A Novel Learning-based Framework for Detecting Interesting Events in Soccer Videos

Nisha Jain[1] Santanu Chaudhury[1] Sumantra Dutta Roy[1] Prasenjit Mukherjee[2] Krishanu Seal[2] Kumar Talluri[2]
[1] Electrical Engineering Department, IIT Delhi
{nisha.iitd,schaudhury, sumantra.dutta.roy}@gmail.com
[2] AOL
{P.Mukherjee, Krishanu.Seal,Kumar.Talluri}@corp.aol.com

Abstract

We present a novel learning-based framework for detecting interesting events in soccer videos. The input to the system is a raw soccer video. We have learning at three levels - learning to detect interesting low-level features from image and video data using Support Vector Machines (hereafter, SVMs), and a hierarchical Conditional Random Field- (hereafter, CRF-) based methodology to learn the dependencies of mid-level features and their relation with the low-level features, and high level decisions ('interesting events') and their relation with the mid-level features: all on the basis of training video data. Descriptors are spatio-temporal in nature - they can be associated with a region in an image or a set of frames. Temporal patterns of descriptors characterise an event. We apply this framework to parse soccer videos into Interesting (a goal or a goal miss) and Non-Interesting videos. We present results of numerous experiments in support of the proposed strategy.

1. Introduction

Automatic extraction of highlights and summarization is an important task in video analysis, especially in the sports domain [12], [4], [1], [5], [2], [14]. In general, sports video analysis involves analysing all feed associated with it - ticker-tape text, audio, still images, graphics, video, and so on. Hakeem and Shah [7] present an unsupervised methodology for learning, detection and representation of events performed by multiple agents in videos. Flieschman and Roy [6] present an unsupervised learning framework based on learning grounded language models. The authors associate low-level patterns from videos with words extracted from the closed captioning text using a generalisation of Latent Dirichlet Allocation. This paper presents an alternate point of view, using CRF-based hierarchical modelling, and supervised learning. Often we have enough ground truth to label representative examples - supervised approaches are very commonly used for such a task [9]. While unsupervised learning is more applicable for mining and discovering events, our work is geared to a different problem - that of automatically parsing video streams. We have specific definitions for events, and their description through a set of training examples. Further, the problem involves probabilistic mappings between features (at numerous levels), and their labels (at many levels, again). Such a framework is naturally amenable to analysis using CRFs - our proposed method builds a learning-based methodology to learn this structure automatically from training videos. Further, our system works towards automated generation of highlights - by segmenting out the sequence of frames where interesting events occur. Xu and Chua [14] have a multi-level system, where the probabilistic uncertainty handling and inference process comes from a hierarchical Hidden Markov Model (HHMM). A CRF-based formulation is more general than any other Markovian or Bayesian formulation [11]: more so, in a multi-level version [8], [10]. We have a probabilistic hierarchical modelling of dependencies between events and entities at different levels - unlike the heuristic combination of rule-based and model-based structure in the work of Chu and Wu [4], or the ones in the work of Ariki et al. [1]. This is also a limitation in the work of Duan et al. [5] who proposes a multi-level hierarchy, but has a rule-based system for event detection. The work of Chang et al. [3] reports the only major use of CRFs for video analysis. The basic structure is quite different from our work, and the aim is also quite different - semantic concept detection in consumer video. To the best of our knowledge, there has been no other related work combining different kinds of learning at various levels right from feature detection to a hierarchical CRF-based formulation for parsing videos and detecting interesting events. The organisation of the rest of the paper is as follows: Sec. 2 gives an overview of the com-

plete system. Sec. 3 describes the various feature extractions schemes. Sec. 4 explains the hierarchical CRF design and the parsing process in detail. Sec. 5 presents results of the proposed approach. Sec. 6 includes the conclusion.

