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

# Video Mining Using Combinations of Unsupervised and Supervised Learning Techniques

Ajay Divakaran, Koji Miyahara, Kadir A. Peker, Regunathan Radhakrishnan, and Ziyou Xiong

TR-2004-007 March 2004

**Abstract**

We discuss the meaning and significance of the video mining problem, and present our work on some aspects of video mining. A simple definition of video mining is unsupervised discovery of patterns in audio-visual content. Such purely unsupervised discovery is readily applicable to video surveillance as well as to consumer video browsing applications. We interpret video mining as content-adaptive or blind content processing, in which the first stage is content characterization and the second stage is event discovery based on the characterization obtained in stage 1. We discuss the target applications and find that using a purely unsupervised approach is too computationally complex and generally unmanageable to be implemented on our product platform. We then describe various combinations of unsupervised and supervised learning techniques that help discover patterns that are useful to the end-user of the application. We target consumer video browsing applications such as commercial message detection, sports highlights extraction etc. We employ both audio and video features. We find that supervised audio classification combined with unsupervised unusual event discovery enables accurate supervised detection of desired events. Our techniques are computationally simple and robust to common variations in production styles etc.

**Publication History:**

1. First printing, TR-2004-007, March 2004

# Video Mining using Combinations of Unsupervised and Supervised Learning Techniques

Ajay Divakaran, Koji Miyahara, Kadir A. Peker, Regunathan Radhakrishnan, Ziyou Xiong
Mitsubishi Electric Research Laboratories, Cambridge, MA 02139

## ABSTRACT

We discuss the meaning and significance of the video mining problem, and present our work on some aspects of video mining. A simple definition of video mining is unsupervised discovery of patterns in audio-visual content. Such purely unsupervised discovery is readily applicable to video surveillance as well as to consumer video browsing applications. We interpret video mining as content-adaptive or "blind" content processing, in which the first stage is content characterization and the second stage is event discovery based on the characterization obtained in stage 1. We discuss the target applications and find that using a purely unsupervised approach is too computationally complex and generally unmanageable to be implemented on our product platform. We then describe various combinations of unsupervised and supervised learning techniques that help discover patterns that are useful to the end-user of the application. We target consumer video browsing applications such as commercial message detection, sports highlights extraction etc. We employ both audio and video features. We find that supervised audio classification combined with unsupervised unusual event discovery enables accurate supervised detection of desired events. Our techniques are computationally simple and robust to common variations in production styles etc.

**Keywords:** Video Mining, Multimedia Mining, Summarization, Motion Activity, Audio Classification

## 1. INTRODUCTION

Video Mining can be defined as the unsupervised discovery of patterns in audio-visual content. The motivation for such discovery comes from the success of data mining techniques in discovering non-obvious patterns of shopping for example. Furthermore, surveillance video often consists of events that are not known beforehand, and is hence an obvious target for unsupervised discovery of patterns, which in this case are events. For instance, a video sequence captured by a camera trained at a crowded marketplace would defy analysis through simple motion detection. In such a case, we do not necessarily know what is usual and what is unusual, let alone a finer classification. With video mining we would hope to discover the interesting events in the video without a priori knowledge of what those events are.

In the results we have presented in [1], 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. 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.

In this paper, we discuss the video mining problem further and suggest that from an applications point of view, combination of unsupervised and supervised techniques yields the best results. The rest of paper is organized as follows: In section 2, we describe an unsupervised unusual event discovery technique. In section 3 and 4, we discuss the applications of the technique described in section 2 to sports highlights detection and commercial detection respectively. In section 5, we discuss our

results and propose that for most applications, a combination of unsupervised and supervised techniques gives the best results. In section 6, we present our conclusions and possibilities for further research.

## 2. REQUIREMENT OF A VIDEO MINING SYSTEM

Past work in completely unsupervised structure discovery includes that of Xie et al [2], Foote et al [3] and Peker (see [4]). Xie et al use an unsupervised Hierarchical Hidden Markov Model (HHMM) framework to discover patterns in soccer video. They use low-level features such as motion intensity and dominant color at the lower level of the HHMM and binary labels at the upper level of the HHMM. It turns out that the binary structure discovered is none other than the play-break structure i.e. sections where the ball is in play and is not. While this approach is completely unsupervised, it is computationally complex and operates exclusively on baseball and soccer video in which a strong binary play-break pattern is known to exist, and has been shown to be detectable using similar low level features by Xie et al [5]. Foote et al and Peker use a similarity matrix to detect self-similar patterns in content. Such patterns are again known to exist in musical content in the form of rhythmic cadences and hence are discovered as per [2]. Peker, on the other hand, finds that such a similarity matrix approach, using motion activity, is computationally cumbersome when he extends it to look for self-similarity across multiple resolutions in a time series that consists of motion activity values. Furthermore, Peker finds that even in seemingly well-structured video such as golf video in which the video alternates between very low motion activity and bursts of high motion activity caused by the camera tracking the ball, the similarity matrix does not clearly uncover the underlying pattern. Note that the approach of Xie et al has the advantage of the mathematical elegance and flexibility of the HHMM framework, while Foote and Peker, use computationally simpler techniques that are not equally flexible.

