarly, by replacing *Y*(*i*, *j*) with *G*(*i*, *j*) or *W*(*i*, *j*). Three field color ratios are computed once every 4 frames; let it be a column vector (3x1). So for a shot we can get a matrix (3xm), m=[n/4], where n is the number of frames in the shot. Figure 3-2 shows some typical examples of sports news shot. The ratios of leadin shot varied



(a)

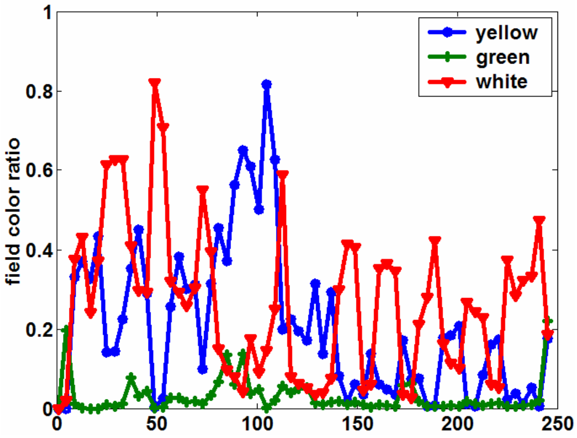
(b)

(c)

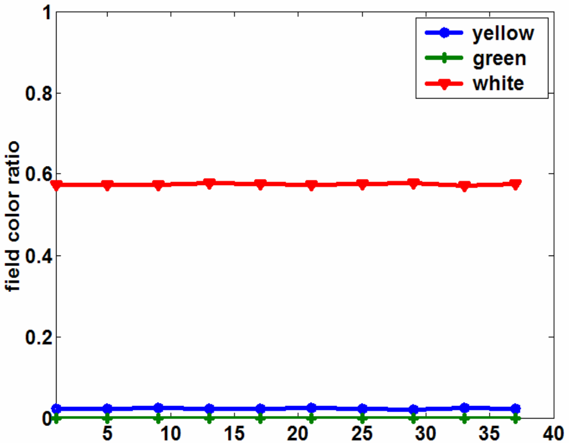
Figure 3-1 Yellow mark extraction binary result, (a) Basketball frame (b) Initial yellow mark (c) Final yellow mark color features of a shot

we can get green color ratio *g_ratio* and white color ratio *w_ratio* too.

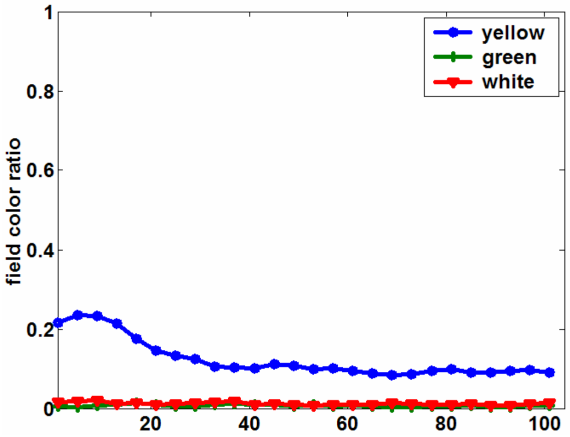
greatly and frequently; on the contrary, ratios of text are very stable. Yellow is the largest ratio color in basketball, but value is about 0.15


1
0.8
0.6
0.4
0.2
0
field color ratio
0
50
100
150
200
250
yellow
green
white

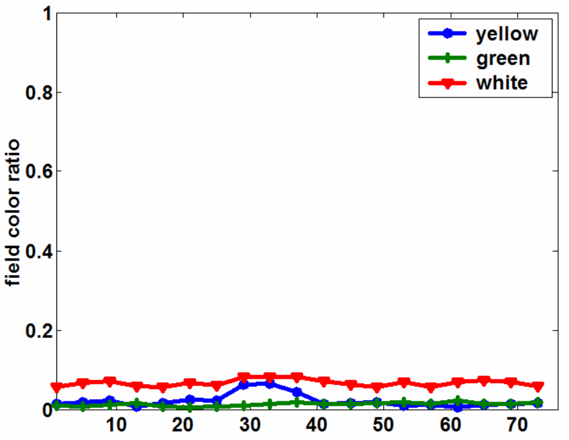
(a)


1
0.8
0.6
0.4
0.2
0
field color ratio
5
10
15
20
25
30
35
40
yellow
green
white

(b)


1
0.8
0.6
0.4
0.2
0
field color ratio
20
40
60
80
100
yellow
green
white

(c)



(d)

Figure 3-2 field color ratios of typical shot. X-axis is frame number
(a) leadin (b) text (c) basketball field (d) non- field shot

when the field is small part of view. No ratio is great than 0.1 in some non-field shot.

3.2.2 Motion features

We extract motion features from P frame in compressed MPEG video streams. We observed that in most of the close-up shots of sports news, movement of foreground is often quite different from that of background; even a slight movement also causes big differences in motion vector. On the other hand, to differentiate some baseball shots with golf shots, background motion is a critical factor. For example, passing ball in baseball field shot also has very high green ratio like most of the shots in golf. But, in order to track the ball, the camera in baseball shot moves quickly which is quite rare in golf shots. Therefore, we proposed a method to extract a new motion feature as follows.

As Figure 3-3 illustrates, we partition macroblocks in a frame into 4 parts. Apart from the lowest 3 rows (Ads. MB) which always show some non-relevant information in CNN headline news and the periphery macroblocks, the neighbor of the periphery macroblocks are background MB because most of the time, most of them belong to the background. The remains are inner MB.

Put all the background MB motion vectors into a set, discard the outlier, we compute the average of remain motion vectors in two directions and denot them as the background motion vectors (mvx_{bg}, mvy_{bg}).

Then the motion consistency ratio of inner MB is computed as follows,

$$mv\_ratio = \sum_{i=1}^{W\_inn} \sum_{j=1}^{H\_inn} cons(i,j) / (W\_inn * H\_inn)$$

$$cons(i,j) = \begin{cases} 1 & if\ non-IntraMB\ and\ \ D < TH \\ 0 & otherwise \end{cases}$$

$$D = [mvx_{inn}(i,j) - mvx_{bg}]^2 + [mvy_{inn}(i,j) - mvy_{bg}]^2$$

where $mvx_{inn}(i,j)$ is the motion vector of an inner MB whose position is the i-th row and the j-th column of the inner MB area. IntraMB is a kind of macroblock which is not coded through motion compensation. *TH* is a threshold, which is set to 8 in out experiment. *W_inn* and *H_inn* are the width and the height of the inner MB area in unit of macroblocks.

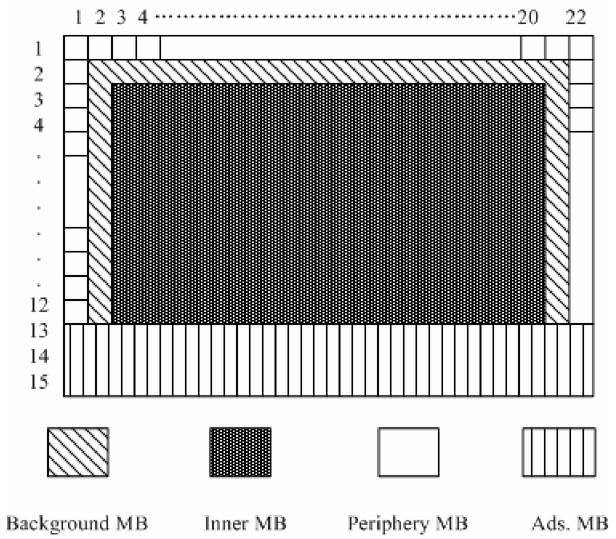
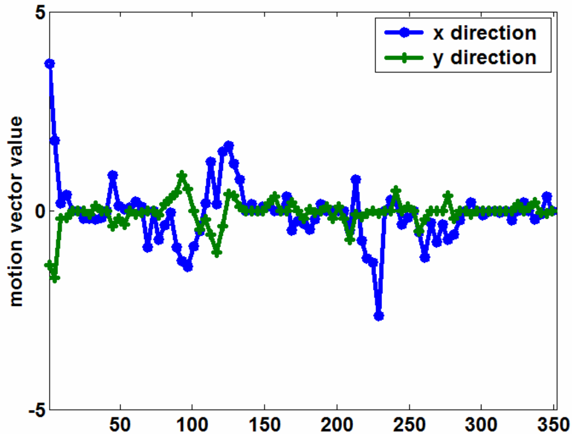


