Unsupervised Content-Based Indexing for Sports Video Retrieval

Michael Fleischman
MIT Media Laboratory
mbf@mit.edu

Humberto Evans
MIT Media Laboratory
hevans@mit.edu

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

ABSTRACT
This demonstration presents an interface to a corpus of broadcast baseball games that have been indexed using an unsupervised content-based method introduced here. The method uses the concept of a grounded language model to motivate a framework in which video is searched using natural language with no reliance on predetermined concepts or hand labeled events. The interface demonstrates the effectiveness of the technique and the ease of use it affords the user.

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

General Terms
Algorithms, Experimentation

Keywords
Video retrieval, grounded language models, sports video.

1. INTRODUCTION
The vast majority of work on sports video retrieval uses supervised methods to index events. As with techniques in the news video domain, these approaches require hand labeled examples of predefined event types (or concepts) for their systems to train on. Further, extra work is required to transform a user's natural language query into something the system can understand (i.e. something that uses the pre-defined concepts). Although powerful in many domains, such supervised approaches to video indexing can be labor intensive both for the system designer, who must build the concept classifiers, and the system users, who may be required to re-write queries in more machine readable terms.

In this demonstration, we present a video retrieval system for broadcast sports that employs an unsupervised approach to content-based video indexing [4]. Our framework extends the language modeling approach of Ponte and Croft [6] by incorporating a grounded language model that links query terms to the non-linguistic context to which they refer. This grounded language model is learned from an unlabeled corpus of baseball games and the paired closed-caption transcripts of the game announcers' speech. Each event in these games is represented using a codebook of automatically mined temporal patterns that describe the relationships between features of the audio/video stream. In the following sections, we give a brief description of the procedure used to learn a grounded language model and then describe the interface and methodology used to enable unsupervised video search of broadcast baseball games.

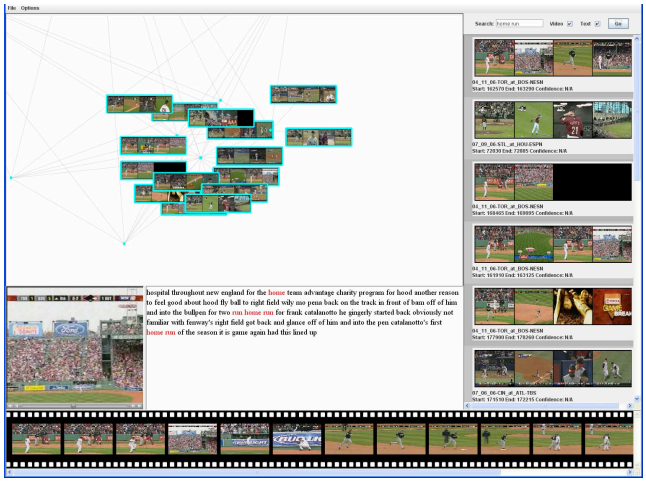

Figure 1. Interface to baseball corpus. Unsupervised indexing simplifies user interaction and facilitates content-based browsing.

2. GROUNDED LANGUAGE MODEL
2.1 Feature Extraction
The first step in learning grounded language models (see Figure 2) is to abstract the very high dimensional raw video data into more semantically meaningful streams of information. We use two types of features here: visual context, and camera motion.

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. After a game is segmented into shots, each shot is automatically categorized into one of three categories: *pitching-scene, field-scene*, or *other*. Categorization is performed using a decision tree trained (with bagging and boosting) using the WEKA machine learning toolkit. Such visual context classification can be easily learned using low level image features (e.g., color histograms, edge detection, motion analysis) extracted from individual key frames chosen from each shot. Performance of the classifier on a held out test set exceeded 96%.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
ACM 978-1-59593-701-8/07/0009.

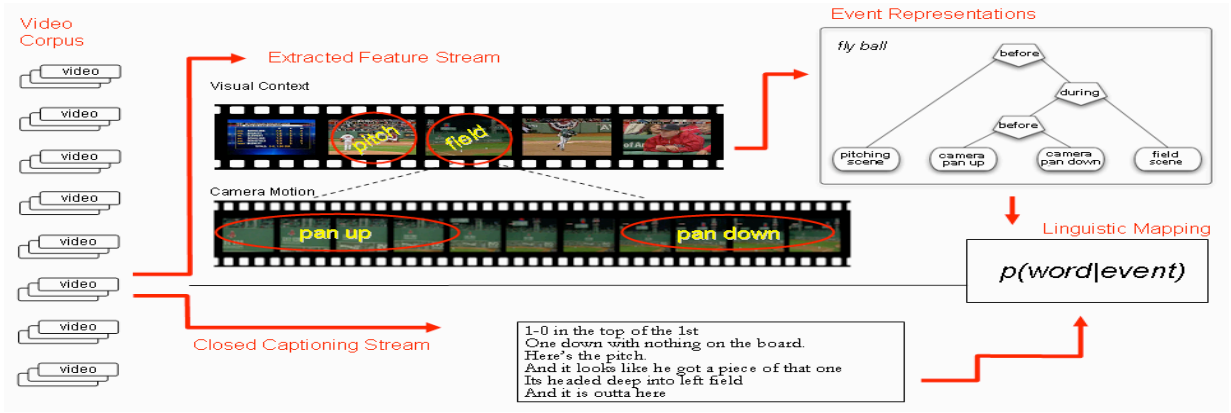

Figure 2. Workflow for learning grounded language models. Feature streams extracted from raw video encode visual context and camera motion. Temporal patterns mined from streams are used to represent events. Representations are mapped to words using EM algorithm.

Detecting camera motion (i.e., pan/tilt/zoom) is a well-studied problem in video analysis. We use the system of [2] 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, using the Graphical Modeling Toolkit.

2.2 Temporal Pattern Mining

Once feature streams have been extracted, we use temporal data mining to discover a codebook of temporal patterns that are used to represent the video events [5]. 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. 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*]]). These discovered patterns make up the codebook used to form the event representations that are then mapped to words.

2.3 Linguistic Mapping

The last step in learning the grounded language model 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 [1], 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 expectation-maximization to estimate the parameters [3]:

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

This paired corpus is created from a corpus of raw video by first abstracting each video into the feature streams described above. The videos are then segmented into events, such that each event starts with a *pitching scene* and ends exactly four shots after. Each of these events is then represented using the codebook of temporal patterns mined in Section 2.3, and then paired with all the words from the closed captioning that occur during that event.

3. DEMONSTRATNG VIDEO RETRIEVAL

Our demonstration will consist of an interface to a corpus of broadcast baseball games that have been indexed using our unsupervised method. The interface allows natural language querying for video events and requires no user interaction or reliance on predefined concepts. Users simply type in their query (e.g. "home run") and can choose to get results based solely on the closed captioning text, or with the grounded model. Further, the interface allows browsing of video events by displaying events not only as ranked lists based on the query, but also using a cluster view, in which events are organized based on their visual similarity (as calculated using the grounded language model).

4. REFERENCES

[1] 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.

[2] 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).

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

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

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

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