In the light of the above and our product platform constraints, we thus set out the following requirements for a video mining technique:

a. It should be unsupervised
b. It should be flexible i.e. should not have any assumptions about the data.
c. It should be computationally simple
d. It should discover interesting events

We see that the aforementioned approaches do not meet all of the above requirements. It is also apparent that there is a potential contradiction between requirement b and requirement d. Xie et al and Foote et al, while using unsupervised techniques, do reveal their knowledge of the domain through their choice of features and strategy for pattern discovery. Furthermore, they work in domains that are strongly structured. On the other hand, Peker's results show that treating the feature sequence as a generic time series without regard to its domain leads to inconclusive or no pattern discovery even when the video is as "clearly structured" as golf video is. Requirement d is in fact a semantic requirement that is unlikely to be met through simplistic computations with low-level features. Furthermore, conventional unsupervised clustering, while satisfyingly non-parametric, has the drawback of ignoring temporal relationships. The success of HMM-based methods for event modeling shows that any method that overlooks temporal relationships ignores an essential aspect of audio-visual content. At the same time, if we consider content such as surveillance content, since the events are diverse and not known a priori, using event models such as HMM's is difficult if at all feasible.

The above discussion motivates us to pursue solutions that attempt to address the above contradictions by relaxing requirements (a) and (b) somewhat so as to finally satisfy requirement d, i.e. find interesting patterns. So we restate the requirements for the video mining technique as follows:

1. It should be as unsupervised as possible.
2. It should have as few assumptions about the data as possible.
3. It should be computationally simple
4. It should discover interesting events.

Let us try to define "interesting events." The definition of interesting is of course highly context dependent. For instance, in a sports video, highlights such as home runs, goals, three-point shots are interesting while the general run of play is not so interesting. In a TV broadcast, the program is interesting but the commercial messages are not. So in the former example, the rare events were interesting while in the latter example, it was the common events that were interesting. Yet another category of interesting events is transition points in the content that mark the semantic boundaries of the content. For example, through principal cast detection, we can discover the topic boundaries in a news video. Note that the transition points are usually not so common, and thus locating the transition points is also essentially in the category of finding rare events. Furthermore,

identifying the rare/unusual events in the video can also be viewed as a part of summarization of the video. Depending on the video, a video summary is nothing but some combination of the common and the uncommon events. For instance, in a soccer video, it is the uncommon events that constitute the summary since the common events are not interesting. On the other hand, in a surveillance video, we would like to know both the common and the uncommon events to summarize the video since we may not know them upfront.

The success and significance of video mining depends on the content genre. Carefully scripted and staged video such as drama, news etc. has extremely strong structure but has tremendous intra-genre variation in production styles that vary from country to country or content-creator to content-creator. Sports content falls somewhere in the middle since it is not scripted but is constrained by rules. Surveillance content is not scripted and often not constrained by any rules other than those imposed by the geometry of the setting and of course the laws of gravity. Each genre therefore poses unique challenges to pattern discovery.

Note that despite the different considerations with each content genre, there is a common theme of discovering unusual events. We are thus motivated to develop a simple technique that discovers uncommon or unusual events, and thus enables video mining across a wide variety of content. Furthermore, we use higher-level features obtained through audio classification with the hope that the higher level of the features would compensate for the simplistic nature of the techniques. Note that this approach meets requirement 2 since it applies a weak condition that is satisfied by a wide variety of content, viz. there are a few classes of audio that are common to and predominant in content that falls in the aforementioned variety.

### 2.1 Generation of Semantic Audio Labels

Training examples of each of the pre-defined semantic audio classes were collected. Mel Frequency Cepstral Coefficients (MFCC) are extracted for each of the training clips. A Gaussian Mixture Model (GMM) is used to approximate the distribution of MFCC features of each sound class. Given a test clip, MFCC features are extracted and the likelihood of observing these features under each sound class's GMM is calculated. The test clip is classified as belonging to the sound class, for which the likelihood score is maximum. Figure 1 illustrates the framework for generation of semantic audio labels. See [3] for a detailed description.

