SIViP (2017) 11:171–178
DOI 10.1007/s11760-016-0916-3

ORIGINAL PAPER

Automatic analysis of broadcast football videos using contextual priors

Rahul Anand Sharma[1] · Vineet Gandhi[1] · Visesh Chari[1] · C. V. Jawahar[1]

Received: 28 January 2016 / Revised: 25 April 2016 / Accepted: 31 May 2016 / Published online: 15 June 2016

Abstract The presence of standard video editing practices in broadcast sports videos, like football, effectively means that such videos have stronger contextual priors than most generic videos. In this paper, we show that such information can be harnessed for automatic analysis of sports videos. Specifically, given an input video, we output per-frame information about camera angles and the events (goal, foul, etc.). Our main insight is that in the presence of *temporal* context (camera angles) for a video, the problem of event tagging (fouls, corners, goals, etc.) can be cast as *per frame* multi-class classification problem. We show that even with simple classifiers like linear SVM, we get significant improvement in the event tagging task when contextual information is included. We present extensive results for 10 matches from the recently concluded Football World Cup, to demonstrate the effectiveness of our approach.

Keywords Sports video analysis · Broadcast video · Event classification · Content-based retrieval

1 Introduction

Modern developments in multimedia creation, storage and compression technologies have paved the way for extensive archiving of video content. Building applications for search, summarization or editing on such large databases require extensive information about the content of these videos. Current sources of such descriptions are only limited to textual content. Since textual descriptions are both inefficient (descriptions are subjective and vary from person to person) and incomplete (it is difficult to describe all contents in a video to facilitate search, summarization or editing), it is important to build tools to automatically analyze video content and identify salient parts, to generate textual and other kinds of descriptions over the timeline. Given the diverse nature of video content on the web, this task is easier said than done. One approach to this problem is to isolate and process videos by genre. Such an approach has a twofold advantage: (1) each genre could be associated with a set of rules for video creation that might make it easier to design video understanding algorithms and (2) it is easier to distinguish between relevant and irrelevant semantic contents when information about genre is given (for example, information about crowds in a football match is rarely searched for and hence can be ignored). Recently, problems related to sports video analysis have particularly received much attention in this direction with many direct applications like automated highlights generation [1] or analysis of team activities and strategies [13], being built upon semantic analysis of video content.

Following these lines, we have looked at the problem of automatic semantic analysis of football broadcast videos. Our work is built on the fact that broadcast videos of football matches are constructed in a very structured manner, thus imposing some useful restrictions on the content. Firstly, there are fixed vantage points in a football field where PTZ cameras are placed for recording such events, thus limiting the number of views. Secondly, the edited video switches between these cameras in strong correlation with events occurring on the field.

Electronic supplementary material The online version of this article (doi:10.1007/s11760-016-0916-3) contains supplementary material, which is available to authorized users.

✉ Rahul Anand Sharma
rahul.anand@research.iiit.ac.in

[1] Center for Visual Information Technology, International Institute of Information Technology, Gachibowli, Hyderabad, Telangana 500 032, India

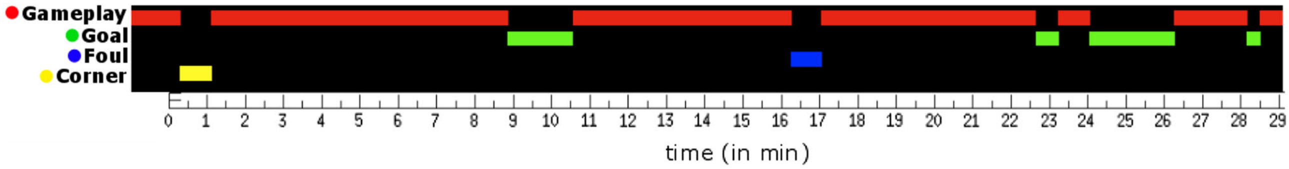

Fig. 1 An example of event tagging using the proposed approach on the first 30 min of the semifinal match between Brazil and Germany in World Cup 2014. The plot shows the occurrence of four different events (goal, foul, corner and gameplay) over the timeline. The proposed method successfully detects all the goal events

This inherent structure of editing in broadcast sports videos motivates us to ask the question, how to define context in such structured settings? We answer this problem based on two key realizations. Firstly, most camera angles associated with the events like goal or corners are predetermined. For example, in the event of a goal, a broadcast video automatically switches to focus in on player huddles/celebrations, which is unlikely to happen during normal game play. Secondly, this strong association also extends temporally since editing rarely switches between unrelated camera angles, which is to say that camera angles vary *smoothly* across time, except when shot changes occur. This strong temporal and event correlation with camera angles prompts us to argue that *contextual* information in sports video analysis can largely be based upon knowledge of camera angles.

