Temporal Feature Induction for Baseball Highlight Classification

Michael Fleischman
MIT Media Laboratory
mbf@mit.edu

Brandon Roy
MIT Media Laboratory
bcroy@media.mit.edu

Deb Roy
MIT Media Laboratory
dkroy@media.mit.edu

ABSTRACT
Most approaches to highlight classification in the sports domain exploit only limited temporal information. This paper presents a method, called temporal feature induction, which automatically mines complex temporal information from raw video for use in highlight classification. The method exploits techniques from temporal data mining to discover a codebook of temporal patterns that encode long distance dependencies and duration information. Preliminary experiments show that using such induced temporal features significantly improves performance of a baseball highlight classification system.

Categories and Subject Descriptors
I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding – *Video Analysis*

General Terms
Algorithms, Experimentation

Keywords
Temporal data mining, discriminative models, sports video, highlight classification, baseball.

1. INTRODUCTION
The desire to search and summarize sports video has sparked a great deal of research focusing on automatic classification of sports highlights. A number of these approaches have used dynamic models, such as Hidden Markov Models, to exploit limited temporal information between low level features in the audio/video stream (e.g., [12]). Recent work in event recognition, however, suggests that accurately classifying many types of video events requires modeling more complex temporal structure than can be encoded in simple dynamic models [9]. In this paper we present a method for *temporal feature induction*, in which complex temporal information is automatically mined from unlabeled video data and used to improve the performance of baseball highlight classification.

Feature induction is a technique used in many applications of machine learning in which low level features are automatically combined to create more complex feature spaces that facilitate classification [5]. We introduce *temporal feature induction* as an analogous methodology in which complex temporal features are automatically mined from time-series data by examining the temporal relations that exist among low-level features of the data.

Incorporating such complex temporal features enable classifiers to exploit high level information that is not easily captured in typical dynamic models, such as the long-distance dependencies that exist between low level features. Further, automatically mining such complex temporal features alleviates the need to hand design complicated dynamic models required by other approaches to event classification [10].

In the following sections, we describe a method for temporal feature induction and evaluate it using a discriminative model of video event classification similar to Gong et al. [8]. We evaluate the approach on a small pilot data set of highlights from broadcast baseball games and demonstrate that the methodology affords a statistically significant improvement in classification performance.

2. RELATED WORK
Sports highlight classification in general, and baseball classification in particular, has received a great deal of attention in the multimodal community [2]. A number of approaches have exploited dynamic probabilistic models such as HMMs that capture patterns in the sequences of low level features from video and/or audio data streams (e.g., [12])

Gong et al. [8] show that using discriminative classification methods (e.g. maximum entropy models) allow for better integration of multimodal features and outperform HMMs in the baseball domain. In this framework, raw video is first segmented into shots (four shots per highlight), and then, low-level audio and video features are extracted from each shot. The multimodal features from each shot are then concatenated into one feature vector, which is then used to train the classifier.

Like traditional HMMs, Gong et al.'s [8] approach captures limited temporal information because each training instance explicitly encodes the sequence of low level features. Further, their approach allows for easy integration of asynchronous features such as color histograms and closed caption keywords. For these reasons, we evaluate the effectiveness of temporal feature induction using this discriminative framework.

3. TEMPORAL FEATURE INDUCTION
Temporal feature induction is a method to automatically expand a set of low level features, such as those described in Gong et al. [8], to improve the performance of a video highlight classifier. Temporal feature induction operates in two phases. In the first phase, low level features are abstracted into multiple streams of