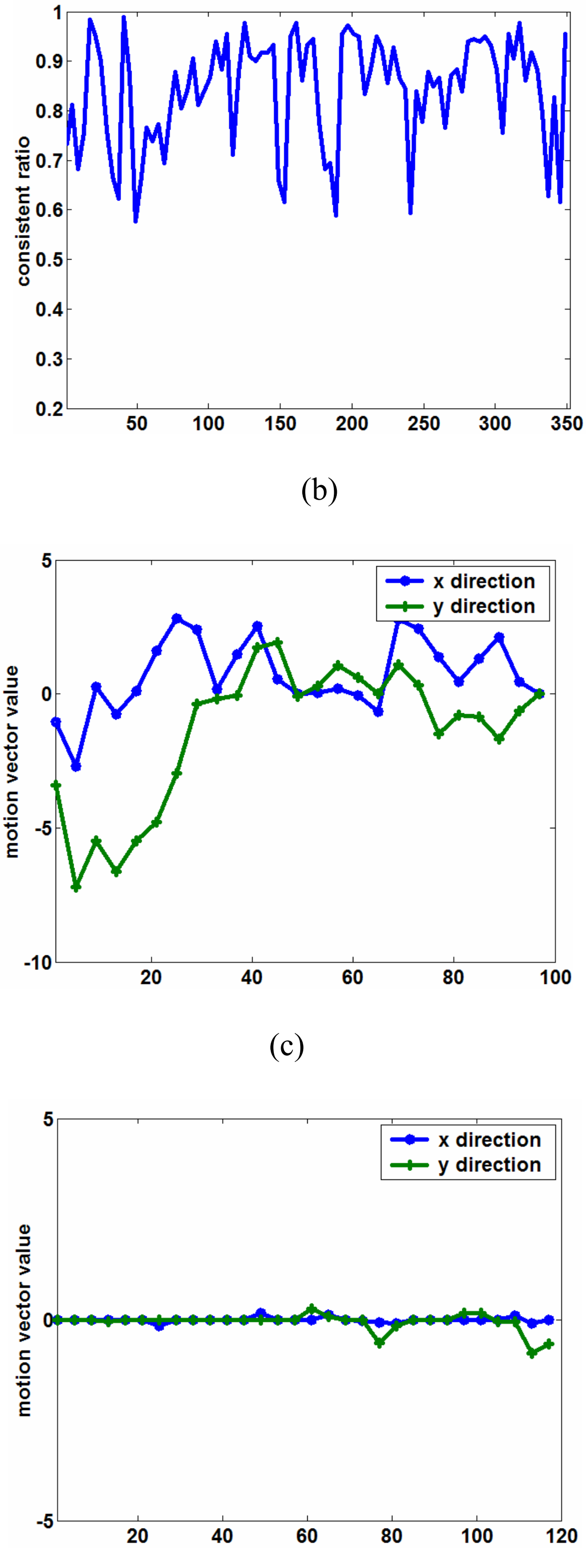
Figure 3-3 Macroblocks partition in a frame

Motion features of a shot

Similar to color features, we can get motion features for a shot. Figure 3-4 shows some examples. Figure 3-4(a) and 4(b) is a shot of close-up of a talking man, sometime he turn his head during the conversation. So we can see that the



(a)



(b)

(c)

(d)

Figure 3-4 motion features of typical shot. X-axis is frame number. (a) and (b) a close-up of a talking man (c) baseball shot (passing ball) (d) golf background motion is relatively small while consistency ratio is lower than 0.8 many times. 4(c) and 4(d) are two shots with high green ratio, 4(c) is baseball shot while 4(d) is golf shot. Obviously, the background motion value of or baseball change more greatly than that of golf.

3.2.3 Compact shot features and classification

Although there are obvious patterns in the field color ratio curves and motion feature curves, the features matrices are not suitable for shot classification since their size (column) change greatly with the shot length and the dimension is still high. So we compute the characteristic parameters of each feature, then a compact features can be figured out for each shot.

Let *y_ratio*(*i*) represents yellow ratio features of a shot, where *i=1,…,m*, the meaning of m is same as Section 3.2.1. We compute the mean and the variance of the yellow ratio features as follows:

$$m_y = \sum_{i=1}^{m} y\_ratio(i) / m$$

$$\mathrm{v}_y = \sum_{i=1}^{m} (y\_ratio(i) - m_y)^2 /(m-1)$$

Similarly, the mean and the variance of green ratio features m_g, v_g, that of white ratio features m_w, v_w, and that of background motion vector m_{vx}, v_{vx} m_{vy}, v_{vy} are also calculated.

Different from above features, for the last motion feature i. e. consistency ratio of inner MB, we do not compute its mean and variance. Observing that in close-up shots, this value is often below a threshold THR, we compute the compact features like following formula:

$$n_{mvr}(i) = \begin{cases} 1 & mv\_ratio(i) > THR \\ 0 & otherwise \end{cases}$$

$$r_{mv} = \sum_{i=1}^{m} n_{mvr}(i) / m$$

where *i*,*m* is the same as above, *mv_ratio*(*i*) is the consistency ratio features of a shot. THR is set to 0.8 in our experiment.

Now for a shot, we get 11 features. These features are used to classify the shots by decision tree.

3.3 Experiment results

We use TRECVID 2003 video to conduct our experiment. The shots in this dataset change greatly both in content and length. The length of the shots ranges from 1 second to 15 seconds. The basketball shots almost include all kinds of NBA field. The baseball shots cover every kind of baseball segments such as pitching, running to base, passing ball etc. Golf shots also contain shots from first stroke to pushing the ball into the hole, tracking the ball etc. The others shots covers the talking people, the surroundings of the sports field, close-up to people in the field or besides the field, wide-angle of audiences etc.

First the sports video are extracted from CNN headline news, and then encoded with TMPEG, the GoP is set to IPPP. Color features are extracted from each I frame and motion features from the first P frame of each GoP.

We separate the dataset into two parts, half of them are for training and validation, while the other for testing. C4.5 decision tree is selected as classification tools through a machine learning tool Weka.

The results are given in Tables 3-1 and 3-2. Table 3-1 is a confusion matrix where each row shows the classification of ground truth. For example, the first row shows that only 1 leadin shot is misclassified into a non-field shot (``Others" in the table) while 17 of them are correctly classified. Table 3-2 shows the classification performance, including Precision (denoted by P), Recall (denoted by R), and F-measure (denoted by F)

Table 3-1 Confusion matrix of test data

	L	T	O	BK	HK	BS	G
Leadin	17	0	1	0	0	0	0
Text	0	11	0	0	0	0	0
Others	0	1	67	0	1	0	0
Basketball	0	0	0	29	0	1	0
Hockey	0	1	2	1	11	0	0
Baseball	1	0	6	1	0	16	1
Golf	0	0	3	0	0	0	3

The tables show that the classification of leadin, text, and basketball shots are quite good. However baseball and golf are easy to be mixed with “others” shots. The macro precision is 0.883, macro recall is 0.822. Moreover, apart from non-field shots, the classification accuracy of field shots is very high.

Table 3-2 Performance of classification

Class	L	T	O	BK	HK	BS	G
P	0.944	0.846	0.848	0.935	0.92	0.94	0.75
R	0.944	1.000	0.971	0.967	0.73	0.64	0.50
F	0.944	0.917	0.905	0.951	0.82	0.76	0.60

From the experiment results, we find that half of golf shots and baseball shots are miss-classified into “others”, a close look to the failed examples indicated that some of the error are caused by the shots content including both the golf field and the surrounded environment. For example in CNN headline news on April 20, 1998, a golf shot from frame 39706th to 39898th covered both sky video segments and field segments, which is not segmented into two shots because of an unidentified gradual transition. On the other hand, we also find our experiment assigned a basketball tag to a baseball shot which is in CNN headline news on April 20, 1998 (frame 38925th to 38984th). After analyzing the feature, we found that the color features for the baseball fields fell into yellow colors area, which is the typical color for basketball field. This phenomenon is caused by the color of the grass becomes nearly yellow in the field.

These failed examples show the importance of shot boundary detection and moreover the further analysis to the visual content. For the later, we gave part of illustration in the next chapter.