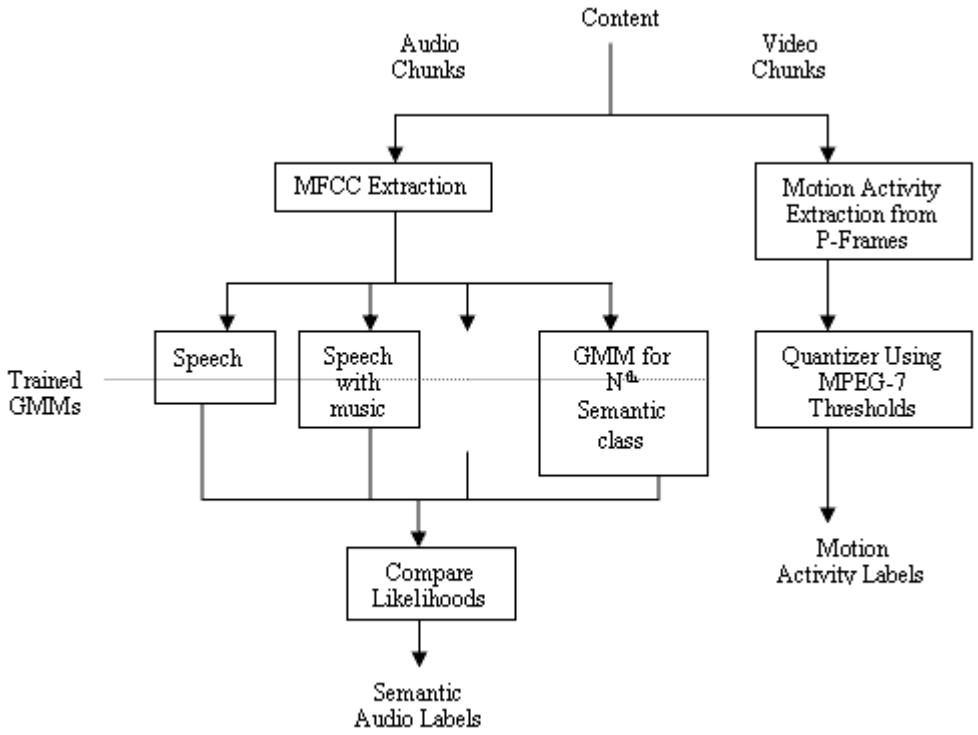



**Figure 1:** Framework for generation of semantic audio & motion labels.

### 2.2 Generation of Motion Activity Labels

The MPEG-7 motion activity descriptor captures the intuitive notion of 'intensity of action' or 'pace of action' in a video segment and can be extracted from motion vectors of each P-frame. The extracted activity values are quantized according to MPEG-7 thresholds to obtain labels representing segments with very low activity to segments with very high activity. See [3] for a more detailed description of the extraction process. Both audio and motion labels are synchronized.

## 3. DETECTION OF UNUSUAL EVENTS USING LOCAL DEPARTURES FROM GLOBAL STATISTICS

An unusual event marks a departure from the usual. By definition, the unusual is much less frequent than the usual is. We can therefore detect the unusual in two steps. First, we should characterize the usual. Second, we should look for departures from the usual. The first step is the key since it has to work even though the location of the unknown events is not known a priori. Since by definition, the unusual event is not frequent, the usual event can be characterized by merely collecting "global" statistics over the whole content. If the unusual event is truly unusual, it will have negligible influence on the global statistics. Hence, the global statistics should be a good approximation of the statistics of the usual event. The more infrequent the unusual event is, the better such a characterization of the usual event will be. The second step would then be to compare local statistics over a small time interval with the global statistics to detect departures from the usual. The unusual events would then reveal themselves as departing significantly from the usual. We use the distribution of semantic audio labels and video labels as our global statistic, since such class composition histograms provide a non-parametric robust characterization.

Once we have obtained semantic labels for both audio and motion from audio & video streams, the problem of "unusual" event detection can be posed as a multimedia-mining problem across these labels for patterns. Figure 2 illustrates the proposed framework for mining audio and motion activity labels. The basic assumption in this approach is that interesting events are "rare" and have different audio and motion characteristics in time, when compared to the "usual" characteristics in a given context. A context is defined by a time window of length $W_L$. In order to quantify what is considered as "usual" in a context, we compute the distribution of labels within it. Then, for every smaller time window $W_S$, we compute the same distribution. We compare the local statistic (computed within $W_S$) with the global statistic (computed within $W_L$) using either an information theoretic measure called relative entropy or a histogram distance metric proposed in [4]. One would expect a large distance value for a $W_S$ with a different distribution compared to what is "usual" within $W_L$. By moving $W_L$ one $W_S$ at a time, we compute ($W_L$/ $W_S$) distance values and find the maximum of this set of values. We associate this maximum value, $M_{ws,}$ with the small window $W_S$. Then, rare events are times when there is a local maximum in the curve of all $M_{ws}$.