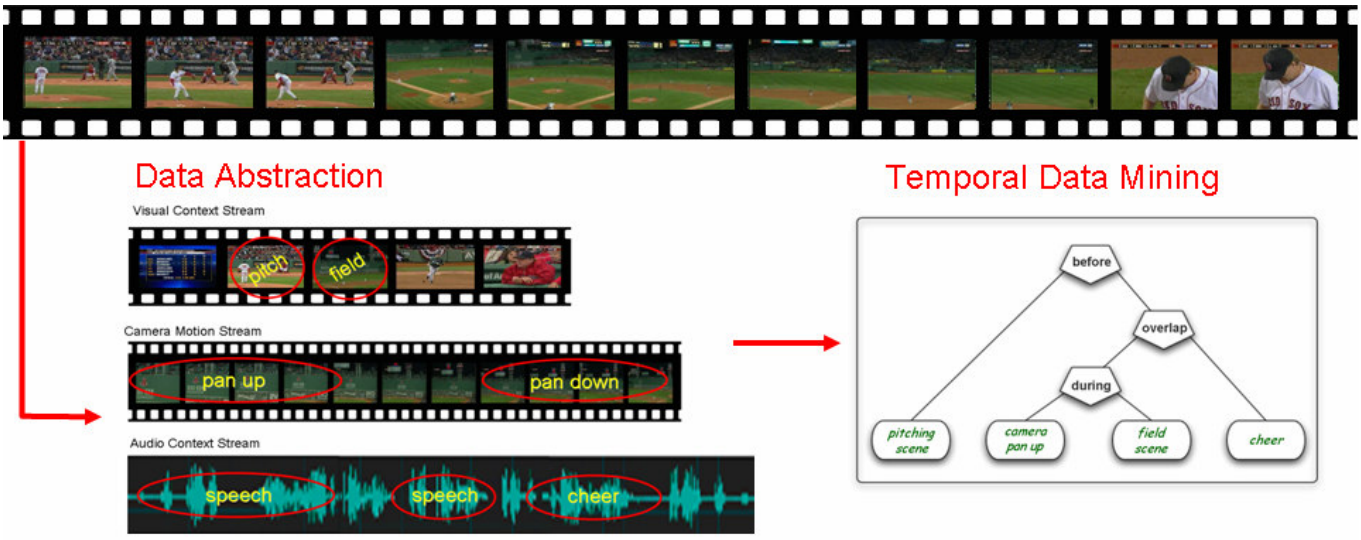

Figure 1. Temporal feature induction operates by first abstracting the raw video into parallel streams corresponding to visual context, camera motion, and audio context features. Temporal data mining then discovers patterns in the streams that are used to improve classification performance.

high level semantic classes. Then, in the second phase, temporal data mining techniques are applied to these abstracted streams in order to discover temporal patterns in the data that can be used to improve highlight classification.

3.1 Data Abstraction

In order to reduce the high-dimensionality of raw audio/video data and facilitate mining complex temporal information, temporal feature induction begins by abstracting the raw video into streams of more semantically meaningful discrete categories (see Figure 1). Each of these streams corresponds to different types of information available in the data. In these experiments, we abstract multiple parallel streams corresponding to visual context, camera motion, and audio context respectively.

3.1.1 Visual Context Stream

Visual context encodes general properties of the visual scene in a video segment. The first step in extracting such features is to split the raw video into "shots" based on changes in the visual scene due to editing [13]. After a game is segmented into shots, each shot is categorized into one of three categories: *pitching scene, field scene*, or *other*. Previous work has shown that these simple classifications can be reliably achieved in the baseball domain by taking advantage of low level features that look at color histograms, motion, line segments, etc. Here, categorization is based on similar features which are extracted from key frames chosen from within each automatically determined shot. A decision tree classifier (with bagging and boosting) is trained using the WEKA machine learning toolkit [14] on hand-labeled example frames and achieves over 96% accuracy on 10 fold cross-validation. This classifier is then used to abstract the shot sequences into a single stream of semantic classes corresponding to the type of visual context seen in each shot of the video.

3.1.2 Camera Motion Stream

Camera motion encodes the amount of pan/tilt/zoom in the video stream. Such information is particularly important for recognizing the actions taking place in a sports highlight. Detecting camera motion is a well-studied problem in video analysis. We use the system of [3] which computes the pan, tilt, and zoom motions using the parameters of a two-dimensional affine model fit to every pair of sequential frames in a video segment. The output of this system is then clustered into characteristic camera motions (e.g. zooming in fast while panning slightly left) using a 1st order Hidden Markov Model with 15 states, implemented using the Graphical Modeling Toolkit [4]. The final output of this procedure is a single data stream of states corresponding to the type of camera motion taking place between every two frames in the video.

3.1.3 Audio Context Stream

Abstracting audio context from raw audio requires both sound classification and segmentation. We employ a sound classification system based on supervised learning algorithms in which binary classifiers for *speech, cheering*, and *music* are built using boosted decision trees [14].

