A Multimodal Approach for Audiovisual Data Segmentation and Annotation

Tong Zhang, Hsuan-Huei Shih, and C.-C. Jay Kuo
Integrated Media Systems Center and Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, CA 90089-2564
Email: {tzhang,cckuo}@sipi.usc.edu, hshih@usc.edu

ABSTRACT

While most approaches for video segmentation and indexing are focused on the pictorial part, there are significant clues contained in the accompanying audio flow. Only by combining the audio and visual information together, a fully functional system for video content parsing can be achieved. Based on the investigation of data structures for different video types, we present in this paper a scheme for the segmentation and annotation of audiovisual sequence which includes tools for both audio and visual content analysis. In the proposed system, the video data is segmented into audio scenes and visual shots by detecting abrupt changes in the audio and visual features, respectively. Then, the audio scene is indexed as one of the basic audio types such as speech, music, song, environmental sound, speech with music background, etc. while the visual shot is represented by keyframes and associated image features. An index table is generated automatically for each video clip through the integration of audio and visual analysis outputs. Experimental results show that the proposed research accomplishes more meaningful and robust indexing of video content compared to previous work.

Keywords. audiovisual data segmentation and indexing, audio content analysis, visual content analysis, video database management, information filtering and retrieval.

1. INTRODUCTION

The proposed research is mainly for the purpose of video annotation, i.e. to automatically generate metadata for video sequences for information filtering and retrieval. There are many applications of this technique, such as delivery of video segments for professional media production, storage and retrieval of video databases, video-on-demand, user agent driven media selection and filtering, semantic multicast backbone video, as well as educational applications and surveillance applications.

Previous work on video segmentation and annotation are primarily based on the visual information. A common framework is to detect video shot changes using histogram differences and motion vectors, and extract keyframes to represent each video shot [6], [3]. However, this visual-based processing often leads to a far too fine segmentation of the audiovisual sequence with respect to its semantic meaning. For example, in one video clip of news program about a congress meeting, there are a dozen of shots including those of the anchorperson, of the speakers, of the audience, and several broad views of the hall. According to the visual information, these shots will be indexed separately. However, according to the audio information, the continuous speech of the anchorperson indicates that they are actually within one news item. Furthermore, there are often significant clues contained in the audio data, e.g. the sounds of song performances in a music show video, the sounds of whistle and applause in a sports video, and sounds of conversation and music in feature movies, which provide enormous help in segmenting and annotating the video content.

A new trend for audiovisual data segmentation and indexing is to combine audio and visual information under one framework. This idea was examined in some recent papers. In [4], break detectors were developed for audio, color, and motion separately. For shot segmentation, results from both color and motion break detections were combined. For scene change detection, one looks for frames where both visual and audio breaks were detected. In [5], audio subband data and color histograms of one video segment were combined to form a "Multiject", and two variations of the hidden Markov model were used to index the Multijects. Experimental results for detecting the events of "explosion" and "waterfall" were reported. In [2], the color histogram differences, the cepstral coefficients of audio data, and the motion vectors were combined by using a hidden Markov model approach to segment video into regions consisting of shots, shot boundaries, and camera movement within shots. Even though it has been demonstrated that these methods can combine the visual and audio information to handle some specific scenarios, their applicability to a generic video type is still yet to be proved. Meanwhile, more delicate audio feature extraction methods should be considered to provide more robust segmentation results.

There are also several systems or structures proposed, which consider the description of audiovisual content by using more than one type of media. Some interesting high-level concepts have been presented, including object-oriented description, multi-level description, multiple modality, metadata dictionary, table of content, analytical index, etc. However, low- to mid-level implementations of audio, visual and textual content analysis and their integration are still lacking in these systems. Proper algorithms have to be developed to support these high-level ideas.

In this research, we propose a scheme for the on-line segmentation and annotation of audiovisual data based on the integration of several kinds of media, namely, audio, visual, and textual information. In the proposed system, the first step is to demultiplex the video stream into the parts of audio, image, and caption. Then, a semantic segmentation of the audio stream based on audio content analysis is conducted. We call such a segmented unit as "au-

dio scene", and index it as pure speech, pure music, song, speech with the music background, silence, etc. based on the audio classification algorithms we developed. Next, the image sequence is segmented into shots based on visual information analysis, i.e. color histogram and motion vectors. Keyframes will be extracted from each shot, and color, shape and texture features of the keyframes will be analyzed to give visual index of each shot. Meanwhile, keywords will be detected from the closed caption in a video sequence to form the textual index. An index table is built for each video sequence to contain the time interval of each segment, as well as the audio, visual and textual indexes. The structure of the index table is designed to be hierarchical and easy to access which represents the relationship of multi-modal feature descriptors. The automatically generated metadata are synchronized with the video sequence for transmission and storage. A filtering and retrieval mechanism is then built to select the video segments which match the user's interest according to the annotation, and to present these segments to the user.

