MERL – A MITSUBISHI ELECTRIC RESEARCH LABORATORY
http://www.merl.com

VIDEO SUMMARIZATION USING MPEG-7 MOTION ACTIVITY AND AUDIO DESCRIPTORS

Ajay Divakaran, Kadir A. Peker, Regunathan Radhakrishnan, Ziyou Xiong and Romain Cabasson

TR-2003-34   May 2003

Abstract

We present video summarization and indexing techniques using the MPEG-7 motion activity descriptor. The descriptor can be extracted in the compressed domain and is compact, and hence is easy to extract and match. We establish that the intensity of motion activity of a video shot is a direct indication of its summarizability. We describe video summarization techniques based on sampling in the cumulative motion activity space. We then describe combinations of the motion activity based techniques with generalized sound recognition that enable completely automatic generation of news and sports video summaries. Our summarization is computationally simple and flexible, which allows rapid generation of a summary of any desired length.

Publication History:

1. First printing, TR-2003-34, May 2003

Chapter 1

VIDEO SUMMARIZATION USING MPEG-7 MOTION ACTIVITY AND AUDIO DESCRIPTORS

A Compressed Domain Approach to Video Browsing

Ajay Divakaran, Kadir A. Peker, Regunathan Radhakrishnan, Ziyou Xiong and Romain Cabasson
Mitsubishi Electric Research Laboratories,
Cambridge,MA 02139
{ajayd,peker,regu,zxiong,romain}@merl.com

Abstract We present video summarization and indexing techniques using the MPEG-7 motion activity descriptor. The descriptor can be extracted in the compressed domain and is compact, and hence is easy to extract and match. We establish that the intensity of motion activity of a video shot is a direct indication of its summarizability. We describe video summarization techniques based on sampling in the cumulative motion activity space. We then describe combinations of the motion activity based techniques with generalized sound recognition that enable completely automatic generation of news and sports video summaries. Our summarization is computationally simple and flexible, which allows rapid generation of a summary of any desired length.

Keywords: MPEG-7, Motion Activity, Video Summarization, Audio-Visual Analysis, Sports Highlights, News Video Browsing

Introduction

Past work on video summarization has mostly employed color descriptors, with some work on video abstraction based on motion features. In this chapter we present a novel approach to video summarization using the MPEG-7 motion activity descriptor [1]. Since our motivation is computational simplicity and easy incorporation into consumer system hardware, we focus on feature extraction in the compressed domain.

We first address the problem of summarizing a video sequence by abstracting each of its constituent shots. We verify our hypothesis that the intensity of motion activity indicates the difficulty of summarization of a video shot. We do so by studying the variation of the fidelity of a single key-frame with change in the intensity of motion activity as defined by the MPEG-7 video standard. Our hypothesis motivates our proposed key-frame extraction technique that relies on sampling of the video shot in the cumulative intensity of motion activity space. It also motivates our adaptive playback frame rate approach to summarization. We then develop a two-step summarization technique by first finding the semantic boundaries of the video sequence using MPEG-7 generalized sound recognition and then applying the key-frame extraction based summarization to each of the semantic segments.

The above approach works well for video content such as news video in which every video shot needs to be somehow represented in the final summary. In sports video however, all shots are not equally important since key events occur only periodically. This motivates us to develop a set of sports highlights generation techniques that rely on characteristic temporal patterns of combinations of the motion activity and other audio-visual features.

1. Background and Motivation

1.1 Motion Activity Descriptor

The MPEG-7 [1] motion activity descriptor attempts to capture human perception of the “intensity of action” or the “pace” of a video segment. For instance, a goal scoring moment in a soccer game would be perceived as a “high action” sequence by most human viewers. On the other hand, a “head and shoulders” sequence of a talking person would certainly be considered a ”low action” sequence by most. The MPEG-7 motion activity descriptor has been found to accurately capture the entire range of intensity of action in natural video. It uses quantized standard deviation of motion vectors to classify video segments into five classes ranging from very low to very high intensity.

1.2 Key-frame Extraction from Shots

An initial approach to key-frame extraction was to choose the first frame of a shot as the key-frame. It is a reasonable approach and works well for low-motion shots. However, as the motion becomes higher, the first frame is increasingly unacceptable as a key-frame. Many other subsequent approaches (see [4] for a survey) have built upon the first frame by using additional frames that significantly depart from the first frame in addition to the first frame. Another category consists of approaches that rely on clustering and other computationally intensive analysis. Neither category makes use of motion features and are computationally intensive. The reason for using color is that it enables a reliable measure of change from frame to frame. However, motion-compensated video also relies on measurement of change from frame to frame, which motivates us to investigate schemes that use motion vectors to sense the change from frame to frame in a video sequence. Furthermore, motion vectors are readily available in the compressed domain hence offering a computationally attractive avenue. Our approach is similar to Wolf's (see [4]) approach in that we also make use of a simple motion metric and in that we do not make use of fixed thresholds to decide which frames will be key-frames. However, unlike Wolf, instead of following the variation of the measure from frame to frame, we propose that the simple shot-wide motion metric, the MPEG-7 intensity of motion activity descriptor, is a measure of the summarizability of the video sequence

1.3 The Fidelity of a Set of Key-Frames

The fidelity measure [3] is defined as the Semi-Hausdorff distance between the set of key-frames S and the set of frames R in the video sequences. A practical definition of the Semi-Hausdorff distance d_{sh} is as follows: Let the key frame set consist of m frames $S_i, i = 1..m$, and let the set of frames R contain n frames $R_i, i = 1..n$. Let the distance between two frames S_i and R_i be $d(S_i, R_i)$. Define d_i for each frame R_i as

$$d_i = min(d(S_j, R_i)), j = 1, m$$

Then the Semi-Hausdorff distance between S and R is given by

$$d_{sh}(S, R) = max(d_i), i = 1, n$$

Most existing dissimilarity measures satisfy the properties required for the distance over a metric space used in the above definition. In this chapter, we use the color histogram intersection metric proposed by Swain and Ballard (See [3]).

2. Motion Activity as a Measure of Summarizability

We hypothesize that since high or low action is in fact a measure of how much a video scene is changing, it is a measure of the "summarizability" of the video scene. For instance, a high speed car chase will certainly have many more “changes” in it compared to say a news-anchor shot, and thus the high speed car chase will require more resources for a visual summary than would a news-anchor shot. Unfortunately, there are no simple objective measures to test such a hypothesis. However, since change in a scene often also involves change in the color characteristics as well, we first try to investigate the relationship between color-based fidelity as defined in 2.2, and intensity of motion activity. Let the key frame set for shot A be S_A and that for shot B be S_B. If S_A and S_B both contain the same number of key frames, then our hypothesis is that if the intensity of motion activity of shot A is greater than the intensity of motion activity of shot B, then the fidelity of S_A is less than the fidelity of S_B.

2.1 Establishing the Hypothesis

We extract the color and motion features of news video programs from the MPEG-7 test-set, which is in the MPEG-1 format. We first segment the programs into shots. For each shot, we then extract the motion activity features from all the P-frames by computing the standard deviation of motion vector magnitudes of the macro-blocks of each P frame, and a 64 bin RGB Histogram from all the I-frames, both in the compressed domain. Note that intra-coded blocks are considered to have zero motion vector magnitude. We then compute the motion activity descriptor for each I-Frame by averaging those of the previous P-frames in the Group of Pictures (GOP). The I-Frames thus all have a histogram and a motion activity value associated with them. The motion activity of the entire shot is got by averaging the individual motion activity values computed above. From now on, we treat the set of I-frames in the shot as the set of frames R as defined earlier. The simplest strategy for generating a single key frame for a shot is to use the first frame, as mentioned earlier. We thus use the first I-frame as the key frame and compute its fidelity as described in 2.2. We find empirically that a key frame with Semi-Hausdorff distance at most 0.2 is of satisfactory quality, by analyzing examples of ”talking head” sequences. We can therefore classify the shots into two categories, those with key frames with d_{sh} less than or equal to 0.2 i.e. of acceptable fidelity and those with key frames with d_{sh} greater than 0.2, i.e. unacceptable fidelity. Using the MPEG-7 motion activity descriptor, we can also classify the shots into five categories ranging from very low to very high activity. We then find the percentage duration of shots with d_{sh} greater than 0.2 in each of these categories for the news program news1 (Spanish News) and plot the results in Figure 1.1. We can see that as the motion activity goes up from very low to very high, the percentage of unacceptably summarized shots also increases consistently. In other words, the summarizability of the shots goes down as their motion activity goes up. Furthermore, the fidelity of the single key frame is acceptable for 90 percent of the shots in the very low intensity of motion activity category. We find the same pattern with other news programs. We thus find experimental evidence that with news program content, our hypothesis is valid. Since news programs are diverse in content, we would expect this result to apply to a wide variety of content. Since we use the MPEG-7 thresholds for motion activity, our result is not content dependent.