Based on the above discussion, we argue that for the problem of automated sports video analysis two main tasks gain prominence. Firstly, classification of images into different camera views forms the initial basis of *contextual* understanding of a sports video. Secondly, the analysis of *events* in such videos gives an almost complete summary of the match (a motivating example is illustrated in Fig. 1). Accordingly, we focus on these two aspects in the current paper. Furthermore, as evidence of the usefulness of such information, we present an application of context reliant targeted spatial segmentation. Formally, we make the following contributions

- We present an automatic approach to first identify the camera view type for each frame in a video.
- Using predicted view types, we then propose an algorithm to accurately predict salient events like goal, foul, corner, substitution, etc.
- We also present an application of spatial segmentation that benefits from such contextual information.

2 Related work

Previous work on semantic segmentation of videos has looked into two different directions: first is segmentation based on shots and second is based on events. Shot-based temporal segmentation divides the video into smaller segments by identifying transitions from a camera to another. Most existing methods [11,14,18] perform shot segmentation by detecting transitions (cut, fade, dissolve, etc.) based on difference of features in consecutive frames. The segmented shots are then used as minimal units for further tasks.

In [25], Xu et al. classified football video shots into the views of global, zoom-in and close-up using the grass area ratio. In [24], Gu et al. used motion vectors in addition to dominant color to classify different view types in football broadcasts. Duan et al. [4] used a more elaborate set of features fusing motion vectors, color, texture, shape and shot length information for a similar task. These approaches suffer from the drawback of strong reliance on detecting transitions, which may not be robust due to strong correlation between different shots (for example, a camera change may not result into difference in frame color histograms, which are commonly used features in these approaches). To suppress the negative effects of inaccurate shot boundaries, it is also common to manually label the transition frames [25]. Our work addresses this issue in a more elaborate manner, by first training detectors for different camera types (as illustrated in Fig. 2) and then assigning each frame an unique camera label. By merging consistent camera labels over time, our method in turn can be used to obtain shot-based segmentation.

On the other hand, event-based segmentation divides the video into shorter clips where each clip contains only one event. The granularity of event is lower than that of shot (an event can compass several shots), and there is no standard relation between an event and the number of shots in it. In [17], Qian et al. define an event clip as a set of sequential video frames, which begins with a global view (far zoom-out view) and ends with non-global views. They further use hidden conditional random fields to classify event clips into five different categories like goal, shoot, normal, etc. But such a hard-coded definition of an event may not always hold. Sigari et al. [18] segment the video based on camera motion (estimated using block matching between consecutive frames) and use this information to detect event like counter-attack. Xie et al. [22] used a sliding window approach to classify the football video into play vs break segments.

Heuristic rule-based approaches [1,22] have also been proposed for the detection of event boundaries. These approaches perform low-level analysis to detect marks (field,

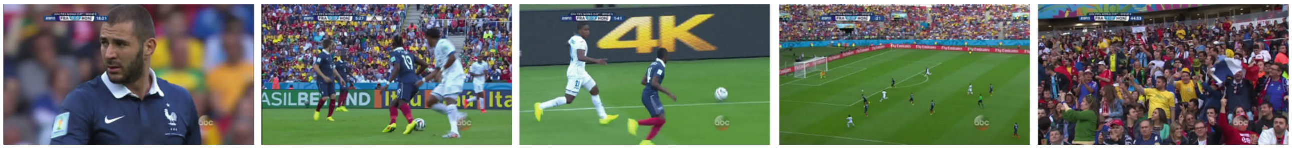

Fig. 2 We classify camera viewpoints into five different categories, namely (from *left* to *right*) ground zoom-in, ground zoom-out, top zoom-in, top zoom-out and miscellaneous (covering mainly the crowd view)

lines, arcs and goalmouth), player positions, ball position, etc. A set of hard-coded rules based on these low-level features is then used to detect corresponding events. Assfalg et al. [1] used such a rule-based approach for identification of salient events like goal, corner, kickoff, etc.