2. MODELING OF AUDIOVISUAL DATA

2.1. Models for Different Video Types

The common model used for video data is a hierarchical structure consisting of scenes, shots, and frames [1]. However, different video types may have different associated semantics, and the video content model should be built according to features of each video type individually. In this work, we observed video data belonging to the following five types: news bulletins, documentaries (such as scientific videos and educational videos), feature movies (including TV drama series), variety shows, and sports video. The characteristics of each video type are summarized below.

A data model for news video was examined in [7]. It is a simple sequence of news items that are possibly interleaved with commercials. A news item in most news programs always starts with an anchorperson shot, followed by a sequence of shots which illustrate the news story. Frames of the anchorperson shots have a well defined spatial structure. Thus, the scene concept in news video is relatively simple. It is composed of news items, and each news item has one or more shots. We also observe that, in terms of both audio and visual properties, each news item normally starts with not only the image but also the speech of the anchorperson. Then, there may be shots of other sights with the anchorperson's voice as the offscene sound.

In one such example, there are eight shots in the news item. The first shot is the anchorperson shot, and the voice of the anchorperson runs through the whole item. Sometimes there are also voices of on-site reporters, speech in interviews, as well as environmental sounds in one news item. However, the picture and the voice of the anchorperson will be back when the next news item begins. Therefore, by recognizing the voice and the image of the anchorperson, one can detect breaks between news items easily. The voice of an anchorperson is usually fast, stable and clear. Sometimes, the video shot may not change over two consecutive news items and simply stays at the anchorperson. However, there is obvious pause in the anchorperson's speech between two news items, by which one still can detect breaks of news items. After news item segmentation, each item can be represented by keyframes of visual shots as well as the first one or two sentences of the anchorperson.

Similar to news bulletin, the variety show video does not have complicated scenes either. It is mainly composed of a sequence of performances. There are normally music and/or songs during one performance. A performance usually begins with some pure music. At the end of each performance, there are the pause of music, the applause and acclaim from the audience, and the speech of the host. Sometimes, the music does not stop between two performances. However, there is a change of the melody and the rhythm of the music which can be detected. By looking for these audio clues, we can segment the video program into individual performances and transitional parts (e.g. interviews with the actor or the audience by the host). Each performance can be indexed by keyframes of the actor(s) as well as the audio semantics (such as pure music, song, etc.).

In ball games (such as basketball, soccer, volleyball) of sports video, sounds of whistle may represent the start and/or end of one episode. Bursts of applause and acclaim from the audience may indicate some exciting moments. Also, the color of the floor (normally yellow in basketball and volleyball games) or the grass may indicate shots of the game (rather than shots of the audience, of a particular player or of some other sights). This dominant color feature can be easily detected from color histograms. Thus, combining audio and visual information, we can detect breaks of episodes as well as exciting moments within one game.

In documentary movies and videos, there are the structures of semantic scenes which are difficult to define simply by using audio and visual features. However, with the help of audio clues, segmentation results can be much more improved than using the visual information alone. In such kind of video, audio parts are accompanied with the pictorial parts off-line. Normally, there is music all through the program with commentary speech appearing from time to time. There are often obvious pauses or changes of audio content at scene breaks. For example, the long pause between two speech segments, the variation or stop of music may all indicate change of scenes. Let us give an example below. There is a 33-second long clip in a documentary movie which provides a scene of the life style in Lancaster. There are seventeen shots in the scene displaying various aspects of the life in Lancaster. However, the accompanying music episode, which is continuous during the period and isolated from the former and the latter episodes by obvious pauses, shows that these shots are within one semantic unit. Some keyframes in this scene are shown in Figure 1.

In feature movies and TV dramas, the situation is even more complicated because their content structures are of a greater diversity. Nevertheless, there are many scenes of conversations in such video types which can be detected by tracing voices and shots of speakers. Sometimes, there is also music as the offscene sound with multiple shots appearing, and the scene can be defined with such an audio feature. In certain TV drama series, there is a fixed short music clip at each scene break. Most of the time, there are several shots within one audio scene (defined as a unit of video with continuous audio content), but there are also sit-