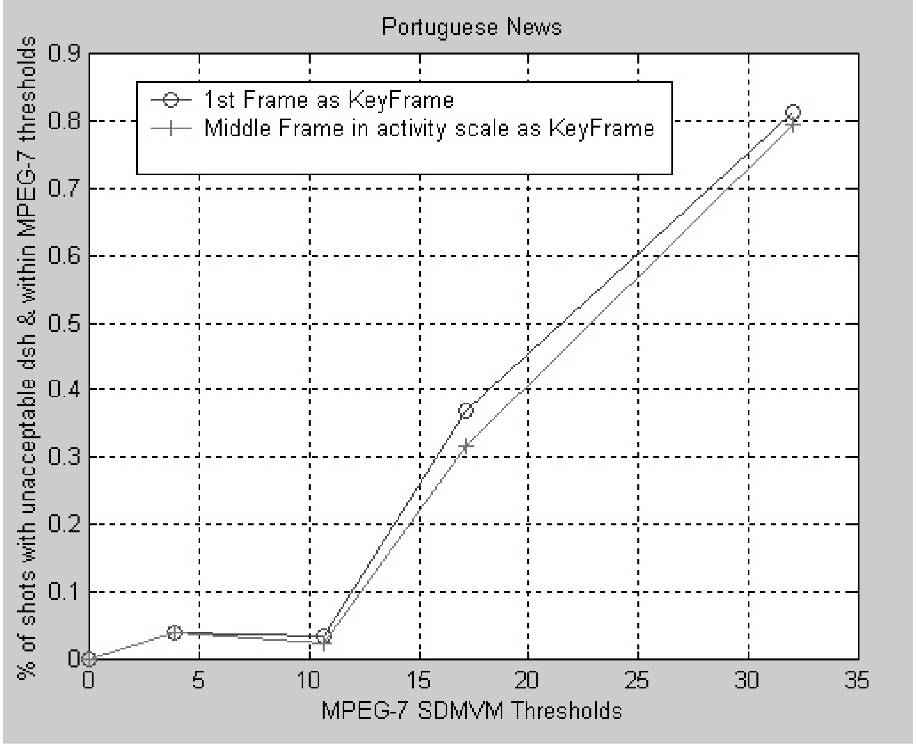

Figure 1.1. Verification of Hypothesis and choice of single key-frame Motion activity (Standard Deviation of Motion Vector Magnitude) Vs percentage duration of unacceptable Shots (Portuguese News from MPEG-7 Test Set jornaldanoite1.mpg)

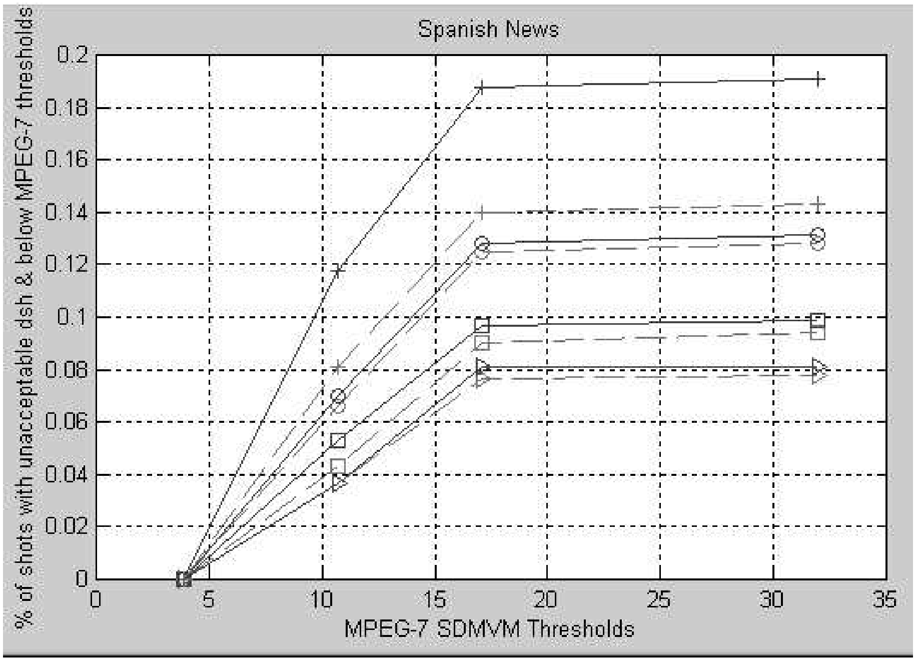

Figure 1.2. Motion activity (Standard Deviation of Motion Vector Magnitude) Vs percentage duration of unacceptable Shots (Spanish News from MPEG-7 Test Set) The firm line represents the "optimal" key-frame strategy while the dotted line represents the progressive key-frame extraction strategy. Each shape represents a certain number of key-frames, the + represents a single frame, the circle two frames, the square three frames and the triangle five frames.

2.2 A Motion Activity based Non-Uniform Sampling Approach to Key Frame Extraction

If as per section 2.1 intensity of motion activity is indeed a measure of the change from frame to frame, then over time, the cumulative intensity of motion activity must be a good indication of the cumulative change in the content. Recall that in our review of previous work we stated that being forced to pick the first frame as a key-frame is disadvantageous. If the first frame is not the best choice for the best first key-frame, schemes that use it as the first key-frame such as those surveyed in [4] start off at a disadvantage. This motivates us to find a better single key-frame based on motion activity. If each frame represents an increment in information then the last frame is at the maximum distance from the first. That

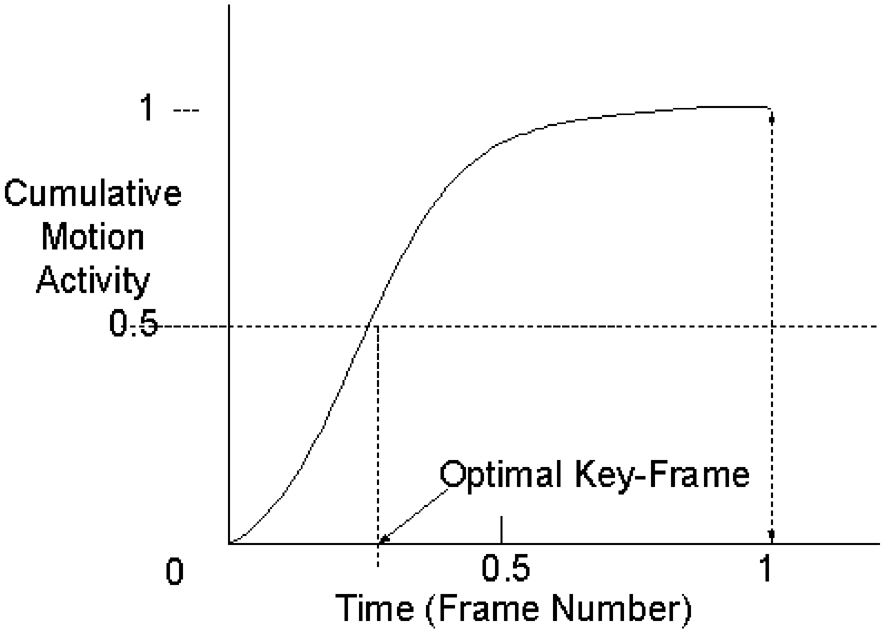

Figure 1.3. Illustration of single key-frame extraction strategy. Note that there is a simple generalization for N key frames

Table 1.1. Comparison with Optimal Fidelity Key-Frame

Motion Activity	Δd_{sh} First Frame	Δd_{sh} Proposed KF	Number of Shots
Very Low	0.0116	0.0080	25
Low	0.0197	0.0110	133
Medium	0.0406	0.0316	73
High	0.095	0.0576	28
Very High overall avg.	0.0430	0.022	16

would imply that the frame at which the cumulative motion activity is half the maximum value is the best choice for the first key-frame. We test this hypothesis by using the frame at which the cumulative motion activity is half its value for the entire shot as the single key-frame instead of the first key frame for the Spanish News sequence and repeating the experiment in the previous section. We find that the new key-frame choice out-performs the first frame, as illustrated in Figure 1.1. Since previous schemes have also improved upon using the first frame as a key-frame, we need to compare our single key-frame extraction strategy with them. For each shot, we compute the optimal single key-frame as per the fidelity criterion mentioned in section 2.2. We compute it by finding the fidelity of each of the frames of the video, and then finding the frame with the best fidelity. We use the fidelity of the aforementioned optimal key-frame as a benchmark for our key-frame extraction strategy by measuring the difference in $d_s h$ between the optimal key-frame obtained through the exhaustive computation mentioned earlier and the key-frame obtained through our proposed motion-activity based strategy. We carry out a similar comparison for the first-frame based strategy as well. We illustrate our results in Table 1.1. Note that our strategy produces key-frames that are nearly optimal in fidelity. Furthermore, the quality of the approximation degrades as the intensity of motion activity increases. In other words, we find that our strategy closely approximates the optimal key-frame extraction in terms of fidelity while using much less computation. This motivates us to propose a new nearly optimal strategy, which is very similar to the activity-based sampling proposed in the next section [5] as follows. To get n key-frames, divide the video sequence into n equal parts on the cumulative motion activity scale. Then use the frame at the middle of the cumulative motion activity scale of each of the segments as a key-frame, thus getting n key-frames. Note that our n key-frame extraction strategy scales linearly with n unlike the exhaustive computation described earlier, which grows exponentially in complexity because of the growth in the number of candidate key-frame combinations. It is for this reason that we do not compare our n-frame strategy with the exhaustive benchmark. We illustrate our strategy in Figure 1.3.

