Event Recognition in Broadcast Soccer Videos

Himangi Saraogi Rahul Anand Sharma Vijay Kumar
Center for Visual Information Technology, IIIT Hyderabad
{himangi.saraogi@students, rahul.anand@research, vijay.kumar@research}.iiit.ac.in

ABSTRACT

Automatic recognition of important events in soccer broadcast videos plays a vital role in many applications including video summarization, indexing, content-based search, and in performance analysis of players and teams. This paper proposes an approach for soccer event recognition using deep convolutional features combined with domain-specific cues. For deep representation, we use the recently proposed trajectory based deep convolutional descriptor (TDD) [1] which samples and pools the discriminatively trained convolutional features around the improved trajectories. We further improve the performance by incorporating domain-specific knowledge based on camera view type and its position. The camera position and view type captures the statistics of occurrence of events in different play-field regions and zoom-level respectively. We conduct extensive experiments on 6 hour long soccer matches and show the effectiveness of deep video representation for soccer and the improvements obtained using domain-specific cues.

CCS Concepts

•**Computing methodologies** → **Activity recognition and understanding;** *Video segmentation;*

Keywords

Soccer event recognition, deep convolutional features, playground registeration, view-shot estimation.

1. INTRODUCTION

There is an explosion in the amount of videos generated in the recent years. This proliferation has increased the need for developing video analysis techniques that can understand, summarize and analyze the video content. Such analysis also help in applications such as content-based search in videos, human-computer interaction, video surveillance, etc.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICVGIP, December 18-22, 2016, Guwahati, India
© 2016 ACM. ISBN 978-1-4503-4753-2/16/12…$15.00
DOI: http://dx.doi.org/10.1145/3009977.3010074

Sports video analysis [9–11, 19], particularly soccer videos is getting a lot of attention due to the massive popularity of the game. Soccer video broadcasters often want to improve the viewing experience and provide interesting game analysis to millions of viewers. However, dealing with sports videos which are often lengthy containing large portions of uninteresting events is a non-trivial task. There are additional difficulties seen in sports videos due to intra-class variations, background clutter, viewpoint change, camera motion, blur, resolution and zoom-level. It thus becomes necessary to develop algorithms that can automatically detect and analyze interesting events to enable intelligent game summarization, performance analysis of the players and browsing based on semantic analysis, e.g., highlight detection, tactics analysis, player tactics, etc.

In this paper, we propose an approach for event recognition in soccer videos. Our approach shown in Figure 1 aims at exploiting multiple cues to make a prediction. Specifically, our classifier is a linear combination of three classifiers each of them is trained to learn different aspects of the video to make a prediction. The first classifier is based on the appearance and motion cues. Inspired by the success of discriminatively trained convolutional features over hand-crafted features for action recognition [1, 15, 17, 18], we propose to learn convolutional features for capturing appearance and motion cues. We use the recently proposed trajectory pooled convolutional descriptor (TDD) [1] for its state-of-the-art performance.

The remaining two classifiers capture domain-specific cues based on active playground region and view shot type. The region in the playground where the players are active gives a dominant cue regarding the event. For instance, the corner events usually happen near the corner regions of the playground, and similarly goal event occurs near the goal. Given a video frame, we first register the video frame with static playground template to identify the playground location which is then used to predict the event. The registration approach builds a large dictionary of images, field line images and their homography. In a given frame, non-ground regions including crowd and players are filtered, edge maps are extracted which are searched over the dictionary using nearest neighbor search.