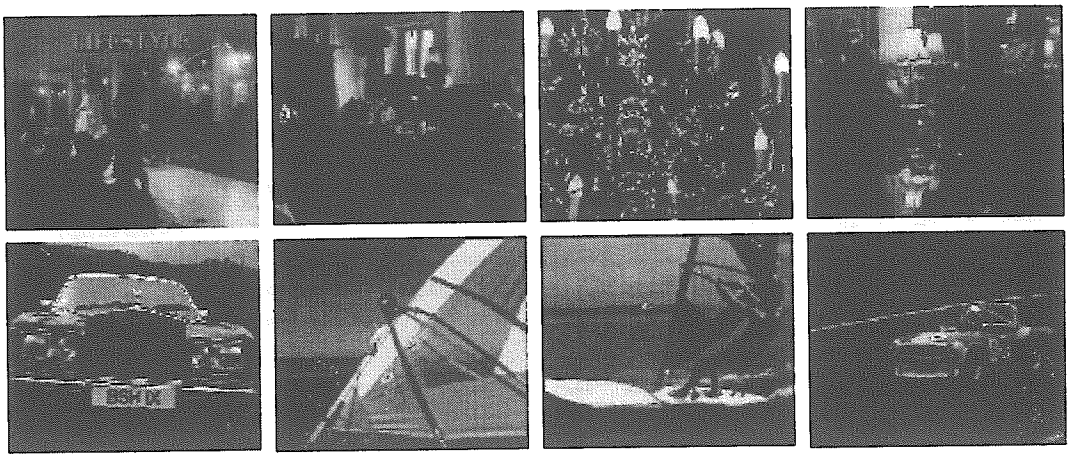

Figure 1: Keyframes within one scene ("lifestyle in Lancaster") in a documentary movie.

uations that the audio content may change (e.g. the start or stop of music, the pause of speech) within one visual shot. Therefore, a scene break can only be defined if both audio scene change and visual shot change occur.

2.2. Proposed Scheme for Video Content Parsing

Based on observations described above, we propose a scheme for video content parsing which is illustrated in Figure 2. The video input is first separated into the audio data stream and the image sequence. Then, audio data are segmented into audio scenes, and each scene is classified as one of the basic audio types. The image sequence is broken into shots, and keyframes are extracted from each shot. Meanwhile, facts about the video data including the video type and other *a priori* knowledges, which are either metadata accompanying with the video data or the input from the user, are used to select a data structure from video models.

Each data structure in the video model base provides a guidance for generating the index table, and for special audio and visual feature analysis based on the particular video model. For example, the detection of the sounds of whistle, applause and acclaim, and the shots with green grass are important for parsing video clips of soccer and football games. Included in the data structure are the syntax of index table, as well as models for characteristic sounds, shots, and keyframes for the specific video type. Finally, The audio scene indexes, the keyframes of each shot, as well as the special audio and visual analysis results are combined to build the index table according to the video model selected for this video sequence. Examples of index table will be shown in Section 5. In the next two sections, we will introduce techniques for generating and indexing audio scenes and visual shots, respectively.

2.3. Design of Index Table for Non-linear Access

The primary index table of video content which is produced automatically during the real-time processing is chronologically ordered. While it is convenient to be generated on-line, it can only provide a linear access which may not be optimal for fast browsing and retrieval. Therefore, it is necessary to design a second kind of content index form which may provide non-linear access to the video sequence. One such approach based on clustering was proposed in [10] where video shots were clustered according to color histogram of keyframes and some other features. However, this structure is constructed only on the basis of low-level features which may not be suitable for many application scenarios.

We propose here preliminary design of a secondary index form for video content which is derived from the primary index table. For each video clip, this index form includes three index trees for the audio, visual, and textual information, respectively. The audio index hierarchy is built based on semantic meanings. For example, the first level of tree may include nodes indexed as "pure speech", "music", "song", "environmental sound", "speech with music background", "silence", etc. And the node of "pure speech" may have a second level hierarchy including "male speech", "female speech", "conversation" and so on. Under each node, there are time intervals of segments within the video clip which belong to the audio type indexed by this node.

The textual index tree may be organized alphabetically or according to a subject list. Each node in the tree is indexed with a keyword, and maintains time intervals of segments in the video clip which may be indexed with this keyword. For visual information, it is quite hard to find some semantic way to index it. One possibility is to use the keyframes, and classify them into categories such as "background image" and "object image". Then, the "object image" may be further separated into clusters including "people", "animal", "building" and so on. And the "background image" may include "scenery", "crowd", etc. In this way, the visual index tree can be built. The user may choose to browse or retrieval video segments according to one type of index or a combination of several types of indexes.

3. AUDIO CONTENT ANALYSIS

3.1. Segmentation and Indexing of Audio Data

The procedure for segmenting and indexing the audio data flow into audio scenes is shown in Figure 3. Four features are extracted from audio data, i.e. the short-time

