A Study of Content-Based Video Classification, Indexing and Retrieval

Master of Philosophy

First-term Research Paper

Supervised by

Professor LYU, Rung Tsong Michael

Prepared by

HOI, Chu Hong

chhoi@cse.cuhk.edu.hk

Department of Computer Science and Engineering

The Chinese University of Hong Kong

Hong Kong S.A.R.

November 28, 2002

Abstract

With the rapid increasing amount of video data, content based video classification and retrieval has been attracted more and more focuses in last decade. Although fruitful results were acquired in recent years, automatic analyzing the semantic content of video is still very challenging at the current state-of-the-art. In order to map the low level feature to high level semantic content, many efforts are lead to the semantic indexing and modeling of video content through semi-automatic approach. In this paper, some recent advances in content based video classification and retrieval are first reviewed. After that, a multimodal framework for video content interpretation is proposed in the paper. A lot of implementation problems and challenges are discussed.

Table of Contents

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 video classification and retrieval have been actively deployed in many research communities in past years. There have emerged a lot of successful paradigms for video parsing, indexing, summarization, classification and retrieval [1][2][3][4][5]. Although fruitful results have been achieved in last decade, more challenging problems need to be addressed and overcome in 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 video data, many problems still need to be addressed and tackled. One of important issue is how to classify and index the video data automatically or semi-automatically by machines. Another crucial and challenging issue is how to bridge the gap between the low level features and high level semantic concepts.

In this paper, we review some advance techniques in content-based video classification and retrieval recently. A lot of previous and recent surveys could be found in [6][7][8][9][10]. This paper is organized as follows. We first introduce some basic background of video structure parsing at section 2. Secondly, we survey some advances of video summarization and indexing at section 3. Then, we survey recent advances of video classification and indexing techniques at section 4. We then review two typical techniques used in video retrieval domain and address the disadvantage of current retrieval system at section 5. After that, we introduce the important concepts of semantic video analysis at section 6. Then we propose a multimodal framework for approaching video content interpretation and discuss a lot of implementation tasks involved the framework at section 7. Finally, we make a conclusion at section 8.

2. Video Structure Parsing

Video structure parsing is an initial step to organize the content of videos. Video data are typically organized in a typical hierarchical structure as shown in Figure 1.

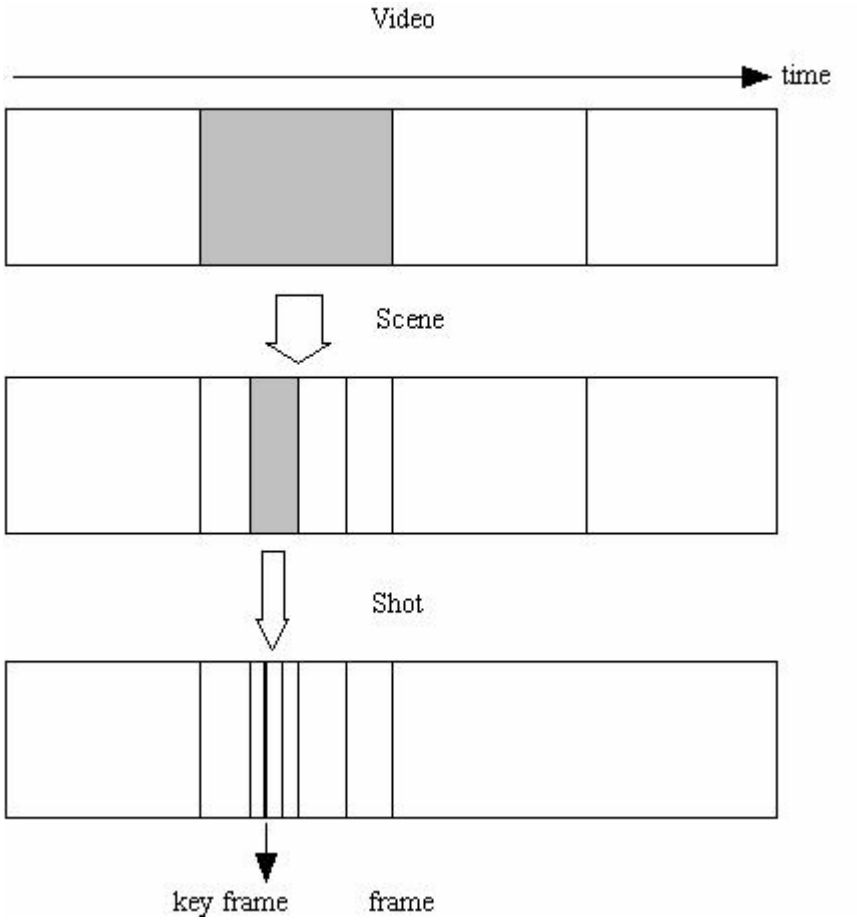

Figure 1. Hierarchical structure of video.

In this step, some elementary units such as scenes, shots, frames, key frame and objects are generated. A successful structure parsing is important for video indexing, classification and retrieval. In past, many works have been done in video structure parsing, especially in shot detection, motion analysis and video segmentation.

2.1 Shot Detection

As discussion above, video data are structured into a lot shot units. Shot changes should be detected before dividing video data into shot units. A shot change can viewed as detection of a camera break. Normally, there are three major editing types of camera breaks: cut, wipe and dissolve. A cut is an immediate change from a shot to another shot; a wipe is a change where first frame of a shot replace with last frame of another shot gradually; a dissolve is a change where one shot gradually appears (fade-in) and another shot slowly disappears (fade-out). A cut can be detected by comparing two adjacent frames. While wipe and dissolve are difficult to detect since they are change gradually.

There are much work for detection of camera breaks in past few years. They can be grouped into two categories: uncompressed and compressed domain. Some typical methods for the detection of camera breaks could be found in [6][7][8][9][10]. Recent published papers for shot change detection could be found in [11]-[17]. Some work for performance evaluation of shot detection could be found in [18][19].

Here we give some recent methods for cut detection. Vasconcelos *et al.* [11] introduced a statistical model for shot duration and activity. He extended the standard thresholding model in an adaptive and intuitive way to improve the performance of detection. Lee *et al.* [12] proposed an approach to partitioning of a video into shots based on image motion information.

2.2 Video Segmentation

Video segmentation is the important step toward to content-based video classification and retrieval. Perfect automatic segmentation of video is hard to achieve in current state-of-the-art. User interaction may need to guide the segmentation procedure. However, some successful examples of automatic video segmentation have been conducted in past few years. In most cases [20][21], video segmentation worked in a hierarchical way. They first segment the video frame into region units based on image segmentation techniques. Then regions are merged into pseudo-objects based on hand crafted heuristics or visual similarity measures [20]. In [20], Chang et al. merge the over segmentation regions using edge information. And Nguyen et al. [21] proposed a motion similarity measure to merge the over segmentation regions. Temporal and spatial information of foreground moving objects are important factors to distinguish video segmentation with image segmentation. Recent work [22]-[25] were proposed to segment the foreground moving object and background region based on motion information and background models.

3. Feature Extraction

Feature extraction is a crucial preprocessing step for a video indexing, classification and retrieval system. Most work on video classification and retrieval can be viewed as the extension of traditional image retrieval techniques. They select key frames in the video shots and extract image features based on selected key frames. However, this approaches neglect the important spatio-temporal information of video. We will review recent work of motion analysis in later section.

3.1 Visual Features

Visual features of Image are widely adopted by most video classification and retrieval system. Undoubtedly, visual features provide important information to recognize content of video. Several visual feature that frequently been adopted are color feature, texture feature, shape feature and sketch feature.

(1) Color features

Dominant color histograms with region information are most frequently adopted since it is acquired easily and seems effective in most cases. Statistical moments of color histograms are computed as following formulas:

$$mean \quad E_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij}$$

$$\text{var}iance \quad \sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N} \left(p_{ij} - E_i\right)^2 \right)^{\frac{1}{2}}$$

$$skewness \quad s_i = \left(\frac{1}{N}\sum_{j=1}^{N} \left(p_{ij} - E_i\right)^3 \right)^{\frac{1}{3}}$$

(2) Texture features

Several methods like Co-occurance matrices, Spatial Frequency, Gabor Functions and Wavelet QMF Filter are classical and effective techniques.

