GEORG WALTNER, THOMAS MAUTHNER, HORST BISCHOF

Improved sport activity recognition using spatio-temporal context

1 Introduction

Activity recognition in sport is an attractive field for computer vision research. Game, player and team analysis are of great interest and research topics within this field emerge with the goal of automated analysis. As the execution of same activities differs between players and activities cannot be modeled by local description alone, additional information is needed. Inspired by the concept of group context ([Choi11], [Lan12], [Zhu13]), we employ contextual information to support activity recognition. Compared to other sport activity recognition systems e.g. proposed by [Bialkowski13], we focus on single player activities rather than on general team activities.

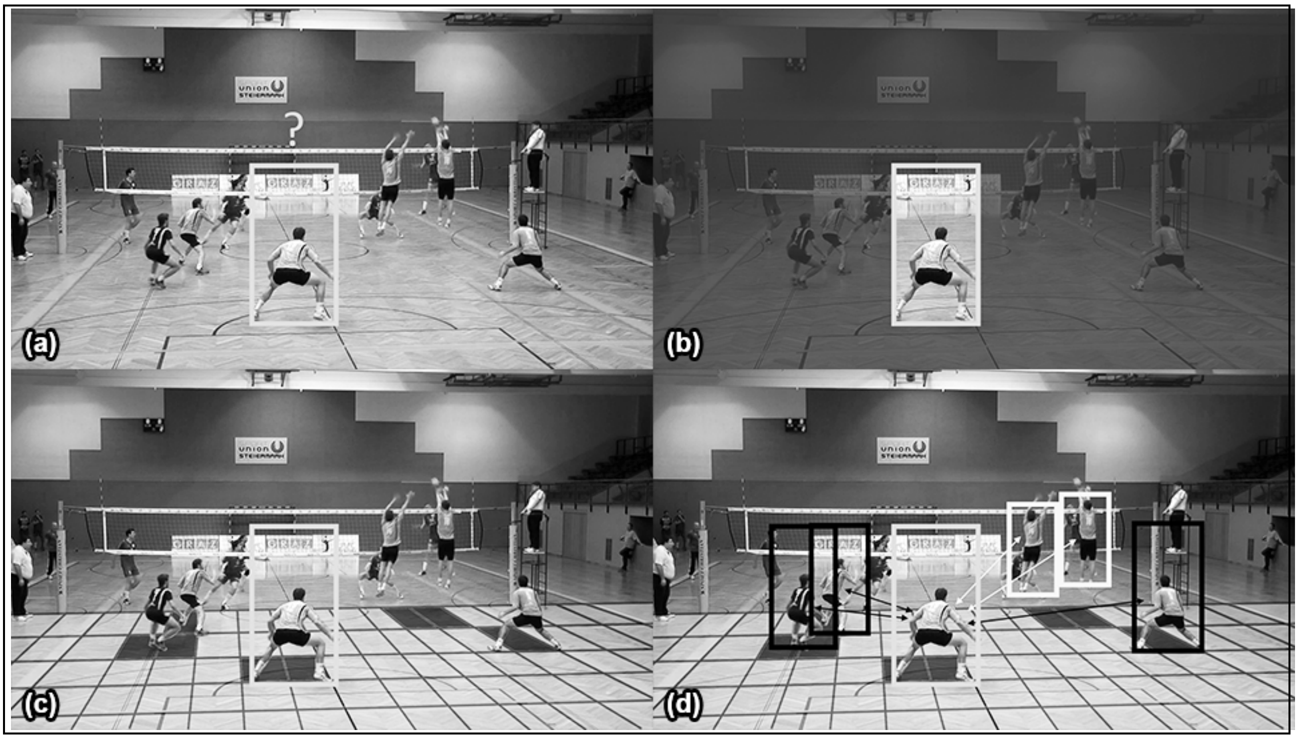


Fig.1: Proposed method. (a) Player under investigation. (b) Local features (HOG/HOF), no context is used. (c) On court positions of the players are incorporated into spatial context. (d) Information about other player activities is used via an activity context descriptor to improve recognition of the investigated player activity. (For simplicity, the AC descriptor is illustrated on one frame only, while in reality multiple frames are used to calculate the spatio-temporal relations of players.)