Chapter 4

A Text Information Retrieval Approach to Automatic Image Annotation

In this section, a novel framework is proposed to address automatic image annotation problem. This framework benefits from the text representation of the image content and multi-class, multi-label maximal figure of merit (MC MFoM) based discriminative classifier learning. Here an image is first tokenized into some sets of the symbols. Each set is extracted from a distinctive visual feature and characterizes the image content from a different point. For example, the color set of symbols can be detected from the color features in the training set and the texture set of symbols is gotten from texture features. It is similar to the proposed models. Unlike the proposed methods, which model co-occurrence of the symbols (often 1 symbol set is used) and keywords, here we can use multiple sets of the feature specific symbols and a set of patterns is further extracted from these symbols. An image can be described using the patterns and their co-occurrence. The patterns may be unigram, bigram for a specific set of symbols or cross-unigram or -bigram. Anyway, the patterns can be detected using any available technology. If we treat the patterns as a visual vocabulary, an image can be viewed as a text document. This view facilitates fusing the distinctive visual features and exploiting cross-relations among them. It is also easy to utilize the high-order statistics of the patterns (e.g. long-term contextual) since many techniques have been developed in text information retrieval to automatically finding semantic relations among the word terms. For each image document, a high dimensional vector is used to represent it.

4.1 Text Representation for Image Content

Image representation has been exhaustively investigated in the communities of image processing and computer vision. A lot of excellent works have been done for feature extraction. A common sense is that there is not any single visual feature that can describe the rich content of an image and differentiate the objects and concepts. Efficiently fusing them together is necessary and useful for many tasks such as object recognition, AIA and CBIR. A fact is that human can discriminate two objects or describe different concepts using the most expressive attributes and ignore the non-informative ones. For example, only color feature will work well for discriminating a red apple and a green one while the shape will not do. Therefore, adaptively selecting the most informative features is important as well as the fusion.

4.1.1 Image Representation

If the objects in the images can be robustly and correctly detected, the image content can be depicted using their combination and relations. Unfortunately, the generic object detection is still an unsolvable problem in computer vision, although some special detectors have been successfully developed such as the face detector. Contrary to the object feature, the mid-level visual features (e.g. the region based features) and low-level feature (e.g. the patch or grid based features) are easily accessed. In these methods, an image is first divided into a set of the regions using the automatic segmentation algorithms (e.g. normalized cut [39]) or a set of patches or grids. Then the statistical visual features are easily extracted from them and an image is described using the set of visual features. These visual features can be in a continuous space [24,25,12] or are further tokenized using the unsupervised clustering techniques [20,22,8]. These are the commonly used representations in the current CBIR and AIA systems.

4.1.2 Text Representation

Text categorization and information retrieval have been deeply studied. Many successful techniques have been developed, e.g. feature selection and reduction algorithms, semantic inferring techniques, robust classifier learning, etc. A good survey can refer to [38]. An obvious benefit from text representation is that it is relatively easy to explore the (e.g. syntactic and semantic) relations among the terms. However, this symbolic representation is naturally in hand like a text for an image. Representing the image at the meaningful level requires the object detection, which is still a hard problem. A coarse and easy method is to quantize the regions or grids to get the symbols. Some of the proposed models learn the co-occurrence statistics between the symbols and keywords. The statistical contextual dependence among the grids or regions can be incorporated such as 2D-HMM [25] and MRF [8].

In this chapter, we will further extend the symbolic image representation and view it as a text document where the contextual dependence is explicitly represented by a pattern. Here the pattern means a symbol sequence such as the n-gram in the language model or a combination of some symbols according to some syntactic rules. A visual lexicon will be constructed using all detected patterns. Figure 4-1 shows a possible way to get the direction specific bigram patterns. For the site X_{22}, its direction specific bigrams (e.g. $X_{22}X_{21}$, $X_{22}X_{23}$, $X_{22}X_{11}$ and $X_{22}X_{33}$) are gotten from its neighboring sites. Here 8 directions are shown. The extracted bigrams will be treated as the distinctive patterns for the image representation. Sometimes these patterns can be clustered to reduce the size of the visual lexicon. In Figure 4-1, for example, the horizontal bigrams such as $X_{22}X_{21}$ and $X_{22}X_{23}$, can be merged. When the visual lexicon is gotten, any image can be viewed as a text document with the terms in the visual lexicon to describe its content. Except for the n-gram patterns, many other patterns can be exploited, e.g. location or position specific pattern, shape pattern, etc.

The benefits from the text representation for the image are obvious. With the text representation the diverse visual features are transformed into the distinctive patterns and are uniformly treated. The importance of each visual pattern can be explored using the feature selection techniques. It divides the image annotation problems into the pattern mining, selecting and modeling, which is relatively easier comparing to learning an overall joint distribution for the image content and labeled keywords. It also facilitates the integration of the state-of-art feature extraction techniques and specific visual feature detectors. Due to the text representation, the high-order statistics among the patterns, e.g. syntactic and semantic, can be extracted using the automatic semantic inferring techniques such as latent semantic indexing (LSI).

The arising question is whether the text representation will work and the symbols and their statistics have sufficient discriminative power. The past works have partially given a positive answer.

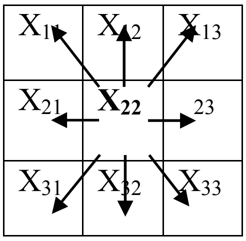


Figure 4-1 An example to show bigram patterns (X_i is the site for the i-th grid)

4.1.3 LSI-based Image Representation

When the visual lexicon is obtained, an image document is often represented by a multidimensional feature vector with the dimension equal to the size of the lexicon. Each component of the vector corresponds to its importance of a pattern occurred in the image document. In many typical real-world applications, there are usually more than ten thousand entries in the lexicon, e.g. there will be 10,000 bigrams patterns if the size of symbols is 100. Many techniques, such as feature selection [5], have been proposed to reduce the dimension. Latent semantic indexing (LSI) [15] is an efficient way to achieve both feature extraction and reduction. The clustered words based on LSI have some semantic relation, which is a long-term dependence among the words. The LSI-based language model is also studied in [15]. Here we also apply LSI for feature reduction and selection for the image representation, although it is difficult to explain the meaning for each component in the LSI space. We expect LSI can capture some long-term contextual dependence among the image patterns. As in [15], singular value decomposition (SVD) based LSI is used to get a lower dimension than the original one by decomposing the term-document matrix H into a multiplication of three

$$H = USV^{T} \tag{4.1}$$

Here U:M×R is the left singular matrix with the rows u_i, 1≤i≤M, U: M×R is the diagonal matrix of singular values S_1≥S_2≥…≥S_R>0 and V: K×R is the right singular matrix with rows v_j, 1≤j≤K. M is the size of the visual lexicon and K is the number of the image documents in the training set.

Both the left and right singular matrices are column-orthogonal. If only the top P singular values are remained in matrix S and other (R-P) components are zeroed out, we can effectively reduce the LSI feature dimension to P that could be much smaller than R. By doing so, three matrices are much smaller in the size than those in Eq. (4.1) and the computation requirements are greatly reduced.

In matrix H, its (i,j)-th element, H(i,j), describes the association between the i-th visual pattern and the j-th image document. It is defined as

$$H(i,j) = (1-\varepsilon_i)\cdot c_{i,j}/n_j \tag{4.2}$$

where $c_{i,j}$ is the number of times of the i-th visual pattern occurred in the j-th image document, n_j the total number of visual patterns which appear in the j-th image document, ε_i the normalized entropy of the i-th term in the training set which is further defined as

$$\varepsilon_i = -\frac{1}{\log K}\sum_{j=1}^{K}\frac{c_{i,j}}{t_i}\log\frac{c_{i,j}}{t_i} \tag{4.3}$$

Where $t_i = \sum_j c_{i,j}$ denotes the occurrence times of the i-th visual patterns in the training set.

Using the above definitions, any image document, which is represented by a vector d^T in the original space with a dimension M, can be characterized using an R dimensional vector, *v*, in the LSI space as in Eq. (4.4).

$$v = d^T U S^{-1} \tag{4.4}$$

4.2 Model Estimation With MC MFoM Learning Algorithm

Many statistical models have been proposed for AIA [7,10,20,24,22,30,25,8,12]. Some of them, such as MRF, CMRM, CMRM, translation model, etc., are directly to model the joint distribution between the set of visual feature (continuous or discrete symbols) and the annotated keywords, while others such as 2D-HMM model the keyword conditioned distribution of the visual features. The contextual dependence is also explored for MRF and 2D-HMM while CMRM, CMRM and translation model ignore it. As discussed in Section 4.1, we describe the image using a high dimensional vector with each component corresponding to its contribution for classification. This let us have more choice for the classifier learning. Many classifier learning algorithms have been proposed and studied for text information retrieval [5] such as SVM, naïve Bayesian, decision tree. Here we just use a linear classifier for the image annotation problem because of its simplicity and meaningful explanation for its weights, although others can also be exploited.