The final classier is based on the shot view type information. The game is recorded at different zoom-levels in order to give better viewing experience. We observe that certain shot types dominate in particular events. For example, corner event is often shot in long view. We use view-type as an additional cue for predicting the events. We train an SVM

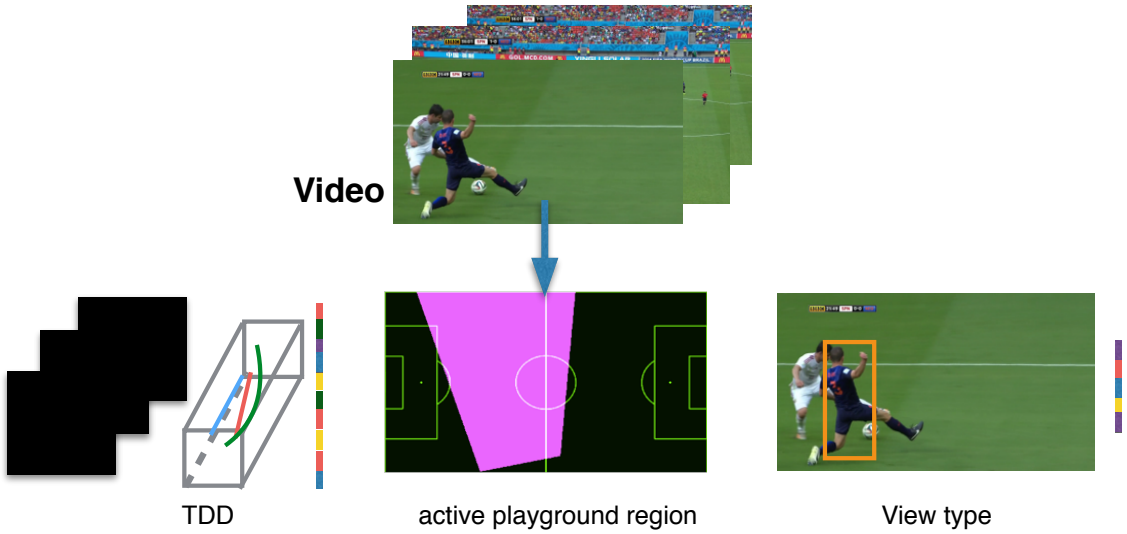


Figure 1: Our approach combines information from different classifiers to capture multiple semantic cues. Our baseline classifier is based on discriminatively trained deep convolutional features trained to learn appearance and motion cues. The performance of the baseline classifier is boosted by domain-specific cues based on playground position and view type.

classifier using multiple features including average channel colour, average size of the players, and the ratio of play-field region to non play-field region to predict the view type. A naive Bayes classifier is then trained to predict the event label. As we show in the experiments, these two domain-specific cues provide complementary information and greatly help in improving the performance of appearance and motion based classifier.

To summarize, we make the following contributions:

- We show the effectiveness of deep convolutional features for the task of event detection in sport videos.
- Our approach exploits domain specific information from active playground region, camera zoom type to improve the performance.
- Finally, we conduct extensive experiments on 6-hour long videos and show the effectiveness of our approach.

2. RELATED WORK

There are many sport analysis techniques proposed in the literature [7–12]. We briefly review few approaches. Jia *at al.* [19] propose an approach for player detection, tracking, and team labeling of the players. Qian *et al.* [23] use hidden markov models to classify event clips into five different categories like goal, shoot, normal etc. Previous approaches include both feature learning and heuristic rule based systems to detect events [10, 11]. These approaches perform low level analysis to detect marks (field, lines, logo, arcs, and goalmouth), player positions, ball position etc and then derive mid level features using these cues. In the end they learn a rule based system to detect salient events like goal, corner etc.

In [2], a technique for classifying the view shot is proposed. They consider the frame-wise color values of each pixel in the HSV color space along with the object size within the segmented play-field region. In [8], an approach for replay detection in soccer is proposed. They first perform a shot classification and then a scene transition structure analysis on the generated shot label sequence to extract the replay scenes. In [3], a technique for multiple soccer analysis such as shot boundary, goal, referee and penalty-box detection is proposed. The approach uses both low-level and high level features for detection. The low-level features include dominant color region detection, robust shot boundary detection, and shot classification while the high level information is obtained through goal detection, referee detection, and penalty-box detection. In [6], only goals are detected. The goal events are inferred by estimating the ball position and comparing it with respect to the location of the goalpost. The performance of algorithm depends on ball detection and is not suited for detecting events such as pass, fall which are not related to goal post.