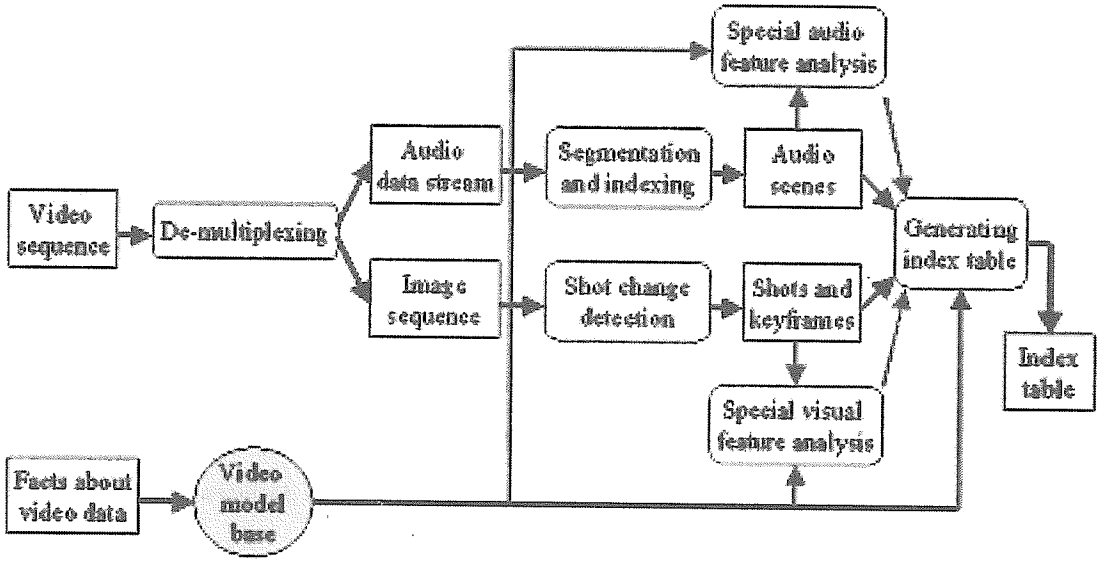

Figure 2: Video content parsing based on combined audio and visual information analysis.

energy function, the short-time average zero-crossing rate (ZCR), the short-time fundamental frequency, and the spectral peak tracks. For details of the characteristics of these audio features and feature calculation, we refer to [8]. A method for estimating the fundamental frequency from the AR model computed power spectrum was developed. It is used to represent the harmony property of the sound. Peak tracks in the spectrogram of an audio signal may often reveal some features of a sound type. We extract spectral peaks for the purpose of characterizing sound segments of song and speech. It is done by detecting peaks in the power spectrum generated by the AR model parameters. For online segmentation of audio data stream, the short-time values of the energy function, the average zero-crossing rate, and the fundamental frequency are computed on the fly with incoming audio data. Whenever there is an abrupt change detected in any of these three features, a segment boundary is set. Then, each segment is classified into one of basic audio types according to the following procedure.

The first step is to check whether the audio segment is silence or not, based on both energy and ZCR measures. As observed from movies and TV programs, music is an important type of audio component frequently appearing, either alone or as the background of speech or environmental sounds. Therefore, the next step is to separate non-silent audio segments into two categories, i.e. with or without music components. It is done by detecting continuous frequency peaks from the power spectrum generated from AR model parameters of the audio signal. If there are peaks detected in consecutive power spectra which stay at about the same frequency level for a certain period of time, this period of time is indexed as having music components. Based on a threshold for the ratio of such indexed periods within the audio segment together with some other rules, audio segments can be separated into two categories. The first category contains harmonic and stable environmental sound, pure music, song, speech with the music background, and environmental sound with the music background. In the second category, there are pure speech and non-harmonic environmental sound. Further classification is conducted within each category.

Harmonic and stable environmental sounds are separated out by checking the temporal curve of the short-time fundamental frequency. Then, pure music segments are distinguished based on the average zero-crossing rate and the fundamental frequency properties. For other audio segments in the first category, we extract the spectral peak tracks. The song segments may be characterized by one of the three features: ripple-shaped harmonic peak tracks (due to the vibration of vocal chords), tracks which are of longer durations compared to those in speech, and tracks which have a fundamental frequency higher than 300Hz. Next, if there are spectral peak tracks concentrating in the lower to middle frequency bands (with fundamental frequency between 100 to 300 Hz) and having lengths within a certain range, the segment is indexed as "speech with music background". Finally, what left in the first category is indexed as "environmental sound with music background".

To detect pure speech, five conditions are checked, i.e. the relation between temporal curves of the energy function and the ZCR, the shape, the variance, and the range of amplitudes of the ZCR curve, as well as the property of the short-time fundamental frequency. Then, what left in the second category is classified into one type of the non-harmonic environmental sound, including "periodic or quasi-periodic", "harmonic and non-harmonic mixed", "non-harmonic and stable", and "non-harmonic and irregular" environmental sound, based on the three short-time audio features. A post-processing step is applied after the indexing of audio segments to reduce possible segmentation errors. Then, each segment will be considered as an audio scene. The time interval and index of each scene will be integrated into the index table of the video clip.