(3)Shape features

In analysis of shape features, edge detection need to be done before analysis. After edge detection, some parameters like circularity and eccentricity are computed.

Papers for detailed visual feature extraction can be found in image retrieval surveys [26][27].

3.2 Audio Features

Sometimes, audio feature is an additional information for video. In some cases, audio is an important channel to convey video content such as broadcast news videos. How to integrate the audio information is a crucial issue for content-based video analysis. Some frequent used audio features include loudness (It is RMS of audio signal measured in dB), pitch (It is the common divisor of peak in Fourier spectra.), brightness (It is the centriod of short-time in magnitude spectra.), bandwidth (It is the magnitude-weight average of difference between spectral components and the centroid.) and harmonicity (It is the deviation of the sound's spectrum from the

harmonic spectrum.), etc.

Although there are large advances in speech recognition and related techniques in past few years, processing nonspeech audio has little progress. In some cases, extraction of audio features is based on speech recognition in noise-free environment. Speech transcript is generated after speech recognition. Some examples can be found in [1][2][28][29]. Recently, there are a lot of attempts for segmentation and classification audio streams from video [30]-[38]. The result of audio segmentation and classification can be integrated into video classification and retrieval system as an important factor.

3.3 Motion Features

Recently, more works are proposed to exploit the spatio-temporal relation of video frames, as to extract the motion information among the video streams [20], [39]-[42]. Some survey of motion features can be found in [9]. From the survey, motion features include the motion trajectories and motion trails of objects [20], principle components of MPEG motion vectors and temporal texture. Motion trajectories and motion trails of objects are employed to describe the relation of moving objects across time. Since principle components of MPEG motion vectors contain the motion information of a video sequence, they could be imported as vectors in a video clustering and classification system. Temporal textual are applied for modeling more complicated dynamic motions [42].

Besides the motion features of moving object, it still need to address the camera motion features in a video sequence. In previous section, we have discussed that camera motion is important for detection of video shots. It is also viewed as an important feature for video content indexing and classification. Camera motions are normally divided into six main categories: panning (horizontal rotation), tilting (vertical rotation), zooming (focal length change), tracking (horizontal transverse movement), booming (vertical transverse movement) and dollying (horizontal lateral movement) [6].

4. Video Classification and Indexing

Video classification is useful for content-based video browsing, filtering and searching. If video content can be categorized into different genres, domain-specific content-based analysis algorithm can be applied to different genres. For examples, if

we can categorize the TV program into sports program and news program, then we may apply sports domain knowledge to classify the sports program into football, basketball and baseball program, etc. However, topic classification of video content is still a challenging problem. There are many works to approach this difficult task in recent years [43]-[49]. Here, we survey several recent examples.

4.1 Video Classification

Different methods have been introduced to categorize the video into predefined genres automatically or semi-automatically recently. Some research work carried out the classification by using domain methods and exploiting the temporal information along with the visual information in the video [43]. Some work are based on multimodal framework. In [43], Dimitrova *et al.* presented a method for video classification using face and text trajectories based on Hidden Markov Model (HMM). In their experiments, they integrated the information of face and text, including their size, movement and duration, to classify four type TV programs: News, Commercial, Sitcom and Soap. And the HMM framework could grasp the temporal information along with the video. However the method can be applied in limited domain since they did not fully exploit other video information. In [44], Reaaijmakers *et al.* proposed a multimodal topic segmentation and classification of news video. They presented a fully automated feedback modal for the interaction between visual, auditory and textual resources. They introduced domain-specific knowledge to segment and classify the video content based on visual, audio and textual information. For example, they segmented the different news topics by detecting the appearance of anchorperson since news topics are normally opened and closed by the anchorperson. For audio features, they detected silent period as the boundary of two news topic since anchorperson usually ends news topic with a noticeable silence. Although the idea of their framework is fine, they did not point out how to combine the visual and textual features in their framework which would significantly affect the performance of classification.

4.2 Indexing and Summarization

Besides topic classification of video, effective video indexing and summarization are important for efficient and effective browsing, searching and handling of video documents [10]. Manual indexing is not suitable for large video collections. Many automatic or semi-automatic methods were proposed to facilitate the indexing of video content in past [10][50]-[53][8]. Most methods for automatic video indexing are

based on unique modal, such as visual, auditory or textual feature. A survey of unique modal approach could be found in [8]. Instead of unique modal, multimodal video indexing is employed to classify a video by fusing multimodal information. A review of multimodal indexing can be found in [53][10]. A recent review [10] of semantic index hierarchy of video documents is shown in figure 2.

Multiple modality indexing usually adopts Hidden Markov Model (HMM) to grasp the temporal information. A multimodal learning network can be illustrated in figure 3. In the figure, input feature vector is capture from multiple modalities, such as visual, audio and textual, etc. After receiving the input feature, each modality has a corresponding HMM classifier. Each HMM classifier outputs its probability parameter as the input of a 3-layer perception neural network. Once the neural network finished some training process, it can work to output classification result for testing data.

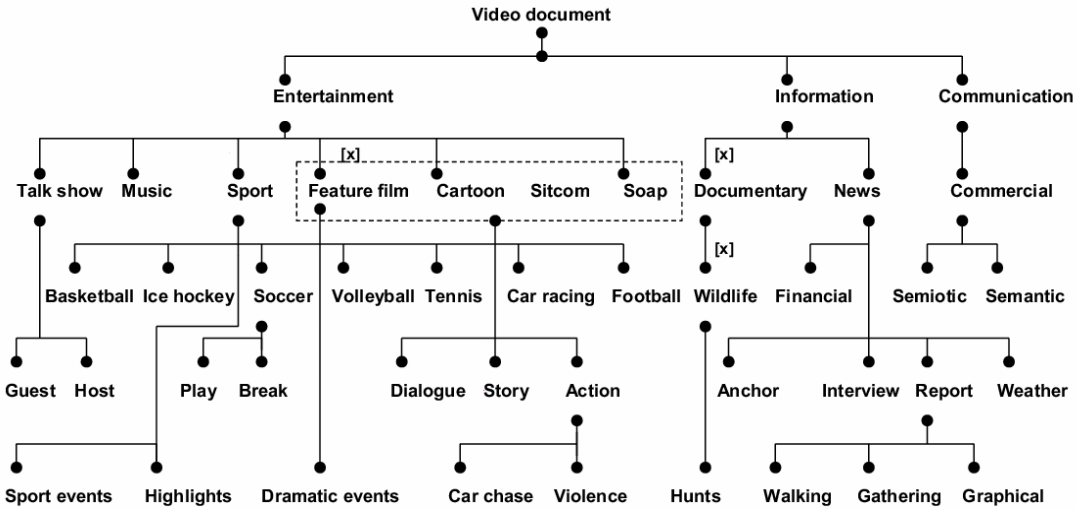

Figure 2. Semantic index hierarchy of video documents

4.3 Content Description Interface — MPEG-7

In order to solving the difficult problem of organizing and searching problem of multimedia content, Moving Picture Experts Group proposed MPEG-7 standard in 2001. MPEG-7 is a standard for describing the multimedia content data that supports some degree of interpretation of the information's meaning, which can be passed onto, or accessed by, a device or a computer code [72]. Figure 3 [72] shows the scope of MPEG-7. From the figure, we may see that MPEG-7 just focuses on multimedia content description. How to generate the description and how to consume the description is out of range of MPEG-7. In state-of-the-art, automatic generating the

raw video data into MPEG-7 format is still impossible. Most current techniques are based on semi-automatic indexing methods. Recently, a lot of works are performed to facilitate the automatic description tools of MPEG-7. Here, we will brief review the overview of MPEG-7. A lot of detailed survey and overviews can be found in [73]-[80].

