Understanding Sport Activities from Correspondences of Clustered Trajectories

Francesco Turchini
University of Florence
francesco.turchini@unifi.it

Lorenzo Seidenari
University of Florence
lorenzo.seidenari@unifi.it

Alberto Del Bimbo
University of Florence
alberto.delbimbo@unifi.it

Abstract

Human activity recognition is a fundamental problem in computer vision with many applications such as video retrieval, automatic visual surveillance and human computer interaction. Sports represent one of the most viewed content on digital tv and the web. Automatically collected statistics of team sports game play represent actionable information for many end users such as coaches and broadcast speakers. Many computer vision methods applied to sport activity classification are often based on multi-camera setups, player tracking and exploit information on the groundplane. In this work we overcome this limitations and propose an approach that exploits the spatio-temporal structure of a video grouping local spatio-temporal features unsupervisedly. Our robust representation allows to measure video similarity making correspondences among arbitrary patterns. We tested our method on two dataset of Volleyball and Soccer actions outperforming previous results by a large margin. Finally we show how our representation allows to highlight discriminative regions for each action.

1. Introduction

Human activity recognition is a fundamental problem in computer vision [15, 7, 25] with many applications such as video retrieval[17], automatic visual surveillance[19, 20] and human computer interaction[26]. Sports represent one of the most viewed content on digital tv and the web. Sports are watched by millions of people and broadcasters are constantly improving user experience by providing real-time statistics of games. Classifying player actions in sports is an extremely relevant task that can provide several commercial and professional applications. Speakers, analysts and directors may obtain in real-time similar plays from the current or other games providing an improved experience for the audience. Head coaches may easily classify all the plays of a certain player to track improvement or to analyze other teams tactics; finally gameplay statistics can be automatically gathered such as the amount of shots on goal and corner kicks a team had in a game or a season.

Recently many computer vision researchers directed their efforts in the automatic analysis of sports videos. Sports video analytics is often performed to collect statistics on player positions during games extracting individual trajectories and team formation patterns [12, 6, 1].

Many action recognition datasets are comprised of just sport videos and there is some, limited, interest in recognizing in a video [21, 14, 8]. More effort has been poured in the analysis of team tactics and activity [4, 3]. Team activities are defined best by player positions in the field, for this reason many works exploit this datum. Many methods are based on multi-camera systems deployed to get full coverage of the court.

There are few methods, apart from generic action recognition systems, that attempt to classify player activities without localizing and tracking individual players [2]. Indeed several techniques require a calibrated fixed view to fuse visual features with geometrical features such as player trajectories or positions in the field.

In this paper we propose an activity recognition method that targets team sports. Our method does not require calibrated views of the field, player track annotations or player tracking, neither is based on player team recognition. Our method automatically groups visual features forming a robust representation of videos. We show how the proposed method can recognize individual player activities as well as collective team activities in two popular sports: soccer and volleyball.

Our method is based on improved trajectories [24] and does not encode explicitly player positions or the temporal sequence of a video. We automatically group trajectories and define a match kernel able to make arbitrary correspondences of spatio-temporal patterns.

Our method is similar to [5, 7] but differently from [5] we do not require a hierarchical partitioning of the features nor use quantized local features that have worst performance with respect to Fisher encoded descriptors; moreover [7] require importance maps obtained processing Hierachical Space-Time Segments [13] while we just rely on our feature grouping method.