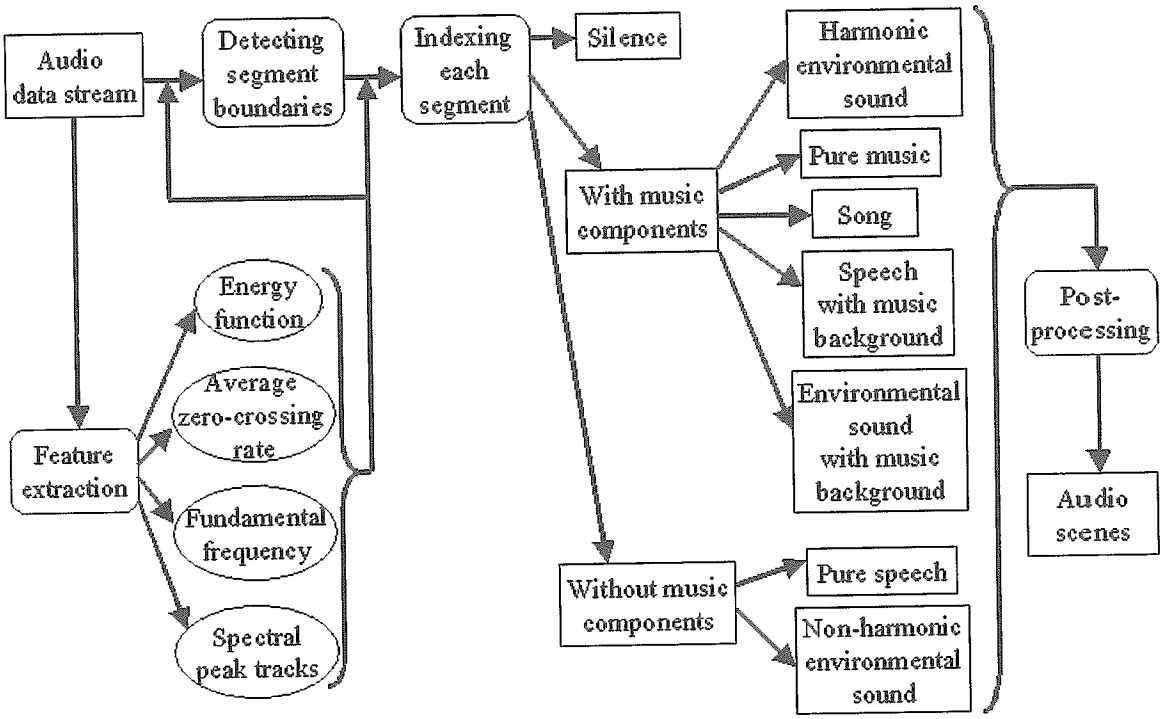

Figure 3: Automatic segmentation and indexing of audio data stream into audio scenes.

3.2. Recognition of Feature Sounds

Some characteristic sounds are of special importance for certain video types, such as the voice of the anchorperson in news bulletin, the sounds of whistle, applause and acclaim in sports video, the separation tune in TV drama series, and the sounds of shooting and explosion in feature movies. Models of these sounds can be built *a prior*, and stored either with the video program (for the cases of anchorperson's voice and separation tune) or in the video model (for the environmental sounds). Sounds in related audio scenes will be matched with these models for recognition of feature sounds. The detection of anchorperson's voice belongs to the problem of speaker identification. The Gaussian mixture model (GMM) has been proposed for robust text-independent speaker identification which proved to have quite good performance. Thus, the voice of a speaker is modeled by GMM parameters. The separation tune may be represented with harmonic lines. As for the recognition of environmental sounds, we proposed a method using the hidden Markov model (HMM) [9].

For classifying and retrieving environmental sounds, we investigate features from the time-frequency representation of audio signals, which might reveal subtle differences among different classes of sounds. We use "timbre" and "rhythm" to describe the perception of environmental sound. For non-harmonic sounds, timbre is largely determined by the spectral envelope of the audio signal. As most environmental sounds are non-harmonic, we take a smoothed version of the short-time power spectrum to represent timbre for a sound at a certain instant. A feature vector is obtained with a dimension of 65, which represents the frequency distribution from 0 to π. The rhythm information (denoting the change pattern of timbre and energy in a sound clip) is reflected by the transition and duration parameters in HMM.

To build HMM for one certain environmental sound class (e.g. applause, explosion, whistle), feature vectors in the training set of this class are clustered into several sets, with each set having distinct energy and spectral shape properties from those of others, and modeled later by one state in HMM. Based on the clustering results, a training process for estimating the HMM parameter set is developed. It achieves about the same precision with the widely used Baum-Welch method but with a much lower complexity. The HMM parameters of critical sounds for each video model are stored as the key accompanying information. In the procedure of video parsing, environmental sounds will be matched with these HMM parameters to determine whether they belong to any of these feature sounds.

4. VISUAL CONTENT ANALYSIS

4.1. Detection of Shot Changes

We investigated methods for detecting shot changes in image sequences of the YUV format. The histogram difference between neighboring frames in a video clip containing 4000 frames is plotted in Figure 4 for the Y coordinate. All abrupt shot changes are reflected by remarkable pulses in this figure. However, problems arise when the transition is gradual (as produced by special camera effects such as fade-in, fade-out, wipe, and dissolve) where the shot does not change abruptly but over a period of a few frames.