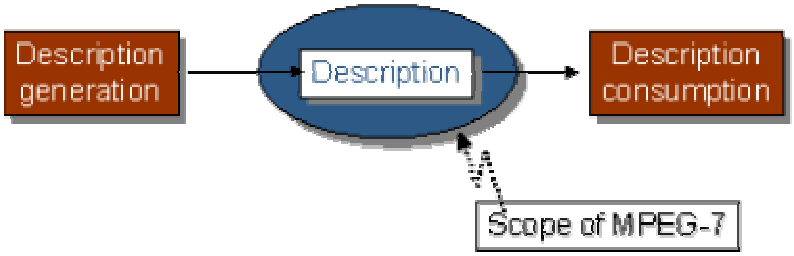

Figure 3. Scope of MPEG-7

In MPEG-7, Descriptor, Description scheme and Description Definition Language (DDL) terminologies are defined. The DDL allows the definition of the MPEG-7 description tools, both Descriptors and Description Schemes [72]. The description tools are described based on textual format (XML format). A figure of MPEG structure is shown in figure 4.

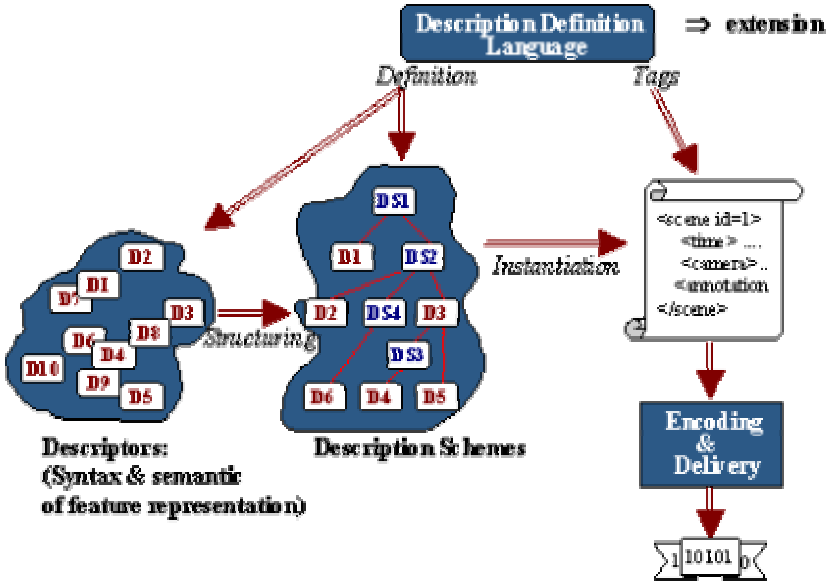

Figure 4. Structure of MPEG-7 Elements

Before establishing the standard of MPEG-7, various proposals for color, texture, shape/contour and motion descriptors were evaluated for their performance. Some of the successful descriptors are adopted into the standards. We list several successful descriptors in follows.

1. Visual Color Descriptor [81]
   Color is widely used in current image and video retrieval. In MPEG-7, several color descriptors are defined. We briefly list the descriptor in follows.
   (1) Scalable Color Descriptor
   It is used to describe the color distribution in image.
   (2) Dominant Color Descriptor
   Dominant Color Descriptor is employed to describe the global color distribution as well as local color distribution of image and video visual information.
   (3) Color Layout Descriptor
   Color layout descriptor aims to describe the spatial distribution of color in any region of image or video data.
   (4) Group-of-Frame/Group-of-Pictures (GoF/GoP) Color Descriptor
   The GoF/GoP Color descriptor is designed to define a structure to represent the color feature of similar video frames.

2. Visual Texture Descriptors
   (1) Homogenous Texture Descriptor
   The Homogenous Texture Descriptor is designed to describe directionality, coarseness and regularity of patterns in images. It is suitable in the quantitative format for describing still image texture.
   (2) Non-Homogenous Texture Descriptor
   Non-Homogenous Texture Descriptor is designed to describe non-homegenous texture images. An Edge Histogram Descriptor is defined in MPEG-7 to capture the spatial distribution of sketch of image and video frames.

3. Visual Shape Descriptors
   (1) 3-D Shape Descriptor
   The descriptor is designed to comparing the nature 3-D objects and virtual 3-D object extracted from multimedia data.
   (2) Region-Based Descriptor
   Region-based descriptor is suitable to describe some regions that can be represented in shape regions rather than in sketch region.
   (3) Contour-Based Shape Descriptor
   It is based on curvature scale-space representation.
   (4) 2-D/3-D Shape Descriptor

4. Motion Descriptor for Video
   (1) Motion Activity Descriptor

(2) Camera Motion Descriptor
(3) Warping Parameters Descriptor
(4) Motion Trajectory Descriptor

5. Audio Descriptor
   (1) Basic Descriptor
   (2) Basic Spectral Descriptor
   (3) Signal Parameters Descriptor
   (4) Timbral Temporal Descriptor
   (5) Spectral Basis Descriptor
   (6) Silence segment Descriptor

Detailed information about MPEG-7 descriptors can be found in references [73]-[81].

5. Video Retrieval

To date, most video retrieval systems are used to retrieval similar video based on low level features. Video retrieval faces the same problem with image retrieval that it lacks a semantic model and effective representation tool to express human perception. There exists a gap between high semantic concept and low level features. How to bridge the gap is the most challenging topic in video classification and retrieval field. In this section, we will survey recent work on similarity measure based and clustering based video retrieval. Exploration of semantic video retrieval will be given in later section.

5.1 Similarity Measure Based Retrieval

In current video retrieval system, there are two method used for retrieval: similarity measure based and cluster-based method [9]. Similarity measure method is employed to retrieval similar video key frame, shot or video scene segment. Similarity measure can be conducted by matching the features locally or globally. In a simple way, similarity measure is based on computing the similarity of related key-frame between two videos. More sophisticated methods are employed the spatio-temporal features of video frames between two videos [20][54] [41]. Chang *et al.* [20] proposed a method to retrieval video object by computing similarity of motion trajectories and trails in the spatial and temporal domains. Chang *et al.* also presented a semantic visual template which can express the semantic concept [59]. Detailed explanation of the idea will be discussed in later section. Dagtas *et al.* [41] presented several motion

descriptors as intermediate motion model for event-based video retrieval. They retrieved the event videos by computing the similarity of different motion models.

5.2 Clustering Based Retrieval

Clustering method is introduced as a solution to organize the content of video collections. It provides efficient method to classify and index the video since similar video are clustered into similar group. Recent work on cluster-based retrieval can be found in [55][56][66]. In [66], Clarkson *et al.* proposed a framework to find the event by clustering the nature input audio/visual data. They developed a system that can cluster the video data into events such as passing through doors and crossing the street [66]. The clustered events can also be clustered into high-level scene.

6. Semantic Modeling and Indexing of Video

Semantic video modeling and indexing of video content is viewed as the most final frontier of computer vision and multimedia. Effective semantic modeling and indexing of video is a way to ultimate multimedia understanding. Figure 5 illustrates the relationship between the video content interpretation and natural language processing. Currently most works are focusing on frame-based structure modeling. Fully automatic multimedia understanding is almost impossible in state-of-the-art. Although it is a very challenging work, there still have some good research work resided on this topic [57]-[69]. Here we will review several novel work in recent years.