Note that our criterion of acceptable fidelity combined with the n-frame strategy enables a simple and effective solution to the two basic problems of key-frame extraction:

- How many key-frames does a shot require for acceptable fidelity?
- How to generate the required number of key-frames?

2.3 A Simple Progressive Modification

Progressive key-frame extraction is important for interactive browsing since the user may want further elaboration of summaries that he has already received. Since our key-frame extraction is not progressive, we propose a progressive modification of our technique. We start with the first frame, and then choose the last frame as the next key-frame because it is at the greatest distance from the first frame. We carry this logic forward as we compute further key-frames by choosing the middle key-frame in cumulative motion activity space as the third key-frame, and so on recursively. The modified version is slightly inferior to our original technique but has the advantage of being progressive. In figure 1.2, we illustrate a typical result. We have tried our approach with several news programs from different sources[7, 6].

3. Constant Pace Skimming Using Motion Activity

3.1 Introduction

In the previous section, we showed that the intensity of motion activity (or pace) of a video sequence is a good indication of its "summarizability." Here we build on this notion by adjusting the playback frame-rate or the temporal sub-sampling rate. The pace of the summary serves as a parameter that enables production of a video summary of any desired length. Either the less active parts of the sequence are played back at a faster frame rate or the less active parts of the sequence are sub-sampled more heavily than are the more active parts, so as to produce a summary with constant pace. The basic idea is to skip over the less interesting parts of the video.

3.2 Constant Activity Sub-Sampling or Activity Normalized Playback

A brute force way of summarizing video is to play it back at a faster-than-normal rate. Note that it can also be viewed as uniform sub-sampling. Such a fast playback has the undesirable effect of speeding up all the portions equally thus making the high motion parts difficult to view, while not speeding up the low motion parts sufficiently. This suggests that a more useful approach to fast playback would be to play back the video at a speed that provides a viewable and constant level of motion activity. Thus, the low activity segments would have to be speeded up considerably to meet the required level of motion activity, while the high activity segments would need significantly less speeding up if at all. In other words, we would speed up the slow parts more than we would the fast parts. This can be viewed as adaptive playback speed variation based on motion activity, or activity normalized playback. Another interpretation could be in terms of viewability or "perceptual bandwidth." The most efficient way to play the video is to make full use of the instantaneous perceptual bandwidth, which is what the constant activity playback achieves. We speculate that the motion activity is a measure of the perceptual bandwidth as a logical extension of the notion of motion activity as a measure of summarizability. Let us make the preceding qualitative description more precise. To achieve a specified activity level while playing back video, we need to modify the activity level of the constituent video segments. We first make the assumption that the intensity of motion activity is proportional to the motion vector magnitudes. Hence, we need to modify the motion vectors so as to modify the activity level. There are two ways that we can achieve this:

- Increasing/Decreasing the Playback Frame-Rate - As per our earlier assumption, the intensity of motion activity increases linearly with frame rate. We can therefore achieve a desired level of motion activity for a video segment as follows:

  Playback frame rate =(Original Frame rate)*(Desired level of motion activity/original level of motion activity)

- Sub-Sampling the Video Sequence - Another interpretation of such playback is that it is adaptive sub-sampling of the frames of the segment with the low activity parts being sub-sampled more heavily. This interpretation is especially useful if we need to summarize remotely located video, since we often cannot afford the bandwidth required to actually play the video back at a faster rate.

In both cases above, the summary length would then be given by

Summary length=(Sum of frame activities)/desired activity

Note that we have not yet specified the measure of motion activity. The most obvious choices are the average motion vector magnitude and the variance of the motion vector magnitude [1, 2]. However, there are many variations possible, depending on the application. For instance, we could use the average motion vector magnitude as a measure of motion activity, so as to favor segments with moving regions of significant size and activity. As another example, we could use the magnitude of the shortest motion vector as a measure of motion activity, so as to favor segments with significant global motion.

The average motion vector magnitude provides a convenient linear measure of motion activity. Decreasing the allocated playing time by a factor of two, for example, doubles the average motion vector magnitude. The average motion vector magnitude $\hat{r}$ of the input video of N frames can be expressed as:

$$\hat{r} = (\frac{1}{N})\sum_{i=1}^{N} r_i$$

where the average motion vector magnitude of frame i is r_i. For a target level of motion activity r_{target} in the output video, the relationship between the length L_{output} of the output video and the length L_{input} of the input video can be expressed as:

$$L_{output} = \frac{\hat{r}}{r_{target}} L_{input}$$

While playing back at a desired constant activity is possible in theory, in practice it would require interpolation of frames or slowing down the playback frame rate whenever there are segments that are higher in activity than the desired level. Such interpolation of frames would be computationally intensive and difficult. Furthermore, such an approach does not lend itself to generation of a continuum of summary lengths that extends from the shortest possible summary to the original sequence itself.

The preceding discussion motivates us to change the sampling strategy to achieve a guaranteed minimum level of activity as opposed to a constant level of activity, so we are able to get a continuum of summaries ranging from the sequence being its own summary to a single frame summary. With the guaranteed minimum activity method, we speed up all portions of the input video that are lower than the targeted minimum motion activity r_{target} so that they attain the targeted motion activity using the above formulations. The portions of the input video that exceed the targeted motion activity can remain unchanged.

At one extreme, where the guaranteed minimum activity is equal to the minimum motion activity in the input video, the entire input video becomes the output video. When the guaranteed minimum activity exceeds the maximum motion activity of the input video, the problem reduces to the above constant activity case. At the other extreme, where the targeted level of activity is extremely high, the output video includes only one frame of the input video as a result of down-sampling or fast play.

The length of the output video using the guaranteed minimum activity approach can be determined as follows. First, classify all of the frames of the input video into two sets. A first set S_{higher} includes all frames j where the motion activity is equal to or higher than the targeted minimum activity. The second set S_{lower} includes all frames k where the motion activity is lower than the targeted motion activity. Then, the length of the input video is expressed by:

$$L_{input} = L_{higher} + L_{lower}.$$

The average motion activity of frames j that belong to the set S_{lower} is

$$\hat{r}_{lower} = \frac{1}{N_{lower}} \sum_{j}^{N_{lower}} r_k$$

and the length of the output converted is

$$L_{output} = \frac{\hat{r}_{lower}}{r_{target}} L_{lower} + L_{higher}$$

It is now apparent that the guaranteed minimum activity approach reduces to the constant activity approach because when L_{higher} becomes zero, the entire input video needs to be processed.

3.3 How Fast Can You Play the Video?

While in theory it is possible to play the video back at infinite speed, the temporal Nyquist rate limits how fast it can be played without becoming imperceptible by a human observer. A simple way of visualizing this is to imagine a video sequence that captures the revolution of a stroboscope. At the point where the frame rate is equal to the rate of revolution, the stroboscope will seem to be stationary. Thus, the maximum motion activity level in the video segment determines how fast it can be played. Furthermore, as the sub-sampling increases, the video segment reduces to a set of still frames or a "slide show." Our key-frame extraction technique of Section 2 can therefore also be seen as an efficient way to generate a slide show since it uses the least feasible number of frames. It is obvious that there is a cross-over point where it is more efficient to summarize the video segment using a slide show instead of with a video or "moving" summary. How to locate the cross-over point is an open problem. We hope to address this problem in our ongoing research.

3.4 Experimental Procedure, Results and Discussion

We have tried our adaptive speeding up of playback using diverse content and got satisfactory results. We find that with surveillance footage of a highway (see Figure 1.4), using the average motion vector magnitude as the measure of motion activity, we are able to produce summaries that successfully skip across the parts where there is insignificant traffic, and focus on the parts with significant traffic. We get good results with the variance of motion vector magnitude as well. We have been able to focus on parts with large vehicles, as well as on parts with heavy traffic. Note that our approach is computationally simple since it relies on simple descriptors of motion activity.