Zhang *et al.* proposed a twin-comparison method by

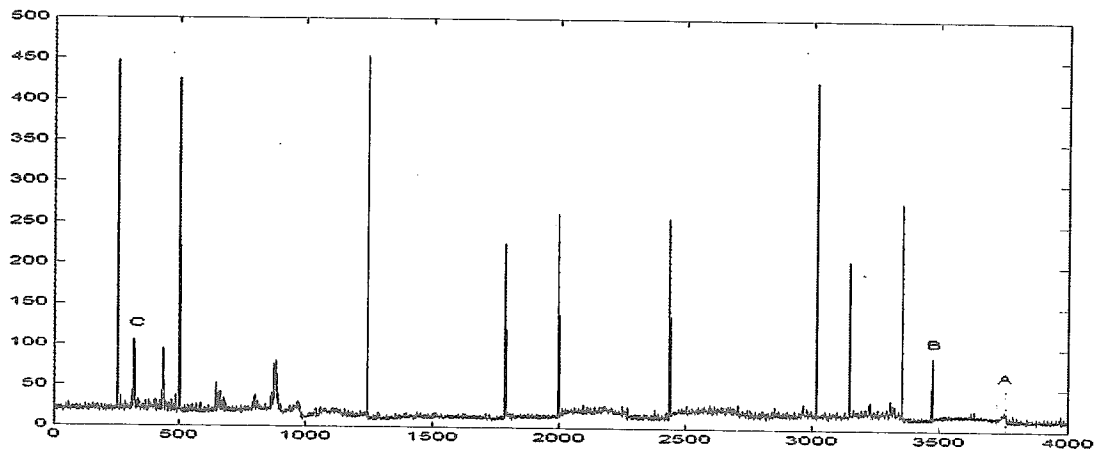

Figure 4: Histogram difference for the Y component between neighboring frames.

taking into account cumulative differences between frames [6]. This method requires two cutoff thresholds. In the first stage a higher threshold is used for the detection of abrupt transitions. In the next stage a lower threshold is used on the rest of the frames; any frame that has the difference more than this threshold is declared as the potential start of a gradual transition. This frame is then compared with subsequent frames and the difference added. Usually this difference value increases and when this value increases to the level of the higher threshold, camera break is declared at that frame. If the value falls between the consecutive frames then the potential frame is dropped and the search starts all over.

Although this method has been proved to be quite effective in many situations, it is sometimes difficult to choose a proper value for the lower threshold. For example, there is a dissolve sequence in Figure 4 between the two dotted lines marked with the letter "A". Some frames within the sequence are shown in Figure 5. We can see that histogram difference values are quite small during the dissolve procedure, especially in the beginning frames. In order to detect this shot transition, the lower threshold should be selected to be a very small value, which turns out to be lower than many histogram differences within one shot. This implies that there will be a lot of computations wasted on the calculation of cumulative histogram differences between frames where there are actually no shot changes.

Nevertheless, we observe that there are high pulses reflecting the dissolves in the histogram difference of the V-component. Therefore, we compute histogram differences for Y- and V- components separately. High pulses occurring in the V-component difference but not in the corresponding Y-component difference require a second check in the Y histogram difference values via calculation of cumulative differences. Although the calculation of V-component histograms requires additional cost, the lower threshold can be raised to a higher value so that a less amount of computation will be spent on determining cumulative differences. Thus, the overall complexity is still reduced.

There is a similar problem with the selection of the higher threshold. The two pulses marked with letters "B" and "C" in Figure 4 have similar heights. However, the former one represents a real shot cut, while the latter one is only the difference between two consecutive frames within one panning motion. For the former case, although the two frames have very different content, their histograms of the Y component have similar distributions. While for the latter case, even though the two frames are quite alike, their histograms have a slight offset to each other. To differentiate such cases, the information revealed by the histogram differences of the V component is again valuable. For example, the pulse in the differences of the V component corresponding to the former case is much higher than that for the latter one. Thus, we may choose the higher threshold to be a bit bigger, and furthermore check positions with remarkable pulses in the V-component difference but not having an amplitude higher than the threshold in the Y-component difference.

4.2. Adaptive Keyframe Extraction

A simple and common way for keyframe extraction is to choose the first and/or the last frame of each shot. Actually, situations are quite diverse among different shots. Some shots may contain frames which are very similar to each other while others may have large-scope camera motions in which the frame content changes dramatically. To handle various situations, we propose to use an adaptive scheme for keyframe extraction. That is, the first frame of one shot is chosen as the first keyframe. Then, its histogram is used as the reference to compare with those of latter frames. When the difference is higher than a predefined threshold, a second keyframe is selected. Next, the second keyframe will be used as the reference and compared with latter frames. This procedure continues until the end of the shot. In this way, one shot may have only one keyframe or have multiple keyframes depending on the complexity involved in it.

