VIDEO MINING: PATTERN DISCOVERY VERSUS PATTERN RECOGNITION

Ajay Divakaran, Kadir A. Peker, Shih-Fu Chang, Regunathan Radhakrishnan, Lexing Xie

Mitsubishi Electric Research Laboratories, USA
Email: {ajayd,peker,regu}@merl.com
Columbia University, USA
Email: {sfchang,xlx}@ee.columbia.edu)

ABSTRACT

We examine the significance of video mining as pattern discovery in multimedia content. We examine the underlying issue of pattern discovery versus pattern recognition, since most past work has not drawn such a sharp distinction. We argue that while the term "pattern discovery" implies a purely unsupervised approach, in practice a mixture of unsupervised and supervised techniques will have to be used. We compare conventional data mining with video mining and observe that a key difference is in the multi-layered semantics of multimedia content. We then identify significant challenges posed by video mining.

1. INTRODUCTION

Video mining can be defined as the unsupervised discovery of patterns in audio-visual content. Past work on video mining, such as that covered in the book edited by Rosenfeld et al [1], has not necessarily emphasized unsupervised pattern discovery. In our contribution [2] to [1] for instance, we have presented results on principal cast detection in news video, as well as sports highlights detection. These are based on detection of known patterns in audio-visual content such as speaker-changes or high motion followed by audience reaction etc. In other words, most past techniques have relied on detection of known patterns in specific content genres. We have recently begun to emphasize pattern discovery and have described a technique for pattern discovery in soccer videos through Hierarchical Hidden Markov Models [3] as well as an unsupervised pattern discovery technique for sports video in [4].

Since so little work has been done on pattern discovery, we need to explain why it is necessary. It is obviously needed when the events we are looking for in a segment of video content are not known. This is often the case in surveillance video in which most of the video consists of long stretches of repetitive or "uninteresting" parts occasionally interrupted by unusual or "interesting" parts, which are too diverse to be anticipated in advance. Even in better -understood genres of video such as news video, there is tremendous variation across content producers. For example, the beginning of a news story is presented in many different styles such as news-anchor with graphic followed by story, graphic followed by news anchor, related video followed by news anchor etc. Therefore, even looking for an event that is known can become unmanageable if it requires an exhaustive search for its possible variations. We therefore require content-adaptive pattern discovery techniques that would adapt to variations in content, and make the event-search tractable. It is hence our view that video mining is useful for all kinds of genres that span the gamut from highly produced, such as news video, to spontaneous but constrained, such as sports video, and further to completely spontaneous, such as surveillance video.

In this paper, we first establish the requirements for a video mining system based on the above discussion. Second, we present two instances of pattern discovery in video based on our work. Third, we compare the two approaches and also compare video mining with data mining in the light of our work. Fourth, we identify significant challenges for further research.

2. REQUIREMENTS OF A VIDEO MINING SYSTEM

The discussion in section 1 leads us to formulate the following requirements for a video mining system:

1. It should be unsupervised.
2. It should not have any assumptions about the data
3. It should uncover interesting events.

Note that requirements 2 and 3 are somewhat contradictory, since the notion of "interesting" is subjective, and highly dependent on knowledge of the content. We therefore relax requirement 2 by aiming for having as few assumptions as possible.

The range between purely unsupervised and purely supervised techniques can be thought of as a continuum that goes from the general to the particular. Our aim is to find out how few assumptions we can make about the

0-7803-8554-3/04/$20.00 ©2004 IEEE.

content without detecting events that are too general to be meaningful. This would help us understand the content-specific heuristics reported in most previous work, and help set up a framework for systematic use of domain knowledge. Note that sports video is in our view a good genre to start with because it has some constraints but is spontaneous as well. It thus provides a tractable test-bed for pattern discovery techniques.

3. HIERARCHICAL HIDDEN MARKOV MODEL-BASED STRUCTURE DISCOVERY IN SOCCER VIDEO

Hidden Markov Models (HMM) are known to successfully model temporally correlated signals such as speech phonemes. In our previous work [4] we showed that we could successfully model play-break patterns in soccer video with supervised HMM's that use low-level motion and color features. Our results motivate us to investigate whether we can discover patterns in soccer video using low-level feature based HMM's in a purely unsupervised fashion.