2. Overall System Description

Our framework for detecting interesting events in videos operates in three phases (see Fig. 1): first, the raw video data is abstracted into multiple streams of discrete features. CRF models are used to classify the low level features into generate mid- level features. The mid -level features are fed to the hierarchical CRF-based parsing model that parses out interesting events from the video. Fig. 2 shows an example

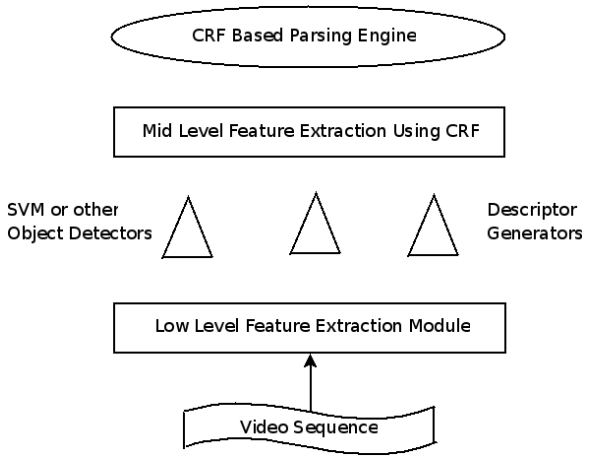

Figure 1. Three phases of the proposed framework, to detect interesting events in videos (Sec. 1).

of a hierarchical CRF in our system. This links low-level features to the overall decision e.g., Goal Sec. 3 explains our trainable feature extraction process, and In Fig. 2, each circle in the centre indicates an intermediate-level feature such as zoom in or zoom in on a goal post. The edges indicate probabilistic dependencies between these objects. The input to our model is the video which is broken into a sequence of frames. We segment each frame in order to generate a sequence of feature nodes at the next level of the model. Our model labels different features in each frame.

3. Trainable Feature Extraction Scheme

The first step in detecting interesting events is to abstract the raw video data into more semantically meaningful streams of information. We divide the video into frames and then find features in each frame of the video sequence. Here, we emphasise that a feature need not be specific

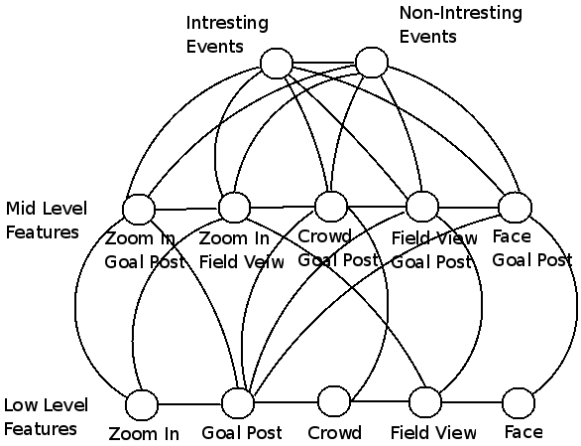

Figure 2. An example of a hierarchical CRF model. This links low-level features at the bottom to mid-level features, and the overall interesting events at the topmost level. Each link is associated with the corresponding conditional probability. This is learnt in the CRF learning phase. Sec. 1 outlines the overall system, and Sec. 4 gives the details of the hierarchical CRF-based scheme.

to data collected from a single frame. One may use any number of low-level features such as audio-based features, image-based features, graphics detection and analysis, and of course, getting spatio-temporal video features. For our experiments with soccer videos, we choose the following low-level features, and outline their detection process in detail in the subsequent sections: Zoom In, Goal Post Detection, Crowd Detection, Field View Detection, and Face Detection.

3.1. Zoom-In Detection

We propose the use of divergence for detecting zoom in across sequence of frames of a video. First we convert the coloured frame to a gray scale image. Then calculate the optic flow across frames of the video sequence. We get measure in both directions (x & y) say u and v. We calculate the divergence of the optical flow value i.e. calculate the partial derivative of the optic flow value (u_x, v_y) for each pixel in the frame. There can be noise at pixels which is removed by applying thresholding. In a range of some divergence value (say 50 to 150) we consider only those values that satisfy the threshold (say values greater than 75 and less than 125). We make different bins for each range and count the valid divergence values in each bin. This is the histogram of the divergence values. In the final step, we feed the histogram values to a Support Vector Machine

(SVM) for classification of zoom and non-zoom sequences of frames. One can use any method for the classification - we choose SVMs for their versatility and universal appeal in being generalised linear classifiers. We train the SVMs in their decision boundary using a large number of representative sequences.
In our experiments, we trained this using 100 representative sequences. For 50 new non-training examples given to the system, it marked a Zoom In correctly in 47 of them, representing an accuracy of 94%.