Many features have been investigated for the representation of image or video content such as those of color, texture, shape and motion. We select only color histograms of keyframes for typical shots as the first step in this work (e.g. anchorperson shots in news bulletin, shots of football or

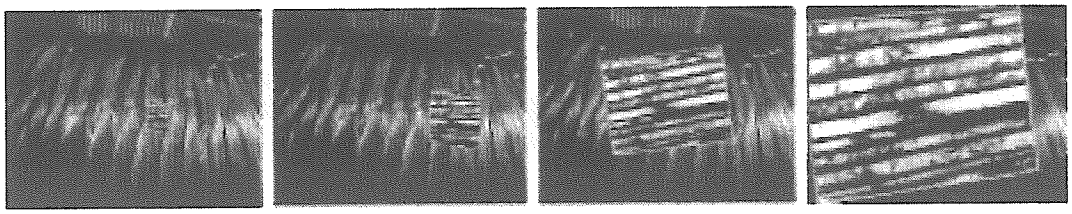

Figure 5: Frames within a dissolve sequence.

basketball game in sports video) to be features included in related video models. Keyframes from corresponding video types will be matched with these histogram patterns to decide whether they belong to typical shots or not.

5. EXPERIMENTAL RESULTS

5.1. Audio Segmentation and Indexing

We developed an audio database containing around 1200 short sound clips (with duration from several seconds to more than one minute) of various audio types, as well as dozens of longer audio clips recorded from movies or TV programs for testing the segmentation and indexing performances. With an extensive experimental test, an accuracy rate of more than 90% was achieved for the classification of audio scenes. One example for the segmentation and indexing of an audio clip based on a documentary video input is shown in Figure 6. In this 49-second long audio recording, there is first the sound of the waterwheel which is indexed as "non-harmonic and irregular environmental sound". Then, comes the music together with the environmental sound which is indexed as "environmental sound with music background". Afterwards, there is a female speech segment which is indexed as "pure speech", followed by the sounds of ox cart and dog bark which is indexed as "non-harmonic and irregular environmental sound". Finally there is a segment of music which is indexed as "pure music". Boundaries between segments are precisely detected and each segment is properly indexed. To test the HMM-based algorithm for classifying environmental sounds, we selected 210 sound clips from 18 classes (including applause, gun shot, whistle, etc.). An accuracy rate of 86% was obtained.

5.2. Visual Information Analysis

We made three kinds of experiment for video content analysis, i.e. shot change detection, keyframe extraction, and search of special shots with color histogram models. The scheme for detecting shot changes based on combining the Y and V value histograms and the twin-comparison method was tested on image sequences excerpted from feature movies and documentaries. It was found that both sensitivity and recall rates around 95% were achieved with real-time processing. The proposed keyframe extraction method was proved to be able to present keyframes showing different views of the shot. An example is shown in Figure 7, where the keyframes extracted from one shot in a documentary video are displayed. As for the detection of special shots, we tried to search for the anchorperson shots in a news video with a pre-computed histogram model of the anchorperson frame and 100% of the shots were found.

5.3. Index Table Generation

We worked on building video models and generating index tables for two types of videos, i.e. news bulletin and documentaries. The index table of a news video clip has a structure containing two levels. The first level is for the description of the whole video which is composed of news items. The time interval and the first 10-second speech of the anchorperson of each news item are ordered chronologically in this level. The news items are separated from each other by detecting the anchorperson shots with color histogram comparison and pauses in the anchorperson's speech. Then, the second level description is for the content within each news item, including time intervals and indices of audio scenes, as well as time intervals and keyframes of shots in the news item. The audio and visual indices are interleaved together in each news item according to the temporal order.

For a documentary video, there are also two levels of structure in the index table: scene and shot. A scene break is defined only when both audio scene change and visual shot change occur. However, the definitions of audio scene changes are slightly different from the segmentation rules as described in Section 3. For example, with the same music episode continuing, the appearance and the disappearance of speech on top of music will be regarded as the break between two audio types (i.e. "pure music" and "speech with music background"), but will not be taken as an audio scene break. Instead, it will be viewed to be within one scene as long as the same music continues. However, if there is change of music (i.e. two different episodes) or long pause between two speech segments, an audio scene break will be claimed. For each scene, the time interval, the first 10-second offscene speech (if there is any) and the first image frame will be included at the first level description. Then, for each shot, there will be the time interval, the audio type index and keyframes contained in the second level description.