First, we set up the unsupervised training framework by setting up a Hierarchical HMM structure illustrated in Figure 1.

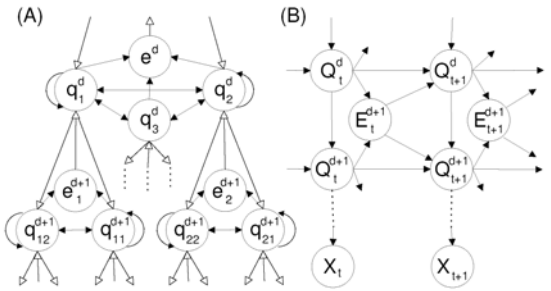


Figure 1: Hierarchical HMM's

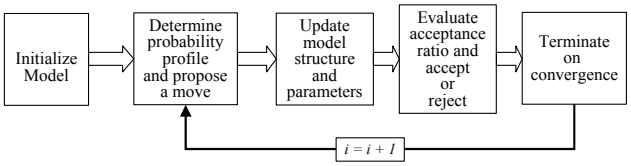
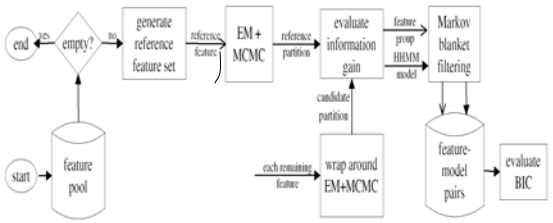


Figure 2: Iterative Unsupervised training of HHMM's

Our results in [4] motivate us to use two levels of hierarchy, and two clusters. Figure 2 illustrates our use of the EM and Monte Carlo sampling techniques to learn the model structures and parameters from the content. Note that we iterate until the optimal trade-off between model complexity and accuracy of it has been achieved through the Bayesian information criterion. We find that when we use the same features as we did in [4], we get two clusters that in fact correspond to the play-break category with

Figure 3: Overview of feature selection algorithm



slightly better accuracy than the approximately 75% accuracy we get with the supervised technique reported in [4], with the Korean soccer content. While this is a surprising result at first glance, it is in fact due to the careful choice of features, which bias the structure discovery towards uncovering play-break structures.

This motivates us to investigate automatic feature choice, which we illustrate in Figure 3. We find that given a feature pool consisting of features from [4] as well as other features such as camera motion parameters, audio features etc., the optimal two-cluster partition still corresponds to the play-break segmentation, with the features being satisfyingly close to the set from [4]. Note that we have presented a highly condensed description of the techniques because of space considerations. For details, please see [2].

Note that the proposed technique is in fact applicable to a wide variety of content since it models first order temporal structures. In further work, we propose to apply it to genres other than soccer video. Note that the technique is best suited to uncover patterns of strong temporal correlation. It will not uncover patterns that do not rely on first order correlation. Moreover, it is computationally complex.

4. COMBINATION OF UNSUPERVISED AND SUPERVISED LEARNING TO EXTRACT SPORTS HIGHLIGHTS AND OTHER EVENTS

In [3] we described a sports-highlights extraction framework based on detection of contiguous stretches of applause/cheering through audio classification. Our results motivate us to discover patterns in the classifier-generated audio labels of a video stream. Furthermore, we are interested in the case in which the video stream consists of long stretches of a typical or usual event, interspersed with rare occurrences of atypical or unusual events. This is commonly the case with surveillance video in which the problem is further complicated by the diversity of the unusual events, which defy description by a common model. We are also interested in computationally simple techniques.