Our work is also to related to previous action recognition approaches. Most of the action recognition approaches are focused on feature design and learning. Some of the features that have been applied for action recognition include Histogram of Gradients (HOG), SURF, Histogram of Optical Flow (HOF) [21], and spatio-temporal features such as dense trajectories [20], improved trajectories [14]. While the success of these approaches depend on the feature design choice, the deep convolutional approaches [15,17,18] on the other hand learn the discriminative features from the training examples.

Inspired by success of deep learning in various vision tasks, we use the deep convolution features for capturing appearance and motion cues which are vital for event recognition in videos. To our best knowledge, ours is the first work that evaluates the convolutional features for sport analysis. Unlike previous soccer analysis approaches that use several heuristics for detecting events, we adopt a learning strategy which provides flexibility to extend to other related sports such as ice hockey. We also show how domain-specific cues when incorporated can improve the recognition performance.

3. PROPOSED APPROACH

We introduce a novel approach for event recognition in soccer videos which combines information of different classifiers. It consists of three components:

- The *generic event recognition classifier* using the discriminatively trained convolutional feature descriptor [1]. The CNN is discriminatively trained on the soccer videos to obtain the convolutional feature maps which are then pooled using the video trajectories.
- The *active play-ground region classifier* using the ground position where the event is occurring.
- The *view shot classifier* using the zoom level information of the video frames.

While generic event recognition classifier captures the discriminative appearance and motion information, view-type and active play-ground position provide domain-specific cues which further boost the performance. Given a video x, we predict the event using a linear combination of the predicted probabilities of the above classifiers

$$s(x, y) = \sum_i w_i P_i(y|x), \quad \forall i \in \{1, 2, 3\} \tag{1}$$

where $P_i(y|x)$ is the normalized probability of the event y given by the i-th classifier and w_i is its associated weight which are estimated using cross-validation. Finally, the event can be predicted using $\hat{y} = \arg\max_y s(x, y)$. Below we give an overview of three classifiers.

3.1 Event Recognition Classifier

The event recognition classifier aims at learning the salient and informative regions of the video that help in event understanding. This is usually achieved using a video representation that capture these informative regions. In principle, any kind of video representation proposed in the literature for action recognition can be used. In our work, we use the recently proposed trajectory-pooled deep convolutional descriptor (TDD) [1] for its superior performance.

The TDD descriptors are derived from a combination of discriminatively trained deep convolutional features and improved trajectories [14]. The CNN features are constraint-pooled using the improved trajectories. Since the proposed approach is based on TDD descriptor, we briefly review its steps below:

- *Trajectories:* TDD representation uses the improved trajectories [14] for trajectory extraction due its superior performance. Improved trajectories generate the trajectories by densely sampling the points at multiple scales on an image grid. The sampled points are then tracked by median filtering of dense flow field [20].

  $$\mathcal{P}_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (\mathcal{M} * \omega_t)|_{(\hat{x_t}, \hat{y_t})}, \tag{2}$$