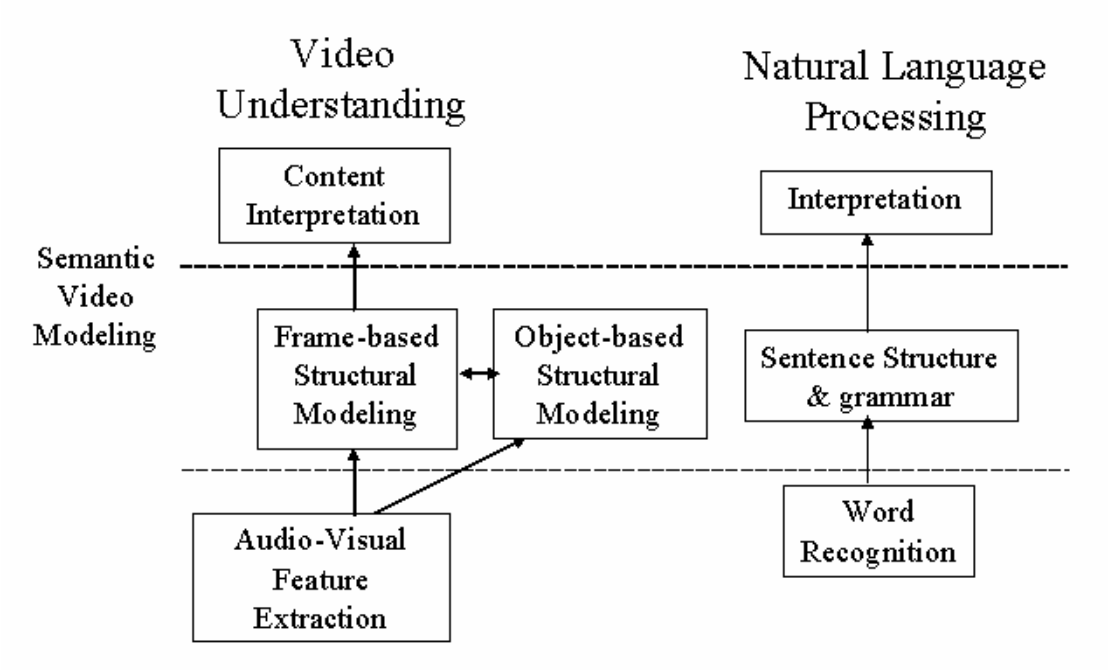

Figure 5. Three layer structure of video modeling comparing with NLP

6.1 A Probabilistic Framework: Multijject and Multinet

In [57], Naphade *et al.* proposed a probabilistic framework for modeling multimedia object called 'Multiject' and modeling semantic concepts called 'Multinet'. Multijects belong to one of the three categories: objects (car, man, building), sites (outdoor, beach, mountain) and events (explosion, man-walking, dancing) [57]. Figure 6 illustrated the concept of a multiject. A multiject can support every level concept. It can represent low-level feature, such as visual features, audio features and textual features. It can also express the intermediate-level meaning such as semantic template [58] and other high-level semantic concepts. In case of present of a multiject, other high level multiject could be formulated if there is some correlation between low-level features and high-level semantic.

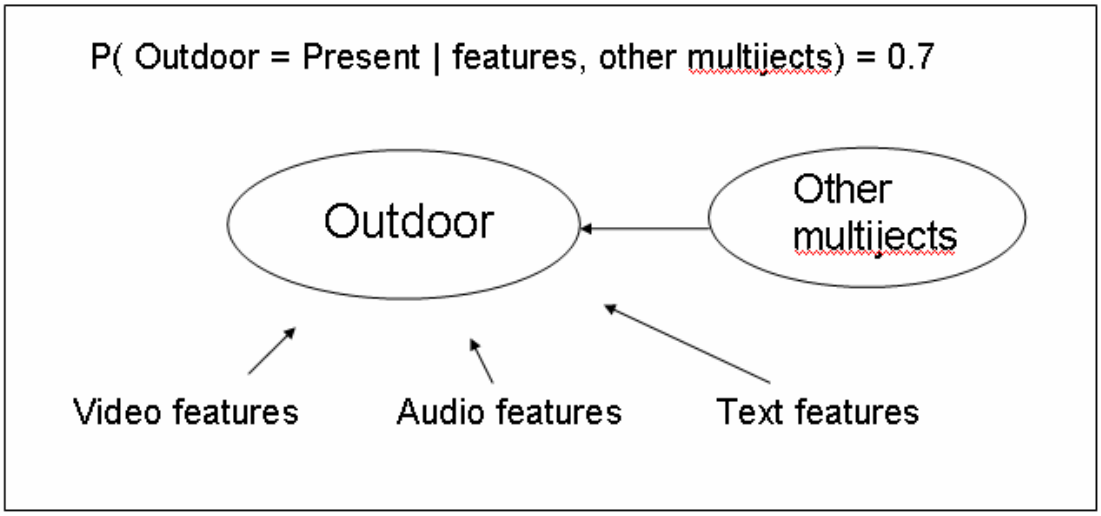

Figure 6. A probabilistic multimedia object

In order to create a multiject, a video sequence should make some preprocessing work to detect the shot boundary. Then spatio-temporal segmentation is made within shot frames. After that, region-based feature extraction should be done. Feature extraction includes color, texture, edge direction and shape, etc. In order to model the semantic concept of different multijects, different distribution models are introduced. For example, in order to model the 'sites', mixture of Gaussian model is employed. And hidden Markov models are introduced to model the events in the multijject system. Not only for modeling simple modal, hierarchical HMMs are employed to modeling multiple modalities. A hierarchical HMM model is shown in figure 7. In the figure, 'AO' is the audio observations of HMM model and 'VO' is the video observations of HMM model. The 'Qi' nodes are the nodes of supervisor HMM.

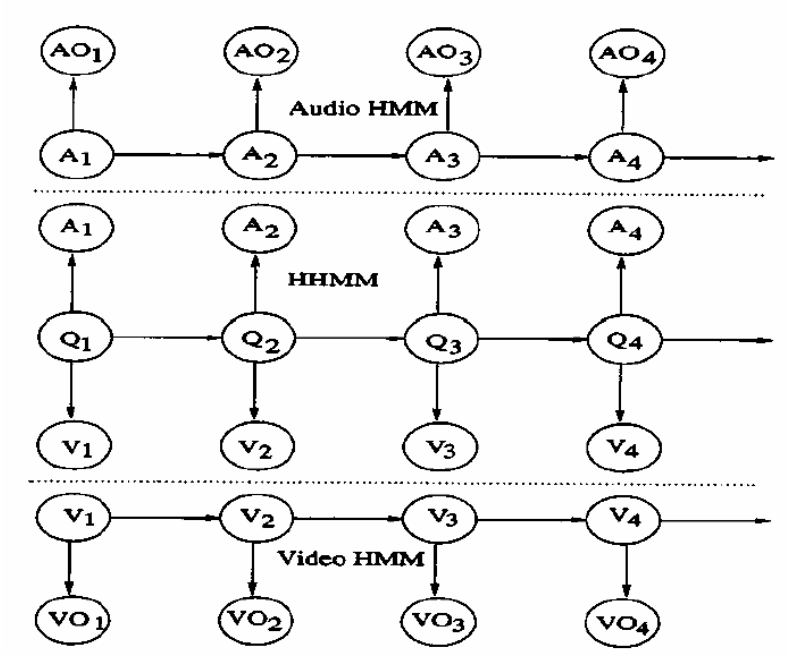

Figure 7. A hierarchical HMM model

In order to model the semantic interaction of multiject, a multinet concept is proposed in [57]. Figure 8 is an interaction network between multimedia object. In the figure,

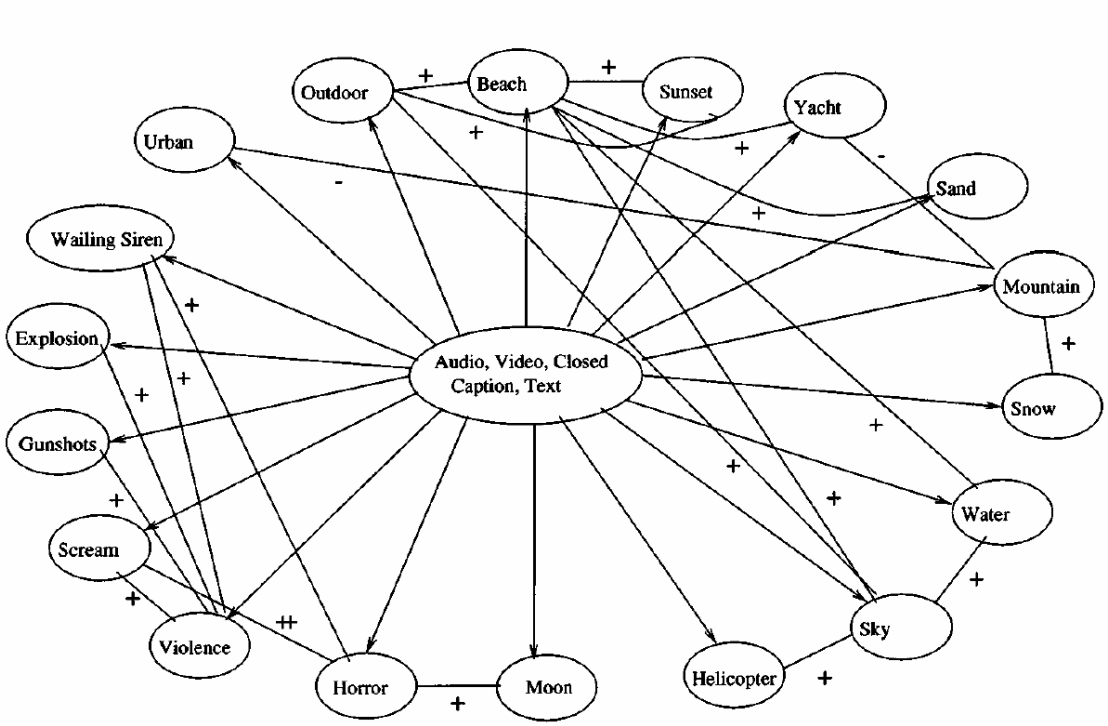

Figure 8. An interaction network of multiject (multinet)

