Automatic Statistics Extraction for Amateur Soccer Videos

Jan C. van Gemert — **University of Amsterdam**
John G.M Schavemaker — **TNO**
Koen Bonenkamp — **University of Amsterdam / TNO**

Abstract

Amateur soccer statistics have interesting applications such as providing insights to improve team performance, individual coaching, monitoring team progress and personal or team entertainment. Professional soccer statistics are extracted with labor intensive expensive manual effort which is not realistic for amateur matches. In this paper we develop a solution that automatically extracts action-related soccer statistics from a static camera pointed at the pitch. We implement a solution to player localization and action classification problem in human action recognition. Our method does not rely on player tracking, sliding windows, super voxels or construction of multiple hypotheses. Our work is developed with actual application in mind and a fully functional recognition pipeline is implemented, specifically tailored to meet the inherent challenges of action-rich soccer video.

Introduction

Professional soccer matches are manually annotated by people watching matches. From these annotated matches statistics are extracted which are used to inform the public with match overviews and fun facts. Bookmakers use these statistics to determine the likelihood of a match outcome. Team managers use statistics of opponents to develop strategies as well as statistics of their own team to improve teamwork and individual players. The majority of such statistics is extracted from video manually and is thus time consuming and expensive. For professional games the value of the data justifies manual labeling of every match. For amateur soccer matches, in contrast, the manual annotation effort is neither profitable nor possible given the huge amount of amateur games for children, adults and veterans. In this paper we propose an automatic method for statistics extraction from amateur soccer matches by placing a fixed inexpensive camera on the side of the pitch. From this video data, our method is able to automatically localize players and recognized the action these players are performing.

Related Work

Dense trajectories Wang et al. [9] reason that the 2D spatial domain and 1D temporal domain in videos show different characteristics and that it is more intuitive to handle them in a different manner than via interest point detection in a joint 3D space [1,2,3]. Wang et al. [9] densely track interest points through video sequences using optical flow. They (re)introduce the motion boundary histogram descriptor [1] that offers invariance in the case of camera motion. Dense trajectories in combination with motion boundary histograms consistently perform well in comparative studies [1,9]. For our own application we closely follow the approach by Wang et al. [9], our work differs however in that we also treat localization of actions.

Localization-by-detection Recent efforts in recognizing actions from longer video sequences evaluate an action classifier at various segments of the video [4]. Such an exhaustive search is computationally expensive because it requires scanning both the spatial and temporal location, as well as various spatial and temporal scales. This yields a huge 5-dimensional search space, or even a 6-dimensional search space if the aspect ratio of bounding boxes is not fixed. Several solutions have been developed that effectively avoid the exhaustive search, by only evaluating a subset of spatio-temporal windows (hypotheses) [2, 4, 5]. Jain et al. [5] generate a set of hypothesis bounding-boxes moving in time, such temporal bounding boxes are denoted *tubelets*. Derpanis et al. [2] implement an efficient template matching algorithm to detect actions. Action bank [6] builds on this template matching technique to create a high-level representation of video by constructing many individual action detectors. Our method differs from these works as we compute a single solution (hypothesis) directly without generating multiple hypotheses

Method

In fig. 1 we show an overview of our approach. For classification our work follows the approach by Wang et al. [9] closely, our work differs however in that we also treat localization of actions.

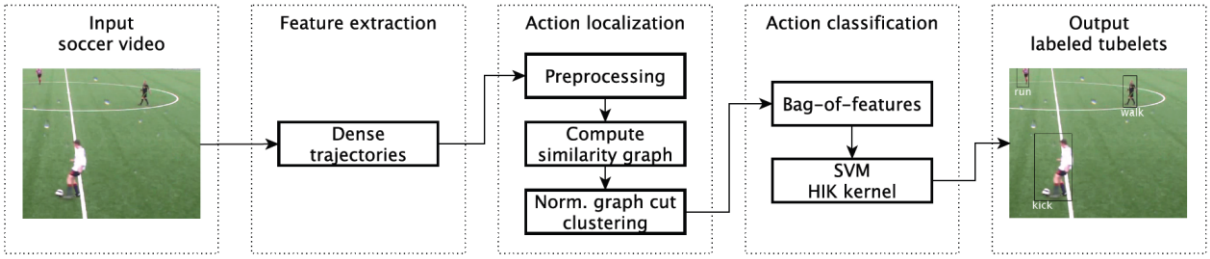

Figure 1: action localization pipeline

In action localization we cluster trajectories into groups that correspond to individual actions. In fig. 2 (left) we show raw trajectories. In fig. 2 (right) we show the ground truth obtained by manual player annotation of the video, the ground truth is the ideal clustering we aim to obtain.