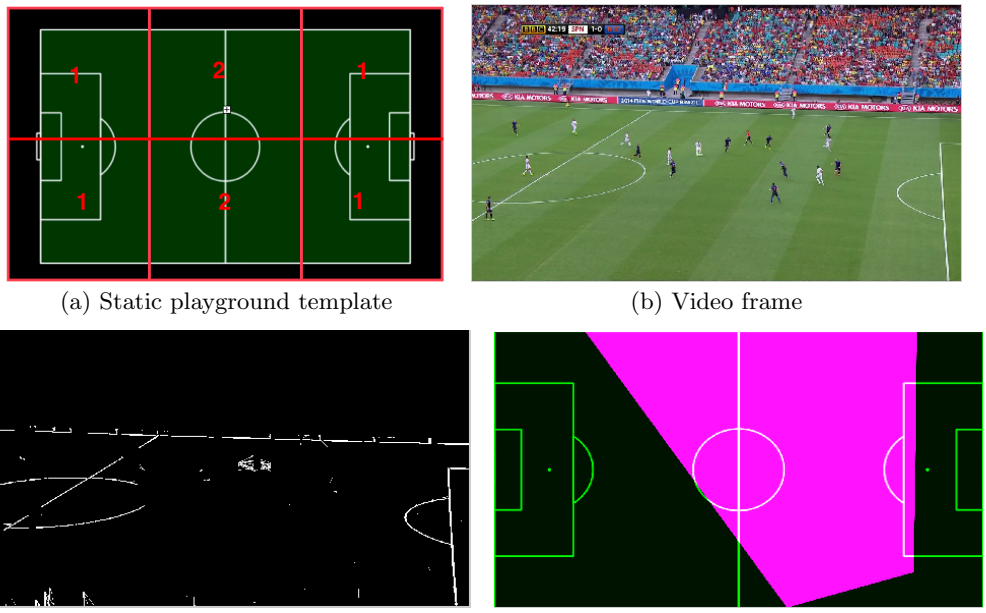
  where $\mathcal{M}$ is the median filtering kernel, $\omega = (u_t, v_t)$ is the dense optical flow field of the t-th frame, and $(\hat{x_t}, \hat{y_t})$ is the rounded position of (x_t, y_t). The improved trajectories also considers the camera motion into account. It uses traditional SURF feature matching technique to find point correspondences between two consecutive frames and estimates the homography matrix. The homography matrix is used to remove the camera motion and re-calculate the optical flow. For efficiency, TDD tracks the points only in the single original spatial scale and extracts multi-scale TDDs around the extracted trajectories. To avoid drifting problem of tracking, the maximum length of trajectory is set as 15 frames.
- *Convolutional feature maps:* TDDs uses two-stream deep convolutional architecture [15] which contain two separate convolutional networks (convNets) namely spatial nets and temporal nets. The spatial nets are designed to capture the static appearance cues by training on individual image frames while temporal nets are designed to learn the dynamic motion information, whose input are volumes of stacked optical flow fields. We use the default architecture described in [1] for training the convNets on soccer videos. The architecture consists of alternating convolution and pooling layers followed by fully connected layers. Once the discriminate training is complete using the labeled video examples, the target layers are discarded and only final convolutional feature maps are retained. Given a video, a set of spatial and temporal feature maps are obtained which are then pooled around the trajectories to produce TDD descriptor for the video.
- *Pooling:* Once the convolutional feature maps are obtained, they are normalized and pooled around the trajectories. Feature maps are normalized across each channel independently to ensure all the feature maps lie in the same interval. Similarly, pixels across the feature channels are normalized to lie in the same range. Finally, the TDD descriptor is obtained by sum pooling of the normalized feature maps over the 3D volume centered around the trajectory.

We refer the soft-max scores of the TDD network which indicate event probabilities as $P_1(y|x)$.

3.2 Camera-position Classifier

During the game the cameras are focused to the region of the playground where the game is active. For instance during a corner event the camera is focused towards the corner region of the playground at which the game is active. The position of the playground can provide a vital cue regarding the on-going event. Given a video clip we identify the playground region by registering its frames to the static playground template and predict the event based on the identified playground region.