the '+' sign means there is a positive relation between the two multijects. By contraries, the '-' sign means there is a negative relation between the two multijects. The reason of using multinet is that semantic concepts do not occur independently or in isolation from each other [57]. There is an important cooccurrence between multimedia objects. It is easy to figure that the present of a multiject will increase the probability of present of another multiject. For example, the sign between 'water' and 'outdoor' is '+' that means the present probability of water increase the present probability of outdoor.

The advantages of a multinet are that it provides a framework for support four aspects of semantic indexes. [57]. First, it may enhance the detection by using the mutual detection. Secondly, it could support inference based on the interaction between multijects. Thirdly, the multinet can provide the mechanism for imposing prior knowledge of multimodalities and enforce context-changes on the structure. Also the multinet can combine multiple classifiers and fuse multiple modalities.

Although the idea of multiject and multinet is a very good framework to modeling the

semantic concepts between multimedia objects, the framework has some disadvantages. One of its disadvantages is that the complexity of the framework will increase exponentially when the range of knowledge is increased. It is hard to handle such a network when the collection of nodes is very large. Some interaction between multijects may not be independent, thus it is difficult to compute the strength of the network in such cases.

6.2 Semantic Visual Template

In [58], Chang *et al.* provide a Semantic Visual Templates (SVT) to modeling the low-level feature and high level semantic object. They introduced a idea of SVT to bridge the gap between the user's information needs and what the systems can deliver. Each template represents a semantic concept, e.g. sunset, meetings, slalom, etc. The differences of SVT distinct from other methods are follows[58]. First, SVT employs a two-way learning method. Machine can improve the retrieval performance by learning the relevant feedback from the user. Secondly, SVT emphasizes the intuitive and understandable method to express the semantic concepts of video and make it easily understood by users. Also several SVTs can be combined to form a more complicated semantic concept.

In their work, they used the VideoQ system to implement the idea of SVT and showed that it is effective for modeling the semantic concepts. Before the SVT-based query can work, the videos in the VideoQ system should be preprocessed. The first preprocessing step is to do the scene and shot detection. Then, automatic object segmentation and tracking are employed in the detected shot. For the segmented objects, visual features and motion features are extracted from them. Also the spatio-temporal information of the object is extracted. In the VideoQ, a semantic visual template is made up of several icons. Each icon represents a semantic concept in the video shots. According to different semantic concept, the icons may emphasize different features. The elements of the icon set can overlap in the templates. On of the goals of the templates is to cover all the concepts by minimizing the number of icons. A template example from [58] is illustrated in figure 9. In the example, we can see that the 'high-jumer' template emphasizes the point of motion feature. However, the 'sunset' template focuses the global color feature.

Although the semantic visual template can express the semantic concept intuitively, however it can only describe some basic and simple semantic concept. It is quite difficult to represent a high-level semantic event concept by sketch an intuitive

template. For example, if I want to semantic concept like ‘Christmas’, user can not know how to express their ideas in the SVT framework. In order to overcome the disadvantage, Chang *et al.* have added Bayesian relevant feedback during the retrieval

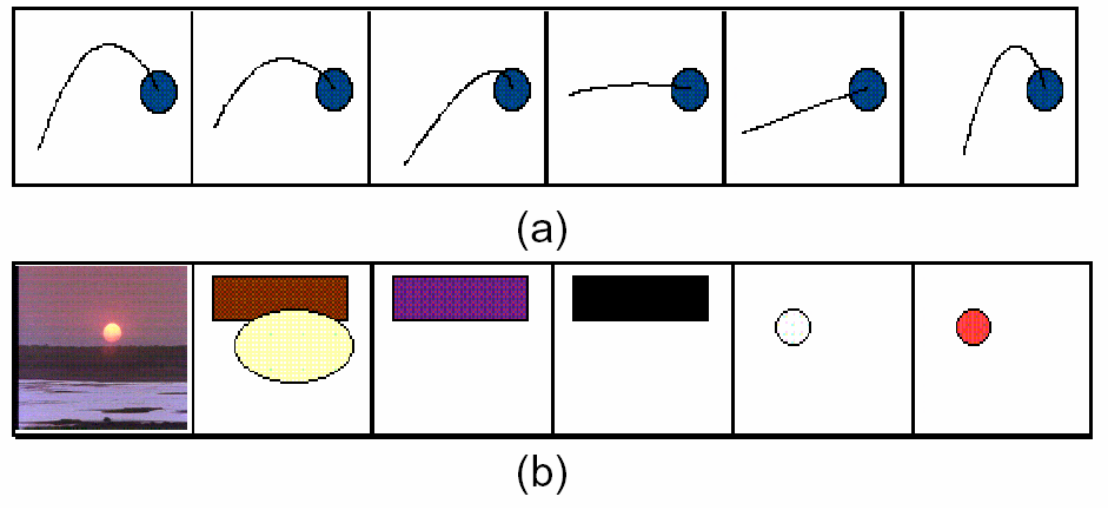

Figure 9. Examples of SVT (a) a ‘high-jumper’ template (b) a ‘sunset’ template

procedure. In the retrieval process, new querying templates are generated from initial sketch by user. User can narrow down the templates that represent their desired concept by through the Bayesian relevant feedback.

6.3 Semantic Video Object-based Abstraction

In past, a lot of works are addressed to extract and abstract the video objects in order to model the semantic concepts of objects and events. In [64], Hwang *et al.* proposed a scheme for object-based abstraction and analysis and semantic event modeling. In their approach, video objects (called VOs) are extracted automatically [64]. In their framework, they adopted two methods to extract the Video objects. One of the methods is change detection which is suitable for stationary background. Another is object tracking method which is applied in moving background [64]. They assumed that the semantic content is focused on extracted VOs. Other regions in the video frame are considered as background. After extraction of VOs, semantic feature modeling based on the low-level features of VOs is performed. They employed Dynamic Bayesian Network (DBN) to modeling the semantic concepts of video objects and events [64]. A block diagram of an object-based analysis procedure [64] is shown in figure 10.

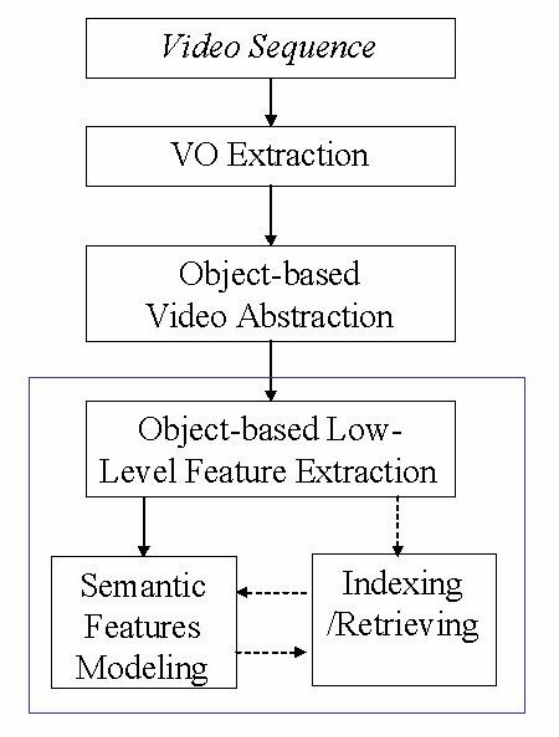

Figure 10. Block diagram of an object-based analysis and interpretation system

Fully automatic video object based analysis and interpretation is a way toward video understanding. However, in state of the art, it is difficult to build such a system since the semantic features modeling depending on demain-specific knowledge. It'd better to introduce some method to facilitate this approach. Maybe the probabilistic framework of 'Multiject' and 'Multinet' [57] could approach this goal.