2 Methods

Figure 1 depicts the main parts of our approach. For recognizing the action of a specific player (Fig. 1 (a)), we start with per-frame player-centered activity recognition by training a *Support Vector Machine (SVM)* using commonly known descriptors for motion (*HOF* [Laptev08]) and shape (*HOG* [Dalal05]) (Fig. 1 (b)) together with position information (*Real World Player Coordinates (RWPC)*) and *Spatial Context (SC)* (Fig. 1 (c)). In a second step, contextual information is incorporated via an *Activity Context (AC)* and a *Radial Activity Context (RAC)* descriptor by evaluating the previously trained SVM on detected players. The activity descriptors collect information about all player's activities over a certain timespan prior to the investigated frame (Fig. 1 (d)).

2.1 Calibration

In this work, we use court coordinates (RWPC) as features. Therefore a homography between the court and the camera plane is estimated in a calibration step, where the user selects four points through a guided graphical user interface. This homography is later used to calculate court positions from player detections in image coordinates.

2.2 Bayes-like framework for player detection and spatial context

For the spatial context descriptor and the player detection we fuse background filtering with color information (represented by *Gaussian Mixture Models (GMM)*) within a Bayes-like detection framework. This framework is illustrated in Figure 2. As a first step we estimate a background model by median filtering randomly sampled images throughout the video (Fig. 2 (a)). Frame differencing between this background model and the actual frame (Fig. 2 (b)) then results in a simple measurement of non-static content (Fig. 2 (c)). Additionally two GMM, trained in advance by annotation of players within five frames for the investigated team and from the background image, are applied on the image (Fig. 2 (d-e)). The combination finally yields probabilities for the non-static content being a player (Fig. 2 (f)). This procedure allows for automatic detection of players and is needed for the SC descriptor as well as for the activity descriptors.

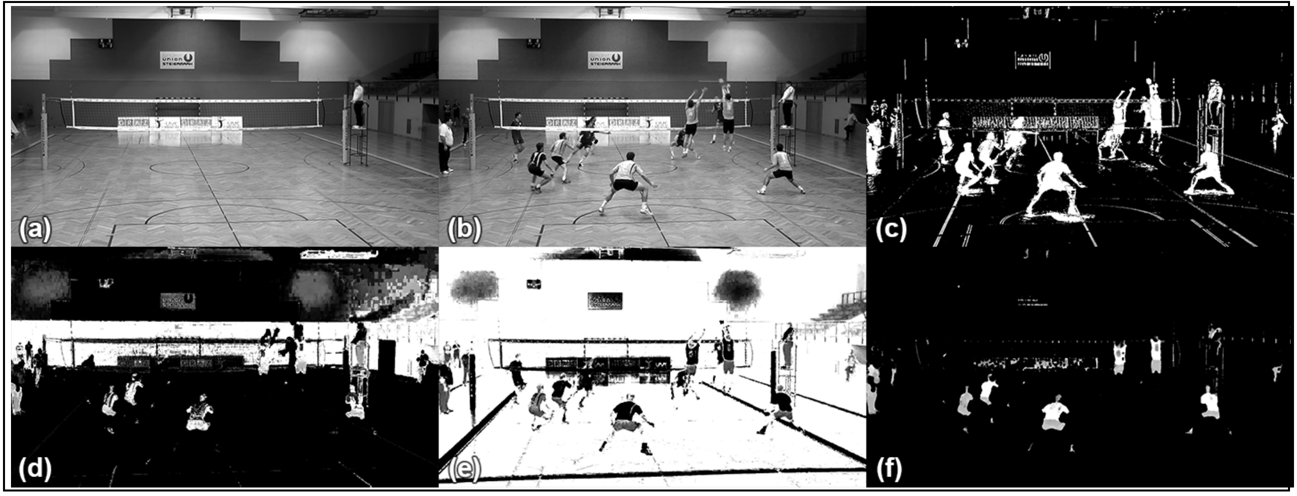


Fig.2: Bayes framework for foreground prediction used for player detection. (a) Median filtered background image. (b) Investigated frame. (c) Dynamic measurement as difference between (a) and (b). (d-e) Frame after applying the two GMM for team and background respectively. (f) Result of foreground prediction, using the torso centroids scaled bounding boxes can be estimated.

The spatial context descriptor estimates the positions and on court distribution of players. Therefore the output of the detection framework is densely sampled by a sliding window approach using about 75000 rectangles. Using the calibration information, these rectangles are scaled according to their on-court position. Within each of the rectangles the player probability is evaluated and these numerous evaluations are binned to 300 (15x20) partitions to reduce the size of the resulting SC descriptor, as shown in Fig. 3.