We use simple edge based features and formulate the registration problem as nearest neighbor search to the closest edge map in a precomputed dictionary with known projective transforms. During the training stage, we prepare a large dictionary of frame images, compute their edge maps and manually label their correspondences with the static template after which homography is computed. In order to reduce the manual labeling effort, we adapt a semi-supervised approach to increase the dictionary size by simulating large number of edge maps by varying the homographies of the labeled examples. Given a test frame, we extract the edge maps and compute the stroke width transform to filter out the strokes with size greater than 10 pixels. We remove the crowd region using dominant playground color, and players using person detector trained on players to obtain the



(a) Static playground template (b) Video frame

(c) Edge map (d) The registered video frame

Figure 2: **Playground registration:** Given a video frame (b), we first obtain the edge map (c) by extracting the edges and removing all the lines whose width is greater than 10 pixels. Non-playground regions including players and audience are filtered out by thresholding using dominant color of the ground. The HoG features are extracted from the edge map and the homography is obtained using a nearest neighbor search over a large dictionary of edge maps with known homography. The homography is then used to register the given frame (d) to the fixed playground template (a). The trained dictionary consists of large number of manually and synthetically generated edge maps and their corresponding homographies.

(a) Long view (b) Medium view (c) Closeup view (d) Out-of-field view

Figure 3: Different view shots in soccer broadcast videos. In order to provide better viewing experience, broadcasters capture different events with different camera zoom levels which provide an useful semantic cue for event recognition.

edge map primarily containing only the field lines. We then search for a nearest neighbor in the dictionary edge maps and assign its homography to the given frame using HoG features. We show the static playground template, input frame, edge map and its the registered region to the fixed template in Figure 2.

For event prediction using the playground region, we divide the entire ground into 6 bins as shown in Fig 2(a). These 6 bins are labeled as either as 1 and 2 due to the playground symmetry. Given a registered frame, we find the number of pixels in each of these 6 bins. We consider the label of top three bins (b_1, b_2, b_3) that have maximum number of pixels as a feature for predicting the event type. Given a training set, we estimate the distribution of each event given all possible bin features as shown in Table 1. During testing, we first compute the bin feature of the registered frame and then compute its event score $P_2(y|x)$ using naive Bayes classifier. The score for entire video is obtained by calculating the scores for every 20th frame and computing their average score.

3.3 View shot classifier

In order to produce better viewing experience, broadcasters often capture the game using different shots with varying camera zoom level. As shown in Figure 3, four different types of shots namely long, medium, close-up, and out-of-field (audience) are commonly seen in broadcast soccer videos. We also observe that certain types of events are captured in particular zoom-level. For example, corner event is captured in long view with greater probability. Similarly a pass or dribble events are mostly shot at medium zoom-level. We show the distribution of 5 different events and their view type in Table 2. Based on this observation, we try to incorporate the semantic information provided by view-shots to improve the event recognition performance.

In order to incorporate view-shot type, we need to identify the view-shot of the given video clip. There are many methods available in the literature for shot classification [2,3,8,9]. In this work, we adapt a simple strategy for estimating the shot type and then use it to predict the event label. We use the average color of each channel, average size of the

Bin feature (b_1, b_2, b_3)	Corner	Dribble	Pass	Goal	Fall
(1,1,2)	0.30	0.18	0.12	0.31	0.09
(1,2,1)	0.37	0.18	0.12	0.14	0.19
(1,2,2)	0.21	0.27	0.21	0.10	0.21
(2,1,1)	0.12	0.10	0.26	0.23	0.29
(2,2,1)	0.09	0.20	0.31	0.26	0.14
(2,1,2)	0.03	0.27	0.22	0.18	0.30

Table 1: The statistics of occurrence of events in different playground region. The bin feature is a $3D$ vector containing the labels (1 or 2) of the top 3 regions (out of 6) of ground that contain the maximum number of pixels after registration.

	Corner	Dribble	Pass	Goal	Fall
long-view	85	50	39	16	68
medium-view	63	107	99	59	62
close-up view	51	43	60	123	70
out-of-field	1	0	2	2	0

Table 2: The statistics of training set showing the occurrence of events in different view shot type.

players, and the ratio of play-field region to non play-field region as features to determine the shot type of each frame. Given training examples, we train an SVM classifier that outputs the shot probability score $p(v|x)$ where v is the shot type and x is the video frame. To determine the event label y, we train a naive Bayes classifier which outputs the class-conditional probabilities $P_3(y|x)$ as,

$$P_3(y|x) = P(y|v)P(v|x).$$

The view type conditional label probabilities $p(y|v)$ are estimated from the training set using $P(y|v) = \frac{P(v|y)P(y)}{\sum_y P(v|y)P(y)}$. For computational simplicity we find the score for every frame and assign their average score for entire video.