Like text categorization, image annotation is also a multi-class, multi-label classification problem. The conventional solution for this classification problem is 1) to learn multiple binary classifiers each corresponding to a class and then independently decides whether a class label should be assigned to a test sample or not, or 2) to learn a multi-class classifier and assign the top-N class labels to a test sample. As discussed in [15], the former method cannot capture discriminative power simultaneously among the multiple classes while the latter has not a flexible decision, i.e. fixed size of labels is not a good choice in most of cases. To overcome these problems, Gao [15] proposed the multi-class, multi-label maximal figure-of-merit (MC MFoM) learning algorithm. This learning algorithm can fully take advantage of both positive and negative training samples. In contrast to the popular binary classification algorithms, e.g. SVM, the MC MFoM can simultaneously learn a multi-class classifier with an embedding multiple-label decision rule and the preferred metric. It was shown on the task of text categorization that the MC MFoM learned classifier is more robust, particularly for small sample training, and work better than the corresponding binary MFoM classifier and linear SVM. Since MC MFoM is a discriminative learning algorithm, most expressive and informative features for classification will be picked out in the learning.

In the next, we will first introduce multiple-label decision rules for automatic image annotation and then MC MFoM learning for AIA.

4.2.1 Multiple-Label Decision Rules

In the image annotation problem, there are multiple ground truth labels for an image. Here we will first discuss the multiple-label decision rule introduced in [15]. The same notations are used. Given N keywords, $C=\{C_j, 1\leq j\leq \mathrm{N}\}$, and an annotated training set, $T=\{(X,Y)\}$, and C_j is the j-th keyword, with X being a sample in a D-dimensional space, Y a set of labels for X and a subset of C. N

classifiers with the model parameter set, Λ are estimated from T. Denote Λ_j as the parameter set for the j-th keyword, C_j, then $\Lambda=\{\Lambda_j, 1\leq j\leq \mathrm{N}\}$. In this chapter, a linear discriminant function, $g_j(X; \Lambda_j)$, is used for the j-th keyword. It is defined as

$$g_j\left(X;\Lambda_j\right)=W_j \cdot X + b_j \tag{4.5}$$

where W_j is a weight vector with an equal dimension to X, and b_j a shift. They are the model parameters for the j-th keyword.

Then a competitive model, named class anti-discriminant function, is defined for each keyword,

$$g_j^-\left(X;\Lambda^-\right)=\log\left[\frac{1}{\left|C_j^-\right|}\sum_{i\in C_j^-}\exp\left(g_i\left(X;\Lambda_i\right)\right)^{\eta}\right]^{1/\eta} \tag{4.6}$$

Where C^-_j is a subset containing the most competitive keywords against C_j, $|\mathrm{C}^-_j|$ is the cardinality of the subset, Λ^- is the parameter set for all the competitive keywords, and η is a positive constant. Eq. (4.6) measures the score from all the competing categories and it functions as a negative model for the j-th keyword, which is different from the binary classifier where the negative model is trained from all negative samples. Based on Eq. (4.5) and (4.6), the decisions rule for multiple classes, multiple labels classification problem is defined as the following:

$$\begin{cases}\text{Accept} \quad \mathrm{X}\in C_j \ \text{if}\, g_j\left(X;\Lambda_j\right)-g_j^-\left(X;\Lambda^-\right)>0 \\ \text{Reject} \quad \mathrm{X}\in C_j, Otherwise\end{cases} \quad 1\leq j\leq N \tag{4.7}$$

For many real-world applications, it is not necessary to verify all keywords to get multiple labels. It is enough to verify only the top N-best keyword candidates according to their confidence rankings estimated from Eq. (4.5).

4.2.2 MC MFoM Learning

To learn multiple linear classifiers defined in Eq. (4.5), the MC MFoM learning is applied. For this learning method, an overall objective function, which approximates an interested metric as well as embedding multiple-label decision rule in Eq. (4.7), is designed. This function should be a continuous and differential function for optimization. To accomplish this, a one-dimensional class misclassification function, $d_j(X; \Lambda)$, is introduced,

$$d_j(X;\Lambda) = -g_j(X;\Lambda_j) + g_j(X;\Lambda_j^-) \tag{4.8}$$

where $d_j(X; \Lambda)<0$ when a correct decision is made and otherwise, $d_j(X; \Lambda)\geq 0$. It is equivalent to Eq. (4.7) but it is a differential and continuous function while Eq. (4.7) is a discrete function. To further normalize the value in Eq. (4.8) and simulate the classification error, a class loss function, $l_j(X; \Lambda)$, for the keyword C_j, is defined,

$$l_j(X;\Lambda) = \frac{1}{1+e^{-\alpha(d_j(X;\Lambda)+\beta)}} \tag{4.9}$$

where α is a positive constant that controls the size of the learning window and the learning rate, and β is a constant measuring the offset of $d_j(X; \Lambda)$ from 0. They are empirically determined. The value of Eq. (4.9) simulates the error count made by the *j*-th classifier for a given test sample *X*.

With the definition in Eq. (4.9), we can approximate the commonly used metric, i.e. precision, recall and F1, for information retrieval. For a class C_i, they are defined respectively,

$$P_j = \frac{TP_j}{TP_j + FP_j} \tag{4.10}$$

$$R_j = \frac{TP_j}{TP_j + FN_j} \tag{4.11}$$

$$F_j = \frac{2P_j R_j}{R_j + P_j} = \frac{2TP_j}{FP_j + FN_j + 2TP_j} \tag{4.12}$$

Where TP_j is the true positive, FP_j is the false positive, and FN_j is the false negative for the j-th keyword. Correspondingly, their approximated functions on the training set T are as follows:

$$FN_j \approx \sum_{X \in T} l_j(X;\Lambda) \cdot 1(X \in C_j) \tag{4.13}$$

$$FP_j \approx \sum_{X \in T} (1 - l_j(X;\Lambda)) \cdot 1(X \notin C_j) \tag{4.14}$$

$$TP_j \approx \sum_{X \in T} (1 - l_j(X;\Lambda)) \cdot 1(X \in C_j) \tag{4.15}$$

Here 1(.) is the indicator function of any logical expression. With the above definition, the overall objective function can be defined according to the chosen metric. Here we will optimize the micro-averaging F1 measure to estimate the linear model parameters. So the objective function is defined as follows to approximate it,

$$L(X;\Lambda) = 2\sum_{i=1}^{N} TP_i \Big/ \left\lfloor \sum_{i=1}^{N} FP_i + \sum_{i=1}^{N} FN_i + 2\sum_{i=1}^{N} TP_i \right\rfloor \tag{4.16}$$

To find its solution, the generalized probabilistic descent (GPD) algorithm is used, which iteratively estimates the model parameters.

4.3 Discriminative fusion scheme

Now we will study the fusion techniques in the unified framework for AIA discussed above. Since the fusion is based on the multi-category discriminative learning, we name it the discriminative fusion schemes. In the next, an early fusion scheme based on the statistical intra- and inter- pattern association is first discussed. Then a late fusion scheme using the model-based transformation is proposed.

4.3.1 Fusing with Ensemble-Pattern Association

As discussed in Section 4.1.2, an image can be tokenized using multiple sets of the visual lexicons, each of which may be extracted from a distinctive visual feature. They are used together for the image representation. For each distinctive visual lexicon, the intra-lexicon statistical association, e.g. unigram and bigram of the visual patterns, can be extracted to describe the co-occurrence of the intra-lexicon visual terms. To characterize the statistical association of the inter-lexicons among the different visual lexicons, the statistical co-occurrence of the cross-pattern can be extracted. Here the cross-pattern means a combination of more than one pattern, each of them from a distinctive visual lexicon. For example, if there are two visual lexicons, A (color) with M patterns (i-th pattern is A_i) and B (texture) with N patterns (j-th pattern is B_j), then the intra-lexicon association may be the count of A_j occurred for unigram or that of (A_i, A_j) for bigram. Similarly, the intra-lexicon association for B can be extracted. If using the unigram and bigram extracted from one feature to describe the image content, the dimension of the feature vector will be M×(M+1) for A and N×(N+1) for B. If only using the cross-pattern to represent the content, its dimension will be M×N.