7. A Framework For Video Content Interpretation

Automatic video content interpretation is a long-term goal toward machine understandable video processing system. In current state-of-the-art, fully automatic video content interpretation is challenging. However, there still exist some ways to approach. In this section, we propose a multimodal framework to approach this challenging topic.

7.1 Motivation of Video Content Interpretation

In near future, digital TV programs will become more and more popular. People may hope to select their interesting programs and filter some dislike topic. Automatic video scout will become an important function in future personal digital video

processing system [71].

In order to approach the video content interpretation, we propose a multimodal framework which can automatically segment and classify different topics of TV programs. Before go through the details of our framework, we may give a brief definition about multimodality[.

Defnition1 (Multimodality) *"The Capacity of an author of the video document to express a predefined semantic idea, by combining a layout with a specific content, using at least two information channels"[75].*
Three modalities will be considered in our framework:
(1) *Still Visual modality*
In this modality, still image features in the video frames are focused. Some traditional image processing techniques could be applied in this modality.
(2) *Motion Visual modality*
In this modality, we focus on motion features of video data. Temporal-Spatio relationship is important information in video frames.
(2) *Auditory modality*
This modality contains the speech, music and environment sounds including noise. It is very useful for us to detect and extract the information from the audio.
(3) *Textual modality*
This modality contains the video transcript that describes the video content. Speech recognition and video optical character recognition will be helped in this modality.

7.2 Main Idea of the Multimodal Framework

In order to integrate all possible content information to our system, we adopt multiple modalities as visual, audio and textual modalities. Not only do the frame-based structure modeling, we also explore the semantic information of video object. We employed Bayesian Network to modeling the classifier. A multimodal framework for video topic classification (VTC) is illustrated in figure 11.

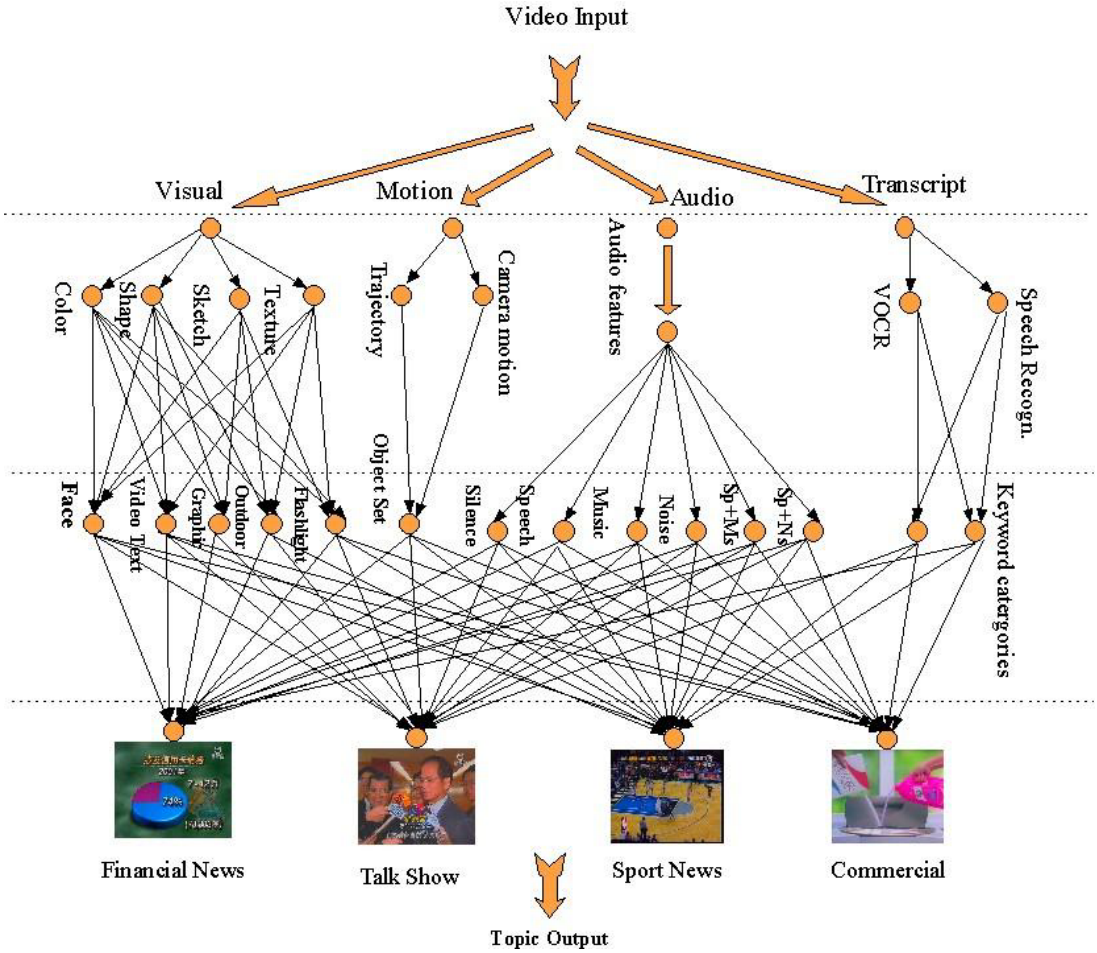

Figure 11. A multimodal framework for Video Topic Classification

There are three levels in the VTC framework. The low-level is frame-based level. In this level, low-level features are extracted. In middle level, basic semantic concepts are constructed. In high-level, final semantic topics are generated.

7.3 Implementation of the Framework

In order to build such a VTC system, a lot of problem should be solved. A several important steps of system scheme are listed at follows.

1. Preprocessing

We need to do the preprocessing on the input video sequence. Firstly, we will do the shot detection and scene segmentation in the video frame. After that, we may need to select the key frame in the video shot. We may also need to do some other preprocessing on the video frames. For examples, Speech Recognition and Video Optical Character Recognition (Video-OCR) can be done in this step. Some features may also can be done in this step, such as camera motion feature extraction.

2. Low-Level Feature Extraction

After preprocessing of the video sequence, we can obtain the key-frame sequences. Then we do the feature extraction on the key frames. We may need to extract the video objects (mainly focus on moving object) and their visual, audio and motion features. The features which are extracted in our frame include follows:

(1). Visual features:

(a) Color

(b) Texture

(c) Shape

(d) Sketch

(2). Audio features:

(a) Average energy / Loudness

(b) Bandwidth

(c) Pitch

(d) Brightness

(e) Harmonicity

(f) me-frequency cepstral coeffiency

(3). Textual features

The text information is the transcript generated from speech recognition and video-OCR. We build a knowledge tree which has a lot of key word categories. Then we use a vote mechanism to process the transcripts. When a key-word is is spotted in the transcript, a vote is increase in the relative category. After processing, we have a multi-dimension textual feature vector.

(4) Motion Object features

For the extracted video object, we need to find their motion trajectories. Moreover, other visual and audio features of the video objects will need to be extracted.

3. Building and Training Bayesian Network

After low-level feature extraction, we need to build and train the network. We use Bayesian Network since it can represent the causal relationship between video objects. Hence we may inject some domain-specific knowledge to improve our classification system.

4. Modeling semantic objects and events

To model the semantic objects and events is an important module in our framework. We will first use hidden Markov model to classify the video objects and then we employ a probabilistic network to modeling the interaction of video objects.

5. Output the topic of video content by the network
   After integrated multiple modalities and building the Bayesian network, we may first training the network. In order to improve the performance, we may include the domain knowledge in the post processing.

7.4 Challenging Problems

Although the scheme seems feasible, we know that there still have several challenging works to be done before building such a VTC system. A lot of challenging problems should be addressed.
(1) Preprocessing is significant in the framework.
To date, accuracy of key-frame selection is still absolutely satisfied. Also the accuracy of speech recognition and VOCR are still not very good in state-of-the-art.
(2) Good feature extraction is important for the performance of classification.
(3) It is difficult to model semantic video objects and events.
(4) How to integrate multiple modalities still need to be well considered.