Recently, purely data-driven approaches have gained importance in sports broadcast, for example [2] learnt directorial styles by training classifiers on a training set of previous broadcasts. They suggest that such an approach could also be useful to determine the boundary of the salient events. Following these lines, we employ a simple bag-of-features [3] representation combined with support vector machines (SVMs) for both view type and event classification. The advantage of our approach is that it is independent of any ad hoc rules or hard-coded definitions, and thus, it is more generalized and can be easily extended to different sports.

Multi-model approaches are also common. Previous work [8] has looked at the problem of event detection using fusion of audio visual features. The main idea in these approaches is that the increased audio activity is a cue for important moments in the game. Textual information has also been used with visual features for event detection [23]. These methods typically use the match report and game log obtained from the web as the text source and require the alignment of the webcasting text with the broadcast sports video. On the contrary, the proposed method in this paper is purely based on visual data.

Recent work [13] has looked at the problem of discovering team behaviors or detecting the locations where the play evolution will proceed [9] by analyzing plan view tracks of the players. The plan view tracks are usually obtained using a set of static cameras manually installed around the corresponding sports field. Although these methods have demonstrated impressive results, they are not applicable for analysis in broadcast videos where only the feed from a single moving camera is available at a given time.

Far zoomed-out shots have been used for the context of sports classification [7] and recognition of group activities in football videos [10]. The work in [7] exploits the fact that the playing surface is largely visible in far zoomed-out viewpoints, and different sports can be classified based on the type of playing surface. In [10], Kong et al. built a local motion descriptor by grouping SIFT keypoint matches into foreground point set and background point set and then used it to classify events like left side attacking or left side defending, etc. Applicability to only far zoomed-out shots limits these approaches in the case of broadcast videos, which require ability to process input from different viewpoints (as illustrated in Fig. 2). Our approach, on the other hand, takes advantage of different viewpoints and presents a comprehensive solution.

3 Method

In this section, we discuss the two key components of the proposed method for automatic analysis of football videos, namely camera viewpoint estimation for each frame of the video and marking salient events on the timeline (event tagging).

3.1 Camera viewpoint estimation

Football broadcast videos are usually edited from a set of source videos recorded from different viewpoints. By analyzing 52 matches of World Cup 2014, we found out that these camera viewpoints can be broadly categorized into five different categories, which are illustrated in Fig. 2. The inclusion of camera angles (ground or top) accounts for the main difference over the previous works, which segregate the camera viewpoints only on the basis of their shot sizes (e.g., medium shot or a close-up). Although most of the game play is covered by top-view cameras located in the crowd area, the ground view cameras (placed at sidelines at field level) play an important role in viewing experience. They are quite handy during breaks and salient events and are used as standard by almost all broadcasters. An example of camera switches during a goal event is illustrated in Fig. 3.

Our goal is to automatically predict the camera viewpoint for each frame of an edited football broadcast video, as this provides a strong context for other higher-level tasks (for example, the pattern of transitions between camera viewpoints is correlated with different salient events as illustrated in Fig. 3). We approach this problem by learning a per-frame multi-class classifier.

Frame representation: Each frame is represented using the classical bag-of-words (BoW) approach [3]. SIFT features are first computed for each frame independently, which are

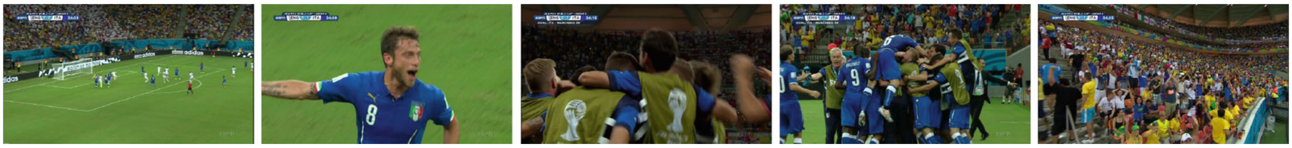

Fig. 3 Typical viewpoint transitions in a goal event (top zoom-out → top zoom-in → ground zoom-in → ground zoom-out → miscellaneous). The camera framing changes with respect to both size and angle