Given an image, its content is represented by the associations of the intra- and inter- lexicon patterns. Since there are multiple visual lexicons, which can be used, this fusion method is named the ensemble pattern association (EnPA) based fusion. All the extracted associations construct a high-dimensional vector for the image representation. Then a classifier is trained based on this representation.

The EnPA method can explicitly describe and embed the spatial contextual information (e.g. bigram) and the relation (inter-pattern co-occurrence) among the different type of features. Since all the features are tokenized, we can treat all visual features using a unified view. This is a good property considering that the visual features maybe extracted using the different scales and techniques. This representation facilitates to exploit more high-order relations (e.g. syntactic and semantic association) among the patterns using the developed techniques for text categorization. For example, the LSI technique can be applied to exploit the semantic relation among the visual patterns.

4.3.2 Model based Transformation

The model based transformation (MBT) fusion can be treated as a supervised mapping from the low-level feature space to the semantic space. It is a later fusion method. In this approach, the first step is to train a classifier for a specific feature. For a N-concept annotation problem, N discriminant functions are learned from the training samples. In this chapter, we apply the MC MFoM learning to train this classifier. If we treat the N discriminant functions as the set of the basis for the transformation, we can obtain a new N-dimensional feature with the similarity between a given sample and a discriminant function as each of its components. In this section, the similarity is the confidence score defined in Eq. (4.8), which is similar to the model based vector representation in [31]. The difference is that our confidence score is derived from the multi-category classifier while theirs are from the binary classifier (i.e. SVM).

Since the MBT based new feature describes the confidence measure and is normalized by the competitive model, it will be more compact with a smaller variance comparing to the raw low-level feature. Using the MBT method, it is easy to fuse multiple distinctive features. For example, if there are K types of features, we can get K N-dimensional features in the model space and then concatenate them into a K*N-dimensional feature to describe the image content in the model space. With the new representation, we can train another classifier using the MC MFoM learning for the final decision.

The MBT method map any type of features (e.g. visual, textual, audio/speech) into a common space, i.e. model space. And then the discriminative MC MFoM learning can automatically weight each type of features. If a visual feature or modality is more powerful and discriminative, it has a heavy weight. Otherwise, its weight will be smaller. It should outperform other heuristic methods such as the maximization / minimization product.

4.3.3 Combinational Scheme

The two fusion methods discussed in the above are totally different. They can be applied separately as well as can be used together. This combinational fusion scheme will provide more flexibility for the semantic concept detection for the TRECVID and AIA.

4.4 Experimental Results and Analysis

To evaluate the proposed AIA framework, we will show the annotation results based on the Corel data set.

4.4.1 Data Set and Baseline

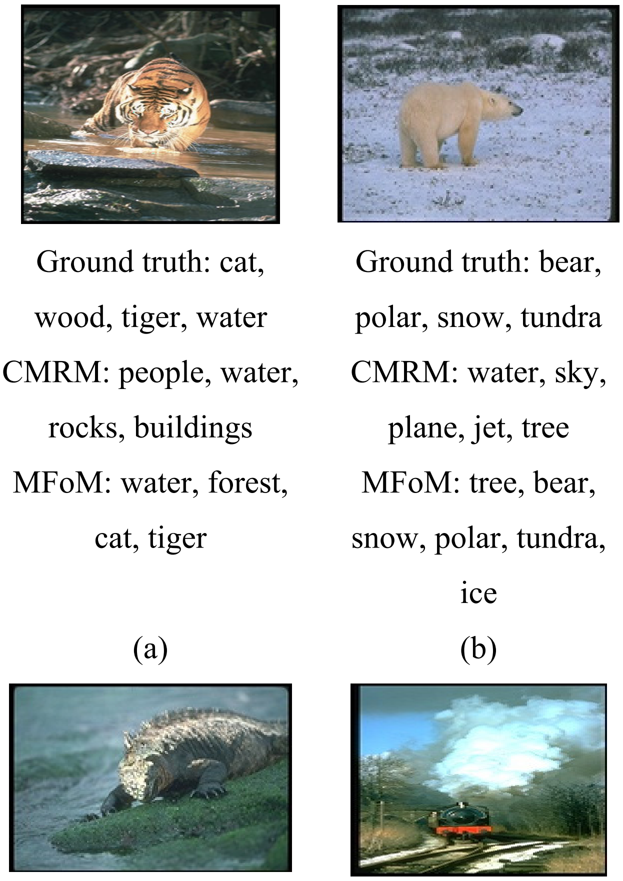
To provide a meaningful comparison with the previously reported results, we also use the dataset provided by Duygulu et al. [10]. The dataset consists of 5,000 images from 50 Corel Stock Photo CDs. Each CD has 100 images on the same topic. Each image contains an annotation of 1-5 keywords. Overall there are 374 keywords. The dataset is divided into a training set with 4,500 images and a test set with 500 images. The baselines are the translation model (TM), CMRM and maximum entropy (ME) and their evaluation results are from [22]. In [22], 125 visterms are clustered from the feature vector consisting of the average LAB color and the output of the Gabor filters in the training set.

In our experiment, an image with the size 128x192 or 192x128 is segmented into a set of grids with the size 16x16. In this experiment, we only use the color feature, which includes the mean and stand deviation of RGB and LAB and its dimension is 12. Then the k-means clustering technique is used to get 64 symbols for the color feature. The visual lexicon includes unigram and bigram of the symbols. Except for 64 symbols, a 'NULL' symbol is added. If a grid is at the boundary of the image, we will count the bigram between the grids with the 'NULL'. Other bigrams are extracted from 4 directions, namely horizontal, vertical, first diagonal and second diagonal. Totally the lexicon size is 4,288. Then an image is represented using a 4,288 dimensional vector extracted as in Section 4.1.3. The LSI-based feature has a full rank 2,293 and we only use 600 dimension based on our preliminary experiment. Other features such as texture, shape, etc. are not tried to explore. The fusion function of this framework is not evaluated here. We will analyze the performance of AIA for only color feature and its adaptively feature selection property.

In this experiment, the settings for the control parameters are set as α=60, β=0.03, η=7. Only the top-20 keyword candidates according the scores calculated from Eq. (4.5) are verified. For each keyword, we use 20 competitive keywords to estimate the scores in Eq. (4.6).

4.4.2 Comparison with Baseline

The overall results for 374 keywords are shown in Table 4-1. Comparing to the translation model, CMRM model and ME model [22], our model is best in term of precision, recall and F1. We get a macro-averaging F1, 0.135, comparing with 0.11 Figure 4-2 illustrates some annotation examples for 4 selected test images,



Ground truth: cat, wood, tiger, water
CMRM: people, water, rocks, buildings
MFoM: water, forest, cat, tiger

(a)

Ground truth: bear, polar, snow, tundra
CMRM: water, sky, plane, jet, tree
MFoM: tree, bear, snow, polar, tundra, ice

(b)

Ground truth: marine, iguana, water
CMRM: water, sky, plane, bear
MFoM: water, sky

(c)

Ground truth: locomotive, railroad, smoke, train
CMRM: water, sky, tree, people
MFoM: moutain, sky, tree, train, locomotive, railroad, aerial

(d)

Figure 4-2 Some annotation examples for CMRM and MFoM

where the CMRM annotation results are from [24]. They clearly show that our MFoM learned for ME, 0.10 for CMRM, and 0.05 for TM. Overall 102 keywords, which have a F1 more than zero, are detected. Of 374 keywords, only 260 in the test set have more than 1 test image. For 260 keywords, our model has a macro-averaging precision, 0.196, recall, 0.193, and F1, 0.193. Considering that only grid-based color feature is used in our experiment and it is simpler than the used in TM, CMRM and ME, this result is surprising. Table 4-2 lists the precision (P), recall (R) and F1 measure for 49 keywords best predicted in terms of F1 metric model is better than CMRM. Even for the false accepted annotations, they still have some semantic relations with the ground truth.

For example, 'Forest' is a false accepts for image (a), while it has a strong relation with the ground truth annotation 'wood'. Similar observations are seen in images (c) and (d).

More importantly, our model can flexibly decide the size of the annotation, where CMRM can only use fixed size.