8. Conclusion

In this paper, we first give a survey of current research of video indexing, classification and retrieval. From the survey, we know that the most challenging topic in content-based video retrieval domain is how to bridge the gap between the low-level features and high level semantic concept. How to achieve automatic video interpretation is the long-term goal in this area. In order to approach this target, video structure modeling and object modeling should be well done before semantic concept can be constructed automatically. Current most works are focused on video structure modeling although there still are a lot of pioneering works. After the survey, we propose a framework for approaching the target of video interpretation. In order to narrow down the range of our experiment, we will apply the framework on TV video. We show that our framework is novel and effective. We also discuss some challenging problem involved in the framework.

References

[1] 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.

[2] 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.

[3] 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.

[4]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 QBIC system. IEEE Computer, 28(9):23-32, 1995.

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

[6] R. Brunelli, O. Mich, and C. M. Modena. A survey on video indexing. IRSTTechnical Report 9612-06, 1996.

[7] G.Ahanger and T. D. C. Little. A survey of technologies for parsing and indexing digital video. Journal of Visual Communication and Image Representation, 7(1):28-43, March 1996.

[8] R. Brunelli, O. Mich, and C. M. Modena. A survey on the automatic indexing of video data. Journal of Visual Communication and Image Representation, 10:78-112, 1999.

[9] C.-W. Ngo, T.-C. Pong and H.-J. Zhang, Recent advances in content based video analysis. International Journal of Image and Graphics, 1(3):445-468, 2001.

[10] C.G.M. Snoek and M. Worring. A State-of-the-art Review on Multimodal Video Indexing. Proceedings of the 8th Annual Conference of the Advanced School for Computing and Imaging pages 194--202, Lochem, The Netherlands, 2002.

[11] N. Vasconcelos and A. Lippman. Statistical models of video structure for content analysis and characterization. Image Processing, IEEE Transactions on , 9(1):3-19, Jan. 2000.

[12] Bouthemy, P., Gelgon, M., and Ganansia, F. A unified approach to shot change detection and camera motion characterization. Circuits and Systems for Video Technology, IEEE Transactions, 9(7):1030-1044, Oct. 1999.

[13] R. Lienhart. Comparison of automatic shot boundary detection algorithm. In SPIE Proc. Storage and Retrieval for Image and Video Database VII, pages 290-301, 1999.

[14] R.A. Joyce and B. Liu. Temporal segmentation of video using frame and histogram space. In Int. Conf. on Image Processing, 2000.

[15] C. W. Ngo, T. C. Pong, and R. T. Chin. A robust wipe detection algorithm. In Asian Conference on Computer Vision, 1:246-251, 2000.

[16] S. W. Lee, Y. M. Kim, and S. W. Choi. Fast scene change detection using direct feature extraction from mpeg compressed videos. IEEE Transaction on Multimedia, 2(4):240-254, 2000.

[17] M. S. Drew, S. N. Li and X. Zhong. Video dissolve and wipe detection via spatio-temporal

images of chromatic histogram differences. In Int. Conf. on Image Processing, 3:929-932, 2000.

[18] U. Gargi, R. Kasturi, S.H. Strayer. Performance characterization of video-shot-change detection methods. Circuits and Systems for Video Technology, IEEE Transactions on, 10(1):1-13, Feb. 2000.

[19] L. F. Cheong. Scene-based shot change detection and comparative evaluation. Journal of Computer Vision and Image Understanding, 79(2):224-235, Aug 2000.

[20] S. F. Chang, W. chen, H. J. Meng, H. Sundaram and D. zhong. A fully automatic content-based vieo search engine supporting multi-object spatio-temporal queryies. IEEE Transaction on Circuits and Systems for Video Technology, 8(5):602-615, 1998.

[21] H.T. Ngyuyen, M. Worring and A. Dev. Detection of moving objects in video using a robust motion similarity measure. IEEE Transaction on Image Processing, 9(1):137-141, Jan 2000.

[22] Y. Tsaig, A. Averbuch. Automatic segmentation of moving objects in video sequences: a region labeling approach. Circuits and Systems for Video Technology, IEEE Transactions on, 12(7):597-612, July 2002

[23] Shao-Yi Chien; Shyh-Yih Ma; Liang-Gee Chen. Efficient moving object segmentation algorithm using background registration technique. Circuits and Systems for Video Technology, IEEE Transactions on , 12(7)577-586, Jul 2002.

[24] Chew Keong Tan, M. Ghanbari. Using non-linear diffusion and motion information for video segmentation. Proceedings. 2002 International Conference on Image Processing , 2:769-772, 2002.

[25] F. Precioso, M. Barlaud. Regular spatial b-spline active contour for fast video segmentation. Proceedings. 2002 International Conference on Image Processing. 2002., 2:761 -764, 2002.

[26] J. Shanbehzadeh, A. Moghadam, and F. Mahmoudi, Image indexing and retrieval techniques: Past, present and next, in Proc. SPIE Storage Retrieval Media Databases, vol. 3972, San Jose, CA, Jan. 2000, pp. 461–470.

[27] Y. Rui, T. S. Huang, and S. F. Chang, Image retrieval: Current techniques, promising directions and open issues. J. Visual Commun. Image Representation, vol. 10, pp. 39–62, Mar. 1999. Ph.D., MIT, Cambridge, MA, 1996.

[28] Audio Search Using Speech Recognition. Compaq Corporate Res. Lab.. [Online]. Available: http://speechbot.research.compaq.com

[29] J. Nam, A. E. Cetin, and A. H. Tewfik, Speaker identification and video analysis for hierarchical video shot classification, in Proc. IEEE Int. Conf. Image Processing, vol. 2, Santa Barbara, CA, Oct. 1997, pp. 550–555.

[30] E. Scheirer and M. Slaney, Construction and evaluation of a robust multifeatures speech/music discriminator, in Proc. IEEE Int. Conf. Accoust., Speech, Signal Processing, vol. 2, Munich, Germany, 1997, pp. 1331–1334.

[31] M. R. Naphade and T. S. Huang, Stochastic modeling of soundtrack for efficient segmentation and indexing of video, in Proc. SPIE Storage Retrieval Multimedia Databases, vol. 3972, Jan. 2000, pp. 168–176.
[32] M. Akutsu, A. Hamada, and Y. Tonomura, Video handling with music and speech detection, IEEE Multimedia, vol. 5, no. 3, pp. 17–25, 1998.
[33] D. Ellis, Prediction-Driven Computational Auditory Scene Analysis, Ph.D., MIT, Cambridge, MA, 1996.
[34] P. Jang and A. Hauptmann, Learning to recognize speech by watching television, IEEE Intell. Syst. Mag., vol. 14, no. 5, pp. 51–58, 1999.
[35] E. Wold, T. Blum, D. Keislar, and J. Wheaton, Content-based classification search and retrieval of audio, IEEE Multimedia, vol. 3, no. 3, pp. 27–36, 1996.
[36] T. Zhang and C.Kuo, An integrated approach to multimodal media content analysis, in Proc. SPIE, I\&T Storage Retrieval Media Databases, vol. 3972, San Jose, CA, Jan. 2000, pp. 506–517.
[37] Z. Liu, Y. Wang, and T. Chen, Audio feature extraction and analysis for scene segmentation and classification, VLSI Signal Processing Syst. Signal, Image Video Technol., vol. 20, pp. 61–79, Oct. 1998.
[38] S. Pfeiffer, S. Fischer, andW. Effelsberg, Automatic audio content analysis, in Proc. ACM Int. Conf. Multimedia, Boston, MA, Nov. 1996, pp. 21–30.
[39] Motion texture: a new motion based video representation Yu-Fei Ma; Hong-Jiang Zhang Pattern Recognition, 2002. Proceedings. 16th International Conference on , Volume: 2 , Page(s): 548 -551, 2002
[40]A new perceived motion based shot content representation Yu-Fei Ma; Hong-Jiang Zhang Image Processing, 2001. Proceedings. 2001 International Conference on , Volume: 2 , Page(s): 426-429 vol.3, 2001
[41]Models for motion-based video indexing and retrieval Dagtas, S.; Al-Khatib, W.; Ghafoor, A.; Kashyap, R.L. Image Processing, IEEE Transactions on , Volume: 9 Issue: 1 , Page(s): 88-101, Jan. 2000
[42] R. Fablet, P. Bouthemy, and P. Perez. Statistical motion-based video indexing and retrieval. In Int. Conf. on Content-based Multimedia Info. Access, pages 602-619, 2000.
[43] G. Wei, L. Agnihotri and N. Dimitrova, TV program classification based on face and text, IEEE multimedia and Expo 2000, New York, 2000
[44] Nevenka Dimitrova, Lalitha Agnihotri and Gang Wei . Video classification based on HMM using text and faces. European Conference on Signal Processing, Finland, September 2000
[45] Multimodal topic segmentation and classification of news video Raaijmakers, S.; den Hartog, J.; Baan, J. Multimedia and Expo, 2002. Proceedings. 2002 IEEE International Conference on , Volume: 2 , Page(s): 33 -36, 2002
[46] A. Hanjalic, R. L. Lagendijk, J. Biemond, Semi-Automatic News Analysis, Indexing, and