As illustrated by our results, the constant pace skimming is especially useful in surveillance and similar applications in which the shots are long and the background is fixed. Note that in such applications, color-based techniques are at a disadvantage since the semantics of the interesting events are much more strongly related to motion characteristics than to changes in color.

We have also tried this approach with sports video and with news content with mixed success. When viewing consumer video such as news or sports, skipping some parts at fast forward and viewing others at normal speed may be preferable to continuously changing the playback speed. For this purpose, we smooth the activity curve using a moving average and quantize the smoothed values (Figure 1.5). In our implementations, we used two-level quantization with the average activity as the threshold. For news and basketball (Figure 1.5 b and d), we used manually selected thresholds.

For the golf video, the low activity parts are where the player prepares for his hit, followed by the high activity part where the camera follows the ball, or closes up on the player. For the news segment, we are able to separate the interview parts from the outdoor footage. For the soccer video, we see low activity segments before the game really starts, and also during the game where the game is interrupted. The basketball game, in contrast with soccer, has a high frequency of low and high activity segments. Furthermore, the low activity parts are when the ball is in one side of the court and the game is proceeding, and the high activity occurs mostly during close-ups or fast court changes. Hence, those low activity parts should be played at normal speed, while some high activity parts can be skipped. In short, we can achieve a semantic segmentation of various content types using motion activity, and use domain knowledge to determine where to skip or playback at normal speed. Then, we can accordingly adapt our basic strategy described in Section 3.1, to different kinds of content.

The foregoing discussion on sports video indicates that the key to summarizing sports video is in fact in identifying interesting events. This motivates us to investigate temporal patterns of motion activity that are associated with interesting events in Section 5.

For news video, it is perhaps better to use a slide show based on our key-frame extraction technique since the semantics of the content are not directly coupled with the motion characteristics of the content. However, it works best when the semantic boundaries of the content are known. In that case, a semantic segment can be segmented into shots and key-frames extracted for each shot so as to produce a set of key-frames for the entire semantic segment. This motivates us to investigate automatic news video semantic, or topic, boundary detection using audio features in Section 4.

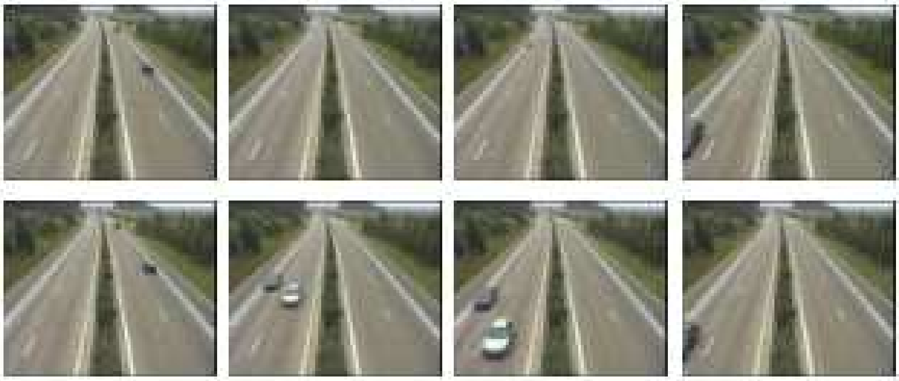

Figure 1.4. Illustration of adaptive sub-sampling approach to video summarization. The top row shows a uniform sub-sampling of the surveillance video, while the bottom row shows an adaptive sub-sampling of the surveillance video. Note that the adaptive approach captures the interesting events while the uniform sub-sampling mostly captures the highway when it is empty. The measure of motion activity is the average motion vector magnitude in this case.

4. Audio-Assisted News Video Browsing

4.1 Motivation

The key-frame based video summarization techniques of Section 2 are evidently restricted to summarization of video shots. Video in general, and news video in particular, consists of several distinct semantic units, each of which in turn consists of shots. Therefore it would be much more convenient to somehow choose the semantic unit of interest and then view its key-frame based summary in real-time, than to generate a key-frame based summary of the entire video sequence and then look for the semantic unit of interest in the summary. If a topic list is available in the content meta-data, then the problem of finding the boundaries of the semantic units is already solved, so that the user can first browse the list of topics and then generate and view a summary of the desired topic.

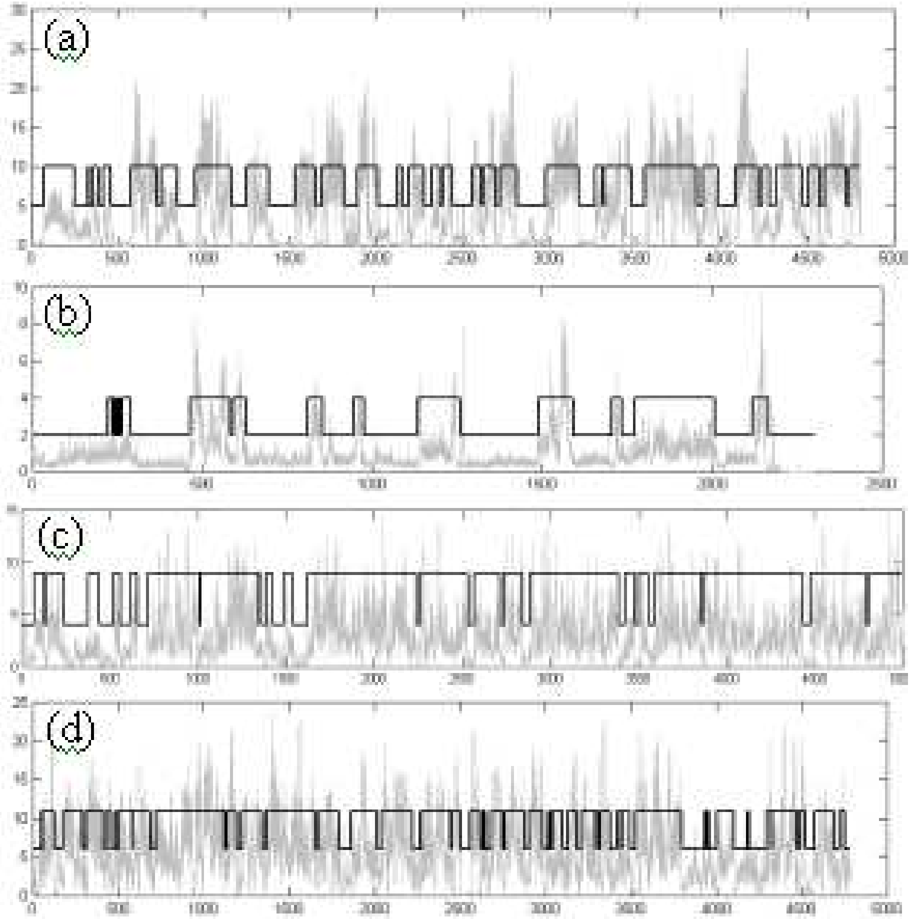

Figure 1.5. Motion activity vs. frame number for four different types of video content, with the smoothed and quantized versions superimposed: a) Golf. b) News segment. c) Soccer. d) Basketball.

However, if the topic list is unavailable, as is the case more often than not, the semantic boundaries are no longer readily available. We then need to extract the semantic/topic boundaries automatically. Past work on news video browsing systems has emphasized news anchor detection and topic detection, since news video is typically arranged topic-wise and the news-anchor introduces each topic at the beginning. Thus knowing the topic boundaries enables the user to skim through the news video from topic to topic until he has found the desired topic, which he can then watch using a normal video player.

Topic detection has been mostly carried out using closed caption information, embedded captions and text obtained through speech recognition, by themselves or in combination with each other (See [8, 9] for example). In such approaches, text is extracted from the video using some or all of the aforementioned sources and then processed using various heuristics to extract the topic(s).

News anchor detection has been carried out using color, motion, texture and audio features. For example, in [14] Wang et al carry out a speaker separation on the audio track and then use the visual or video track to locate the faces of the most frequent speakers or the "principal cast." The speaker separation is carried out by first classifying the audio segments into the categories of speech and non-speech. The speech segments are then used to train Gaussian Mixture Models (GMM's) for each speaker that enable speaker separation through fitting of each speech segment with the different GMM's.

Speaker separation has itself been a topic of active research. The techniques mostly rely on extraction of low level audio features followed by a clustering/classification procedure.

As mentioned in the introduction, speaker separation and principal cast identification provide a solution to the problem of topic boundary detection. Unfortunately, the proposed methods in the literature are highly complex in computation and hence do not lend themselves well to consumer video browsing systems. Furthermore, in a video browsing system, in addition to principal cast identification we would also like to identify further semantic characteristics based on the audio track such as speaker gender, as well as carry out searches for similar scenes based on audio.