3.2. Face Detection

We take faces as the basis of detecting human beings in videos frames. We use a cascade of Haar filters as in the Viola-Jones algorithm [13], and train it with a large number of faces.

3.3. Goal Post Detection

We use a model-based approach for goal post detection, the shape of the goal post is our model which is basically two near-vertical lines and a near-horizontal line between them or a vertical line in combination with a nearly horizontal line joined together. (Often, the perspective distortion does not affect the vertical posts too much, since cameras are usually positioned on the field such that the vertical distortion is minimum.) Morphological operations are applied on each frame to detect the goal post. We convert our coloured frame to a binary image with a high threshold value so that the prominent white areas of the image can be segmented. We apply the closure operation which is dilation followed by erosion. We take the structural elements as lines that satisfy our model.

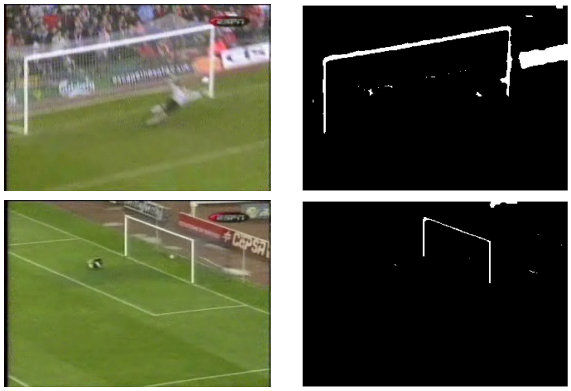

Figure 3. Goal Post Detection Results using morphological operations

Fig. 3.3 shows two examples in which the goal post detected in the frames after applying the morphological operation.The gray thresholding leaves out the whiter portions and after that applying the close operation followed by erosion with structuring element as line (first vertical the horizontal). In some cases the penalty box is also detected but because of perspective distortion those results are few.

We tested this algorithm on a data set of randomly chosen 120 frames; 57 out of 60 frames with goal post were correctly classified, and 50 out of 60 frames without the goal post were correctly classified, an accuracy of 95% and 83.33%, respectively.

3.4. Field View Features

We convert the RGB image to HSV image. We calculate histogram of the hue values of the frames of the video on the bin range of 0 to 255 with each bin of unit size. These histogram values are feeded to the SVM. In this case again, we use SVMs to train the classifier to get an optimal decision boundary.
The training data set had 150 frames information (75 frames of positive examples, and 75 negative examples). The testing data set had 132 frames which did not have any overlap with the training data set. Out of the 132 frames, our SVM-based classifier correctly classified the frame into a field view and a non-field view in 127 cases, giving an accuracy of 96.21%.

3.5. Crowd Detection

We convert the RGB image to HSV image. The input to the SVM in this case is again a histogram of hue values for each frame of the video. In this case also bins of unit length with bin range of 0 to 255 is used.
The training data set had 150 frames (75 frames with a crowd view, and 75 frames without crowd) The testing data set had 100 frames (these did not have anything in common with the training set) - out of which 89 were correctly classified. This gives an accuracy of 89%.

4. Hierarchical Classifier Design using CRFs

We propose a hierarchical approach because different features extracted in the video sequence are interdependent, for example the occurrence of zoom in depends on where exactly the zoom in occurs, on the goal post or field. CRF model exploits the probabilistic conditional dependencies between different features efficiently. Unlike, the HMM output which is dependent on the current state and not on the past, CRF produces outputs based on the current as well as past states. The Baye's method uses directed dependency, our model uses undirected dependency between different

features. The CRF- hierarchical modelling is a discriminative model that gives priorities to observations with more weightage.It discriminates nodes with less weightage and thus the output result is a cumulation of important observations.

