Situated Models of Meaning for Sports Video Retrieval

Michael Fleischman
MIT Media Lab
mbf@mit.edu

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

Abstract

Situated models of meaning ground words in the non-linguistic context, or situation, to which they refer. Applying such models to sports video retrieval requires learning appropriate representations for complex events. We propose a method that uses data mining to discover temporal patterns in video, and pair these patterns with associated closed captioning text. This paired corpus is used to train a situated model of meaning that significantly improves video retrieval performance.

1 Introduction

Recent advances in digital broadcasting and recording allow fans access to an unprecedented amount of sports video. The growing need to manage and search large video collections presents a challenge to traditional information retrieval (IR) technologies. Such methods cannot be directly applied to video data, even when closed caption transcripts are available; for, unlike text documents, the occurrence of a query term in a video is often not enough to assume the video's relevance to that query. For example, when searching through video of baseball games, returning all clips in which the phrase "home run" occurs, results primarily in video of events where a home run does not actually occur. This follows from the fact that in sports, as in life, people often talk not about what is currently happening, but rather, they talk about what did, might, or will happen in the future.

Traditional IR techniques cannot address such problems because they model the meaning of a query term strictly by that term's relationship to other terms. To build systems that successfully search video, IR techniques should be extended to exploit not just linguistic information but also elements of the non-linguistic context, or *situation*, that surrounds language use. This paper presents a method for video event retrieval from broadcast sports that achieves this by learning a *situated* model of meaning from an unlabeled video corpus.

The framework for the current model is derived from previous work on computational models of verb learning (Fleischman & Roy, 2005). In this earlier work, meaning is defined by a probabilistic mapping between words and representations of the non-linguistic events to which those words refer. In applying this framework to events in video, we follow recent work on video surveillance in which complex events are represented as temporal relations between lower level sub-events (Hongen et al., 2004). While in the surveillance domain, hand crafted event representations have been used successfully, the greater variability of content in broadcast sports demands an automatic method for designing event representations.

The primary focus of this paper is to present a method for mining such representations from large video corpora, and to describe how these representations can be mapped to natural language. We focus on a pilot dataset of broadcast baseball games. Pilot video retrieval tests show that using a situated model significantly improves performances over traditional language modeling methods.

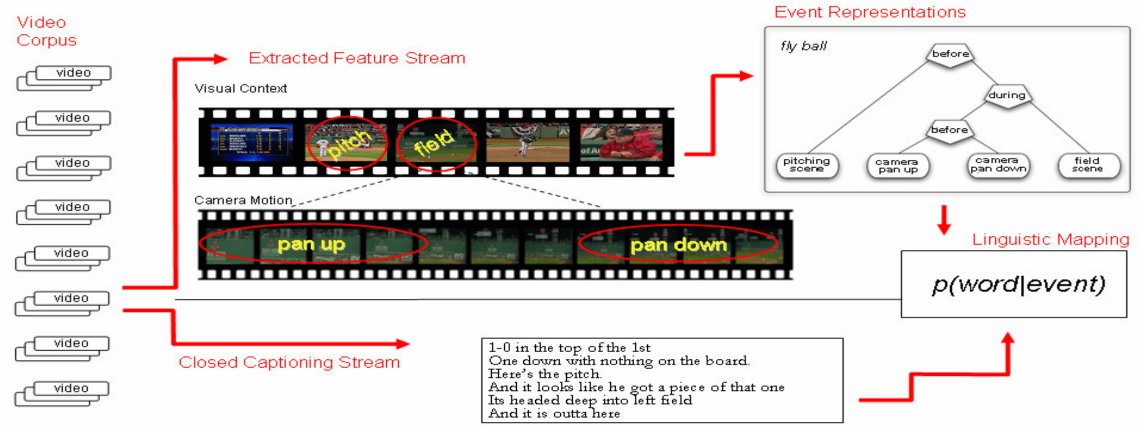
2 Situated Models of Meaning

Building situated models of meaning operates in three phases (see Figure 1): first, raw video data is abstracted into multiple streams of discrete features. Temporal data mining techniques are then applied to these feature streams to discover hierarchical temporal patterns. These temporal patterns form the event representations that are then mapped to words from the closed caption stream.

2.1 Feature Extraction

The first step in representing events in video is to abstract the very high dimensional raw video data into more semantically meaningful streams of information. Ideally, these streams would correspond to basic events that occur in sports video (e.g., hitting, throwing, catching, kicking, etc.). Due to the limitations of computer vision techniques, extracting such ideal features is often infeasible. However, by exploiting the "language of

Proceedings of NAACL HLT 2007, Companion Volume, pages 37–40,
Rochester, NY, April 2007. ©2007 Association for Computational Linguistics



1-0 in the top of the 1st
One down with nothing on the board.
Here's the pitch.
And it looks like he got a piece of that one
Its headed deep into left field
And it is outta here

Figure 1. Video processing pipeline for learning situated models of meaning.

film" that is used to produce sports video, informative features can be extracted that are also easy to compute. Thus, although we cannot easily identify a player hitting the ball, we can easily detect features that correlate with hitting: e.g., when a scene focusing on the pitching mound immediately jumps to one zooming in on the field (Figure 1). While such correlations are not perfect, pilot tests show that baseball events can be classified using such features (Fleischman et. al., in prep).