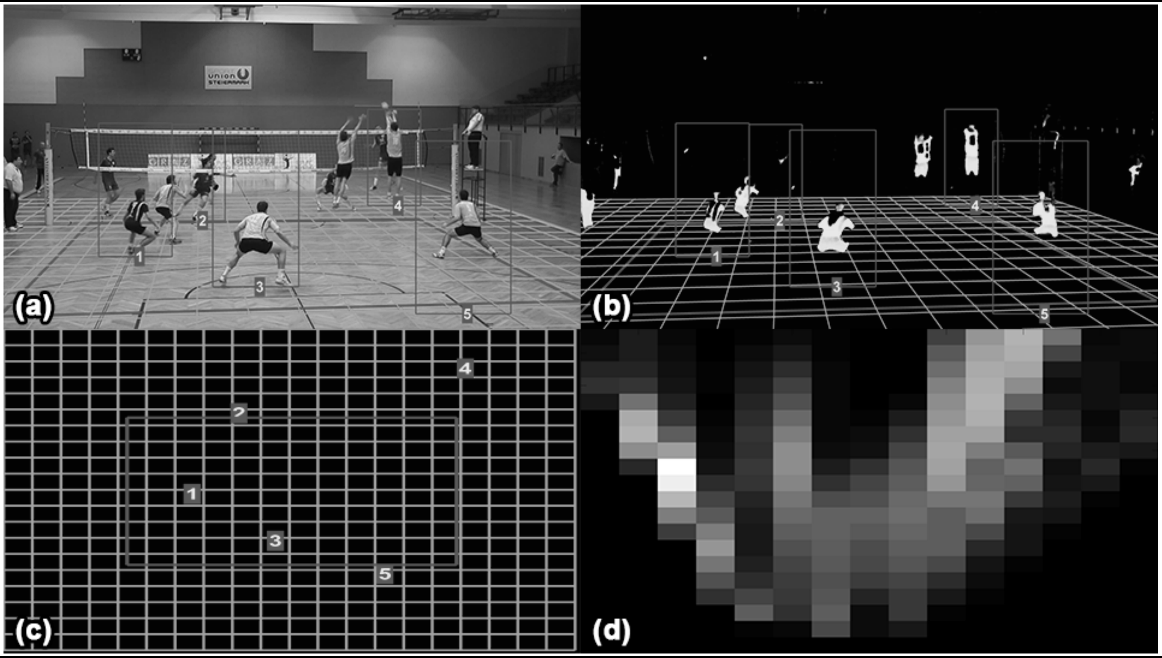


Fig. 3: SC descriptor. (a) Frame with grid overlaid on court and some generic rectangles indicated. (b) Grid and rectangles overlaid on the output of the Bayes framework. (c) Top view of the field with grid and rectangle positions. (d) 15x20 bins SC descriptor indicating player probabilities at specific court locations.

2.3 Activity context

While the previous descriptors (HOG, HOF, RWPC) only model the local appearance, motion and position of a player and the SC descriptor gives a global information in terms of player distribution, the *Activity Context (AC)* descriptor collects all player activities on the court over time. This is motivated by the fact that some activities co-occur within a certain time span (setting after attack) or even at the same time (block of up to three players). On the other hand, some actions should never occur at the same time (block during or after service).

The AC descriptor is an evaluation of the previously trained SVM, evaluated on player detections from the detection framework. Similar to the SC descriptor the activities are binned position-wise with relation to the court plane and averaged over a number of frames to incorporate temporal context.

2.4 Radial activity context

The *Radial Activity Context (RAC)* descriptor is a modified version of the AC descriptor, motivated by the representation of [Choi11]. Imagine one player blocking, there should be a higher probability that at the same time the player next to him is also blocking, but not attacking. Assuming that nearby activities are more significant for the evaluated activity, the binning is done radially with segments becoming larger with increased distance from the player (Fig. 4).