We test our method on two sport activity datasets im-

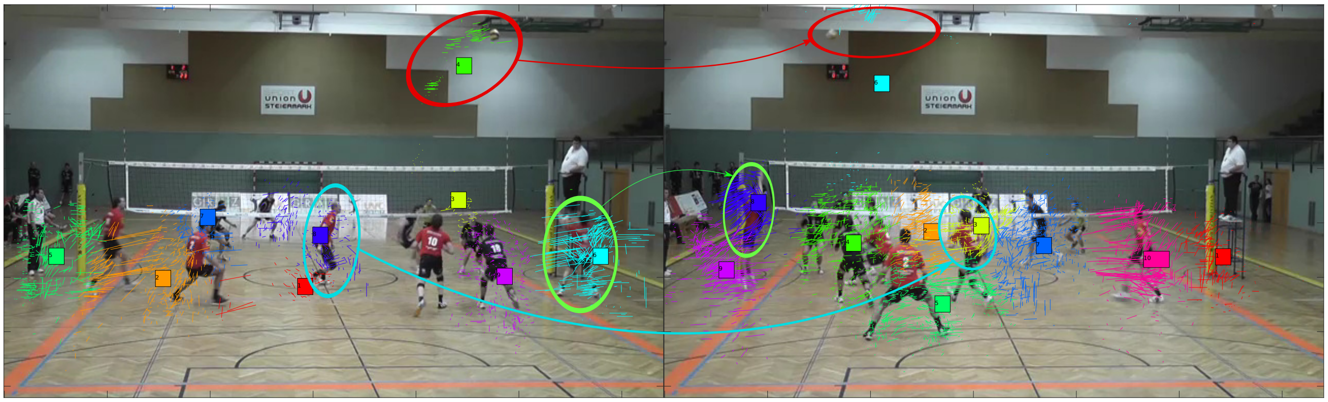


Figure 1: Cluster matching between two videos from the same class of the Volleyball Activity dataset.

proving accuracy with respect to previously published methods by a large margin. We also show state-of-the-art results on UCF-Sports showing how our method is also a viable generic action recognition system.

2. Related Work

We briefly review some recent contributions on automatic sport activity recognition. Atmosukarto *et al.* [1] developed a method to recognize offensive team formation in American football. Their method applies robust video stitching and exploits the localization of the line of scrimmage to compute a feature based on gradient intensity on the offensive side of the line. Bialkowski *et al.* [3] avoid tracking players but apply player detection and team recognition. The method exploits multiple calibrated views of the field to locate players in the field. Team activity is recognized computing team field occupancy maps.

Ballan *et al.* [2] match videos using a kernel for sequences derived from the Needleman-Wunch distance (NWD). The temporal structure of a video is a fundamental cue for recognizing complex events such as sport activities. Their approach is based on the fact that similar actions should share similar appearance in a similar sequence. The main limitation of their method is the use of static features (SIFT) and the fact that NWD is not designed to make arbitrary correspondences between sequences.

Waltner *et al.* [23] propose a method to recognize individual player activities in volleyball. Their method exploits player detection and camera calibration. Single player activities are recognized using a boosting based approach and static and motion local features. They also compute a contextual feature based on player position for which they require player team recognition.

3. Video Representation

Our video representation is designed to capture the spatio-temporal structure of the video. In team sports, activities are often defined only by a subset of the players. Ideally mapping visual features to players or other relevant elements, e.g. the ball, the referee etc., allows to obtain a detailed representation of the scene. Although player tracking and detection is an extremely challenging task that is prone to failure. Failing to track or detect players or other relevant entities breaks the recognition pipeline leading to inconsistent results.

Our method is more robust and consists of two main steps: first we group trajectories in an unsupervised manner with an efficient method allowing to deal with the several thousands of features per frame extracted, and then we propose a cluster match kernel that allows to make correspondences among the grouped trajectories. SVM is used to learn the classifiers.

3.1. Trajectory clustering

We use Improved Dense Trajectories (IDT) [24] as a feature extractor. To cluster trajectories, due to the large amount of features extracted by the IDT algorithm, we used Landmark Based Spectral Clustering (LSC).

Spectral clustering is a relaxation of Normalized Cut algorithm that tries to exploit the connectivity of data. Spectral clustering exploits the eigenvalues of the Laplacian to obtain a better representation that allows to easily separate clusters using k-means.