Table 4-1 Comparison with TM, CMRM and ME

		TM	CMRM	ME	MFoM
Macro-averaging	P	0.06	0.10	0.09	0.136
	R	0.04	0.09	0.12	0.134
	F1	0.05	0.10	0.11	0.135
Micro-averaging	P	NA	NA	NA	0.310
	R	NA	NA	NA	0.374
	F1	NA	NA	NA	0.339

4.4.3 Adaptive Feature Selection

From the linear discriminant function in Eq. (4.5), it is clear that the component of the model parameter, W_j, represent importance of its corresponding component in the feature vector for classification. As MC MFoM trains the classifiers for all keywords simultaneously, it is expected that the MC MFoM will pick out the most expressive and discriminative feature for classification. This property is shown in

Figure 4-3 with selected keywords, ‘sky’, ‘sun’, ‘plane’, and ‘jet’. Here the initial weight is the mean of the feature vectors in the training set for a keyword, and the learned weight is the model parameters after 30 iterations in the GPD training.

Although each component in the feature vector and model parameters has not meaning, we can still learn some good properties for the MC MFoM training. For the initial models of 4 keywords, they have a weak discriminative power. For each model, only a few components have a dominative contribution to the decision. But they are non-discriminative and non-informative for differentiating the distinctive keywords. After the MC MFoM training, the weight distribution is significantly changed. Some important components are enhanced while other non-informative components are suppressed. Even for the semantic-related keywords, e.g. ‘sun’ and ‘sky’, and ‘plane’ and ‘jet’, their weight distribution shows more difference than the initial weight.

4.4.4 Discriminative Feature Fusion

To study the efficiency of the fusion methods, the following experiments are done on the Corel dataset:

E1: only color feature. The color feature is tokenized and the color visual lexicon with 64+652-1 patterns is extracted (one NULL symbol is added to describe the context of a grids at the image boundary). Then a term-document matrix with the size 4,288 * 4,500 (row is the pattern and column is the document) is constructed, LSI is used to reduce the dimension. Its full rank is 2,451 and only 600 eigenvalues are kept. Finally, each image is represented by a 600-dimensional vector in the LSI space. And the linear classifier is trained using the MC MFoM learning.

E2: only texture feature. Similar to E1, we apply LSI to get a 600-dimensional vector (here full rank 2,354) for the image representation.

E3: fusion by concatenating color and texture to get 24-dimensional feature. Then the feature is tokenized. Similar to E1, we apply LSI to get a 600-dimensional vector (full rank 2,219) for the image representation.

E4: EnPA fusion (see Section 4.3.1). Two visual lexicons extracted in E1 and E2 are merged and then cross-pattern between the color pattern and texture pattern is counted. A new visual lexicon with 20,864 patterns is built. Then LSI is applied to get 600-dimensional feature (full rank 2,435).

E5: MBT fusion (see Section 4.3.2). Use the multi-category linear classifier trained on E1 to map the color feature into the 374-dimensional model space. Similarly, the texture is also mapped into the model space using the classifier trained in E2. Then a new 374*2 feature is constructed to represent the image, and a new linear classifier is then trained on this feature using the MC MFoM learning.

To evaluate the performance, we use the mean precision (P), recall (R), and F1(F) measure of the overall concepts as the metric [22,12,24]. In addition, the number of the detected concepts (# of detected) is also compared. The comparison for the above 5 experiments is shown in Table 4-2.

Table 4-2 Performance comparison for Corel dataset

	P	R	F	# of detected
E1	0.166	0.130	0.146	99
E2	0.105	0.102	0.103	90
E3	0.142	0.121	0.131	100
E4	0.153	0.151	0.152	109
E5	0.163	0.179	0.171	126

Table 4-3 Performance comparison for TRECVID 2003

	P	R	F	# of detected
E1	0.238	0.195	0.214	10
E2	0.180	0.222	0.199	10
E5	0.196	0.288	0.233	10

From Table 4-2, it is clearly seen that the proposed fusion methods (E4 & E5) outperform the systems only using one single type of feature and the concatenating method. The simply concatenation fusion even worsens the performance. It is a common phenomenon observed. That is the reason some researchers use various validation method to choose the best concatenation method [2]. Comparing to the best result with the single feature (E1), the EnPA only has a little improvement (~4.1%) in term of F1 measure. But we can see 10% improvement in term of the number of the detected concepts. The best result (E5) is obtained using the MBT fusion. The improvement reaches ~17.1% in term of F1 comparing to the result using the single feature (E1), and ~27.7% in term of the detected concept number. Comparing to the EnPA fusion, the F1 measure is improved ~13.2% and the number of the detected concepts is increased ~15.6%.Similar experiments (E1, E2 and E5) are done on the TRECVID dataset except for the concatenation (E3) and EnPA (E4) fusion.

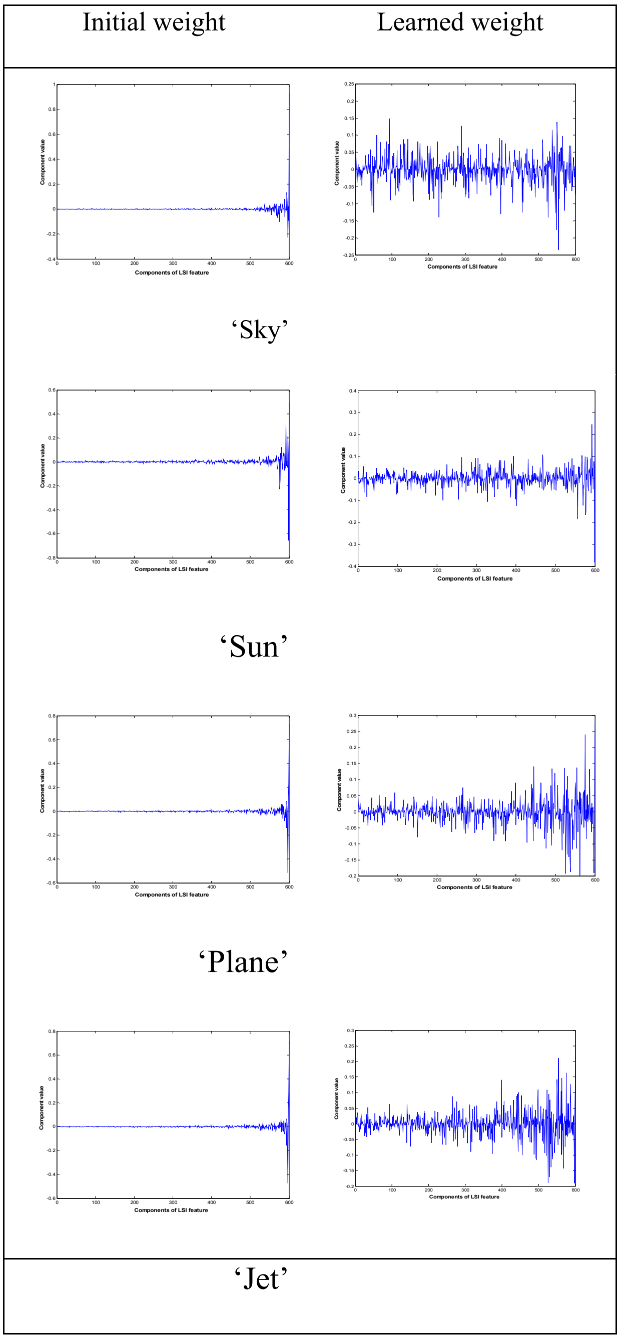


Figure 4-3 Adaptive feature selection for MC MFoM

The results are listed in Table 4-3. Similar observation is seen as for Corel dataset. The MBT fusion gets a best result with an improvement, ~8.9%, in term of F1.

These experimental comparisons clearly show that the proposed fusion methods work well for the easy dataset as Corel and for the challenging dataset as

TRECVID. Table 4-4 shows the performance for best predicted 49 keywords, which give us a roughly idea about the effectiveness of the proposed method.