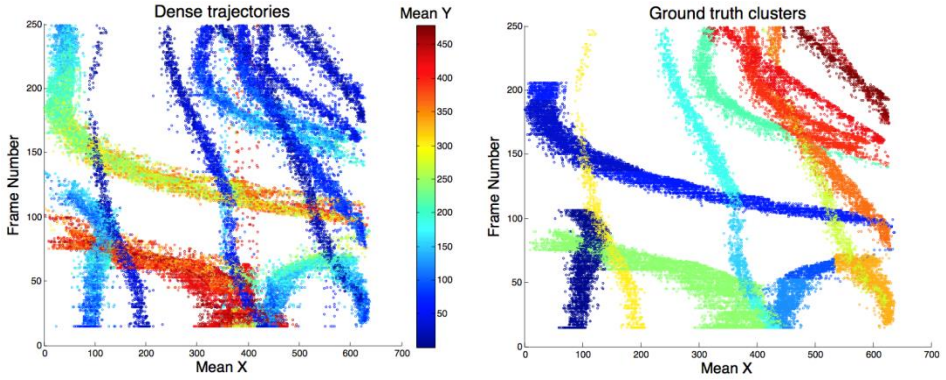

Figure 2: Dense trajectories for a soccer clip. The x-axis is the mean-x position for a trajectory, and the y-axis gives the frame number. **Left**: raw trajectories where colors denote the mean-y position. **Right**: Colors indicate a single player ground truth.

For automatic clustering we use normalized graph cuts [7], as this algorithm naturally allows for non-spherical clusters. The normalized graph cut algorithm [7] is a connectivity-based clustering technique based on a similarity graph. To construct the similarity graph we define the similarity between two trajectories as inversely proportional to the Euclidean distance between their feature vectors: (mean_x, mean_y, frame_nr, dx, dy), where dx and dy denote spatial displacement. In fig. 3 we illustrate results of the clustering on a video from the test set.

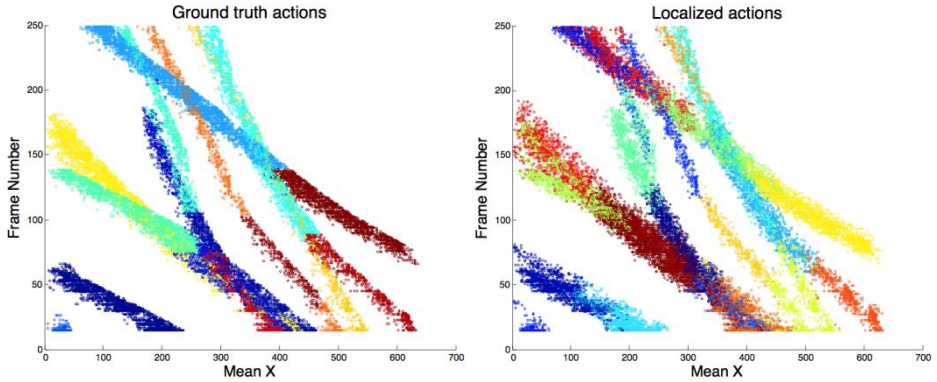

Figure 3: Localization results on a test-video where colors denote clusters. **Left**: Ground truth. **Right**: Predicted.

The normalized graph cut algorithm returns a clustering of trajectories. The final output of our action recognition pipeline should be a set of tubelets (Bounding-boxes over time). To obtain tubelets from clusters we simply create a bounding box for the minimum and maximum x and y per frame number. With these clustered tubelets we automatically label each of these clusters using a Support Vector Machine (SVM) classifier with a Histogram Intersection Kernel (HIK) on a Bag-of-Features (BoF) representation, see [1,3,9]. The BoF approach aggregates information from multiple trajectories into a single histogram of prototypes by allowing each trajectory to add a vote to a prototype. The prototypes are obtained by k-means as standardly done. The features here are standard action classification features [9]: Histograms of Oriented Gradietns (HOG), Histograms of Oriented optical Flow (HOF) and Motion Boundary Histograms (MBH), see [9] for details.

Experiments

Data description The video material is shot in full HD with three fixed cameras that together cover the whole soccer field. We recorded two amateur soccer matches at different locations. One match is used to train our models and the other match we keep for testing purpose only. We spatially crop the videos by taking 640x480 pixels around the middle circle of the field. In fig. 4 we show an example of the pitch. Because many actions repeat temporally [8], we split the soccer video into video clips of 10 seconds (250 frames). For the training set we selected 20 and for the test set 12 videos clips, each contain much activity.

We used Mind's Eye Annotation Software (v28) developed by DARPA to annotate the videos with ground truth tubelets. We identified 3 action labels that capture the majority of what players are doing on the field, those actions are walking, running and kicking (the ball). On average a video contains 20 individual actions (tubelets).

Figure 4: Example of the fixed camera on the amateur soccer pitch.

Results

In fig.5 (left) we show the classification with the confusion matrix showing promising results (79%) between the classes. In fig.5 (right) we illustrate the statistics that can be obtained as the ration between walking and running.