Classification System Based on Topics Preselection, Proc. of SPIE: Electronic Imaging: Storage and Retrieval of Image and Video Databases,January, San Jose, 1999.

[47] Lu, C., Drew, M.S. and Au, J. *Classification of Summarized Videos using Hidden Markov Models on Compressed Chromaticity Signatures*. ACMMultimedia, Ottawa, Canada, 2001.

[48] D. Beeferman, D., A. Berger, A. and Lafferty, J. *Statistical models for text segmentation*. Machine Learning, vol. 34, pp. 1-34, 1999

[49] W.Greiff, A. Morgan, R. Fish, M. Richards, and A. Kundu, Fine-grained hidden markov modeling for broadcast –news story segmentation, in ACM Multimedia 2001, 2001

[50] A. Alatan, A. Akansu, and W. Wolf. Multi-modal dialogue scene detection using hidden markov models for content-based multimedia indexing. Multimedia Tools and Applications, 14(2):137-151, 2001.

[51] N. Babaguchi, Y. Kawai, and T. Kitahashi. Event based indexing of broadcasted sports video by intermodal collaboration. IEEE Transactions on Multimedia, 4(1):68–75, 2002.

[52] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[53] Y. Wang, Z. Liu, and J. Huang. Multimedia content analysis using both audio and visual clues. IEEE Signal Processing Magazine, 17(6):12–36, 2000.

[54] Y. Wu, Y. Zhuang and Y. Pan. Content-based video similarity model. In ACM Multimedia, 2000.

[55] W. Zhou, A. Vellaika, and C. C. Jay Kuo. Rule-based video classification system for basketball video indexing. In international workshop on Multimedia Information Retrieval, 2000.

[56] E. Sahouria and A. Zakhor. *Content analysis of video using principal components*. IEEE Transactions on Circuits and Systems for Video Technology, 9(8):1290--1298, 1999

[57] Extracting semantics from audio-visual content: the final frontier in multimedia retrieval *Naphade, M.R.; Huang, T.S.* Neural Networks, IEEE Transactions on , Volume: 13 Issue: 4 , Page(s): 793 -810, July 2002

[58] S. F. Chang, W. Chen, and H. Sundaram, “Semantic visual templates—Linking features to semantics,” in Proc. IEEE Int. Conf. Image Processing, vol. 3, Chicago, IL, pp. 531–535., Oct 1998,

[59] W. Chen and S. Chang, “Generating semantic visual templates for video databases,” in Proc. IEEE Int. Conf. Multimedia Expo, vol. 3, New York, NY, pp. 1337–1340., July 2000,

[60] Extracting semantic information from news and sport video *Assfalg, J.; Bertini, M.; Colombo, C.; Del Bimbo, A.* Image and Signal Processing and Analysis, 2001. ISPA 2001. Proceedings of the 2nd International Symposium on , Page(s): 4 -11, 2001

[61] Structural and semantic analysis of video *Shih-Fu Chang; Sundaram, H.* Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on , Volume: 2 , Page(s): 687

-690 vol.2, 2000
[62] Changick Kim and Jenq-Neng Hwang, "Object-Based Video Abstraction and an Integrated Scheme for On-line Processing," submitted to IEEE Transactions on Circuits and Systems for Video Technology (CSVT). Oct. 2000.
[63] Region feature based similarity searching of semantic video objects *Di Zhong; Shih-Fu Chang* Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on , Volume: 2 , Page(s): 111 -115 vol.2 , 1999
[64] J. N Hwang, Y. Luo, "Automatic Object based Video Analysis and Interpretation: A Step toward systematic video understanding," invited special session talk in ICASSP, Orlando FL, May 2002
[65] ChangIck Kim, J. N. Hwang, "Object-Based Video Abstraction Using Cluster Analysis," ICIP2001, Greece, Oct. 2001.
[66] Unsupervised clustering of ambulatory audio and video *Clarkson, B.; Pentland, A.* Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on , Volume: 6 , Page(s): 3037 -3040 vol.6, 1999
[67] Changick Kim and Jenq-Neng Hwang, "Fast and Robust Moving Object Segmentation in Video Sequences," IEEE international conference on Image Processing (ICIP'99), Kobe, Japan, Oct. 1999.
[68] D. Zhong and S.-F. Chang, An Integrated Approach for Content-Based Video Object Segmentation and Retrieval, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1259-1268, Dec. 1999.
[69] A computational approach to semantic event detection *Qian, R.; Haering, N.; Sezan, I.* Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., 1999
[70] Statistical models of video structure for content analysis and characterization *Vasconcelos, N.; Lippman, A.* Image Processing, IEEE Transactions on , Volume: 9 Issue: 1 , Page(s): 3 -19, Jan. 2000
[71] Integrated multimedia processing for topic segmentation and classification *Jasinschi, R.S.; Dimitrova, N.; McGee, T.; Agnihotri, L.; Zimmerman, J.; Li, D.* Image Processing, 2001. Proceedings. 2001 International Conference on , Volume: 2 , Page(s): 366 -369 vol.3, 2001
[72] MPEG-7 Overview. José M. Martínez (UPM-GTI, ES), Klangenfurt.
http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm. July 2002
[73]MPEG-7: the generic multimedia content description standard, part 1 *Martinez, J.M.; Koenen, R.; Pereira, F.* IEEE Multimedia , Volume: 9 Issue: 2 , April-June, Page(s): 78 -87 2002
[74] Standards - MPEG-7 overview of MPEG-7 description tools, part 2 *Martinez, J.M.* IEEE Multimedia , Volume: 9 Issue: 3 , Jul.-Sept. , Page(s): 83 -93, 2002
[75] MPEG-7 Systems: overview *Avaro, O.; Salembier, P.* Circuits and Systems for Video

Technology, IEEE Transactions on , Volume: 11 Issue: 6 , Page(s): 760 -764, June 2001
[76] MPEG-7 multimedia description schemes *Salembier, P.; Smith, J.R.* Circuits and Systems for Video Technology, IEEE Transactions on , Volume: 11 Issue: 6 , Page(s): 748 -759, June 2001
[78] Overview of MPEG-7 audio *Quackenbush, S.; Lindsay, A.* Circuits and Systems for Video Technology, IEEE Transactions on , Volume: 11 Issue: 6 , Page(s): 725 -729, June 2001
[79] MPEG-7 visual motion descriptors *Jeannin, S.; Divakaran, A.* Circuits and Systems for Video Technology, IEEE Transactions on , Volume: 11 Issue: 6 , Page(s): 720 -724, June 2001
[80] Semantics of multimedia in MPEG-7 *Benitez, A.B.; Rising, H.; Jorgensen, C.; Leonardi, R.; Bugatti, A.; Hasida, K.; Mehrotra, R.; Tekalp, A.M.; Ekin, A.; Walker, T.* Image Processing. 2002. Proceedings. 2002 International Conference on , Volume: 1 , Page(s): 137 -140, 2002
[81] The MPEG-7 visual standard for content description-an overview *Sikora, T.* Circuits and Systems for Video Technology, IEEE Transactions on , Volume: 11 Issue: 6 , Page(s): 696 -702, June 2001