In training, each example segment is chunked into a sequence of short 30 ms frames. For each frame, a feature vector is computed using Mel Frequency Cepstral Coefficients (MFCCs), energy, the number of zero crossings, spectral entropy, and relative power between different frequency bands.

During classification, features are extracted to produce a sequence of feature vectors, one per 30ms of video. The classifier is then applied to each frame, producing a sequence of class labels. In order to output meaningful segments, a smoothing and segmenting algorithm is applied. Smoothing is performed using a dynamic programming cost minimization algorithm. Each frame class label is treated as an observation of the hidden, true state of that frame. Choosing a hidden state that disagrees with the observed label incurs an observation cost, while switching states incurs a state switching cost. By adjusting these costs, a balance between smoothing too much or too little can be optimized. The output of this smoothing procedure is a single stream of sound

classes that correspond to intervals of *speech, cheering*, and *music* found in the raw audio.

Table 1. Confusion matrix for baseline classifier (above) and classifier using induced temporal features of depth 5 (below)

[hyp→]	Home	Out Hit	In Out	Strike	Out Out	In Hit	Walk	*[rec]*
Home	3	2	1	1	0	0	0	.43
OutHit	1	31	10	3	6	0	1	.60
InOut	0	5	60	3	5	0	0	.82
Strike	0	2	3	31	0	0	10	.67
OutOut	3	16	2	0	11	0	0	.34
InHit	0	2	1	0	0	0	0	0
Walk	0	0	5	13	0	0	6	.25
[prec]	.43	.53	.73	.61	.5	0	.35	

[hyp→]	Home	Out Hit	In Out	Strike	Out Out	In Hit	Walk	*[rec]*
Home	3	2	0	2	0	0	0	.43
OutHit	1	37	5	1	7	0	1	.71
InOut	0	6	60	3	3	1	0	.82
Strike	0	0	5	36	0	0	5	.78
OutOut	0	12	2	0	18	0	0	.56
InHit	0	2	1	0	0	0	0	0
Walk	0	0	0	14	0	0	10	.42
[prec]	.75	.63	.82	.64	.64	0	.63	

3.2 Temporal Data Mining

Temporal patterns are mined from the multiple parallel streams abstracted from the raw video data. Following previous work in video content classification [6], we use techniques from temporal data mining to discover patterns from these multiple asynchronous data streams (see Figure 1).

The temporal data mining algorithm we use is fully unsupervised. It processes feature streams by examining the relations that occur between individual features (across multiple streams) within a moving time window (set to 30 frames). Following Allen [1], any two features that occur within this window must be in one of seven temporal relations with each other (i.e., before, during, overlaps, finishes, starts, identical). The algorithm keeps track of how often each of these relations is observed, and after the entire video corpus is analyzed, uses chi-square analyses to determine which relations are significant. The algorithm iterates through the data, and relations between individual features that are found significant in one iteration (e.g. [BEFORE, *pitching-scene, field-scene*]), are themselves treated as individual features in the next. This allows the system to build up higher-order nested relations in each iteration (e.g. [DURING, [BEFORE, *pitching-scene, field-scene*], *cheering*]]). By changing the number of iterations used, we can control the depth of the mined patterns, and thus, the complexity of the encoded temporal information.

After the algorithm completes, the set of statistically significant patterns discovered is used as a codebook to re-represent video highlights. Given this codebook, highlights can be re-represented by matching each pattern in the codebook against the feature streams of the highlight (much like regular expressions matching to a string). Whenever a pattern from the codebook is found in a highlight, the pattern itself is treated as a feature added to the vector representation of the highlight. The value given to this matched pattern feature is equivalent to the duration that the pattern occurred in the highlight. This duration information, in addition to the higher order temporal information captured by the feature, is not easily encoded in traditional dynamic models, and is a powerful feature in classification.