The main problem with big input data is computing the graph Laplacian and its factorization. Let $\mathbf{L} = [\boldsymbol{l}_1, ..., \boldsymbol{l}_n] \in \mathbb{R}^{m \times n}$ be the data matrix. First we sample the input data to obtain two matrices $\mathbf{U} \in \mathbb{R}^{m \times p}$, the landmarks matrix) and $\mathbf{Z} \in \mathbb{R}^{p \times n}$, the data projected in a smaller space of size $p \ll n$ so that we can approximate $\mathbf{L} \approx \mathbf{UZ}$, thus making Laplacian and eigenvectors computation more

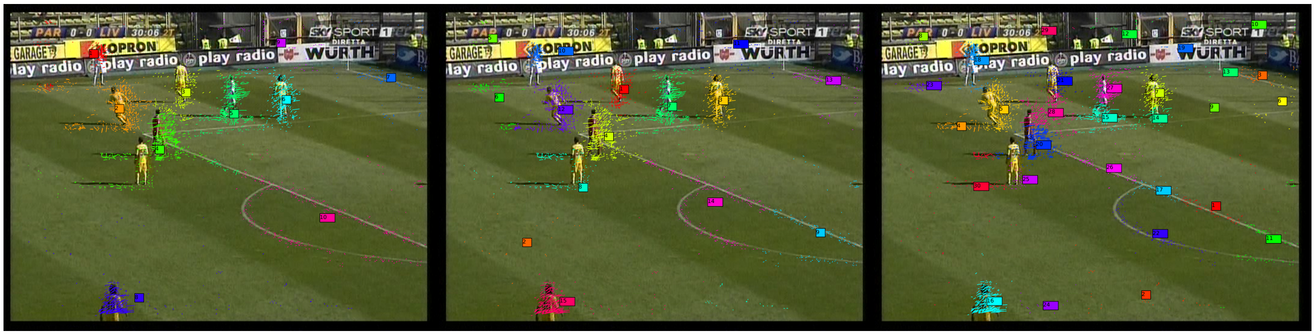


Figure 2: Automatically clustered trajectories on soccer dataset (10, 15 and 30 clusters). Several clusters gather features of a single player. Noisy clusters often capture textured regions of the background.

```
Algorithm 1: LSCClustering
  Data: $n$ data points $\boldsymbol{l}_1, \boldsymbol{l}_2, \ldots, \boldsymbol{l}_n \in \mathbb{R}^m$, Cluster number $k$
  Result: Indices of $k$ Clusters
  Choose $p$ landmarks using a K-Means pass with few iterations
  Compute matrix $\mathbf{Z} \in \mathbb{R}^{p \times n}$ as shown in Equation 1
  Compute the first $k$ eigenvectors of $\mathbf{Z}\mathbf{Z}^T$,
    $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n] \in \mathbb{R}^{k \times n}$
  Apply K-Means to $\mathbf{V}$ to obtain the indices vector $\boldsymbol{I}$ of the $k$ clusters for the $n$ input observations
  return ($\boldsymbol{I}$)
```

lightweight. Preselection of landmarks is performed using K-Means. These samples are the basis vectors used to represent the input data in a reduced space.

Given the samples and matrix $\mathbf{U}$, the elements of the sparse representation matrix $\mathbf{Z}$ can be calculated efficiently as

$$z_{ji} = \frac{K_h(\boldsymbol{l}_i, \boldsymbol{u}_j)}{\sum_{j \in U} K_h(\boldsymbol{l}_i, \boldsymbol{u}_j)} \tag{1}$$

Where $K_h(\cdot)$ is a kernel function, in our case the Gaussian kernel $K_h(\boldsymbol{l}_i, \boldsymbol{u}_j) = \exp(-\frac{||\boldsymbol{l}_i - \boldsymbol{u}_j||^2}{2h})$. We now can compute the eigenvalues and eigenvectors of $\mathbf{Z}\mathbf{Z}^T$, choosing the first k and applying K-Means to obtain the clusters. The clustering pseudocode is outlined in algorithm 1.

3.2. Cluster representation

We represent local features with HoG, HoF, MBH and trajectory descriptors concatenating the normalized spatio-temporal coordinates to the local descriptors. Each cluster is represented with a Fisher Vector encoding of the local descriptors that have been assigned to it.