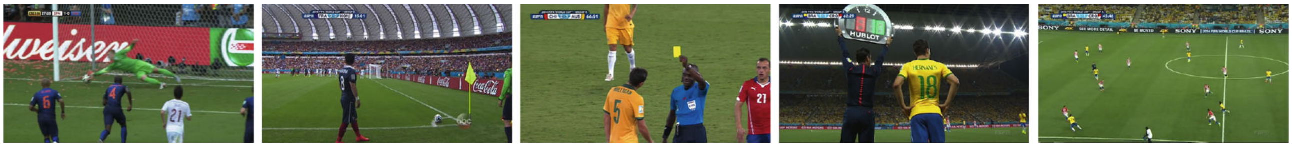

Fig. 4 We classify events into five different categories, namely (from *left* to *right*) goal, corner, foul, substitution and gameplay

then clustered using k-means clustering algorithm to build a visual vocabulary (where each cluster corresponds to a visual word). Each frame is then represented by the normalized count of number of SIFT features assigned to each cluster (BoW histograms). The length of the feature vector is equal to the number of clusters.

Using the BoW feature representation, our method learns a multi-class SVM to classify different camera labels. Training is performed by manually annotating a set of frames with corresponding camera labels. The classification is performed per-frame basis, but for classifying a frame we consider features from a temporal window of 40 frames centered around it.

We further assume that the camera transitions (cuts) are smooth, and each camera viewpoint is maintained for a minimum amount of time before cutting to a different camera. This is a fair assumption, as fast and abrupt camera transitions may appear disturbing to the viewer and are avoided by the expert editors [19]. We benefit from this assumption to smooth out the noise in per-frame camera label prediction, using a Markov random field (MRF) optimization approach.

The optimization method takes as input the multi-class SVM scores for every frame t of the video. It outputs a sequence $\xi = \{s_t\}$ of camera labels $s_t \in [1 : M]$, where M is the total possible camera labels (five in our case) for all frames $t = \{1 : N\}$. We minimize the following global cost function:

$$E(\xi) = \sum_{t=1}^{N} E_d(s_t) + \sum_{t=2}^{N} E_s(s_{t-1}, s_t). \qquad (1)$$

The cost function consists of a data term E_d that measures the evidence of the object state given the SVM scores and a smoothness term E_s which penalizes camera transitions. The data term and the smoothness term are defined as follows:

$$E_d(s_t) = -\log(P(s_t, t)). \qquad (2)$$

here $P(s_t, t)$ is the SVM classification score for camera label s_t at frame t. And,

$$E_s(s_{t-1}, s_t) = \begin{cases} 0 & \text{if } s_{t-1} = s_t, \\ \lambda & \text{otherwise.} \end{cases} \qquad (3)$$

where λ is a constant, which is determined empirically. Finally, we use dynamic programming (DP) to solve the optimization problem presented in Eq. 1. The DP algorithm constructs a graph with $M \times N$ nodes (M rows, N columns) and computes the minimum cost to reach each node. Finally, we backtrack from the minimum cost node in the last column.

3.2 Event tagging

Given a video clip, the goal of event tagging is to mark all the salient events on the timeline. It is an important problem in sports analysis as it can produce activity description and other high-level results (summarization, highlights generation, etc.). In our work, we consider four different salient events (goal, corner, substitution and foul) and a gameplay event (which broadly covers rest of the possible events). The selection was motivated by the online textual commentaries where these four salient events are distinctly marked. An instance for of each of the five classes is illustrated in Fig. 4.

Similar to camera label estimation, we design event tagging problem as a per-frame classification task. For classifying a frame, the feature vector is built using a total of 40 frames centered at that frame, where each individual frame is represented using the BoW histograms, camera label and motion features. The BoW histograms capture the correlation between the events and the distribution of features, and they are obtained in a similar manner as in previous section. The camera label per frame is defined as a five-dimensional Boolean vector and captures the correlation of an event with camera transitions.

The motion features represent the correlation of player movements with different events. They are computed from the player tracks obtained using the combination of discriminative trained deformable part models (DPMs) proposed by Fezenswalb et al. [5] and the data association approach of Pirsiavash and Ramanan [16]. The player detections are performed per-frame basis using a DPM model specifically trained for the case of football videos.