For instance, in a news program the onset of commercials would cause the distribution of semantic audio labels to peak around music and speech-with-music whereas the global distribution in the current context, would peak around speech. Therefore, comparison of the local and global distribution of labels would signal the occurrence of something "unusual" in that context. Points of unusual events detected by the above algorithm only signal a change in characteristics in terms of the semantic audio and/or motion labels. For instance, an anchor person shot followed by a segment of outdoor shot of high activity in a news program would also be signaled as an unusual event even though there was no change in terms of audio characteristics.
Note that unlike past approaches to commercial detection, the proposed algorithm to detect unusual events is content adaptive and does not assume any general rule such as the occurrence of mono-chrome frames, appearance of text etc. The algorithm defines what is usual within a context and then measures deviations from the usual in an adaptive way. Also, the proposed scheme works from features extracted in the compressed domain, which makes it a good candidate approach to be incorporated in consumer electronic devices.

However, note that this approach fails to meet the requirement 4 mentioned earlier in section 2. We will get a different set of unusual events for each distinct feature set we use to compute the statistics. There is no guarantee that these unusual events are in fact interesting. We find therefore that the generality and computational ease of the technique has been achieved at the expense of getting events that are guaranteed to be interesting. However, we can hope that with the right feature choice, our technique would at the very least provide a list of candidate events. We could then use a subsequent stage that uses a priori knowledge about the desired events to detect the desired events in the list from the previous stage. Note that thus, to finally uncover interesting patterns, we are forced to resort to a supervised or a priori rule-based approach at the final stage. In the following two sections, we describe applications of such an approach to commercial message detection and sports highlights extraction.