We apply PCA retaining the first 80 components of all histogram features and 20 of the trajectory descriptors; we concatenate the normalized spatio-temporal coordinate of each trajectory center to the PCA-compressed local feature in order to retain information about the global location. We learn a codebook of 256 Gaussians using GMM. PCA and GMM codebook are learned on a random sample of 200K training features.

Fisher Vectors are calculated using the Improved algorithm (L_2-normalization and power normalization are applied) and we use linear kernels according to [16], summing kernel scores computed from different local features as in Eq. 8.

Given a Gaussian Mixture Model with parameters $\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n, \boldsymbol{\omega}_n$ and given soft-assignments $\gamma_m^{(n)}$ for each of the M augmented local feature $\boldsymbol{x}_m \in \boldsymbol{X}$, the Fisher vector is computed concatenating the likelihood gradients:

$$\Psi(\boldsymbol{X}) = [\mathcal{G}_n^{\mu}(\boldsymbol{X})\ \mathcal{G}_n^{\sigma}(\boldsymbol{X})] \tag{2}$$

where

$$\mathcal{G}_n^{\mu}(\boldsymbol{X}) = \frac{1}{\sqrt{\boldsymbol{\omega}_n}} \sum_{m=1}^{M} \gamma_m^{(n)} \left(\frac{\boldsymbol{x}_m - \boldsymbol{\mu}_n}{\boldsymbol{\sigma}_n^2} \right), \tag{3}$$

$$\mathcal{G}_n^{\sigma}(\boldsymbol{X}) = \frac{1}{\sqrt{2\boldsymbol{\omega}_n}} \sum_{m=1}^{M} \gamma_m^{(n)} \left(\frac{(\boldsymbol{x}_m - \mu_n)^2}{\boldsymbol{\sigma}_n^2} - 1 \right), \tag{4}$$

and

$$\gamma_m^{(n)} = \frac{\boldsymbol{\omega}_n p_n(\boldsymbol{x}_m)}{\sum_{j=1}^{D} \boldsymbol{\omega}_j p_j(\boldsymbol{x}_m)}, \tag{5}$$

4. Video Matching

Consider the set of augmented local features $\boldsymbol{X}$ extracted from a video, the clustering yields a partition $\mathcal{P}(\boldsymbol{X})$ of set $\boldsymbol{X}$ such that

$$\bigcup_{X_i \in \mathcal{P}(\boldsymbol{X})} \boldsymbol{X}_i = \boldsymbol{X} \text{ and } \bigcap_{X_i \in \mathcal{P}(\boldsymbol{X})} \boldsymbol{X}_i = \emptyset \tag{6}$$

We define a kernel inspired by Match Kernels [22] that exploits trajectory grouping to reduce the matching complexity and to compute correspondences among coherent subset of video features.

4.1. Cluster Set Kernel

Given a pair of videos and their respective feature sets $\boldsymbol{X}$ and $\boldsymbol{Y}$, after applying clustering we can compute our cluster set kernel by computing

$$K(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{|\mathcal{P}(\boldsymbol{X})|} \sum_{X_i \in \mathcal{P}(\boldsymbol{X})} \max_j \Psi(\boldsymbol{X}_i)^T \Psi(\boldsymbol{Y}_j) + \tag{7}$$

$$\frac{1}{|\mathcal{P}(\boldsymbol{Y})|} \sum_{Y_j \in \mathcal{P}(\boldsymbol{Y})} \max_i \Psi(\boldsymbol{X}_i)^T \Psi(\boldsymbol{Y}_j)$$

In this way, we obtain a symmetric kernel matrix. Our kernel takes into account the similarity scores both of $\boldsymbol{Y}$ respect to $\boldsymbol{X}$ and of $\boldsymbol{X}$ respect to $\boldsymbol{Y}$. If two videos are similar, we should obtain high scores from both operations, and thus a high combined score.

Even though our kernel can not formally satisfy the Mercer property it has been shown that this is not a strict requirement for an SVM classifier to learn an accurate solution. In practice the kernel matrices we have computed were always positive definite so far.