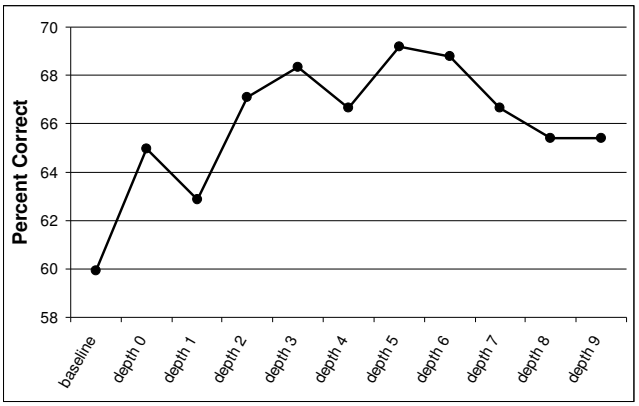

Figure 2. Comparison of classifiers using temporal feature induction. Depth level corresponds to complexity of temporal features. Baseline uses only low level features.

4. EXPERIMENTAL RESULTS

We evaluate the effectiveness of temporal feature induction for highlight classification using a small pilot data set of six broadcast baseball games. The dataset contains 237 distinct highlights hand detected by the experimenters (automatic highlight detection is not a focus of this work). The highlights are comprised of mpg 1 video (29.9 frames/sec) from nine teams, in four stadiums, on four US television stations. Following Gong et al [8], each highlight was hand labeled into one of seven categories: *homerun, outfield hit, infield hit, strikeout, outfield out, infield out,* and *walk.* As stated in section 2, we evaluate performance in a discriminative framework, training a decision tree with bagging and boosting using leave on out cross-validation [14]. We use boosted decision trees because of their relative speed and non-linear capabilities, as well as their high performance compared to other state of the art discriminative classifiers [11].

In order to examine the effect of temporal feature induction, we set up a baseline classification system which follows Gong et al. [8] (see section 2) and is trained only on the low level features used to generate the abstraction streams described in section 3.1. Temporal feature induction is evaluated by iteratively adding pattern features mined by the algorithm to these baseline features.

Figure 2 shows the accuracy of the baseline system compared to systems using temporal feature induction set to varying levels of complexity. Here the level refers to the maximum depth of the temporal pattern mined (i.e. the number of iterations used by the data mining algorithm), where depth 0 refers to only using the duration of the semantic categories described in section 3.1 (e.g. *pitching-scene, cheering,* etc.), depth 1 refers to relations between two categories (e.g., [BEFORE, pitching-scene, field-scene]), etc.

These results demonstrate statistically significant improvement ($p<0.05$; n=237, one-tail) using temporal features of depth two and greater, with a peak performance at depth five. In order to

understand the nature of this performance increase, we show the confusion matrices for these two systems (baseline and depth five) in Table 1. These tables show that temporal feature induction improves precision and recall for all classes, with the most notable increases coming from the categories *walk* and *outfield out*. As can be seen in the confusion matrices, these two categories are often confused in the baseline system with the visually similar categories *strikeout* and *outfield hit,* respectively. The system using temporal features is less prone to such confusions, because of the finer grained temporal information that it captures.

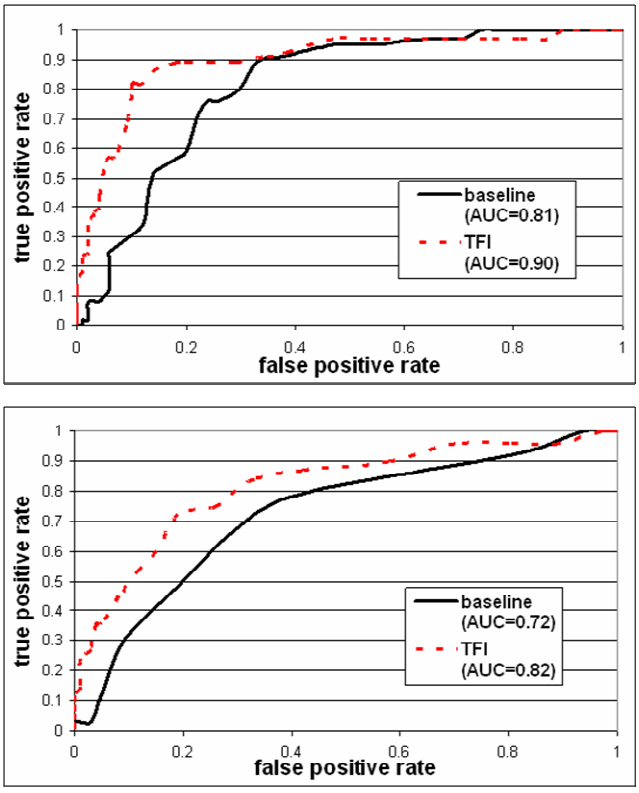

Figure 3. ROC curve for classification of *left field* highlights (above) and *fly ball* highlights (below). Baseline is compared to classifier using temporal feature induction. AUC reports area under the curve for each classifier.