4. EXPERIMENTS

In this section, we first describe the dataset, implementation details and then give the experimental results of our approach.

4.1 Dataset

Since there is no publicly available soccer dataset for event recognition, we created a dataset for 5 events namely *corner*, *dribble*, *pass*, *goal* and *fall*. Our training set consists of 1000 video clips collected from various FIFA 2014 world cup matches. We collected a total of 200 samples for each event. When collecting the clips, we ensured that dataset contains clips from diverse playing conditions. The average length of a clip is about 5 secs. For testing, we created a separate dataset consisting of two complete soccer matches totaling about 4 hours of video. All the video clips have a resolution of 1280×716 pixels.

4.2 Implementation details

We use the implementation of [1] for extracting the TDD descriptors. We use their default architecture which is trained on UCF101 [22] dataset with 101 action classes each containing atleast 100 examples. We pre-train the network with their model which are then fine-tuned on our video dataset. We use their default parameters for descriptor generation. For playground registration, we selected 200 images from the top zoom-out images, manually label the four point correspondences and synthesized a large dictionary of 60000 edge maps. We compute the homography matrices using RANSAC algorithm. We perform registration and class score computation for every 20-th frame and then compute their average scores. Similarly we perform view-type estimation and class score computation on every frame and compute their average scores.

Window size	Overlap	Accuracy
15	1	56.27%
15	5	64.12%
30	1	62.77%
30	5	61.91%
100	1	66.23%
100	10	68.12%
100	50	65.45%
200	1	74.1%
200	**10**	**74.45%**
200	50	71.2%
500	10	68.29%

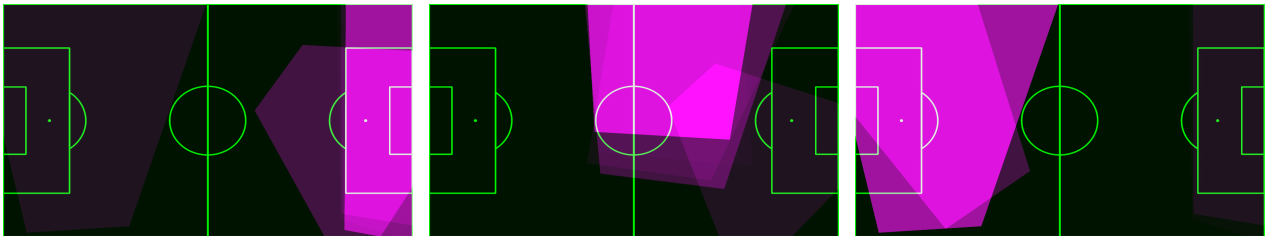
Table 3: The performance for our approach for different sliding window sizes and overlap. Our top performance is achieved for window size of 200 and overlap of 10 frames.

The distribution of events in different playground regions represented by top 3 bins with maximum number of pixels is shown in Table 1. It is used to compute the probability $P_2(y|x)$. We show an heat map that shows the occurrence of different events in different regions of the playground in Figure 4. Higher the intensity, greater is the probability of event occurring in that region. Note how goal and corner events occur near the goal and corner regions of the playground, and events such as dribble and pass mostly occur near the mid-portion of the playground. Similarly, the distribution of the training set for different view types which is used to compute the probabilities $P(v|y)$ and $P(y|v)$ is shown in the Table 2.

4.3 Results

We show the results of our approach in Table 4. We compare our approach with previous features proposed for action recognition such as HOF [21], improved dense trajectories (IDT) [14] and two-stream convnets [15]. We also measure the accuracies using conv4 and conv5 features of spatial nets and conv3 and conv4 for temporal nets in TDD. It is clear that the baseline classifier which uses TDD descriptor outperforms all the previously proposed features for event recognition in sports. The performance is further boosted by the inclusion of domain-specific cues based on view-type and playground region information achieving an accuracy of 74.5%.