Given different groupings $\mathcal{P}(\boldsymbol{X}^f_n)$ for each feature f and the respective kernels $K(\mathcal{P}(\boldsymbol{X}^f_n), \mathcal{P}(\boldsymbol{Y}^f_n))$ our final kernel can be computed as

$$K(\boldsymbol{X},\boldsymbol{Y}) = \sum_f \sum_n K(\mathcal{P}(\boldsymbol{Z}^f_n), \mathcal{P}(\boldsymbol{Y}^f_n)) \tag{8}$$

thus integrating different local representation and spatio-temporal structures.

4.2. Action Localization

To gain insight on the potential localization capability of our approach, we show a method for salient cluster mining. We would like to find which are the clusters that better help the classifier to discriminate. Given a video feature set $\boldsymbol{Z}$, and the learned kernel SVM classifier for an action defined by Eq. 8, the weights α_k and training sample labels y_k we greedily search for the cluster $\boldsymbol{Z}_i$ that, if removed, causes the higher classification score drop:

$$\boldsymbol{Z}_i = \arg\max_i \sum_k \alpha_k y_k \left[K_{\boldsymbol{X}^k}(\mathcal{P}(\boldsymbol{Z})) - K_{\boldsymbol{X}^k}(\mathcal{P}(\boldsymbol{Z}) \setminus Z_i)\right] \tag{9}$$

where $K_{\boldsymbol{X}}(\boldsymbol{Z}) = K(\boldsymbol{X},\boldsymbol{Z})$. In Figure 3 we plot the accumulation of salient clusters bounding boxes generating a heat map highlighting the most salient areas in the scene.

It can be seen that for the "Service" and "Attack" action the serving and attacking players are respectively highlighted. While in the examples of "Reception" and "Setting" multiple players are highlighted. In the "Setting" action, both spikers, the middle-blocker and the opposite player run-up are localized.

This behavior can be expected, as some actions need global contextual information to be recognized, which means that clustering is not able to understand their dynamic nature. It may also happen that some actions are correctly classified by our clustering approach, but identifying as meaningful some parts of the scene which are not intuitively descriptive of the action.

5. Experiments

5.1. Dataset

To test our framework, we performed some tests on a generic sport dataset (UCF Sports), and two specific sport datasets, namely MICC-SOCACT4 and Volleyball Activity Dataset 2014.

UCF Sports UCF Sports is comprised of 10 actions selected from various sports and recorded from TV broadcast (Diving, Golf Swing, Kicking, Lifting, Riding Horse, Running, Skateboarding, Swing-Bench, Swing-Side, Walking). There are 150 scenes at 720x480 resolution. For action recognition, we use the Leave-One-Out (LOO) cross-validation scheme.

MICC-SOCACT4 This dataset is composed by 100 MPEG-2 videos at PAL resolution (720x576). These videos represent 4 soccer actions: "Goal Kick", "Throw In", "Placed Kick", "Shot on Goal" and were recorded from 5 different matches of the Italian "Serie A". We picked a match as the test set and the other 4 as the training set, performing a 5-fold cross-validation.

Volleyball Activity Dataset 2014 This dataset is composed by 6 full volleyball matches of the Austrian Volley League originally recorded in full HD resolution. They were annotated with 7 classes, 5 specific volley classes ("Serve", "Reception", "Setting", "Attack", "Block"), and 2 more general classes ("Stand", "Defense/Move"). We take in exam the tracklets, which represent the continuous player activities lasting about 1-2 seconds. We cut the original videos according to the tracklets, obtaining about 900 videos, that we used as the actual classification dataset. The cut script provided by the authors crops the area around annotated players, while we take the entire frame, reducing its resolution to 640x360, and adding 15 frames at the beginning and 15 at the end of the tracklet. Data was partitioned in 50% for training and 50% for testing, according to [23].