4.2 MPEG-7 Generalized Sound Recognition

The above discussion motivates us to try the sound recognition framework proposed by Casey [13] and accepted by the MPEG-7 standard. In this framework, reduced rank spectra and entropic priors are used to train Hidden Markov Models for various sounds such as speech, male speech, female speech, barking dogs, breaking glass etc. The training is done off-line with training data so that the category identification is done by using the Viterbi algorithm on the various HMM's, which is computationally inexpensive. For each sound segment, in addition to the sound category identification, a histogram of percentage duration spent in each state of the HMM is also generated. This histogram serves as a compact feature vector that enables similarity matching.

4.3 Proposed Principal Cast Identification Technique

Our procedure [10] is illustrated in Figure 1.7. It consists of the following steps:

1 Extract motion activity, color and audio features from the News video.

2 Use the sound recognition and clustering framework as shown in figure 1.7 to find speaker changes.

3 Use motion and color to merge speaker clusters and to identify principal speakers. The locations of the principal speakers provide the topic boundaries.

4 Apply the motion based browsing described in sections 2 and 3 to each topic.

In the following section, we describe the speaker change detection component of the whole system.

4.3.1 Speaker Change Detection Using Sound Recognition and Clustering. The input audio from broadcast news is broken down into sub-clips of smaller durations such that they are homogenous. The energy of each sub-clip is calculated so as to detect and remove silent sub-clips. MPEG-7 features are extracted from non-silent sub-clips and are classified into one of the three sound classes namely male, female and speech with music.

At this point, all male and female speakers are separated. Median filtering is performed to eliminate spurious changes in speakers. In order to identify individual speakers within male and female sound class, an unsupervised clustering step is performed based on the MPEG-7 state duration histogram descriptor. This clustering step is essential to identify individual male and female speakers after classification of all sub-clips into one of three sound classes. Each classified sub-clip is then associated with a state duration histogram descriptor.

The state duration histogram can also be interpreted as a modified representation of GMM. Each state in the trained HMM can be thought of, as a cluster in feature space, which can be modeled by a Gaussian. Note that the state duration histogram represents the probability of occurrence of a particular state. This probability can be interpreted as the probability of a mixture component in a GMM. Thus, the state duration histogram descriptor can be considered as a reduced representation of GMM, which in its unsimplified form is known to model a speaker's utterance well[6]. Note, since the histogram is derived from the HMM, it also captures some temporal dynamics which a GMM cannot. We are thus motivated to use this descriptor to identify clusters belonging to different speakers in each sound class.

The clustering approach adopted was bottom-up agglomerative dendrogram construction based. In this approach, a distance matrix is first obtained by computing pairwise distance between all utterances to be clustered. The distance metric used is a modification of Kullback-Leibler distance to compare two probability density functions (pdf). The modified Kullback-Leibler distance between two pdfs H and K is defined as below:

$$D(H,K) = \Sigma h_i log(\frac{h_i}{m_i}) + m_i log(\frac{k_i}{m_i})$$

where $m_i = \frac{h_i+k_i}{2}$ and $1 \leq i \leq Number\ of\ bins\ in\ the\ histogram$

Then a dendrogram is constructed by merging two closest clusters according to the distance matrix until there is only one cluster. Then, the dendrogram is cut to obtain the clusters of individual speakers (See 1.8).

4.3.2 Second level of clustering using motion and color features. Since clustering is done only on contiguous male/female speech segments, we achieve speaker segmentation only in that portion of the whole audio record of the news program. A second level of clustering is required to establish correspondences between clusters from two distinct portions. Motion and color cues extracted from the video can be used for the second level of clustering. Once the clusters have been merged, it is easy to identify principal cast and hence semantic boundaries. Then, the combination of the principal cast identification and motion-based summary of each semantic segment enables quick and effective browsing of the news video content.

4.4 Experimental Procedure and Results

4.4.1 Data-Set. Since broadcast news contains mainly three sound classes viz, male speech, female speech and speech with music, we collected training examples for each of the sound classes from three and a half hours of news video from four different TV channels manually. The audio signals are all mono-channel, 16 bits per sample with a sampling rate of 16 KHZ. The database for training HMMs is partitioned into 90%-10% training/testing set for cross-validation. The test sequences for speaker change detection were two audio tracks from TV broadcast news: 'news 1' with duration 34 minutes and 'news 2' 59 minutes respectively.

4.4.2 Feature Extraction. The input audio signal from the news program is cut into segments of length 3 seconds and silent segments are removed. For each non-silent 3s segment, MPEG-7 features are extracted as follows. Each segment is divided into overlapping frames of duration 30ms with 10ms overlapping for consecutive frames. Each frame is then multiplied by a hamming window function: $w_i = (0.5 - 0.46cos(2\pi i/N)), 1 = i = N$, where N is the number of samples in the window. After performing a FFT on each windowed frame, the energy in each of the sub-bands is computed and the resulting vector is projected onto the first 10 principal components of each sound class.

We also extract from compressed domain, the MPEG-7 intensity of motion activity for each P-frame and a 64 bin color histogram for each I-frame from the video stream of the news program.

4.4.3 Classification and Clustering. The number of states in each of the HMM was chosen to be 10 and each state is modeled by a single multi-variate Gaussian. Note that the state duration histogram descriptor relates to GMM only if the HMM states are represented by a single gaussian. Viterbi decoding is performed to classify the input audio segment by picking the model for which the likelihood value is maximum which is followed by median filtering on the labels obtained for each 3s segment so as to impose time continuity. For each contiguous set of labels, agglomerative clustering is performed using the state duration histogram descriptor to obtain a dendrogram as shown in figure 1.8. The dendrogram is then cut at a particular level relative to the maximum height of the dendrogram to obtain individual speaker clusters.

The accuracy of the proposed approach for speaker change detection depends on the following two aspects: the classification accuracy of the trained HMMs for segmenting the input audio into male and female speech classes and the accuracy of the clustering approach to identify individual speakers in a contiguous set of male/female speech utterances. Tables 1.2a and 1.2b show the classification performance of the HMM on each of the test broadcast news sequences 'without' any post processing on the labels.

Tables 1.2a and 1.2b indicate that many male and female speech segments are classified as speech with music. These segments actually correspond to outdoor speech segments in the broadcast news and have been mis-classified due to the background noise.

Since clustering was done only on contiguous male or female speech segments instead of the whole audio record, the performance of the system is evaluated as a speaker change detection system even though we achieve segmentation in smaller portions. We compare speaker change positions output by the system against ground truth speaker changes, and count the number of correct speaker change positions. Table 1.2c summarizes the performance of the clustering approach on both the test sequences.

The accuracy of the proposed algorithm for speaker change detection is only moderate for both the news programs for the following reasons. The dendrogram pruning procedure adopted to generate clusters was the simplest one and hence results would improve if multi-level dendrogram pruning procedure was adopted. Some of the speaker changes were missed by the system because of mis-classification of the outdoor speech segments into speech with music class. Moreover, there was no post-processing on the clustering labels to incorporate some domain knowledge. For example, a cluster label sequence such as s1, s2, s1, s2 in which speakers alternate frequently, is highly unlikely in a news program and would simply mean that s1 and s2 belong to the same speaker cluster. However, even with such a moderate accuracy in audio analysis it is shown below that by combining motion and color cues from video, the principal cast from news program can be obtained.

In order to obtain correspondence between speaker clusters from distinct portions of the news program, we associate each speaker cluster with a color histogram, obtained from a frame with motion activity less than a threshold. Obtaining a frame from a low-motion sequence increases the confidence of its being one from a head-shoulder sequence. A second clustering step is then performed based on color histogram to merge clusters obtained from pure audio analysis. Figure 1.9 shows the second level clustering results. After this step, principal cast clusters can be identified as either the clusters that occupy significant period of time or clusters that appear at different times, throughout the news program. Due to copyright issues, we unfortunately cannot display any of the images corresponding to the clusters. In future work, we hope to use public domain data such as news video from the MPEG-7 video test-set.

4.5 Future Work

Our future work will focus on first improving the audio classification using more extensive training. Second, we will further improve clustering of audio by allowing multi-level dendrogram cutting. Third, we will further refine the combination of motion activity and color to increase the reliability of principal cast identification.