4.1. Conditional Random Fields

Our goal is to develop a probabilistic temporal model that can extract high-level activities from sequence of frames. Discriminative models such as conditional random fields (CRF) have recently shown to outperform generative techniques in areas such as natural language processing, web page classification and computer vision. CRF are undirected graphical models used for Relational Learning. CRF's directly represent the conditional distribution over hidden states given the observations. CRF's are thus especially suitable for classification tasks with complex and overlapped observations. Similar to hidden Markov models(HMM's) and Markov random fields, the nodes in CRF's represent a sequence of observations, denoted as $x =< x_1, x_2, ..., x_t >$, and corresponding hidden states(e.g.,mid level features), denoted as $y =< y_1, y_2, ..., y_t >$. These then define the conditional distribution $p(y|x)$ over the hidden states y. The conditional distribution over the hidden states is written as:

$$p(y|x) = \frac{1}{Z(x)} \prod \phi_c(x_c, y_c)$$

where $Z(x) = \sum_y \prod_{c \in C} \phi_c(x_c, y_c)$ is the normalizing partition function.and c is a collection of subsets. Without loss of generality $\phi_c(x_c, y_c)$ are described by log linear combinations of feature functions $f_c()$ i.e.,

$$\phi_c(x_c, y_c) = exp(w_c^T f_c(x_c, y_c))$$

where w_c^T is the transpose of a weight vector w_c and $f_c(x_c, y_c)$ is a function that extracts vector of features from the variable values. The log linear feature representation is very compact guarantees the non-negativeness potential values. We can write conditional distribution as :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in C} exp\{w_c^T f_c(x_c, y_c)\}$$

$$= \frac{1}{Z(x)} exp \left\{ \sum_{c \in C} w_c^T f_c(x_c, y_c) \right\}$$

In this step mid-level features are mined from the low level features abstracted from the low level feature modules in Step 1.We find temporal combinations of different low level features to better identify the events in the frames of the video sequence. We use linear CRF's to classify the mid level features.

The mid-level features that are classified are zoom in on the goal post, zoom in on the field, goal post and simultaneous crowd detection, goal post detection with field view and goal post and face detection simultaneously. These mid level features develop a probabilistic temporal model that can extract high level interesting events from sequence of frames of the in video sequence. These mid level features form the intermediate nodes of the hierarchical CRF model (see Fig. 2). A combination of field view with goal post and zoom in on the goal post has higher probability that the sequence of frames is a goal than just a field view and goal post. This conditional probabilistic model is generated using conditional random fields.

4.2. CRF based Parsing Engine

CRF based parsing engine parses out interesting or non interesting sequence from the complete frames of the video.In our case the interesting event includes a goal and goal miss. Non interesting events include field play in general. We feed the mid level features extracted from the above step to another set of linear CRF's that classify whether the sequence of frames in the video is interesting or not. The complete set of frames are now classified as interesting or non-interesting event. Each event has a set of sequence frames. The sequence of frames is outputted as interesting or non-interesting event.This can be used to create summaries of soccer videos primarily with all the interesting events of the game. It also has an application in mobiles for generation of highlights for the consumers.

5. Video Parsing Results

We trained our system to detect Interesting Events (goal or goal miss) and Non-interesting events. We trained our hierarchical system on a set of 30 videos, and had a set of 170 other videos for testing. First three examples, that are explained below describe how on the basis of low level and mid-level features a result is deduced for the type of event.

In Fig. 4, we show an example of a video correctly detected as an Interesting Event (in this case a goal). The figure shows some frames from the video, and the corresponding low-level features, mid-level features, and of course, the final decision - all with the relation between the entities clearly outlined.The low level features, zoom, goal post, crowd and face detection are combined according to their respective dependency to deduce mid level features, zoom in on the goal post, a crowd detection along with goal post detection and goal post detection followed by a face detection. These features results in the final decision of an interesting event(goal).