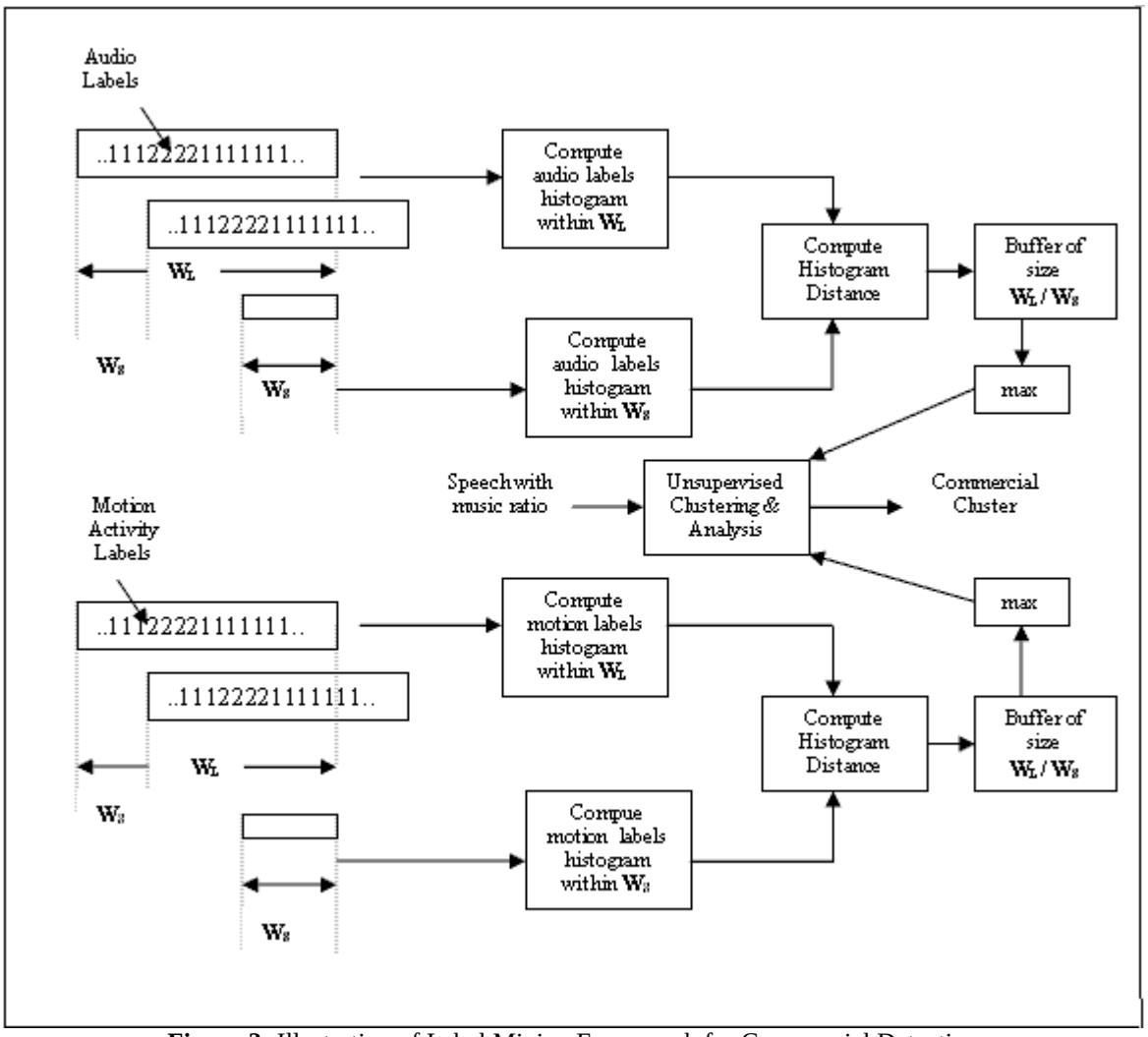



**Figure 2:** Illustration of Label Mining Framework for Commercial Detection

## 4. COMMERCIAL MESSAGE DETECTION

The input content was de-multiplexed into audio and video. The audio stream was divided into chunks of size 0.5s. MFCC features extracted every 30ms with 10ms overlap were tested with pre-trained GMMs with 10 mixture components to classify each 0.5s clip into one of the following classes: speech, music, speech-with-music, audience noise (cheering,applause). The audience noise class was selected to deal with commercial detection in sports video as well. If not for the inclusion of this class, all audience noise would be classified as music or speech-with-music class due to the wideband nature of audience noise class.

From the video stream corresponding to every 1.3s, MPEG-7 intensity of motion activity was extracted and quantized to five levels of perceived activity ranging from very low to very high activity. Since audio labels were extracted for a smaller time duration (0.5s), the motion labels were repeated for atleast two audio chunks to achieve synchronization.

After synchronizing audio and motion labels, label mining was performed on these two label streams using a large window size of 10min ($W_L$ = 10min) and a small window size of 1min ($W_S$ = 1min). Label distribution was computed as a statistic to compare characteristics within $W_L$ with characteristics within $W_S$. The distance metric used was a histogram comparison method proposed in [4].

The proposed algorithm for commercial detection was tested with content from diverse content taken from Japanese and American TV programs. Figure 3 shows the performance of the proposed scheme against ground truth for all these content. Note that we have high accuracy for most of the programs except for DVD2. The false alarms were due to the presence of commercial-like segments in the program itself i.e. segments with high motion and music in the background. See [] for further details on the algorithm and results.