The benefit of this finer grained information is even more pronounced when finer grained classifications are required. For highlight classes focusing on specific types of hits (e.g. *fly balls*) or specific locations of hits (e.g. *left field*), using temporal features becomes increasingly useful. Figure 3 show ROC curves for these example classes. In each figure, the tradeoff between the true positive and false positive rate is graphed as the threshold used for classification is changed. A comparison of the baseline system to one using temporal feature induction shows that using more temporal info enables better fine-grained classifications.

5. CONCLUSIONS

We have presented a method for automatically incorporating complex temporal information into models for baseball highlight classification. The method uses techniques from temporal data mining to discover a codebook of hierarchical temporal relations that are used to represent events occurring in a baseball highlight. The patterns encode complex temporal information, such as long-distance dependencies and durations, not easily captured in traditional dynamic models. Preliminary results indicate that training classifiers using these induced features gives significantly better performance than low level features alone.

While there are clear benefits to using temporal feature induction for supervised classification, our current work is exploring ways to employ such complex temporal information to support other applications, such as video retrieval. [7] explores using temporal feature induction to allow for unsupervised content-based indexing of video events. In this work, the codebook of mined patterns is automatically mapped to words from the closed captioning transcripts of announcers; describing events in the video. By exploiting the rich temporal information produced by temporal feature induction, sports video can be indexed and searched without predefining a large number of classes and hand labeling many examples.

6. REFERENCES

[1] Allen, J.F. (1984). A General Model of Action and Time. Artificial Intelligence. 23(2).

[2] Kokaram, A., Rea, N., Dahyot, R., Tekalp, A., Bouthemy, P., Gros, P., Sezan I. (2006). Browsing Sports Video. IEEE Signal Processing Magazine. 47.

[3] Bouthemy, P., Gelgon, M., Ganansia, F. (1999). A unified approach to shot change detection and camera motion characterization. IEEE Trans. on Circuits and Systems for Video Technology, 9(7).

[4] http://ssli.ee.washington.edu/~bilmes/gmtk/

[5] Della Pietra, S.; Della Pietra, V.; Lafferty, J. (1997). Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence. Pg: 380-393.

[6] Fleischman, M., DeCamp, P. Roy, D. (2006). Mining Temporal Patterns of Movement for Video Content Classification. ACM Workshop on Multimedia Information Retrieval.

[7] Fleischman M, Roy, D. (2007). Situated Models of Meaning for Sports Video Retrieval. HLT/NAACL. Rochester, NY.

[8] Gong, Y., Han, M., Hua, W., Xu, W. (2004). Maximum entropy model-based baseball highlight detection and classification. Computer Vision and Image Understanding. 96(2).

[9] Hongen, S., Nevatia, R. Bremond, F. (2004). Video-based event recognition: activity representation and probabilistic recognition methods. Computer Vision and Image Understanding. 96(2).

[10] Intille, S. and A.F. Bobick, (2001). Recognizing planned, multi-person action", Computer Vision and Image Understanding 81.

[11] Niculescu-Mizil, A. and Caruana, R. (2005) Predicting good probabilities with supervised learning Full text . Proceedings of the 22nd International Conference on Machine Learning

[12] Rui, Y., Gupta, A. and Acero, A., (2000) Automatically extracting highlights for TV Baseball programs ACM Multimedia. Marina del Rey, CA.

[13] Tardini, G. Grana C., Marchi, R., Cucchiara, R., (2005). Shot Detection and Motion Analysis for Automatic MPEG-7 Annotation of Sports Videos. In 13th International Conference on Image Analysis and Processing.

[14] Witten, I. and Frank, E. (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.