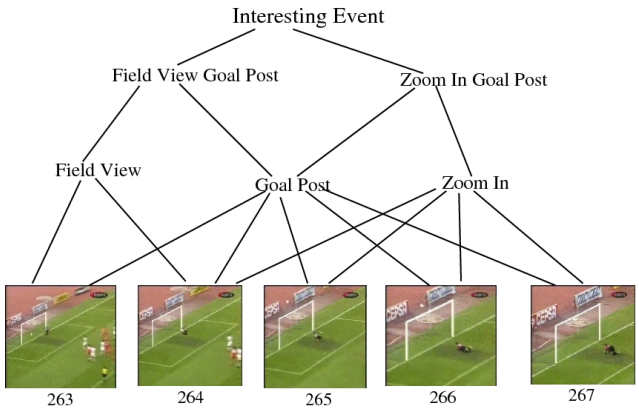

Figure 4. Example 1, Sec. 5: an example of a video correctly detected as an Interesting Event (in this case a goal). The figure shows the corresponding parsing using the hierarchical CRF model.

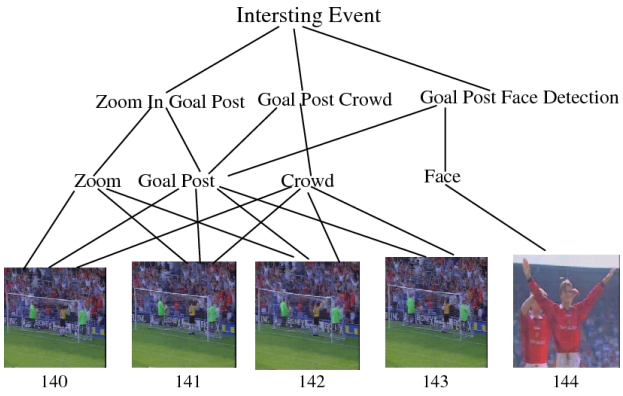

Figure 5. Example 2, Sec. 5: another example of a video correctly detected as an Interesting Event (in this case a goal). The figure shows the corresponding parsing using the hierarchical CRF model.

Fig. 5 shows a different scenario corresponding to a successful classification of an interesting event. Fig. 6 shows an example of a successful detection of a Non-Interesting event. Again, the figure shows some frames from the video, and the corresponding low-level features, mid-level features, and of course, the final decision - all with the relation between the entities.In this example across the sequence of frames the only low level feature detected is field view. So the mid level features, that are deduced donot give any indication of a feature that indicates a goal. Thus the result from the mid level features is an non-interesting event.

Next two examples show the working of the parsing engine. The features of the frames of the video are extracted and in the sequence of the frames the portions of interesting and non-interesting events are parsed.The boundary of the parsed regions are based on the mid-level features. The size or duration of the event varies according to the value of the features.The CRF based parsing engine segments out the events on the basis of the mid level features. Conditional dependency between mid level features is exploited by the CRF's that results in the parsing of interesting and non-interesting events.

Fig. 7 and Fig. 8 shows that the frames of the video that are parsed as interesting and non- interesting events based on the mid level features. In the first example the interesting event is parsed accurately and in the second case the first interesting event is parsed with one additional frame rest all portions are parsed accurately by the CRF based parsing engine.The numbers below the frames correspond to the frame number in the video sequence. The following table describes the global results of CRF- methodology.

Table 1. Compiled Results using Hierarchical CRF: (Details in Sec. 5).

Label	True	Marked	Actual	Precision
Interesting Event	8023	8100	8238	99.049
Non-Interesting Event	8017	8232	8094	97.388
Overall	16040	16332	16332	98.212

The columns in the table 1, the actual values describe the actual number of cases, the marked values describes the number of cases marked correctly by the model and the true values accounts for the number of cases that were marked correctly. Precision is the fraction of marked values that match the actual values. In case of non-interesting events marked are more than actual because some frames are marked non-interesting instead of interesting as explained in the parsing engine example.

The next table Table 2 gives experimental results for comparison of the overall hierarchical CRF- based model and non-hierarchical model. It gives results for the 50 videos which were tested. It shows the global results of the complete computation.