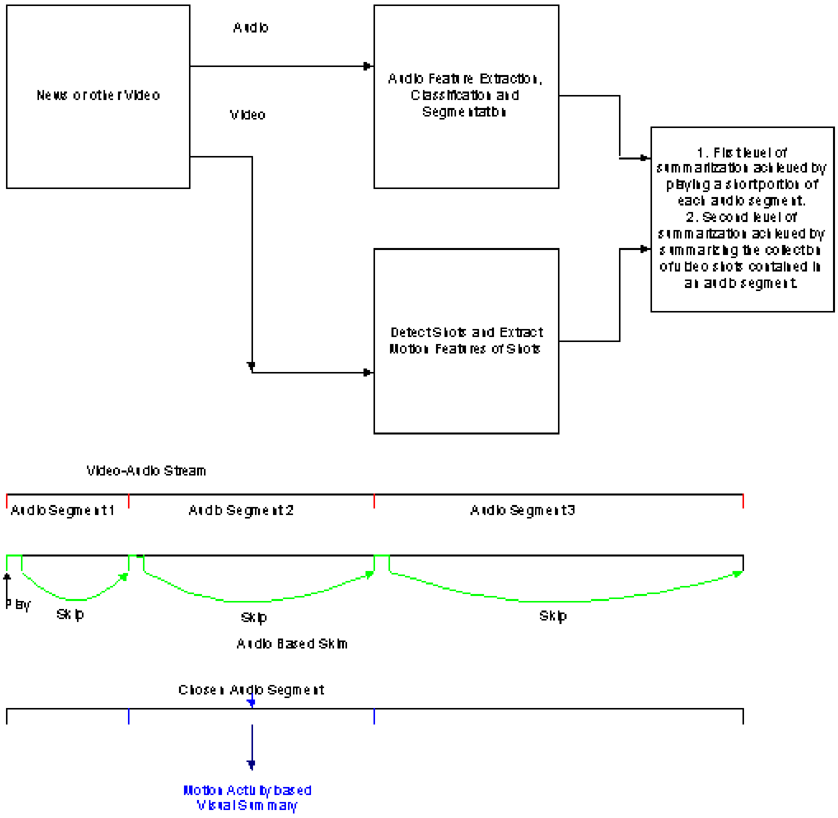
News or other Video
Audio
Video
Audio Feature Extraction, Classification and Segmentation
Detect Shots and Extract Motion Features of Shots
1. First level of summarization achieved by playing a short portion of each audio segment.
2. Second level of summarization achieved by summarizing the collection of video shots contained in an audio segment.
Video-Audio Stream
Audio Segment 1
Audio Segment 2
Audio Segment 3
Play
Skip
Skip
Skip
Audio Based Skim
Chosen Audio Segment
Motion Activity based Visual Summary

Figure 1.6. Audio-Assisted Video Browsing

5. Sports Highlights Detection

Most sports highlights extraction techniques depend on the camera motion, and thus require accurate motion estimation for their success. In the compressed domain, however, since the motion vectors are noisy, such accuracy is difficult to achieve. Our discussion in section 3.4 motivates us to investigate temporal patterns of motion activity as a means for event detection since they are simple to compute. We begin by devising strategies for specific sports based on domain knowledge. We find that using motion activity alone gives rise to too many false positives for certain sports. We then resort to combining simple audio and video cues

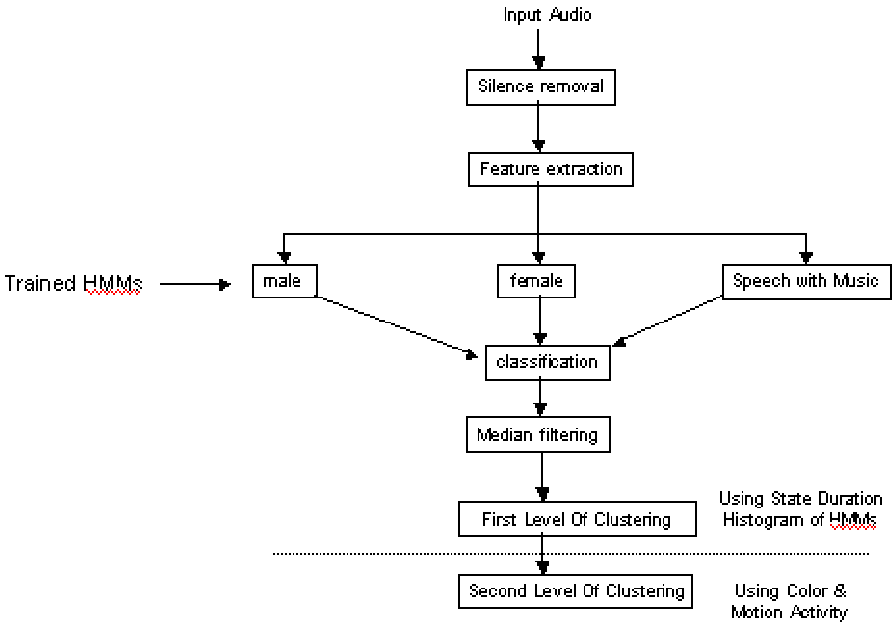
Input Audio
Silence removal
Feature extraction
Trained HMMs
male
female
Speech with Music
classification
Median filtering
First Level Of Clustering
Using State Duration Histogram of HMMs
Second Level Of Clustering
Using Color & Motion Activity

Figure 1.7. Audio Feature Extraction, Classification and Segmentation for Speaker Change Detection

Table 1.2a. Classification Results on News1 with Average Recognition Rate = 80.384%

	Female	Male	Speech and Music
Female	116	52	13
Male	6	184	2
Speech and Music	15	46	264

Table 1.2b. Classification Results on News2 with Average Recognition Rate = 84.78%

	Female	Male	Speech and Music
Female	370	50	8
Male	34	248	1
Speech and Music	47	49	391

to eliminate false positives. Our results with the combination as well as the results from section 4 motivate us to apply the generalized sound recognition framework to a unified highlights extraction framework for soccer, golf and baseball. Our current challenge is therefore to combine visual features with the sound recognition framework. We hope to build upon our experience with combining low-level cues.

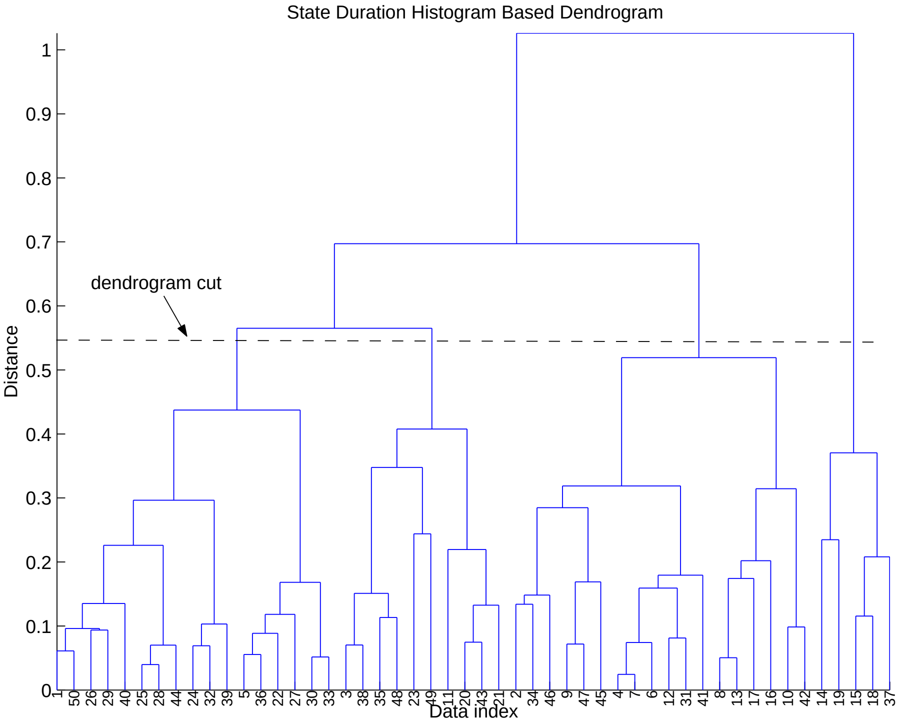

Figure 1.8. Example Dendrogram Construction and Cluster generation for a contiguous set of female speech segments

Table 1.2c. Speaker Change Detection Accuracy on two test news sequences **A** Number of speaker change time stamps in ground truth; **B** Number of speaker change time stamps obtained after clustering step; **C** Number of 'TRUE' speaker change time stamps; **D** Precision = [C]/[A] in %; **E** Recall = [C]/[B] in %

	A	**B**	**C**	**D**	**E**
News1	68	90	46	67.64	51.11
News2	173	156	87	50.28	55.77

Note that the problem has two parts viz. detecting an interesting event and then capturing its entire duration. In this chapter we focus on the first part since in our target applications, we are able to solve the second part by merely using an interactive interface that allows conventional fast-forward and rewind.