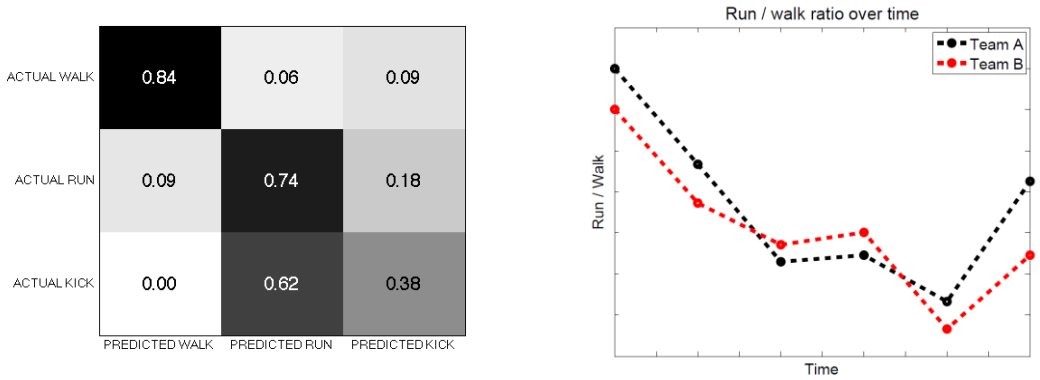

Figure 5: **Left**: Confusion Matrix, avg classification rate: 79%. **Right**: run/walk ratio statistics for the two teams.

In fig.6 we show a different type of statistics indicating the location of the performed actions as predicted.

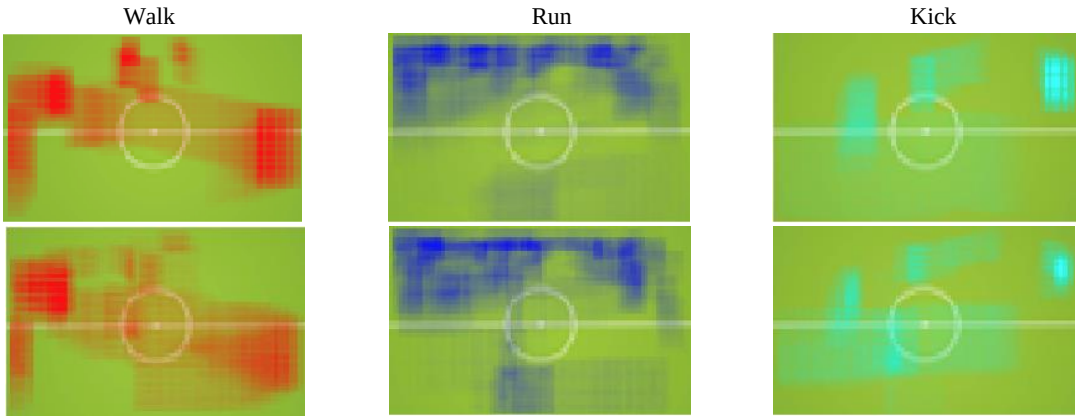

Figure 6: Heatmap of action locations. **Top**: groundtruth actions. **Bottom**: predicted actions.

Conclusions

We extract dense trajectories from video clips and cluster the trajectories using normalized graph cuts. The clusters are represented by bag-of-features histograms over the descriptors of trajectories. An SVM classifier predicts action labels for the histograms with an average accuracy of 79% on the test set. Our work is a first step towards fully automatic statistics extraction in amateur soccer.

References

[1] Dalal, N., Triggs, B., Schmid, C., (2006) Human detection using oriented histograms of flow and appearance. *European Conference on Computer Vision (ECCV)*, 2006

[2] Derpanis, K., Sizintsev, M., Cannons, K., Wildes, R., (2010) Efficient action spotting based on a spacetime oriented structure representation. Computer Vision and Pattern Recognition (CVPR), 2010

[3] Everts, I., van Gemert, J.C., Gevers, T. (2013) Evaluation of Color STIPs for Human Action Recognition. *Computer Vision and Pattern Recognition (CVPR), 2013*.

[4] Hoai, M., Lan, Z., and De la Torre, F., (2011) Joint segmentation and classification of human actions in video. *Computer Vision and Pattern Recognition (CVPR), 2011*.

[5] Jain, M., van Gemert, J.C., Bouthemy, P., Jegou, H., Snoek, C., (2014) Action Localization by Tubelets from Motion. *Computer Vision and Pattern Recognition (CVPR), 2014*.

[6] Sadanand, S., Corso, J., (2012) Action bank: A high-level representation of activity in video. Computer Vision and Pattern Recognition (CVPR), 2012

[7] Shi, J., Malik, J., (2000) Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[8] Van Gemert, J.C., Veenman, C. J., Geusebroek, J. M., (2009) Episode-constrained cross-validation in video concept retrieval. *IEEE Trans. Multimedia*, 11(4):780-785, 2009.

[9] Wang, H., Klaser, A., Schmid, C., Liu, C.-L.(2013) Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013