Tables 2 clearly shows the former to out perform the latter by a large margin.The actual values are the total frames in 50 videos that include interesting and non-interesting event. The marked values are the number of frames of the videos that were correctly parsed as interesting or non-interesting event. The accuracy gives the fraction of frames marked correctly as interesting or non-interesting.

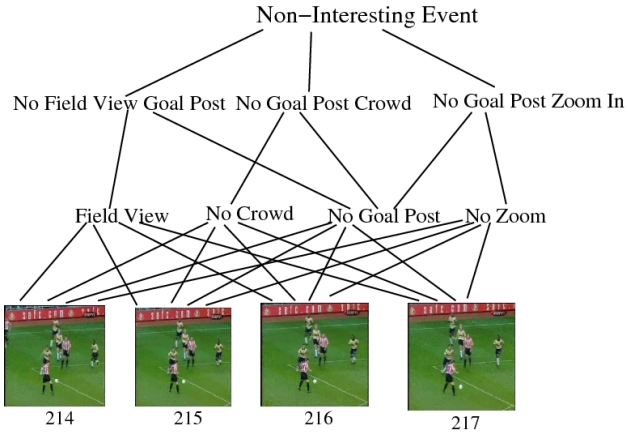

Figure 6. Example 3, Sec. 5: an example of a video correctly detected as a Non-Goal Attempt. The figure shows the corresponding parsing using the hierarchical CRF model.

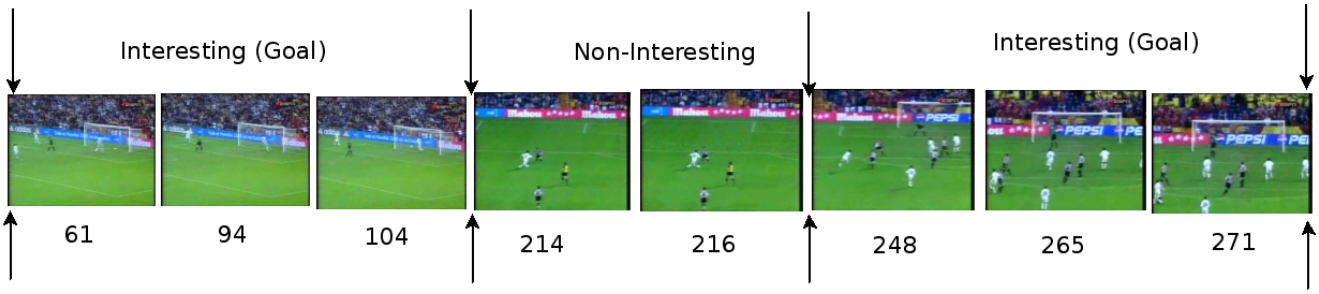

Figure 7. CRF Parsing Example 1 Sec. 5: In this the frames of the video are parsed as interesting and non- interesting events based on the mid level features.Here it parses out two interesting (goal) events and a non-interesting event

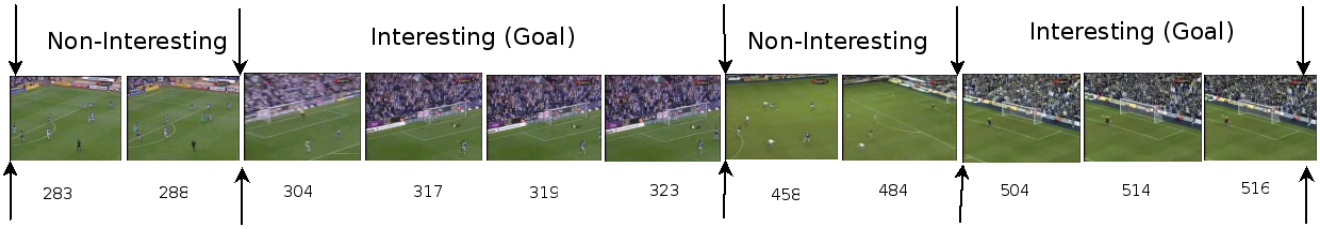

Figure 8. CRF Parsing Example 2 Sec. 5: In this the frames of the video are parsed as interesting and non- interesting events based on the mid level features. The engine parses out two interesting and two non-interesting events

Table 2. Compiled Results using Hierarchical CRF: A comparison of hierarchical and non hierarchical methodology (Details in Sec. 5).