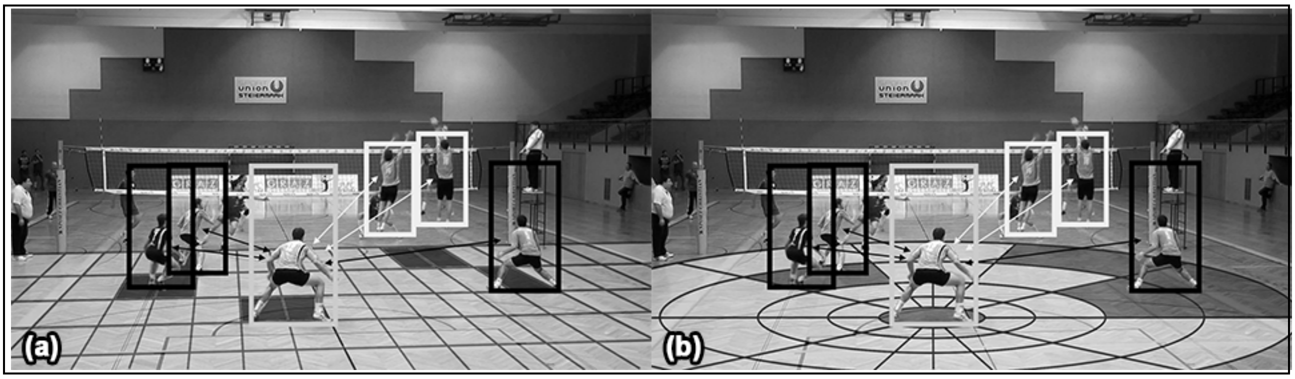


Fig. 4: Comparison of context descriptors. (a) AC descriptor: Actions are binned spatially on a global court plane grid. (b) Instead of using a global court binning, a player-centric binning is applied.

3 Dataset

Along with this paper, we present a new dataset for volleyball activity recognition. The videos were recorded from matches in the AVL[1] in HD resolution (1920x1080) at 25fps, compressed with the DivX codec. 6 video clips from 3 different games with duration of approximately 2.5 hours were processed, resulting in 7973 manual annotations divided into seven classes. These annotations were interpolated resulting in a total of 36k annotations.

Due to the immanent game structure, the activity occurrences differ. Also some activities like "Block" or "Stand" can be executed by multiple players simultaneously. Still, the number of activities is quite balanced. Table 1 shows a list of all activities and their quantities. While the classes "Service", "Reception", "Setting", "Attack" and "Block" are specific volleyball activities, the two other classes "Stand" and "Defense/Move" are more general classes.

Tab. 1. *Volleyball activity dataset. This table contains an overview over the activity quantities in our published dataset. The top row shows the activity names. Tracklets are continuous player activity clips lasting about 1-2 seconds, activities denote a manual annotation in a video frame (every 5-10 frames) and interpolated activities denote the total number of annotations.*

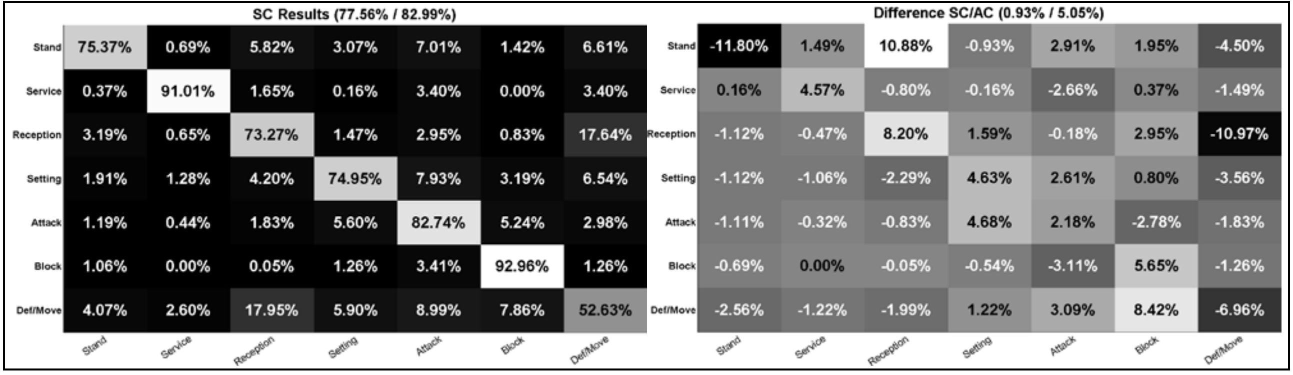
	Stand	Service	Reception	Setting	Attack	Block	Defense/Move	total
tracklets	127	106	83	119	130	214	125	904
activities	1313	868	767	891	1157	1847	1130	7973
activities (interpolated)	6781	3917	3488	3903	5233	8332	5124	36178

[1] Austrian Volley League (http://www.volleynet.at)

The dataset, along with some auxiliary code (e.g. tracklet generation, visualization GUI, homography calculation example), has been made publicly available[2].