Training DPM : While training DPMs, the choice of negative examples strongly impacted the quality of the detector. Using a fixed set of 200 images with manually annotated bounding boxes for players (2120 instances of players) as positive examples and varying the negative examples, we trained three different versions of DPM:

1. Using randomly sampled windows from frames of football videos as negative examples (not overlapping with players and similar in size). Approximately 20,000 negative examples were used.
2. Using entire VOC 2007 dataset for negative samples (all classes except pedestrians)
3. Using both VOC 2007 dataset and randomly sampled windows from football videos

We then compared the detection performance of these three cases with the pre-trained pedestrian detector on a set of 75 test images. The DPM model trained with only VOC dataset as negative examples gave best results and was finally used for obtaining the player detections.

The per-frame detections are then combined into player tracks using [16]. The result is a set of bounding boxes represented by center (x, y) and height (h) and their corresponding labels (representing different tracks). Player tracks shorter than four frames are ignored. Using the corresponding bounding boxes in consecutive frames, we compute a nine-dimensional motion feature vector (mean xy-motion, median xy-motion, average h, median h, min-h, max-h and number of corresponding bounding boxes).

The BoW histograms, camera label and motion features are then concatenated for each frame individually. Finally, we perform experiments on event classification using both the mean and concatenation of features from 40 individual frames. The classification is performed using a five-class SVM, which is trained using manually annotated ground truth data. The per-frame classification results are then temporally smoothed in a similar way as in Eq. 1, penalizing frequent event transitions.

4 Experiments

We perform experiments on broadcast video sequences from 10 matches of Football World Cup 2014. We present results on the camera label estimation and the salient event classification. We make quantitative comparisons in each case using manually annotated ground truth data. Furthermore, using an application of spatial segmentation, we show that even a simple algorithm with the knowledge of camera viewpoint can bring as much as 20 % improvement over the state of the art. Each of these experiments is described with detail in the proceeding sections:

4.1 Camera label estimation

We manually annotated two 45-min videos (324,000 frames), from two different matches with camera labels (with each frame assigned an unique label) for the quantitative analysis of camera label estimation. One part was used for training and another was used for testing.

We compare our method with two commonly used approaches from the previous work based on color [15,25] and the motion vectors [24]. The color-based approaches classify shots into different classes based on the ratio of the green color or the color histogram analysis. We implemented a similar color-based algorithm using ten bin histograms and classified shots into different categories using ratio of the dominant color (we used dominant color instead of a fixed shade of green to bring further robustness). The ratios for each class were learnt from the training data.

For motion vector-based classification, we learnt a SVM classifier based on optical flow between consecutive frames. The optical flow was computed by down-sampling the image to $p \times p$ pixels, where p denotes the bin size. We tested different bin sizes, and only the best result is presented in the paper (using $p = 20$).

We obtained an average accuracy of 45, 61 and 85 % using color-based approach, optical flow-based approach and our method, respectively, using independent per-frame classification. The accuracy improved to 56, 65.4 and 92.2 % respectively, after the MRF smoothing. The confusion matrices of results obtained after MRF smoothing are given in Tables 1, 2 and 3.

The dominant color ratio-based approach (Table 1) fails to do the classification accurately. Strong misclassification between top zoom-out and top zoom-in shots can be observed. This occurs due to the large ratio of green color in both top zoom-in (usually close-up shots from top angle) and top zoom-out shots. The color-based approach completely fails to classify miscellaneous (crowd) shots and also heavily confuses among ground zoom-out and ground zoom-in shots. The optical flow-based approach also gives noisy results and strongly confuses top zoom-out shots with other view points.

On the other hand, our method provides highly accurate results (Table 3). It almost perfectly classifies top zoom-out shots, which typically cover major part of the football game. The confusion of other classes with top zoom-out shots is

Table 1 Camera label estimation results using dominant color ratio with five different camera viewpoints (percentages)

	Top zout	Top zin	Ground zout	Ground zin	Misc
Top zout	**81.4**	6.9	8.2	3.4	0.1
Top zin	68.7	**15.9**	9.3	6.1	0
Ground zout	3	7	**79.7**	10.3	0
Ground zin	3.3	5.7	70.4	**20.6**	0
Misc	2.5	1.8	33.6	55.7	**6.4**

Bold values denote per class classification accuracies

Table 2 Camera label estimation results using optical flow

	Top zout	Top zin	Ground zout	Ground zin	Misc
Top zout	**88.8**	2.3	5.9	2.3	0.7
Top zin	36.2	**25.9**	17.3	17.7	2.9
Ground zout	33.6	8.9	**38.4**	16.2	2.9
Ground zin	28.7	11.6	24.4	**32.4**	2.9
Misc	30	14	22	16	**18**