6. CONCLUSION AND FUTURE WORK

We presented in this paper an approach for video content parsing based on audiovisual data modeling as well as the combination of audio and visual information analysis. Since different video types may have different semantic structures, we built a video model for each video type individually, and used such models to guide the segmentation and annotation of video data. We proposed methods for both audio and

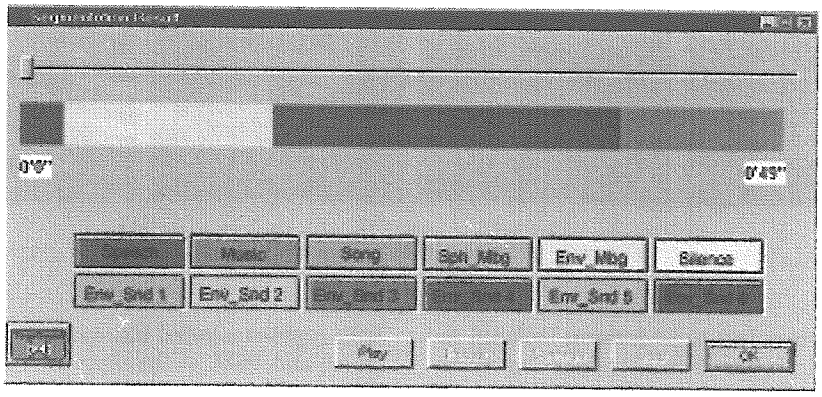

Figure 6: Segmentation and indexing of an audio clip recorded from documentary video.

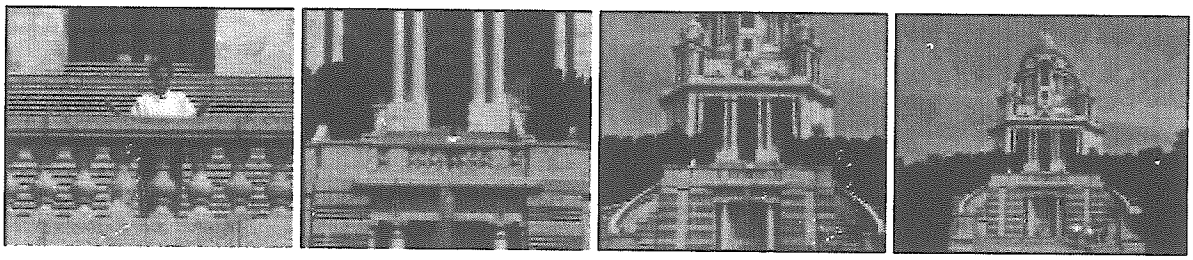

Figure 7: Keyframes extracted from one shot in a documentary video with camera zooming-out motion.

visual content analysis, including tools for audio scene segmentation and indexing, recognition of sound effects, shot change detection, keyframe extraction, and so on. Experimental results showed that these tools were effective in achieving their goals. An index table was generated by integrating analysis results with these tools. The index tables have hierarchical structures and are designed for particular video types, which may provide enormous help to users in filtering and retrieving video segments of their interest.

In future research, we will investigate the segmentation and annotation of video data in compressed formats. Work will be done on demultiplexing the video bitstream, feature extraction in the compression domain for both audio and visual content, synchronization of video data and the index table, and mechanisms for information filtering and retrieving based on the generated metadata.

7. REFERENCES

[1] G. Ahanger and T. D. C. Little, "A survey of technologies for parsing and indexing digital video," *Journal of Visual Communication and Image Representation*, Vol.7, No.1, pp.28-43, 1996.

[2] J. S. Boreczky and L. D. Wilcox, "A hidden Markov model framework for video segmentation using audio and image features," *Proceedings of ICASSP'98*, pp.3741-3744, Seattle, May 1998.

[3] S.-F. Chang, W. Chen, H. J. Meng, *et al.*, "A fully automated content based video search engine supporting spatio-temporal queries," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.8, No.5, pp.602-615, 1998.

[4] J. Huang, Z. Liu, and Y. Wang, "Integration of audio and visual information for content-based video segmentation," *Proceedings of IEEE Conference on Image Processing*, Chicago, Oct. 1998.

[5] M. R. Naphade, T. Kristjansson, B. Frey, *et al.*, "Probabilistic Multimedia Objects (MULTIJECTS): a novel approach to video indexing and retrieval in multimedia systems," *Proceedings of IEEE Conference on Image Processing*, Chicago, Oct. 1998.

[6] H. J. Zhang, A. Kankanhalli, and S. W. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, Vol.1, No.1, pp.10-28, 1993.

[7] H. J. Zhang and S. W. Smoliar, "Developing power tools for video indexing and retrieval," *Proceedings of IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases II*, SPIE Vol.2185, pp.140-149, San Jose, Feb. 1994.

[8] T. Zhang and C.-C. Kuo, "Heuristic approach for generic audio data segmentation and annotation," *ACM Multimedia Conference*, pp.67-76, Orlando, Nov. 1999.

[9] T. Zhang and C.-C. Kuo, "Hierarchical classification of audio data for archiving and retrieving," *Proceedings of IEEE International Conference On Acoustics, Speech, and Signal Processing*, Vol.6, pp.3001-3004, Phoenix, Mar. 1999.

[10] D. Zhong, H. J. Zhang, and S.-F. Chang, "Clustering methods for video browsing and annotation," *Proceedings of SPIE Conference on Storage and Retrieval for Image and Video Database*, San Jose, Feb. 1996.