4 Results

Using solely HOG/HOF features results in an overall recognition rate of 69.74%, while incorporation of spatial context and position information improved the result to 77.56%. Adding the activity context finally improved results by up to 8.20% on specific classes (Fig. 5). For the 5 volleyball classes a recognition rate of 88.03% was achieved. For the two general classes the recognition accuracy deteriorated due to the fact that the spatio-temporal context is of no use for arbitrarily occurring activities compared to activities which are in a strong temporal correlation. Detailed results and extensive parameter evaluations for all features have been published in [Waltner14][3]. Evaluation of the RAC descriptor did result in only small performance improvement of about 0.5% overall compared to AC. We believe that for the AC descriptor the global binning grid represents the spatial relationship of players with similar descriptive power compared to the player centered RAC descriptor. However, as delineated in Figure 5 (right subfigure), results improve in comparison to the local description of activities, showing that context is of assistance for sport activity recognition.



SC Results (77.56% / 82.99%)

	Stand	Service	Reception	Setting	Attack	Block	Def/Move
Stand	75.37%	0.69%	5.82%	3.07%	7.01%	1.42%	6.61%
Service	0.37%	91.01%	1.65%	0.16%	3.40%	0.00%	3.40%
Reception	3.19%	0.65%	73.27%	1.47%	2.95%	0.83%	17.64%
Setting	1.91%	1.28%	4.20%	74.95%	7.93%	3.19%	6.54%
Attack	1.19%	0.44%	1.83%	5.60%	82.74%	5.24%	2.98%
Block	1.06%	0.00%	0.05%	1.26%	3.41%	92.96%	1.26%
Def/Move	4.07%	2.60%	17.95%	5.90%	8.99%	7.86%	52.63%

Difference SC/AC (0.93% / 5.05%)

	Stand	Service	Reception	Setting	Attack	Block	Def/Move
Stand	-11.80%	1.49%	10.88%	-0.93%	2.91%	1.95%	-4.50%
Service	0.16%	4.57%	-0.80%	-0.16%	-2.66%	0.37%	-1.49%
Reception	-1.12%	-0.47%	8.20%	1.59%	-0.18%	2.95%	-10.97%
Setting	-1.12%	-1.06%	-2.29%	4.63%	2.61%	0.80%	-3.56%
Attack	-1.11%	-0.32%	-0.83%	4.68%	2.18%	-2.78%	-1.83%
Block	-0.69%	0.00%	-0.05%	-0.54%	-3.11%	5.65%	-1.26%
Def/Move	-2.56%	-1.22%	-1.99%	1.22%	3.09%	8.42%	-6.96%

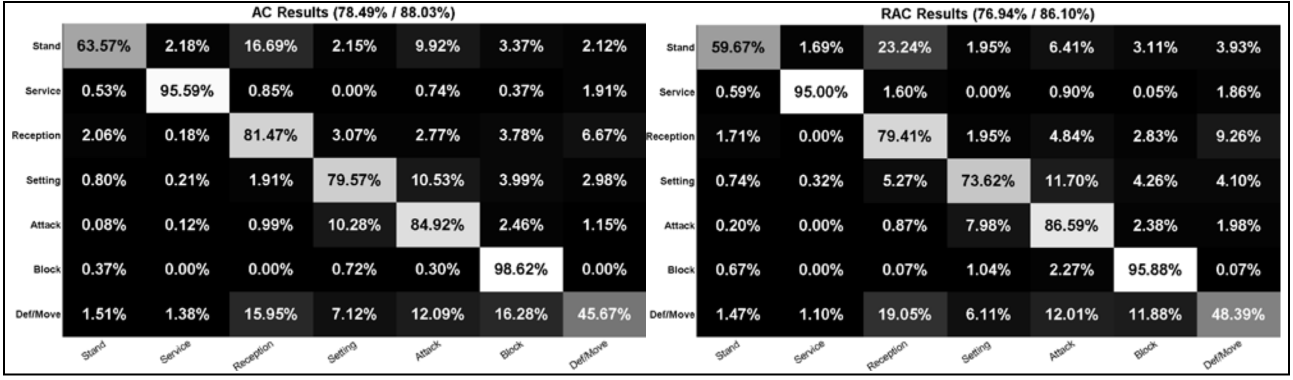
Fig. 5: SC Results. Left: An average accuracy of 77.56% is achieved with the combination of local features (HOG, HOF, RWPC) and SC. Right: Volleyball specific classes profit from supplemental spatio-temporal context (AC) by 5%, while general classes degenerate due to the missing temporal correlation. Also the neutral class stand is more often confused with reception, as these activities are partially similar in appearance/motion and occur in close temporal context.