Bold values denote per class classification accuracies

Table 3 Camera label estimation results using our method

	Top zout	Top zin	Ground zout	Ground zin	Misc
Top zout	**98**	0.7	1.2	0.1	0
Top zin	0.8	**65.9**	8.1	25.1	0.1
Ground zout	0.4	3.4	**79.6**	15.1	1.5
Ground zin	0.1	1.4	19	**78.5**	1
Misc	0	0.1	0.2	4	**95.7**

Bold values denote per class classification accuracies

also negligible (column one in Table 3). In general, we give high accuracy for most classes (near 80 %) and the confusion, when occurs, is understandable. For example, in tight close-up shots, it is difficult to distinguish between top zoom-in and ground zoom-in camera angles (the background is blurred in both cases and only the player profile contributes to the features).

4.2 Event tagging

For event tagging experiment, we created a dataset of 176 clips encompassing all the five events. The dataset includes 43 goal sequences, 30 corner sequences, 42 substitution sequences, 27 fouls sequences and 34 gameplay sequences. The clips were extracted from the video by manually annotating the start and the end of the event. The length of each sequence is different and is defined based on the knowledge of the game. For example, a goal event starts from the point the ball enters the goal post and ends at the new kickoff (so the goal event includes all the celebrations). Similarly, a corner event starts just before the corner kick and ends at the first deflection. The sequences for gameplay event were of random length. We used 60 % of the data for training/validation and 40 % for testing. Example videos of each kind of event with qualitative results are provided in the supplementary material. Another qualitative result on an interesting 30-min sequence (with five goals in quick succession) is shown in Fig. 1.

For quantitative analysis, we sampled test data clips from different classes and joined them in random order to create a large video sequence. The event classification task was then performed per-frame basis on this large sequence. We tested classification task using both the mean (taking mean of features from 40 frames) and concatenation (concatenating features from 40 frames). The average accuracy of around 80 % was obtained in both the cases, which improved to 85.5 % after MRF smoothing.

The confusion matrix of results after MRF smoothing (using mean features from 40 frames) is given in Table 6. The goal event and gameplay event are predicted with an accuracy over 85 %. The confusion, when occurs, is understandable; for example, some corner events lead into goals or near misses and in such instances the corner event is misclassified as goal event. Similarly, substitution event is often followed with the crowd viewpoint (cheering the player) and replay of goals (or assists) by substituted player, which leads to confusion between substitution and goal event.

Comparing results in Tables 4, 5 and 6, we can observe that both context(camera label information) and motion features bring significant improvements in event recogni-

Table 4 Event tagging results using only the BoW histograms considering five different events (percentages)

	Corner	Foul	Goal	Gameplay	Subst.
Corner	**43.4**	4.9	31.3	9.8	10.6
Foul	0	**48.5**	32.9	13.7	4.9
Goal	2.5	8.9	**82**	4.5	2.1
Gameplay	3.9	4.2	5.8	**85.2**	0.9
Subst.	1.6	25.5	41.1	0.7	**31.1**

Bold values denote per class classification accuracies

Table 5 Event tagging results using the combination of camera label information with the BoW histograms

	Corner	Foul	Goal	Gameplay	Subst.
Corner	**71.2**	0	28.8	0	0
Foul	0.5	**43.8**	40.2	8	7.5
Goal	0.6	8.7	**81.2**	4.2	5.3
Gameplay	3.4	3	2.4	**91.2**	0
Subst.	0	12.7	32.7	0	**54.6**

Bold values denote per class classification accuracies

Table 6 Event tagging results using the combination of BoW histograms, camera label information and the motion features

	Corner	Foul	Goal	Gameplay	Subst.
Corner	**75.1**	5.4	18	0.6	0.9
Foul	0.2	**57.4**	22	10.4	10
Goal	2.7	3.4	**84.6**	4.2	5.1
Gameplay	0.9	1.5	1.5	**94.6**	1.5
Subst.	1.3	9.6	24	2.7	**62.4**

Bold values denote per class classification accuracies

tion task. For example, the accuracy of predicting corner improves by almost 30 % after including camera label information (Tables 4, 5). Similarly, including motion information brings almost 15 % improvement in predicting foul event (Tables 5, 6). Overall, the average accuracies increased from 77 to 81.8 % after adding camera context. The average accuracy further improved to 85.5 % after the inclusion of motion features. In terms of time complexity, the initial method using only BoW histograms runs at around 1 fps in our current implementation on a single-core Intel-i5 CPU with 16GB memory. Adding camera label information does not bring any significant change, but including motions cues increases computational time to around 0.4 fps.