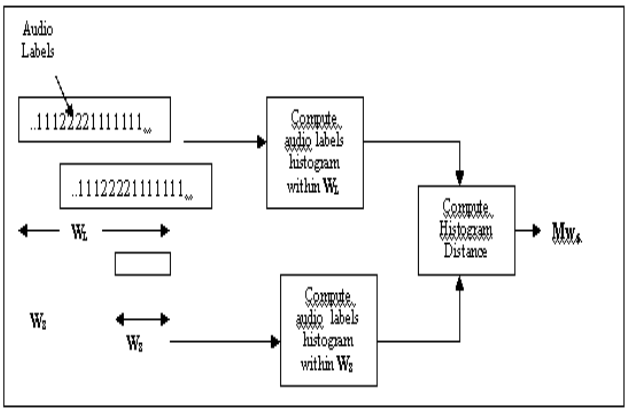


Figure 4: Unsupervised Label Mining Framework.

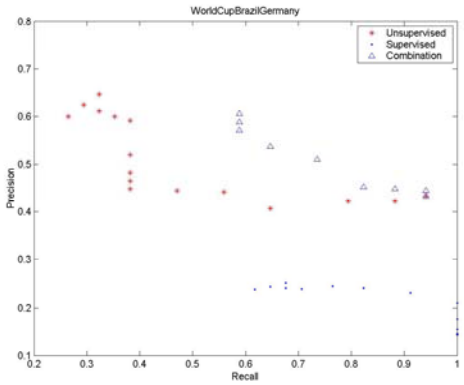


Figure 5: Precision-recall for the combination of unsupervised and supervised techniques.

In Figure 4, we illustrate our unsupervised label-mining framework. We model the stream as a sequence of events mostly from the usual class C_1 occasionally interspersed with the unusual class C_2. Since we do not know the nature of either of the classes, we are attempting to discover if such a binary pattern of usual-unusual exists. We have to proceed in the following steps:

1. Find a statistical model for the usual class C_1. We are helped by the dominance of this class, since the statistics of the entire stream should be close to those of C_1. Hence we do not need to know where the unusual class is located a priori to get an estimate of the statistics.
2. Test each segment to find out if it belongs to C_1. If it does not, we assign it to C_2.

In this paper, we use the audio-class composition histogram to model the statistics, and for each segment compare its statistics with that of its surrounding context i.e. a large window around it, which is not necessarily as long as the entire content since too long a context may add extraneous information. In forthcoming work, we will describe a systematic method for choosing the window sizes, but in this paper we chose them empirically.

Such an approach does in fact discover unusual events in a soccer video for example. We find that as expected, the unusual events uncovered do not belong to any single class. In the case of soccer video, it turns out that they belong to two large classes, highlights and commercial messages. To further classify the unusual events, our technique has to be supplemented by a subsequent stage that can detect known event categories in the unusual events. We illustrate our combination in Figure 6. We find that we are able to train soccer highlight HMM's with professionally created sports highlights video. Figure 5
shows that the unsupervised technique actually does better than the supervised technique, but the combination vastly improves the accuracy of the unsupervised technique. We
think that the reason is that the supervised technique places too strong a structure assumption on the highlights, while the histogram based technique collects gross
statistics that do not apply the same kind of assumption. Once again, this is a highly condensed description of the technique, and for details please see [5]. We have also been able to detect commercial messages using the same technique.

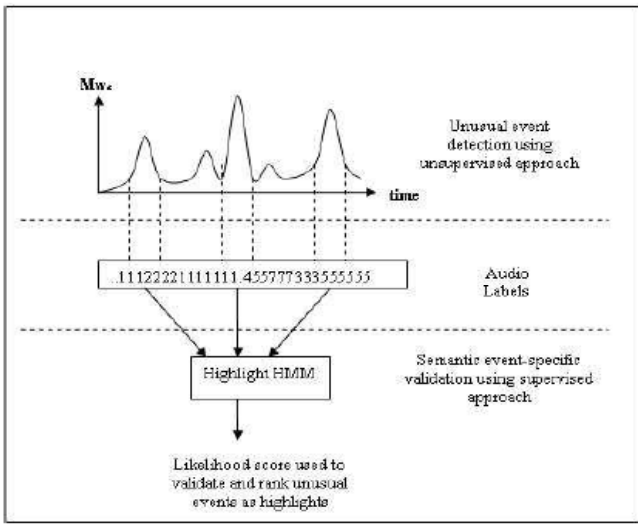


Figure 6: Combination of Unsupervised and Supervised approach for highlights extraction