Since the testing videos are full matches, the choice of window size and stride affects the performance. In Table 3, we show accuracies obtained for different window sizes and overlap length. The window size of 200 frames with an overlap of 10 gives the best accuracy. Finally, we visualize the events over time for a test video segment of 20 min duration in Figure 5. Our approach based on TDD descriptor lo-



(a) Goal (b) Dribble (c) Corner

Figure 4: The heat map shows the occurrence of different events in different regions of the playground. Higher the intensity greater is the probability of event occurring in that region. While goal and corner events occur near the goal and corner regions of the playground, events such as dribble and pass mostly occur near the mid-portion of the playground. Note that since the playground is symmetric, we map the distribution to only one side of the playground. (best seen in colour)

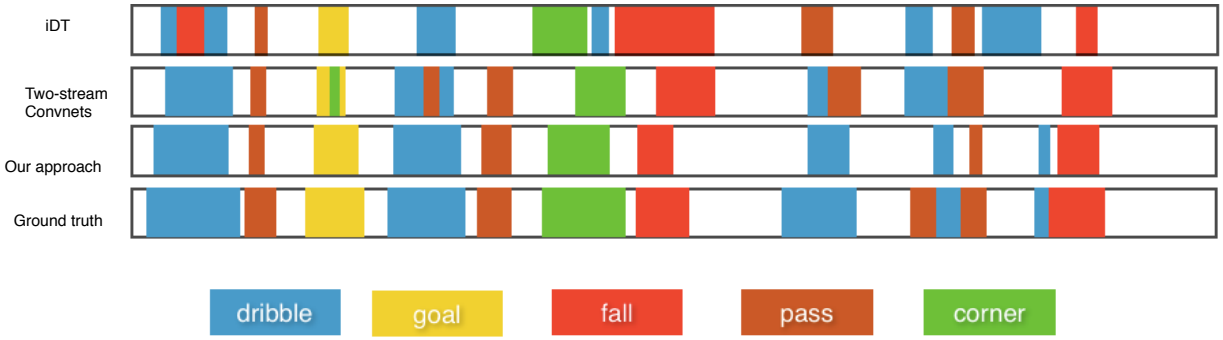


Figure 5: Visualization of events on a 20 minute video segment. It is clear that our approach achieves better localization and recognition compared to previous action recognition approaches - Improved trajectories [14] and Two-stream ConvNets [15].

Features	Baseline (B)	(B) + View type (V)	(B) + (V) + Playground region
HOG [14]	53.20%	55.27%	56.92%
HOF [21]	58.76%	60.48%	61.43%
MHB [14]	64.43%	65.74%	67.23%
HOF+MHB [14]	65.93%	69.32%	69.43%
IDT [14]	64.30%	71.33%	72.10%
Spatial net	56.21%	59.32%	61.45%
Temporal net	62.82%	64.45%	65.30%
Two-stream ConvNets [15]	65.33%	68.55%	71.34%
Spatial conv4	60.22%	62.37%	63.25%
Spatial conv5	59.27%	61.25%	61.88%
Spatial conv4 and conv5	61.87%	65.99%	67.43%
Temporal conv3	68.21%	71.88%	73.08%
Temporal conv4	63.32%	67.43%	68.35%
Temporal conv3 and conv4	71.12%	72.14%	74.05%
TDD	71.76%	73.59%	**74.45**%

Table 4: Comparison of our proposed approach with previous action recognition approaches. Our baseline classifier based on TDD performs better than improved trajectories [14] features and the two-stream ConvNets [15]. The performance is further boosted using domain-specific cues based on view-type and playground information. The table also shows the performance of the features obtained from conv3, conv4 and conv5 layers of convNet.

calizes better than two-stream convnets [14] and improved dense trajectory features [15]. In order to reject the non-event windows, we use the final SVM scores as a confidence score. If the ratio of two maximum class scores is less than 0.35, we consider the window as non-event class.