We further compare our method with a hidden conditional random field (HCRF)-based method [17] and a recent action recognition using trajectory-pooled deep-convolutional descriptors [21] (with combined spatial and temporal features). For comparison with [17], we use our own implementation for computing features with sequence classification toolbox from [20]. The results with precision for predicting each event are presented in Table 7. We can clearly observe that our approach outperforms the generic state-of-the-art action recognition method [21], which shows the importance of the context. Our method also improves the precision over a more sophisticated (using ad hoc domain-specific features) sequence classification method [21] on most of the events.

Table 7 Comparison of our method with TDD [21] and HCRF [17]

	Corner	Foul	Goal	Gameplay	Subst.
TDD	21.05	12.82	77.45	77.08	25.80
HCRF	7.6	15.5	6.6	77.13	**86.44**
Ours	**75.1**	**57.4**	**84.6**	**94.6**	62.4

Bold values denote highest accuracy per class across the three algorithms

4.3 Spatial segmentation

In this section, we demonstrate the usefulness of camera label information with an application of spatial segmentation. Given an individual frame, the task is to assign each pixel with an unique class label. We consider three different classes, i.e., players, crowd and playing field. The experiments are performed on a dataset of 50 images with top zoom-out camera labels and 50 images with zoom-in camera labels (sampled both from ground zoom-in and top zoom-in cameras). The images are equally sampled from 10 different matches to cover different challenges in segmentation like shadows, different color player jerseys, etc. The ground truth labels for all the 100 images are created manually.

We investigate two different segmentation approaches. First, we follow the class-based image segmentation approach of Ladicky et al. [12], which casts the segmentation problem as graph cut-based inference on conditional random fields. Second, we propose a segmentation approach specific to top zoom-out views. Our main insight is that knowing the context can help to design minimal segmentation algorithms bringing significant improvement in terms of both time and labeling accuracy. Knowing that the camera label is a top zoom-out, we can assume that the large part of the frame will be covered by playing field, which can be efficiently segmented based on color. We use a variation of Heckbert's [6] median-cut algorithm to estimate dominant color and segment out the playing field in normalized RGB space. The segmentation is performed using a threshold on Euclidean distance from the dominant color. The holes in the playing field are then labeled as players, and the rest is labeled as crowd.

We then trained three instances of automatic labeling environment (ALE) [12]: first for the top zoom-out camera labels, second for the zoom-in camera labels and third for the combined set. Half of the total images were used for training in each case. The ALE segmentation results for top zoom-out and zoom-in are illustrated in Table 8. We can observe

Table 8 Results of average per class recall measure, defined as $\frac{\text{True Positives}}{\text{True Positives + False Positives}}$ on ALE [12]

	ALE top zoom-out	ALE zoom-in
Players	44.4	80.8
Field	99.31	93.3
Crowd	99.8	91.5

The recall measure for class *Players* is low in top zoom-out viewpoints

that ALE fails to segment players accurately in images with top zoom-out camera labels. Interestingly when we trained ALE by taking frames with from both zoom-out and zoom-in viewpoints, we obtained nearly same results. This suggests that ALE is not taking full advantage of the context. We then performed segmentation using the second context-aware approach (based on dominant color) and the recall for the class *Players* improved to 64.2 % maintaining nearly the same recall for *Field* and *Crowd* classes. The results from the dominant color-based approach were obtained in real time compared to 30-h training (for 50 images with image resolution of 720p) and 3-h testing (4–5 min per frame) in case of ALE. This clearly shows that knowing context (like camera labels) can bring significant improvements for the task of spatial segmentation in terms of both performance and recall.

5 Conclusions

In this paper, we have investigated the problem of automatic analysis of football broadcast videos. We have shown that this problem can be partitioned into two smaller problems, namely camera viewpoint estimation and event tagging. We have demonstrated that since the input videos are already edited, the camera viewpoint information provides a natural context, which could be exploited to improve the other task of event tagging.