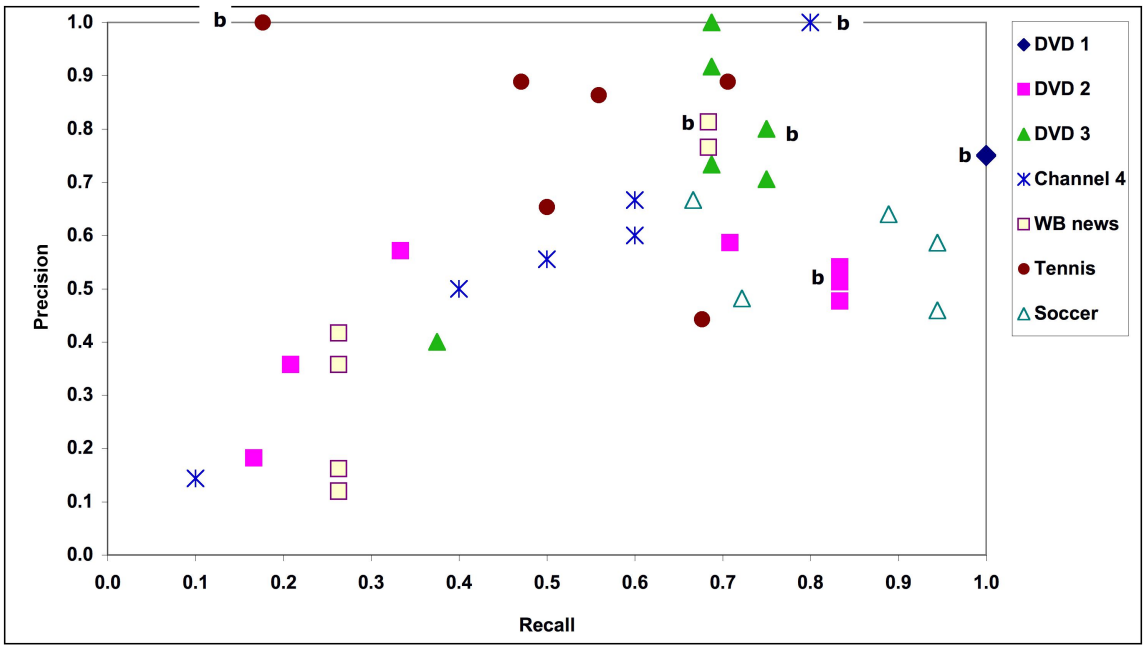



**Figure 3:** Performance of proposed scheme for commercial detection with content from news and sports from the U.S. and Japan. DVD 1 is news content, DVD 2 is a talk show and DVD 3 is an animated feature film.

## 5. EXTRACTION OF SPORTS HIGHLIGHTS

### 5.1 Highlights Validation by Trained HMM

Points of unusual events detected by the above algorithm only signal a change in characteristics in terms of the semantic audio labels. For instance, in a golf game, the onset of commercials would also be signaled as an unusual event. Therefore, in order to capture only semantic events of interest, we use a trained HMM of highlights to validate the candidate "unusual" events signaled by the previous step. This step is analogous to human participation in data mining to validate the unusual patterns output by the data-mining algorithm.

Another advantage of this scheme is that we can use the likelihood values output from the HMM to rank the highlights. This avoids the simple heuristic in [5] that uses the duration of applause/cheering segments for ranking and summary length modulation. Figure 3 illustrates the combination of unsupervised and supervised learning approaches for highlight extraction.

## 5.2 Experimental Results

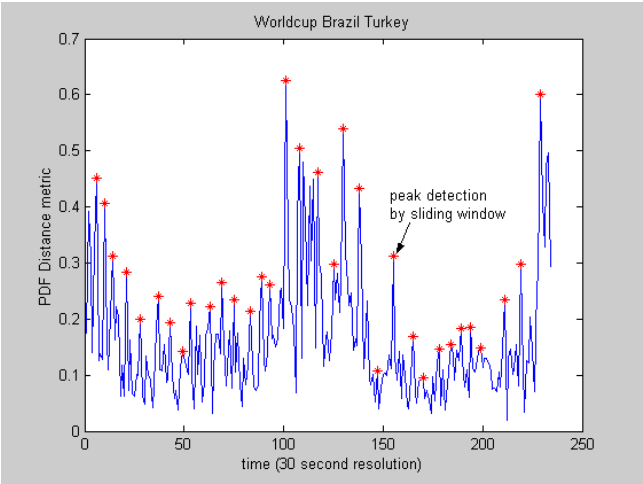
Training data was collected for each of the low-level sound classes from three and a half hours of MPEG video for three different sports namely baseball, soccer and golf. Twelve dimensional MFCC features extracted from every 30ms frame, were used to train a 10 component GMM for each sound class.

Once models are trained, input audio is divided into chunks of 1s segments. MFCC features are extracted from each frame in the 1s segment. Each frame is then classified as belonging to one of the sound classes using the maximum likelihood criterion. We use a majority voting scheme from all the frames to decide the sound class of the 1s segment.

After assigning audio labels to 1s segments, unsupervised label mining was performed to detect unusual events as shown in Figure 1. The value of large window ($W_L$) was chosen to be 4 minutes and the value of was small window ($W_S$) chosen to be 0.5 minutes. The Kullback-Leibler distance metric was used to compare the distribution of audio labels within these chosen windows. The sliding window peak detection algorithm was used on the recorded distance values ($Mw_s$) to detect peaks. Figure 3 shows the peak detection results for two soccer games each of duration two hours.

After identifying candidate "interesting" events from the unsupervised label mining step, we use a Hidden Markov Model (HMM) to validate and rank the "interesting" events. The highlight HMM was trained from the labels extracted from DVD data that contains half an hour of professionally edited highlights of the 2002 soccer world cup. There were 81 "interesting" clips in this training data including mainly attempts at scoring a goal and successfully scored goals. The number of states for the HMM was empirically chosen to be 4. After training, it was observed that one of the states corresponded to the Cheering label and the corresponding state had a high self-transition probability. This implies that the HMM was indeed modeling the occurrence of contiguous cheering segments. This kind of modeling is better than simply using the length of duration of cheering segments as in [5]. To illustrate this we include the corresponding precision-recall figures for the same soccer games in Table 2. The threshold based scheme in [5] to detect highlights, is based on a fixed notion of highlights and is less flexible. Here, we let the HMM learn from the label sequences of training data what a highlight is. Such a data driven approach does not have a fixed notion of highlights and may capture other audio cues apart from cheering/applause. It also opens the possibility of information fusion with other modalities.