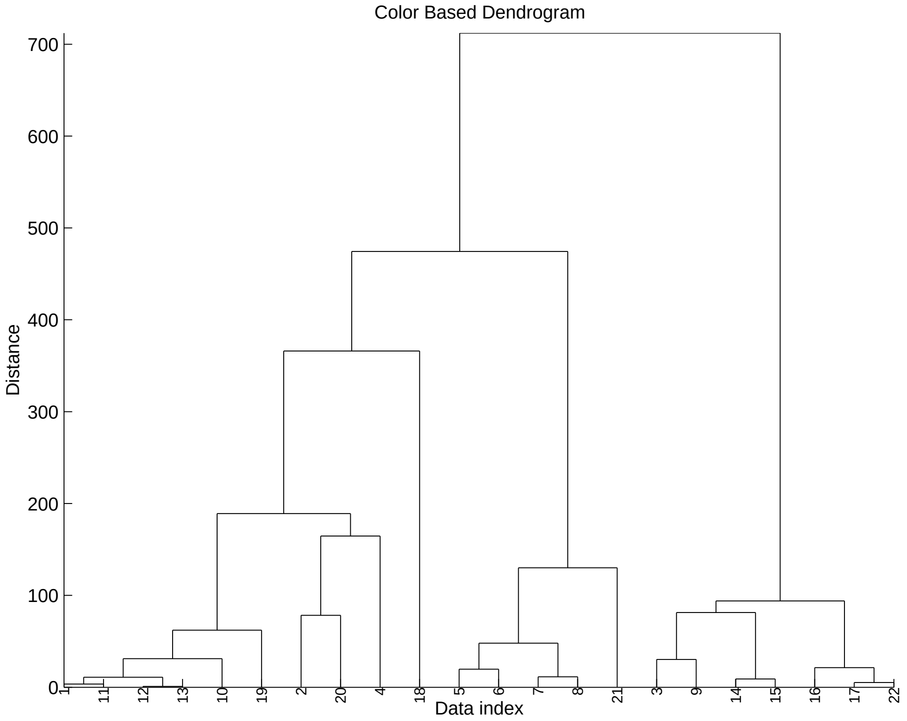

Figure 1.9. Second level of clustering based on color histograms of frames corresponding to male speaker clusters

5.1 Highlights Extraction for Golf

In [11] we describe a simple technique to detect Golf highlights. We first carry out a smoothing of the motion activity of the video sequence as described in [11]. In the golf video, we look for long stretches of very low activity followed by high activity. These usually correspond to the player working on his shot, then hitting, followed by the camera following the ball or zooming on the player. We mark the interesting points in the golf video in this way. We generate a highlight sequence by merely concatenating ten-second sections that begin at the interesting points marked. We get interesting results but miss some events, most notably the putts since they are often not associated with rapid camera motion.

5.2 Extraction of Soccer Video Highlights

Since our scope is restricted to soccer games, we can capitalize on domain specific constraints that would help us locate the highlights. Our basic intuition is that an interesting event always has the following associated effects. The game stops and stays stopped for a non-trivial duration, and the crowd noise goes up either in anticipation of the event, or after the event has taken place

This suggests the following straightforward strategy to locate interesting events or highlights. Locate all audio volume peaks. The peaks correspond to increase in crowd noise in response to the interesting event. At every peak, find out if the game stopped before it and stayed stopped for a non-trivial duration. Similarly find out if the game stopped after the peak and stayed stopped. The concatenation of the stops before and after the audio peak, if valid, forms the highlight associated with that audio peak. We describe the details of the computation of the audio peaks and the start and stop patterns of motion activity in [12]

5.2.1 Experimental Results. We have tried our strategy [12] with 7 soccer games from Korea, Europe and the United States of America including a women's soccer game. We find that we miss only one goal and capture all the other goals in all the games. We also capture several other interesting parts of the game that do not lead to a goal being scored such as attempts at goals, major injuries etc. Despite its success with diverse content, this technique has a significant drawback which is its reliance on a low-level feature like audio volume, which may not always be a good indicator of the content semantics. We are thus motivated once again to resort to generalized sound recognition.

5.3 Audio Events Detection based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework

We describe an audio-classification based approach in which we explicitly identify applause/cheering segments, and use those to identify highlights. We also use our audio classification framework to set up a future investigation of fusion of audio and video cues for sports highlights extraction.

5.3.1 Audio Classification Framework. The system constraints of our target platform rule out having a completely distinct algorithm for each sport and motivate us to investigate a common unified highlights framework for our three sports of interest, golf, soccer and baseball. Since audio lends itself better to extraction of content semantics, we start with audio classification.

We illustrate the audio classification based framework in Figure 1.10. In the audio domain, there are common events relating to highlights across different sports. After an interesting golf or baseball hit or an exciting soccer attack, the audience shows appreciation by applauding or cheering. The duration of the applause/cheering is an indication of the "significance" of the moment. Furthermore, in sports broadcast video, there are also common events related to commercial messages that often consist of speech or speech and music. Our observation is that the audience's cheering and applause are more general across different sports than is the announcer's excited speech. We hence look for robust audio features and classifiers to classify and recognize the following audio signals: applause, cheering, ball hits, music, speech and speech with music. The former two are used for highlights extraction and the latter three are used to filter out the uninteresting segments. We employ a general sound recognition framework based on Hidden Markov Models (HMM) trained for each of the classes. The HMM's operate on Mel Frequency Cepstral Coefficients (MFCC). Each segment is 0.5 seconds long while each frame is 30 ms long. We show that our classification accuracy is high and thus we are motivated to extract the highlights based on the results of the classification.

We collect the continuous or uninterrupted stretches of applause/cheering. We retain all the segments that are a certain percentage of the maximum duration of the applause/cheering. Our default choice is 33%. Note that this gives us a simple threshold with which to tune the highlights extraction for interactive browsing. Finally, we add a preset time cushion to both ends of each selected segment to get the final presentation time stamps. The presentation then consists of playing the video normally through a time-stamp pair corresponding to a highlight and then skipping to the next pair.

Note that the duration of the applause/cheering also enables generation of sports highlights of a desired length as follows: We can sort all the applause/cheering segments in a descending order of duration. Then given a time budget, we can spend it by playing each segment down the list until the budget is exhausted. While the above technique is promising, we find that it still has room for improvement as can be seen in Table 1.3. First, the classification accuracy needs to be improved. Second, using applause duration alone is probably simplistic. Its chief strength is that it uses the same technique for three different sports. Since we do not expect a high gain from increased classification accuracy alone, we

	[A]	[B]	[C]	[D]	[E]	[F]	[G]
[1]	58	47	35	60.3%	74.5%	151	23.1%
[2]	42	94	24	57.1%	25.5%	512	4.7%
[3]	82	290	72	87.8%	24.8%	1392	5.2%
[4]	54	145	22	40.7%	15.1%	1393	1.6%

Table 1.3. Classification Results of the 4 games. [1]: golf game 1; [2]: golf game 2; [3] baseball game; [4] soccer game. [A]: Number of Applause and Cheering Portions(NACP) in Ground Truth Set; [B]: NACP by Classifiers WITH Post-processing; [C]: Number of TRUE ACP by Classifiers; [D]: Precision $\frac{[C]}{[A]}$; [E]: Recall $\frac{[C]}{[B]}$ WITH Post-processing; [F]: NACP by Classifiers WITHOUT Post-processing; [G]: Recall $\frac{[C]}{[F]}$ WITHOUT Post-processing.

are motivated to combine visual cues with the audio classification with the hope that we may get a bigger gain in highlight extraction efficacy.

5.4 Future Work

In ongoing research, we propose to combine the semantic strength of the audio classification with the computational simplicity of the techniques described in sections 5.1 and 5.2.

We are thus motivated to investigate combination of audio classification with the motion activity pattern matching. We illustrate our general framework in Figure 1.10. Note that the audio classification and the video feature extraction both produce candidates for sports highlights. We then propose to use probabilistic fusion to choose the right candidates. Note also that the proposed video feature extraction goes well beyond the motion activity patterns that we described earlier.

Our proposed techniques have the advantage of simplicity and fair accuracy. In ongoing work, we are examining more sophisticated methods for audio-visual feature fusion.

6. Efficacy of Summarization

Using the capture of goals in soccer as a measure of the accuracy of highlights has the big advantage of zero ambiguity but also has the disadvantage of incompleteness since it ignores all other interesting events that could arguably be even more interesting. We are currently working on a framework to assess the accuracy of a sports highlight in terms of user satisfaction, so as to get a more complete assessment. Such a framework would require a carefully set up psycho-visual experiment that creates a ground truth for the "interesting" and "uninteresting" parts of a sports video.

More structured content such as news lends itself to easier assessment of the success of the summarization. However, note that our fidelity based computations for example, did not address semantic issues. The assessment of the semantic success of a summary is still an open problem although techniques such as ours provide part of the solution.