Based on thorough quantitative analysis on variety of tasks in 10 football matches, we have justified our claims. Our method obtains an overall accuracy of 92.2 % for camera viewpoint estimation and 85.5 % accuracy for event classification. We have also demonstrated that the contextual approach can outperform state-of-the-art deep-learning-based action recognition approaches. We further demonstrate that the accuracy of tasks like spatial semantic segmentation can be improved by as much as 20 % using the context.

References

1. Assfalg, J., Bertini, M., Colombo, C., Bimbo, A.D., Nunziati, W.: Semantic annotation of soccer videos: automatic highlights identification. CVIU **92**, 285–305 (2003)
2. Chen, C., Wang, O., Heinzle, S., Carr, P., Smolic, A., Gross, M.: Computational sports broadcasting: automated director assistance for live sports. In: ICME (2013)
3. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV, vol. 1, pp. 1–2 (2004)
4. Duan, L.Y., Xu, M., Tian, Q., Xu, C.S., Jin, J.S.: A unified framework for semantic shot classification in sports video. Multimed. IEEE Trans. **7**(6), 1066–1083 (2005)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. TPAMI **32**(9), 1627–1645 (2010)
6. Heckbert, P.: Color image quantization for frame buffer display. In: SIGGRAPH (1982)
7. Jain, V., Singhal, A., Luo, J.: Selective hidden random fields: exploiting domain-specific saliency for event classification. In: CVPR (2008)
8. Kapela, R., McGuinness, K., Swietlicka, A., OConnor, N.E.: Real-time event detection in field sport videos. In: Computer vision in Sports, pp. 293–316 (2014)
9. Kim, K., Grundmann, M., Shamir, A., Matthews, I., Hodgins, J., Essa, I.: Motion fields to predict play evolution in dynamic sport scenes. In: CVPR (2010)
10. Kong, Y., Hu, W., Zhang, X., Wang, H., Jia, Y.: Learning group activity in soccer videos from local motion. In: ACCV (2010)
11. Koprinska, I., Carrato, S.: Temporal video segmentation: a survey. Signal Process.: Image Commun. **16**(5), 477–500 (2001)
12. Ladicky, L., Russell, C., Kohli, P., Torr, P.H.: Graph cut based inference with co-occurrence statistics. In: ECCV (2010)
13. Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., Sheikh, Y.: Representing and discovering adversarial team behaviors using player roles. In: CVPR (2013)
14. Ma, Z., Yang, Y., Cai, Y., Sebe, N., Hauptmann, A.G.: Knowledge adaptation for ad hoc multimedia event detection with few exemplars. In: ACM Multimedia, pp. 469–478 (2012)
15. Nguyen, N., Yoshitaka, A.: Shot type and replay detection for soccer video parsing. In: Multimedia (ISM), 2012 IEEE International Symposium on, pp. 344–347 (2012)
16. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR (2011)
17. Qian, X., Liu, G., Wang, Z., Li, Z., Wang, H.: Highlight events detection in soccer video using hcrf. In: Proceedings of the Second International Conference on Internet Multimedia Computing and Service, pp. 171–174 (2010)
18. Sigari, M.H., Soltanian-Zadeh, H., Kiani, V., Pourreza, A.R.: Counterattack detection in broadcast soccer videos using camera motion estimation. In: AISP, pp. 101–106 (2015)
19. Thompson, R., Bowen, C.: Grammar of the Edit. Focal Press, Massachusetts (2009)
20. Walecki, R., Rudovic, O., Pavlovic, V., Pantic, M.: Variable-state latent conditional random fields for facial expression recognition and action unit detection. In: Automatic Face and Gesture Recognition (2015)
21. Wang, L., Qiao, Y., Tang, X.: Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR (2015)
22. Xie, L., Chang, S.F., Divakaran, A., Sun, H.: Structure analysis of soccer video with hidden Markov models. In: ICASSP (2002)
23. Xu, C., Wang, J., Wan, K., Li, Y., Duan, L.: Live sports event detection based on broadcast video and web-casting text. In: ACM Multimedia, pp. 221–230 (2006)
24. Xu, G., Ma, Y.F., Zhang, H.J., Yang, S.Q.: An hmm-based framework for video semantic analysis. Circuits Syst. Video Technol. IEEE Trans. **15**(11), 1422–1433 (2005)
25. Xu, P., Xie, L., Chang, S.F., Divakaran, A., Vetro, A., Sun, H.: Algorithms and system for segmentation and structure analysis in soccer video. ICME **1**, 928–931 (2001)