Importantly, this is the only phase of our framework that is domain specific; i.e., it is the only aspect of the framework designed specifically for use with baseball data. Although many feature types can be extracted, we focus on only two feature types: visual context, and camera motion.

Visual Context

Visual context features encode 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 (e.g., jumping from a close up of the pitcher to a wide angle of the field). Shot detection is a well studied problem in multimedia research; in this work, we use the method of Tardini et al. (2005) because of its speed and proven performance on sports video.

After a game is segmented into shots, each shot is categorized into one of three categories: *pitching-scene, field-scene*, or *other*. Categorization is based on image features (e.g., color histograms, edge detection, motion analysis) extracted from an individual key frame chosen from that shot. A decision tree is trained (with bagging and boosting) using the WEKA machine learning toolkit that achieves over 97% accuracy on a held out dataset.

Camera Motion

Whereas visual context features provide information about the global situation that is being observed, camera motion features afford more precise information about the actions occurring in the video. The intuition here is that the camera is a stand in for a viewer's focus of attention. As action in the video takes place, the camera moves to follow it, mirroring the action itself, and providing an informative feature for event representation.

Detecting camera motion (i.e., pan/tilt/zoom) is a well-studied problem in video analysis. We use the system of (Bouthemy et al., 1999) 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 (GMTK).

2.2 Temporal Pattern Mining

In this step, temporal patterns are mined from the features abstracted from the raw video data. As described above, ideal semantic features (such as hitting and catching) cannot be extracted easily from video. We hypothesize that finding temporal patterns between scene and camera motion features can produce representations that are highly correlated with sports events. Importantly, such temporal patterns are not strictly sequential, but rather, are composed of features that can occur in complex and varied temporal relations to each other. For example, Figure 1 shows the representation for a fly ball event that is composed of: a *camera pan-*

ning up followed by a *camera pan down*, occurring during a *field scene,* and before a *pitching scene*.

Following previous work in video content classification (Fleischman et al., 2006), we use techniques from temporal data mining to discover event patterns from feature streams. The algorithm we use is fully unsupervised. It processes feature streams by examining the relations that occur between individual features within a moving time window. Following Allen (1984), any two features that occur within this window must be in one of seven temporal relations with each other (e.g. *before, during, etc.*). 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, *camera panning up, camera panning down*]), 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, *camera panning up, camera panning down*], *field scene*]]). The temporal patterns found significant in this way are then used as the event representations that are then mapped to words.

2.3 Linguistic Mapping

The last step in building a situated model of meaning is to map words onto the representations of events mined from the raw video. We equate the learning of this mapping to the problem of estimating the conditional probability distribution of a word given a video event representation. Similar to work in image retrieval (Barnard et al., 2003), we cast the problem in terms of Machine Translation: given a paired corpus of words and a set of video event representations to which they refer, we make the IBM Model 1 assumption and use the expectation-maximization method to estimate the parameters (Brown et al., 1993):

$$p(word \mid video) = \frac{C}{(l+1)^m} \prod_{j=1}^{m} p(word_j \mid video_{a_j}) \qquad \textbf{\textit{(1)}}$$

This paired corpus is created from a corpus of raw video by first abstracting each video into the feature streams described above. For every shot classified as a *pitching scene*, a new instance is created in the paired corpus corresponding to an event that starts at the beginning of that shot and ends exactly four shots after. This definition of an event follows from the fact that most events in baseball must start with a pitch and usually do not last longer than four shots (Gong et al., 2004).

For each of these events in the paired corpus, a representation of the video is generated by matching all patterns (and the nested sub-patterns) found from temporal mining to the feature streams of the event. These video representations are then paired with all the words from the closed captioning that occur during that event (plus/minus 10 seconds).

3 Experiments

Work on video IR in the news domain often focuses on indexing video data using a set of image classifiers that categorize shots into pre-determined concepts (e.g. *flag*, *outdoors*, *George Bush, etc.).* Text queries must then be translated (sometimes manually) in terms of these concepts (Worring & Snoek, 2006). Our work focuses on a more automated approach that is closer to traditional IR techniques. Our framework extends the language modeling approach of Ponte and Croft (1998) by incorporating a situated model of meaning.

In Ponte and Croft (1998), documents relevant to a query are ranked based on the probability that each document generated each query term. We follow this approach for video events, making the assumption that the relevance of an event to a query depends both on the words associated with the event (i.e. what was said while the event occurred), as well as the situational context modeled by the video event representations:

$$p(query \mid event) = \prod_{word}^{query} p(word \mid caption)^{\alpha} * p(word \mid video)^{(1-\alpha)} \qquad \textbf{(2)}$$

The *p(word|caption)* is estimated using the language modeling technique described in Ponte and Croft (1998). The *p(word|video)* is estimated as in equation 1 above. α is used to weight the models.