5. DISCUSSION

Let us compare the two techniques against the requirements we described in section 2. The HHMM

based technique is completely unsupervised and assumes the presence of strong temporal correlation. The technique of section 4 assumes the presence of a dominant class but then uses simple statistical models that are non-parametric and do not assume a strong temporal correlation. It is also predicated on supervised audio classification at the moment. However, the basic methodology is applicable to lower level features as well, which would then make the comparison more reasonable. Note that both methods cluster the video stream into events without specifying the semantics a priori. Whether the events are interesting or not has to be determined subjectively. However, note that in both cases, the feature choice ensures that a certain kind of interesting event is picked. It suggests that we should try the method of section 4 with a feature selection methodology that parallels that proposed in section 3, with a very large feature set. We also need to substantially enlarge the feature set used in section 3 to find out what events will then be discovered in the soccer video.

Another way to compare our two techniques is to recognize that both are carrying out unsupervised binary classification into play-break in one case and highlight-non-highlight in the other case. We find that even supervised HMM's do not classify highlights well, and so it is likely that a HHMM will not capture highlights either with audio labels. On the other hand, the evidence from [2] would indicate that play-break structures require a strong temporal correlation model and would not therefore be detected by the method of section 4. Furthermore, the unusual-usual model is not applicable in this case, since plays and breaks are of similar frequency. In short, the class of patterns discovered is determined by the essential nature of the statistical models.

That again raises the third criterion of whether the discovered events are interesting. That brings up what we believe is an essential difference between conventional data mining and video mining. Since multimedia data has many layers of semantics, patterns can be discovered at many layers as well. This makes multimedia data a treasure trove of patterns, and hence a much bigger variety of tools have to be developed to uncover all possible patterns. Our results with combining supervised and unsupervised learning indicates that for a given task, a common stage of generic processing can be developed that uncovers unusual events that can be further identified in a domain specific manner. This common stage can consist of a limited subset of possible techniques based on the task at hand.

6. CONCLUSION AND FUTURE WORK

We presented and compared two approaches to pattern discovery in video content. We found that they both satisfy the basic requirements of pattern discovery and find complementary patterns. Our results suggest the following avenues for further research:

1. Making the subjective judgment on the significance of the pattern uncovered manageable. Current techniques quickly overwhelm the human judge with too many events.
2. Exploration of various statistical models to capture diverse aspects of content semantics, especially those that take the multi-layered nature of content into account. Investigation of a wide variety of genres and feature choice from a wide pool of features.
3. Interpretation of pattern discovery results, so as to systematically evaluate the utility of domain knowledge. A common processing stage for various event detection techniques could then be developed.

7. REFERENCES

[1] Video Mining, eds. A. Rosenfeld, D. Doermann and D. DeMenthon, Kluwer Academic Publishers, 2003.

[2] A. Divakaran, K. A. Peker, R. Radhakrishnan, Z. Xiong, R. Cabasson, "Video Sumarization using MPEG-7 Motion Activity and Audio Features ," Video Mining, eds. A. Rosenfeld, D. DoDoermann and D. DeMenthon, Kluwer Academic Publishers, 2003.

[3] 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.

[4] L. Xie, S. Chang, A. Divakaran, and H. Sun, "Structure Analysis of Soccer Video with Hidden Markov Models," *Proc. ICASSP, IEEE Int'l Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, FL, May 2002.

[5] R.Regunathan, Z. Xiong, A. Divakaran, Y. Ishikawa, "Generation of sports highlights using a combination of supervised and unsupervised learning in the audio domain," ICICS-PCM Conference, Singapore 2003.

[6] A. Divakaran, K. Miyahara, K. A. Peker, R. Radhakrishnan, Ziyou Xiong, "Video Mining using combinations of unsupervised and supervised learning techniques," Special Session on Video Mining, SPIE Electronic Imaging Conference on Storage and Retrieval for Media Databases, San Jose, CA, January 2004.