Table 1 shows the performance in the two soccer games.



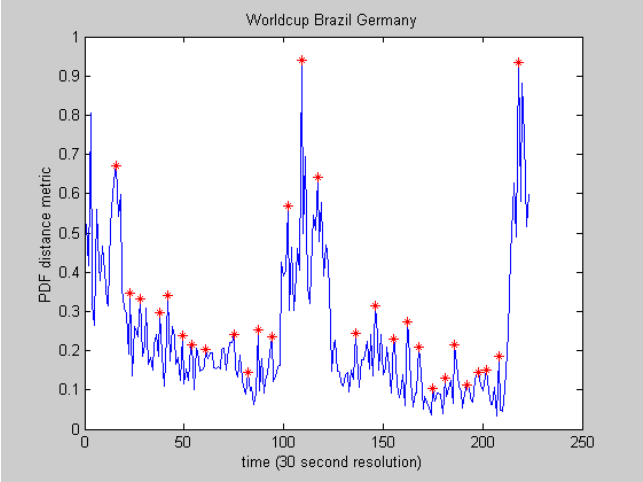



**Figure 3:** Unsupervised label mining results on soccer games.

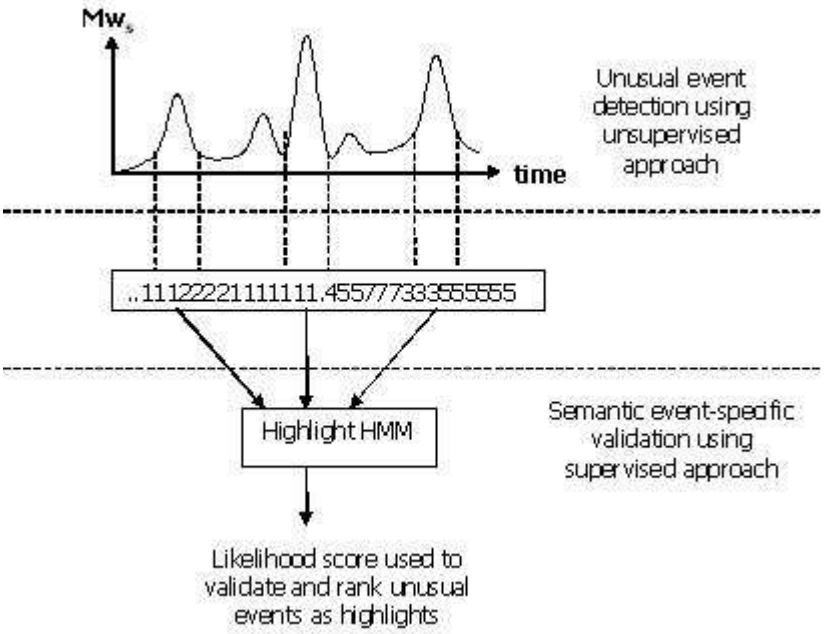



**Figure 4:** Combination of Unsupervised and Supervised learning for highlight extraction.

| | [A] | [B] | [C] | Precision | Recall |
|---|---|---|---|---|---|
| Game 1 | 34 | 22 | 32 | 40.74% | 64.71% |
| Game1 | 34 | 6 | 8 | 42.86% | 17.65% |
| Game 1 | 34 | 16 | 24 | 40% | 47.06% |
| Game 1 | 34 | 11 | 20 | 35.48% | 32.35% |
| Game 2 | 45 | 25 | 35 | 41.67% | 55.55% |
| Game 2 | 45 | 5 | 0 | 100% | 11.11% |
| Game 2 | 45 | 20 | 14 | 58.82% | 44.44% |
| Game 2 | 45 | 9 | 4 | 69.23% | 20% |