The results in Figures 5 and 6 are displayed in the form of confusion matrices. The true labels are plotted vs. the predicted labels. Each row shows the result for a class and the mix-ups with other classes. An optimal result would have all diagonal elements at 100% accuracy and the off-diagonal values at 0%. The right plot in Figure 5 shows the changes in recognition accuracies when using AC on top of the

[2] http://lrs.icg.tugraz.at/download.php#vb14

[3] Due to added/corrected annotations in the time between these publications results in this paper differ slightly from those published before.

other features, while the overall improvement is 1%, for volleyball specific classes results significantly improve by 5%.



AC Results (78.49% / 88.03%)

	Stand	Service	Reception	Setting	Attack	Block	Def/Move
Stand	63.57%	2.18%	16.69%	2.15%	9.92%	3.37%	2.12%
Service	0.53%	95.59%	0.85%	0.00%	0.74%	0.37%	1.91%
Reception	2.06%	0.18%	81.47%	3.07%	2.77%	3.78%	6.67%
Setting	0.80%	0.21%	1.91%	79.57%	10.53%	3.99%	2.98%
Attack	0.08%	0.12%	0.99%	10.28%	84.92%	2.46%	1.15%
Block	0.37%	0.00%	0.00%	0.72%	0.30%	98.62%	0.00%
Def/Move	1.51%	1.38%	15.95%	7.12%	12.09%	16.28%	45.67%

RAC Results (76.94% / 86.10%)

	Stand	Service	Reception	Setting	Attack	Block	Def/Move
Stand	59.67%	1.69%	23.24%	1.95%	6.41%	3.11%	3.93%
Service	0.59%	95.00%	1.60%	0.00%	0.90%	0.05%	1.86%
Reception	1.71%	0.00%	79.41%	1.95%	4.84%	2.83%	9.26%
Setting	0.74%	0.32%	5.27%	73.62%	11.70%	4.26%	4.10%
Attack	0.20%	0.00%	0.87%	7.98%	86.59%	2.38%	1.98%
Block	0.67%	0.00%	0.07%	1.04%	2.27%	95.88%	0.07%
Def/Move	1.47%	1.10%	19.05%	6.11%	12.01%	11.88%	48.39%

Fig. 6: AC Results. An average accuracy of 88.03% is achieved by adding spatio-temporal activity context. The results for the RAC descriptor are quite similar compared to AC, showing that both methods are equally useful for recognizing the volleyball-specific activity classes.

5 Discussion, Conclusion and Outlook

We presented an evaluation of single player activity recognition on our newly published indoor volleyball dataset. Starting the classification from standard local features (HOG/HOF), we show that spatial and temporal context information can supply essential information for recognition and also helps resolving ambiguities that can hardly be sorted out with local information alone. A further enhancement of the annotations is thinkable, in terms of finer subdivisions of the classes. By splitting the activities spatially into regions, the discriminative power of the RWPC descriptor could be further improved. Also parting the activities temporally into sequences of short basic actions like “jump”, “hit”, etc. might improve recognition as the intra-class variability would be diminished. One might also investigate higher level relations for specific plays (team activities such as attack or defense). Furthermore other features (velocities, trajectories, ball information) and methods might bring performance improvement by exploit of additional information.

References

Bialkowski, A., Lucey, P., Carr, P., Denman, S., Matthews, I., Sridharan, S. (2013). Recognizing team activities from noisy data. In Proceedings CVPR workshops

Choi, W., Shahid, K., Savarese, S. (2011). Learning context for collective activity recognition. In Proceedings CVPR

Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. In Proceedings CVPR

Lan, T., Wang, Y., Mori, G., Robinovitch, S. N. (2010). Retrieving actions in group contexts. In Proceedings ECCV workshops

Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B. (2008). Learning realistic human actions from movies. In Proceedings CVPR

Waltner G., Mauthner T., Bischof H. (2014). Indoor Activity Detection and Recognition for Sport Games Analysis. In Proceedings AAPR

Zhu, Y., Nayak, N. M., Roy-Chowdhury, A. K. (2013). Context-aware modeling and recognition of activities in video. In Proceedings CVPR