Data

The system has been evaluated on a pilot set of 6 broadcast baseball games totaling about 15 hours and 1200 distinct events. The data represents video of 9 different teams, at 4 different stadiums, broadcast on 4 different stations. Highlights (i.e., events which terminate with the player either *out* or *safe*) were hand annotated, and categorized according to the type of the event (e.g., *strikeout vs. homerun)*, the location of the event (*e.g., right field vs. infield*), and the nature of the event (e.g., *fly ball vs. line drive*). Each of these categories was used to automatically select query terms to be used in testing. Similar to Berger & Lafferty (1999), the probability distribution of terms given a category is estimated using a normalized log-likelihood ratio (Moore, 2004), and query terms are sampled randomly from this distribution. This gives us a set of queries for each annotated category (e.g., *strikeout*: “miss, chasing”; *flyball:* “fly, streak”). Although much noisier than human produced queries, this procedure generates a large amount of test queries for which relevant results can easily be determined (e.g., if a returned event for the query “fly, streak” is of the *flyball* category, it is marked relevant).

Experiments are reported using 6-fold cross validation during which five games are used to train the situated model while the sixth is held out for testing. Because data is sparse, the situation model is trained only on the hand annotated highlight events. However, retrieval is always tested using both highlight and non-highlight events.

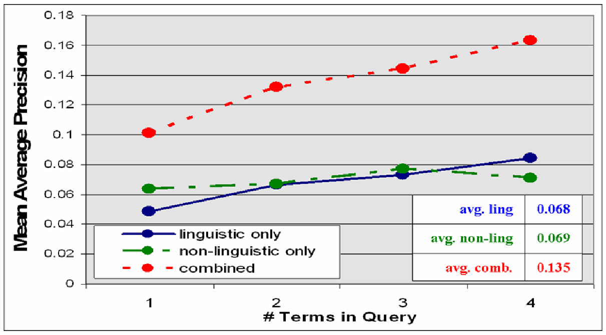


Figure 2. Effect of situated model on video IR.

Results

Figure 2 shows results for 520 automatically generated queries of one to four words in length. Mean average precision (MAP), a common metric that combines elements of precision, recall, and ranking, is used to measure the relevance of the top five results returned for each query. We show results for the system using only linguistic information (i.e. α=1), only non-linguistic information (i.e. α=0), and both information together (i.e. α=0.5).

The poor performance of the system using only non-linguistic information is expected given the limited training data and the simple features used to represent events. Interestingly, using only linguistic information produces similarly poor performance. This is a direct result of announcers’ tendency to discuss topics not currently occurring in the video. By combining text and video analyses, though, the system performs significantly better ($p<0.01$) by determining when the observed language actually refers to the situation at hand.

4 Conclusion

We have presented a framework for video retrieval that significantly out-performs traditional IR methods applied to closed caption text. Our new approach incorporates the visual content of baseball video using automatically learned event representations to model the situated meaning of words. Results indicate that integration of situational context dramatically improves performance over traditional methods alone. In future work we will examine the effects of applying situated models of meaning to other tasks (e.g., machine translation).

References

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

Barnard, K, Duygulu, P, de Freitas, N, Forsyth, D, Blei, D, and Jordan, M. (2003), "Matching Words and Pictures," Journal of Machine Learning Research, Vol 3.

Berger, A. and Lafferty, J. (1999). Information Retrieval as Statistical Translation. In Proceedings of SIGIR-99.

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):1030-1044.

Brown, P., Della Pietra, S., Della Pietra, V. Mercer, R. (1993). The mathematics of machine translation: Parameter estimation. Computational Linguistics, 19(10).

Fleischman, M. and Roy, D. (2005). Intentional Context in Situated Language Learning. In Proc. of 9th Conference on Comp. Natural Language Learning.

Fleischman, M., DeCamp, P. Roy, D. (2006). Mining Temporal Patterns of Movement for Video Content Classification. The 8th ACM SIGMM International Workshop on Multimedia Information Retrieval.

Fleischman, M., Roy, B., Roy, D. (in prep.). Automated Feature Engineering inBaseball Highlight Classification.

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).

Hongen, S., Nevatia, R. Bremond, F. (2004). Video-based event recognition: activity representation and probabilistic recognition methods. Computer Vision and Image Understanding. 96(2). pp: 129 - 162

Moore, Robert C. (2004). Improving IBM Word Alignment Model 1. in Proc. of 42nd ACL.

Ponte, J.M., and Croft, W.B. (1998). A Language Modeling Approach to Information Retrieval. In Proc. of SIGIR’98.

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.

Worring, M., Snoek, C.. (2006). Semantic Indexing and Retrieval of Video. Tutorial at ACM Multimedia