Approach	Marked	Actual	Accuracy
Hierarchical	2634	2682	98.5
Non Hierarchical	2243	2682	83.67

6. Conclusion

We have presented novel learning scheme for detecting interesting events in soccer videos using a hierarchical Conditional Random Field approach. The model exploits the dependencies between low level features and mid level features extracted from each frame of the video sequence. The hierarchical methodology clearly outperforms the non hierarchical approach as it uses the probabilistic conditional dependencies amongst the features detected. The methodology proposed can be extended to other games like basketball, cricket by adding few more features like action recognition along with the existing features discussed in the paper.

7. Acknowledgments

The authors wish to thank AOL India for support towards this work.

References

[1] Y. Ariki, M. Kumano, and K. Tsukada. Highlight Scene Extraction in Real Time from Baseball Live Video. In *ACM Int'l Conf. on Multimedia Information Retrieval (MIR)*, pages 209 – 214, 2003.

[2] M. Bertini, A. Del Bimbo, and C. Torniai. Automatic Annotation and Semantic Retrieval of Video Sequences using Multimedia Ontologies. In *ACM Int'l Conf. on Multimedia (MM)*, pages 679 – 682, 2006.

[3] S.-F. Chang, D. Ellis, W. Jiang, K. Lee, A. Yanagawa, A. C. Loui, and J. Luo. Large-Scale Multimodel Semantic Concept Detection for Consumer Video. In *ACM Int'l Conf. on Multimedia Information Retrieval (MIR)*, pages 255 – 264, 2007.

[4] W.-T. Chu and J.-L. Wu. Explicit Semantic Events Detection and Development of Realistic Applications for Broadcasting Baseball Videos. *Multimed. Tools Appl.*, 38:27 – 50, 2008.

[5] L.-Y. Duan, M. Xu, T.-S. Chua, Q. Tian, and C.-S. Xu. A Mid-level Representation Framework for Semantic Sports Video Analysis. In *ACM Int'l Conf. on Multimedia (MM)*, pages 33 – 44, 2003.

[6] M. Flieschman and D. Roy. Unsupervised Content-Based Indexing of Sports Video. In *ACM Int'l Conf. on Multimedia Information Retrieval (MIR)*, pages 87 – 94, 2007.

[7] A. Hakeem and M. Shah. Learning, Detection and Representation of Multi-Agent Events in Videos. *Artificial Intelligence*, 171:586 – 605, 2007.

[8] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale Conditional Random Fields for Image Labeling. In *Proc. IEEE CVPR*, pages 695 – 702, 2004.

[9] A. Kokaram, N. Rea, R. Dahyot, A. Tekalp, P. Bouthemy, P. Gros, and I. Sezan. Browsing sports video. *IEEE Signal Processing Magazine*, 47, 2006.

[10] L. Liao, D. Fox, and H. Kautz. Hierarchical Conditional random Fields for GPS-based Activity Recognition. In *Proc. Int'l. Symp. Robotics Research*, pages 487 – 586, 2007.

[11] C. Sutton and A. McCallum. An Introduction to Conditional Random Fields for Relational Learning. In *Introduction to Statistical Relational Learning*. MIT Press, 2007.

[12] B. T. Truong and S. Venkatesh. Video Abstraction: A Systematic Review and Classification. *ACM Transactions on MultimediaComputing, Communications and Applications*, 3(1):1 – , 2007.

[13] P. Viola and M. J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137 – 154, 2004.

[14] H. Xu and T.-S. Chua. Fusion of AV Features and External Information Sources for Event Detection in Team Sports Video. *ACM Transactions on MultimediaComputing, Communications and Applications*, 2(1):44 – 67, 2006.