Table 4-4 Best predicted 49 keywords in terms of F_1

	P	R	F_1
FESTIVAL	1.000	1.000	1.000
WHALES	1.000	1.000	1.000
POOL	0.909	0.909	0.909
PILLAR	1.000	0.700	0.824
NEST	0.750	0.857	0.800
FORMULA	0.750	0.750	0.750
TURN	0.600	1.000	0.750
JET	0.765	0.684	0.722
SWIMMERS	0.667	0.750	0.706
TRACKS	1.000	0.546	0.706
POLAR	0.727	0.615	0.667
CORAL	0.667	0.667	0.667
PATH	0.500	1.000	0.667
TIGER	0.750	0.600	0.667
BENGAL	0.667	0.667	0.667
LAWN	1.000	0.500	0.667
MOOSE	0.500	1.000	0.667
MARSH	0.500	1.000	0.667
OUTSIDE	0.500	1.000	0.667
PROTOTYPE	0.600	0.750	0.667
MAUI	1.000	0.500	0.667
PLANE	0.778	0.560	0.651
OCEAN	0.600	0.667	0.632
FOALS	0.444	0.889	0.593
CAT	0.833	0.455	0.588
SKY	0.447	0.800	0.573
BLACK	0.400	1.000	0.571
CARIBOU	0.667	0.500	0.571
PYRAMID	0.500	0.667	0.571
HORSES	0.429	0.750	0.546
MARE	0.412	0.778	0.539
CARS	0.778	0.412	0.539
FLOWERS	0.565	0.482	0.520
RUNWAY	0.333	1.000	0.500
RESTAURANT	0.500	0.500	0.500

SLOPE	0.333	1.000	0.500
ZEBRA	0.500	0.500	0.500
CANYON	1.000	0.333	0.500
GOAT	0.500	0.500	0.500
BIRDS	0.667	0.353	0.462
PLANTS	0.546	0.400	0.462
RAILROAD	0.600	0.375	0.462
SUNSET	0.333	0.714	0.455
SNOW	0.546	0.387	0.453
SCOTLAND	0.350	0.636	0.452
DEER	0.400	0.500	0.444
BEAR	0.700	0.318	0.438
WATER	0.330	0.621	0.431
LOCOMOTIVE	0.600	0.333	0.429

Chapter 5

Conclusion and future work

5.1 Conclusion

In this thesis, two parts of research work on semantic concept detection from visual content are introduced, namely sports news video genre identification and automatic image annotation.

For the former, a novel feature extraction method has been proposed. Sports field ratio based on pre-determined field colors for specific types of sports, and background motion, ratio consistent with the background in motion are proposed features, they are applied to classify the sports news video shots into 7 predefined video shot classes including 4 sports field namely basketball, baseball, ice hockey and golf, using C4.5 decision tree. The advantage of our method is they are extracted from every frame rather than one key frame in a shot, more over they are compact and have some semantic meaning. The effectiveness of the method is demonstrated by our experiments conducted with challenging dataset from TRECVID 2003.

For the latter, a multi-class text categorization framework for automatic image annotation is proposed. The proposed approach benefits from a text representation for image content and MC MFoM multi-class discriminative classifier learning. The image-text representation applies the ensemble visual lexicons detected with the various techniques to tokenize a given image and then represent its content using multiple symbolic documents. This method makes it feasible to exploit the statistical associations among the different features and more high-order statistics. In this framework, two fusion methods, i.e. ensemble-pattern association and model-based transformation, are discussed. Based on the representation, the MC MFoM learning is used to train robust multi-class classifiers jointly. In the training stage, the weighting coefficients can be automatically adjusted according to their importance for annotation. Finally, the proposed framework is evaluated on the Corel dataset and TRECVID 2003 dataset. Our experimental results show that this framework supplemented with the model-based transformation fusion achieves a high performance for image annotation.

For the Corel dataset, we obtain a macro-averaging F1 of 0.179 for the 374 concepts, which outperforms the state-of-art results using the MBRM model.

5.2 Summary of the major work

- A novel feature extraction method was proposed, the feature can capture the temporal pattern of sports news video shot and classify them into some predefined classes. The effectiveness of the method is demonstrated by very challenging dataset.
- A novel image representation technique was proposed. The novelty is the contextual information of the images can be represented as bigrams in a text document, so many text document techniques can be employed to address image annotation problem.

- A discriminative multi-class classifier method was employed to conduct AIA. With this framework two flexible information fusion methods for fusing diverse visual features were proposed. Experiments conducted on CorelCD and TRECVID show above methods outperform the state of the arts AIA techniques.

5.3 Future work

Although promising results were achieved by the above work, however, the pattern extracted from images has no obvious meaning, which impairs the effectiveness of the framework. To tackle the SCD problem more effectively, we should find some semantic mid-level features to bridge gap between non-meaningful low-level features and high meaningful semantic concepts. These mid-level features or patterns are intermediate representations of the content, which can facilitate the description of the content. We will conduct research on employing semi-supervised learning techniques to detect patterns from large-scale image/video database. We expect that this approach can make the detected patters have semantic meanings to some degree. On one hand, because these patterns or mid-level features are clustering result from low-level features, their dimensions should be far smaller than the original low-level features; moreover, having some semantic meaning they can be a good intermediate to bridge the gap between low-level features and high-level semantic concepts. On the other hand, we only need to label very small training data to carry the prior knowledge; it obviously lessens the boring and error-prone manually labeling burden.

So we can summarize the three main objectives of coming research:

- To bridge the gap between low-level features and high-level semantic concepts we employ semi-supervised clustering method to exploit intermediate features. The key research points are how to model the prior knowledge of semantic concept and design an objective function to integrate them. These constraints can be formulated as an instance [50,23], a statistical model [49], or some other properties of the data (e.g. locally lineal, [36]). In Wanjun Jin ACMMM'04 [51], negative constraint are formulated according to co-occurrence based correlation and thesaurus (WordNet) based correlation. On the other hand, due to imperfect segmentation or small size grids, the regions are often over-segmented. How to group those regions is a challenging problem. We can label a few images as a positive soft constraint to improve the clustering result.
- Employ our semantic concept detection method to more challenging data such as images collected from internet, combine the detected concept with other visual features to improve the performance of current image retrieval system.
- Extending the above method to spatial-temporal signals to detect concepts from video shots. Some method to cluster temporal signals need to be explored. HHMM is successfully employed to extract patterns from video signals (Lexing Xie et al, ICME'03 [58]). We need to find some method to improve temporal signal clustering.

Bibliography

[1] W. H. Adams, A. Amir, C. Dorai, S. Ghoshal, G. Iyengar, A. Jaimes, C. Lang, C. Y. Lin, M. R. Naphade, A. Natsev, C. Neti, H. J. Nock, H. Permutter, R. Singh, S. Srinivasan, J. R. Smith, B. L. Tseng, A. T. Varadaraju, and D. Zhang, IBM research TREC-2002 video retrieval system, in Proc. Text Retrieval Conference (TREC), Gaithersburg, MD, Nov 2002, pp. 289–298.

[2] A Amir, M. Berg, S. F. Chang, G. Iyengar, C. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock, W. Hsu, I. Sachdev, J. Smith, B Tseng, Y. Wu, and D. Zhang, IBM research trecvid-2003 video retrieval system, Nov 2003, NIST TRECVID 2003.

[3] J. Assfalg, M. Bertini, C. Colombo, and A. D. Bimbo, Semantic Annotation of Sports Videos, IEEE Multimedia 9(2): 52-60, 2002.

[4] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan, Matching words and pictures, Journal of Machine Learning Research, 3:1107-1135, 2003.

[5] J.R. Bellegarda, Exploiting latent semantic information in statistical language modeling, Proc. of the IEEE, Vol.88, No.8, pp.1279-1296. 2000.

[6] S. Belongie, J. Malik and J. Puzicha, Shape matching and object recognition using shape contexts, IEEE PAMI, vol. 24, no. 24, April 2002.

[7] D. Blei, and M. I. Jordan, Modeling annotated data. In Proceedings of the 26th Intl. ACM SIGIR Conf., pages 127–134, 2003.

[8] P. Carbonetto, N. D. Freitas and K. Barnard, A statistical model for general contextual object recognition. In Proc. Of ECCV'04.

[9] S.F. Chang, W. Chen, H. Meng, H. Sundaram, and D. Zhong, VideoQ: an automatedcontent based video search system using visual cues. In Proceedings of ACM Multimedia 1997, Seattle, November 1997.

[10] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, In Seventh European Conf. on Computer Vision, pages 97-112, 2002.

[11] L. Fei-Fei, R. Fergus, and P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, CVPR 2004, Workshop on Generative-Model Based Vision.

[12] S. L. Feng, R. Manmatha and V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, In Proc. of CVPR'04.