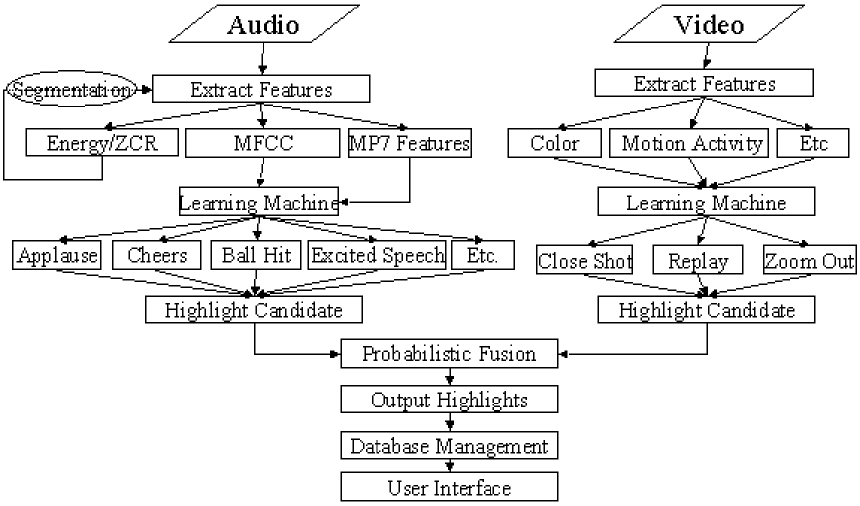

Figure 1.10. Highlights Extraction Framework: We have partially realized the video and the probabilistic fusion.

7. Data Mining vs Video Mining: A Discussion

Finally we consider the video mining problem in the light of existing data mining techniques. The fundamental aim of data mining is to discover patterns. In the results we have presented here, we have attempted to discover patterns in audio-visual content through principal cast detection, sports highlights detection, and location of "significant" parts of video sequences. Note that while our techniques do attempt to satisfy the aim of pattern discovery, they do not directly employ common data mining techniques such as time-series mining or discovery of association rules. Furthermore, in our work, the boundary between detection of a known pattern and pattern discovery is not always clear. For instance, looking for audio peaks and then motion activity patterns around them could be thought of as merely locating a known pattern, or on the other hand, could be thought of as an association rule formed between the audio peak event and the temporal pattern of motion event, through statistical analysis of training data.

Our approach to video mining is to think of it as content adaptive or blind processing. For instance, using temporal association rules over multiple-cue labels could throw up recurring patterns that would help locate the semantic boundaries of the content. Similarly, we could mine the time series stemming from the motion activity values of video frames. Our experience so far indicates that techniques that take advantage of the spatio-temporal properties of the multi-media content are more likely to succeed than methods that treat feature data as if it were generic statistical data. The challenge however is to minimize the content dependence of the techniques by making them as content adaptive as possible. We believe that this is where the challenge of video mining lies.

8. Conclusions

We presented video summarization techniques based on sampling in the cumulative intensity of motion activity space. The key-frame extraction works well with news video and is computationally very simple. It thus provides a baseline technique for summarization. It is best used to summarize distinct semantic units, which motivates us to identify such units by using MPEG-7 generalized sound recognition. We also addressed the related but distinct problem of generation of sports highlights by developing techniques based on the MPEG-7 motion activity descriptor. These techniques make use of domain knowledge to identify characteristic temporal patterns of high and low motion activity along with audio patterns that are typically associated with interesting moments in sports video. We get promising results with low computational complexity. There are a few important avenues for further improvement of our techniques. First, the audio-assisted video browsing can be made more robust and further use made of the semantic information provided by the audio classification. Second, we should develop content-adaptive techniques that adapt to variations in the content, from genre to genre or within a genre. Third, we should investigate incorporation of visual semantics such as the play-break detection proposed in [17]. The main challenge then is to maintain and enhance our ability to rapidly generate summaries of any desired length.

References

[1] Jeannin, S., and A. Divakaran. *MPEG-7 Visual Motion Descriptors,* IEEE Transactions on Circuits and Systems for Video Technology,

Vol 11, No. 6, pp. 720-724, June 2001.

[2] Peker K.A. and A. Divakaran *Automatic Measurement of Intensity of Motion Activity of Video Segments* , Proc. SPIE Conference on Storage and Retrieval for Media Databases, January 2001.

[3] Chang, H.S., S. Sull and S.U. Lee, *Efficient video indexing scheme for content-based retrieval*, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1269-1279, December 1999.

[4] Hanjalic A., and H. Zhang, *An Integrated Scheme for Automated Video Abstraction Based on Unsupervised Cluster-Validity Analysis*, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, December 1999.

[5] Peker K. A., A. Divakaran and H. Sun, *Constant pace skimming and temporal sub-sampling of video using motion activity*, Proc. IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece, October 2001.

[6] Divakaran A., K.A. Peker and R. Radhakrishnan, *Video Summarization with Motion Descriptors*,Journal of Electronic Imaging, October 2001.

[7] Divakaran A., K.A. Peker and R. Radhakrishnan, *Motion Activity-based Extraction of Key-Frames from Video Shots*, Proc. IEEE International Conference on Image Processing (ICIP), Rochester, NY, USA, October 2002.

[8] A. Hanjalic, G. Kakes, R.L. Lagendijk, and J. Biemond, *Dancers: Delft advanced news retrieval system,*" in SPIE Electronic Imaging 2001: Storage and retrieval for Media Databases, San Jose, USA., 2001.

[9] R.S. Jasinschi, N. Dimitrova, T. McGee, L. Agnihotri, J. Zimmerman, and D. Li, *Integrated multimedia processing for topic segmentation and classification,* in ICIP-2001, Thessaloniki, Greece, 2001, pp. 366-369

[10] Divakaran A., R. Radhakrishnan, Z. Xiong and M. Casey *A Procedure for Audio-Assisted Browsing of News Video using Generalized Sound Recognition*, Proc. SPIE Conference on Storage and Retrieval for Media Databases, January 2003.

[11] Peker K.A., R. Cabasson and A. Divakaran *Rapid Generation of Sports Highlights using the MPEG-7 Motion Activity Descriptor,* Proc. SPIE Conference on Storage and Retrieval for Media Databases, January 2002.

[12] Cabasson R. and A. Divakaran *Automatic Extraction of Soccer Video Highlights using a combination of motion and audio features*, Proc. SPIE Conference on Storage and Retrieval for Media Databases, January 2003.

[13] Casey M. *MPEG-7 Sound Recognition Tools,* IEEE Transactions on Circuits and Systems for Video Technology, Vol 11, No. 6, June 2001.

[14] Wang Y., Z. Liu and J-C. Huang, *Multimedia Content Analysis*, IEEE Signal Processing Magazine, November 2000.

[15] Y. Rui, A. Gupta, and A. Acero, *Automatically extracting highlights for TV baseball programs,"* Eighth ACM International Conference on Multimedia, pp. 105–115, 2000.

[16] W. Hsu, *Speech audio project report,* Class Project Report, 2000, www.ee.columbia.edu/∼winston.

[17] L. Xie, S.F. Chang, A. Divakaran, and H. Sun, *Structure analysis of soccer video with hidden markov models,* Proc. Interational Conference on Acoustic, Speech and Signal Processing, (ICASSP-2002), May 2002, Orlando, FL, USA.

[18] P. Xu, L. Xie, S.F. Chang, A. Divakaran, A. Vetro, and H. Sun, *Algorithms and system for segmentation and structure analysis in soccer video,* Proceedings of IEEE Conference on Multimedia and Expo, pp. 928–931, 2001.

[19] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition,* Prentice Hall, 1993.

[20] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T.S. Huang, *Audio Events Detection based Highlights Extraction from Baseball, Golf and Soccer Games in A Unified Framework,* ICASSP 2003, April 6-10, 2003.

Acknowledgments

The authors would like to thank Padma Akella and Pradubkiat Bouklee for carrying out the basic design and implementation of the video browsing demonstration system. We would also like to thank Anthony Vetro for numerous discussions and suggestions. We would like to thank Dr. Huifang Sun for his guidance and encouragement. We would like to thank Dr. Tommy Poon for his enthusiastic support and comments. We would also like to thank Shih-Fu Chang for many helpful discussions and suggestions. We would like to thank Michael Casey for providing his audio classification expertise and software. We would like to thank our colleagues Dr. Tokumichi Murakami, Mr. Takashi Kan and Mr. Kohtaro Asai for their steady support and encouragement over the years. We would like to thank our colleagues Dr. Masaharu Ogawa, Mr. Kazuhiko Nakane, Mr. Isao Otsuka and Mr. Kenji Esumi, for their valuable application oriented comments and suggestions.