**Table 1:** Soccer Game 1: World cup Brazil Germany; Soccer Game 2: World cup Brazil Turkey; [A]: Number of true highlight segments; [B]: Number of true highlight segments output by the algorithm; [C]: Number of False Alarms

| | [A] | [B] | [C] | Precision | Recall |
|---|---|---|---|---|---|
| Game 1 | 34 | 22 | 59 | 27% | 64.71% |
| Game 1 | 34 | 6 | 11 | 35% | 17.65% |
| Game 1 | 34 | 16 | 39 | 29% | 47.06% |
| Game 1 | 34 | 11 | 30 | 27% | 32.35% |
| Game 2 | 45 | 25 | 44 | 36% | 55.55% |
| Game 2 | 45 | 5 | 10 | 32.5% | 11.11% |
| Game 2 | 45 | 20 | 30 | 40% | 44.44% |
| Game 2 | 45 | 10 | 14 | 42% | 20% |

**Table 2:** Results for the same two games using the approach in [4].

In order to compare the current approach with the threshold based scheme in [4], we compare the precision values for the same recall values in two soccer games. For the current approach, different points on the precision-recall curve were generated by changing the likelihood threshold of the highlight HMM.

Note that the proposed approach outperforms the simple threshold based scheme in [4] for the recall values. The number of false alarms has been reduced by the inclusion of the peak detection and a validation step in place of threshold-based highlight detection. Please see [7] for a detailed explanation of the technique described here.

## 6. CONCLUSIONS & FUTURE WORK

We presented a new approach to commercial detection and sports highlight based on mining of semantic audio and motion activity labels, combined with a subsequent supervised classifier. The approach makes use of the fact that unusual events tend to be different from the program in a given context. A global distribution of labels within the chosen context is computed to quantify what is "usual" and deviations from the usual are computed within a smaller window. This makes the proposed approach to detect unusual events, content adaptive. After detecting points of unusual events we look for high percentage of speech-with-music or music audio chunks to single out commercials. This rule might not work for content with already high percentage of music (Music Video) and would cause the accuracy to suffer. With our current approach it is also possible to predict this by looking at the global distribution of semantic audio labels. Our future work is to adaptively determine which feature would be a best discriminator based on global label distribution which characterizes the content. Our results also show that a model of sports highlights works well after we have obtained a list of candidates using the unusual event discovery technique.

Our experience therefore shows that for practical applications, we cannot afford to be completely unsupervised. Note that in both of our example applications, we used supervised classification to classify the audio, and then used an unsupervised method to get the candidates and then used a supervised method to finally get the desired interesting events. A purely unsupervised approach that gets the interesting events is probably not feasible except when the feature choice has been made with a certain event in mind. An alternative in a practical system could be to be completely unsupervised in the unusual event detection but to involve a human being in pruning the list of candidates. We are currently examining such an approach to surveillance video. Thus generality and flexibility are always obtained at the expense of the semantic "interest" of the event. We need to choose the appropriate mix of supervised and unsupervised depending on the application. In the two applications we showed, we showed the advantage of having a common base technique that feeds two different detectors. In ongoing work we are finding that even unsupervised clustering with the results from the global vs local departure detectors works remarkably well for both sports highlights and commercial detection (See [6]).

In short, our future challenge is to reconcile the requirements of section 2 so as to devise techniques that maintain the common base unusual event detector and then go on to detect various interesting events.

## 6.REFERENCES

1. A. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong and R. Cabasson, “Video Summarization using MPEG-7 Motion Activity and Audio Descriptors,” Video Mining, eds. A. Rosenfeld, D. Doermann and D. DeMenthon, Kluwer Academic Publishers, 2003.
2. L. Xie, S-F. Chang, A. Divakaran and H. Sun, “Unsupervised Mining of Statistical Temporal Structures in Video ,” Video Mining, eds. A. Rosenfeld, D. Doermann and D. DeMenthon, Kluwer Academic Publishers, 2003.
3. J. Foote, M. Cooper, and Unjung Nam, "Audio Retrieval by Rhythmic Similarity," in *Proc. Third International Symposium on Musical Information Retrieval (ISMIR),* September 2002, Paris.
4. http://dimacs.rutgers.edu/Workshops/Video/Slides/ajay.ppt
5. Z. Xiong, R. Radhakrishnan, A. Divakaran, “Generation of Sports Highlights Using Motion Activity in Combination with a Common Audio Feature Classification Framework” to appear in International Conference on Image Processing, Barcelona, Spain 2003.
6. K-S. Goh, K. Miyahara, R. Radhakrishnan, Z. Xiong and A. Divakaran, “Commercial Detection based on Mining of Audio-Visual Labels,” SPIE Storage and Retrieval from Media Databases, San Jose, CA, 2004.
7. R. Radhakrishnan, Z. Xiong, A. Divakaran and Y. Ishikawa, “Generation of Sports Highlights using a Combination of Supervised and Unsupervised techniques in the Audio Domain,” IEEE PCM, Singapore, 2003.