[13] R. Fergus, P. Perona and A. Zisserman, Object class recognition by unsupervised scale-invariant learning, CVPR, 2003.

[14] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, B. Dom Q. Huang, M. Gorkani, J. Hafner, D.Lee, D. Petkovic, D. Steele, and P. Yanker, Query by image and video content: the QBICsystem, IEEE Computer, 28(9):23-32, 1995.

[15] S. Gao, W. Wu, C-H Lee, T.-S. Chua, A MFoM learning approach to robust multiclass multi-label text categorization. In Proc. of ICML'04.

[16] A. Hamrapur, A. Gupta, B. Horowitz, C.F. Shu, C. Fuller, J. Bach, M. Gorkani, R. Jain, Virage Video Engine SPIE Proceedings on Storage and Retrieval for Image and Video Databases V, pages 188-97, San Jose, Feb. 1997.

[17] Hauptmann, A., Thornton, S., Houghton, R., Qi, Y., Ng, D., Papernick, N., Jin, R., Video Retrieval with the Informedia Digital Video Library System, Proceedings of the Tenth Text Retrieval Conference (TREC-2001), Gaithersburg, Maryland, November 13-16, 2001. 25.

[18] A. Hauptmann, R. Baron, M Chen, M Christel, P Duygulu, C Huang, R jin, W Lin, T Ng, N Moraveji, N Papernick, C. Snoek, G Tzanetakis, J. Yang, R. Yan, and H Wactlar, Informedia at TRECVID 2003: Analyzing and searching broadcast news video, Nov 2003, NIST TRECVID 2003.

[19] X. Huang, G Wei, and V Petrushin, Shot boundary detection and high-level features extraction for the TREC video evaluation 2003, Nov 2003, NIST TRECVID 2003.

[20] J. Jeon, V. Lavrenko and R. Manmatha, (2003) Automatic Image Annotation and Retrieval using Cross-Media Relevance Models, In Proceedings of the 26th Intl. ACM SIGIR Conf., pages 119–126, 2003.

[21] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, Pinar Duygulu, GCap: Graph-Based Automatic Image Captioning, MDDE 2004.

[22] Jiwoon Jeon, R. Manmatha, Using Maximum Entropy for Automatic Image Annotation. CIVR 2004: 24-32.

[23] D. Klein, S.D. Kamvar, and C.D. Manning, From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering, ICML'02.

[24] V. Lavrenko, R. Manmatha and J. Jeon, A Model for Learning the Semantics of Pictures. In Proceedings of the 16th Annual Conference on Neural Information Processing Systems, NIPS'03.

[25] J. Li and J. Z. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on PAMI, 25(10): 14, 2003.

[26] Ling-Yu Duan, Min Xu, Tat-Seng Chua, Qi Tian, A mid-level representation framework for semantic sports video analysis. ACM Multimedia 2003: 33-44.

[27] D. Lowe, Distinctive image features from scale-invariant keypoints, IJCV, vol. 60, issue 2, November 2004.

[28] M.R. Lyu, E. Yau, and K.S. Sze. iVIEW: An Intelligent Video over Internet and Wireless Access System, in Proc. 11th International World Wide Web Conference (WWW2002), Practice and Experience Track, Hawaii, May 7-11, 2002.

[29] Milind R. Naphade and John R. Smith, A hybrid framework for detecting the semantics of concepts and context, in Lecture Notes in Computer Science: Image and Video Retrieval, M. Lew, N. Sebe, and J. Eakins, Eds. Springer, 2003.

[30] F. Monay and D. Gatica-Perez, On image auto-annotation with latent space models, In Proc. of ACM Multimedia, 2003.

[31] A. Natsev, M. R. Naphade, J. R. Smith, Semantic representation, search and mining of multimedia content, Proc. ACM KDD'04.

[32] C.W. Ngo, T.C. Pong, and H.J. Zhang, On Clustering and Retrieval of Video Shots, In Proc. of ACM Multimedia 2001, pp. 51-60, 2001.

[33] G. Quenot, D. Moraru, L Besacier, and P Muthem, Clips at trec 11: Experiments in video retrieval, in The Eleventh Text Retrieval Conference, TREC 2002, Gaithersburg, MD, Nov 2002, pp. 181–187.

[34] M Rautiainen, J Pebttila, P. Peterila, K Noponen, M Hosio, T Koskela, S Makela, J Peltola, J Liu, T Ojala, and T Seppanen, TRECVID 2003 experiments at mediaTeam Oulu and VTT, Nov 2003, NIST TRECVID 2003.

[35] E. Sahouria and A. Zakhor, Content analysis of video using principal components, IEEE Transactions on CSVT, 9(8):1290-1298, 1999.

[36] L.K. Saul and S.T. Toweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, Journal of machine learning research 4 (2003), 119-155.

[37] D. D. Saur, Y. P. Tan, S. R. Kulkarni and P. J. Ramadge, Automated Analysis and Annotation of Basketball Video, SPIE Vol. 3022, Sep. 1997.

[38] F. Sebastiani, Machine learning in automated text categorization, In ACM Computing Surveys, Vol.34, No.1, pp.1-47, March 2002.

[39] J. Shi and J. Malik, Normalized cuts and image segmentation, IEEE Trans. on PAMI, 22(8): 888–905, 2000.

[40] Shinobu Hattor, Shin'ichi Takagiz, Akihisa Kodatez, Hideyoshi Tominaga, A content based video classification semantic description extraction, SCI2003.

[41] J. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C. Lin, M. Naphade, D. Ponceleon, and B. Tseng, Integrating features, models, and semantics for content-based retrieval, NIST video-TEC notebook, 2001.

[42] J. R. Smith and S. F. Chang, Visualseek: A fully automated contentbased image query system, in Proc. ACM Multimedia, Boston, MA, Nov. 1996.

[43] C.G.M. Snoek and M. Worring，Multimodal Video Indexing: A Review of the State-of-the-art, Multimedia Tools and Applications, 2004 (in press).

[44] F. Souvannavong, B. Merialdo, and B Huet, Latent semantic indexing for video content modeling and analysis, Nov 2003, NIST TRECVID 2003.

[45] F. Souvannavong, B. Merialdo, and B Huet, Semantic feature extraction using mpeg macro-block classification, in The Eleventh Text Retrieval Conference, TREC 2002, Gaithersburg, MD, Nov 2002, pp. 227–231.

[46] A. Torralba, K. P. Murphy, and W. T. Freeman, Sharing visual features for multiclass and multiview object detection, CVPR'2004.

[47] B.-L. Tseng, Lin, C.-Y., Naphade. M, Natsev. A, Smith. JR, Normalized classifier fusion for semantic visual concept detection, Proc. ICIP'03.

[48] P. Viola and M. Jones, Robust real-time object detection, ICCV, 2001.

[49] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, Constrained k-means clustering with background knowledge, ICML'01.

[50] K. Wagstaff and C. Cardie, Clustering with instance-level constraints, ICML'00.

[51] Wanjun Jin, Rui Shi, Tat-Seng Chua, A Semi-Naive Bayesian Method Incorporating Clustering with Pair-wise Constraints for Auto Image Annotation, accepted by ACM Multimedia 2004 (short paper), New York, USA, Oct. 2004.

[52] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In Proc. CVPR, June 2000.

[53] M. Weber, M. Welling, and P. Perona, Unsupervised learning of models for recognition, In Proc. ECCV, pages 18–32, 2000.

[54] L. Wu, Y Guo, X Qiu, Z Feng, J Rong, W Jin, D Zhou, R Wang, and M Jin, Fudan university at TRECVID 2003, Nov 2003, NIST TRECVID 2003.

[55] Yi Wu, Edward Y. Chang, Optimal multimedia fusion for multimedia data analysis. ACM Multimedia'04: 572-579.

[56] http://www-nlpir.nist.gov/projects/tv2003/tv2003.html.

[57] Xavier Gilbert, Huiping Li and David Doermann, Sports video classification using HMM, ICME2003.

[58] L. Xie, S.-F. Chang, A. Divakaran and H. Sun (2003), Unsupervised Discovery of Multilevel Statistical Video Structures Using Hierarchical Hidden Markov Models, ICME 2003, Baltimore, MD, July 2003.

[59] R. Yan and A. G. Hauptmann, The combination limit in multimedia retrieval, ACM Multimedia, 2003.

[60] D. Zhong, S.-F. Chang, Structure Analysis of Sports Video Using Domain Models, In Proc. of ICME 2001.