5. CONCLUSION

In this paper, we propose a novel approach for event recognition in soccer videos that combines information from different classifiers. The first classifier is based on deep convolutional features trained to learn appearance and motion cues. The remaining two classifier provide soccer specific

cues based on camera view type and playground region. We conduct extensive experiments on 4 long hour soccer videos and show the effectiveness of the baseline CNN based classifier and the drastic improvements obtained by incorporating domain-specific cues.

Acknowledgment

The authors would like to thank C. V. Jawahar for helpful suggestions. Vijay Kumar is supported by TCS PhD fellowship.

6. REFERENCES

[1] L. Wang, Y. Qiao, and Xiaoou Tang, Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors, In CVPR, 2015

[2] A. A. Halin, M. Rajeswari and D. Ramachandram, Shot view classification for playfield-based sports video,In ICSIPA, 2009.

[3] Ekin A., Tekalp A.M., and Mehrotra R., Automatic Soccer Video Analysis and Summarization. IEEE Transactions on Image Processing, 2003.

[4] Zhao Z., Jiang S., Huang Q., and Ye Q., Highlight Summarization in Soccer Video Based on Goalmouth Detection. Asia-Pacific Workshop on Visual Information Processing, 2006.

[5] Pan H., Li B., and Sezan M., Automatic Detection of Replay Segments in Broadcast Sports Programs by Detection of Logos in Scene Transitions. In ICASSP, 2002.

[6] Ancona N., Cicirelli G., Branca A., and Distante A., Goal Detection in Football by Using Support Vector Machines for Classification, In IJCNN, 2001.

[7] D'Orazio T. and Leo. M., A Review of Vision-based Systems for Soccer Video Analysis. Pattern Recognition, 2010.

[8] Jinjun Wang, Engsiong Chng and Changsheng Xu, Soccer replay detection using scene transition structure analysis, In ICASSP, 2005.

[9] Chung-Lin Huang, Huang-Chia Shih and Chung-Yuan Chao, Semantic analysis of soccer video using dynamic Bayesian network, In MM, 2006.

[10] A. Ekin, A. M. Tekalp, and R. Mehrotra, Automatic soccer video analysis and summarization, IEEE Transactions on Image Processing, 2003.

[11] D. W. Tjondronegoro, and Y. P. Chen, Knowledge-discounted event detection in sports video, IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2010.

[12] Rui Y., Gupta A., and Acero A.,Automatically Extracting Highlights for TV Baseball Programs. In MM, 2000.

[13] Leonardi R. and Migliorati P., Semantic Indexing of Multimedia Documents. In MM, 2002.

[14] Heng Wang, Cordelia Schmid. Action Recognition with Improved Trajectories. In ICCV 2013.

[15] Karen Simonyan, Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. *NIPS*, 2014.

[16] Lauer, Fabien, and Gerard Bloch. Incorporating prior knowledge in support vector machines for classification: A review. Neurocomputing, 2008.

[17] Ji, Shuiwang, Wei Xu, Ming Yang, and Kai Yu. 3D convolutional neural networks for human action recognition. In PAMI, 2003.

[18] Karpathy, Andrej, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In CVPR,2014.

[19] Liu, Jia, Xiaofeng Tong, Wenlong Li, Tao Wang, Yimin Zhang, and Hongqi Wang. Automatic player detection, labeling and tracking in broadcast soccer video. Pattern Recognition Letters, 2003

[20] Wang, Heng, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In CVPR, 2011.

[21] Laptev, Ivan, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In CVPR, 2008.

[22] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[23] Qian, Xueming, Huan Wang, Guizhong Liu, and Xingsong Hou. HMM based soccer video event detection using enhanced mid-level semantic. Multimedia Tools and Applications, 2012.