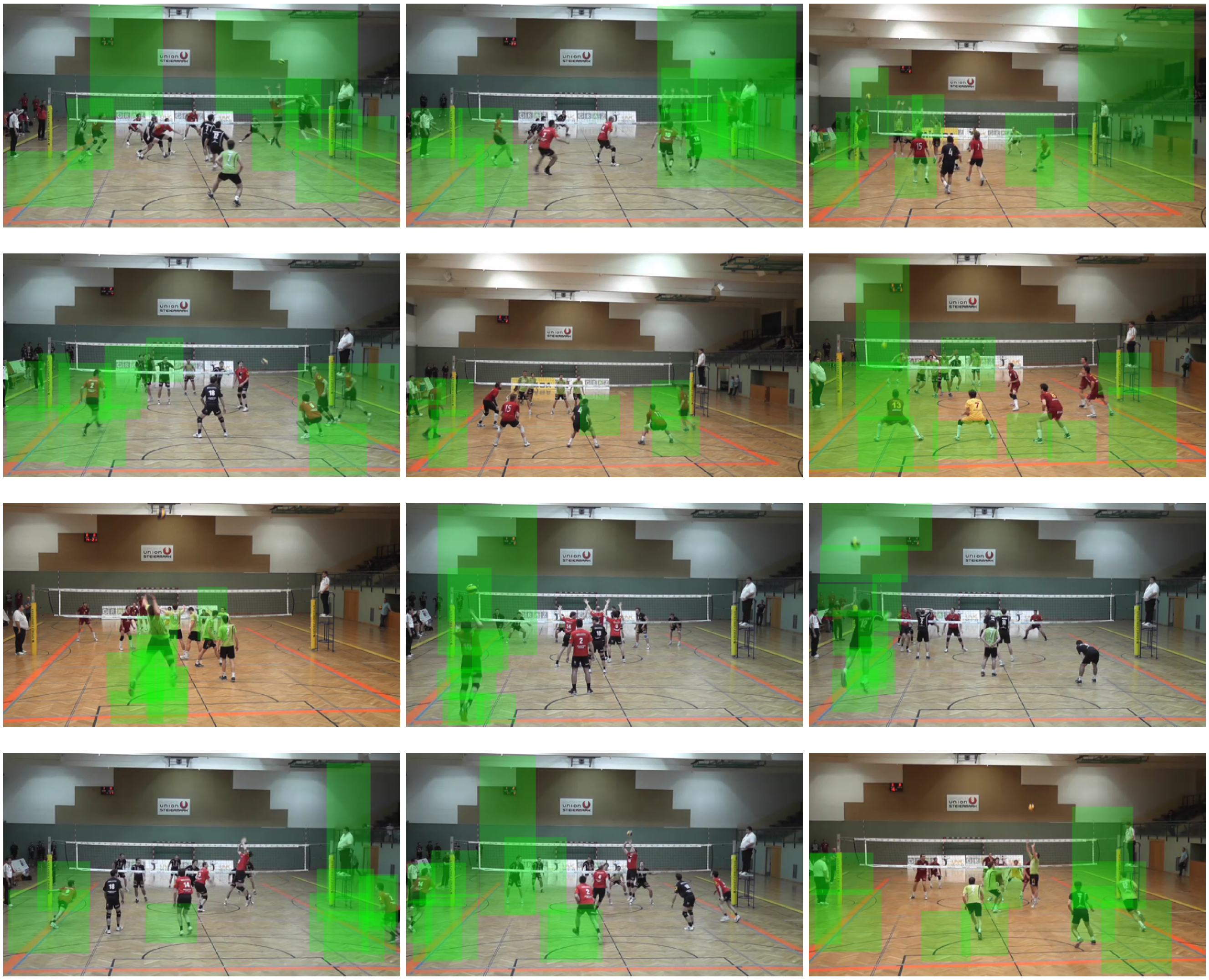


Figure 3: Heatmap of the most relevant clusters for collective actions “Attack”, “Reception” and “Service” (from top to bottom).

5.2. Action Recognition

In all experiments we extracted dense trajectories descriptors with the default parameters. For SOCACT and UCF Sports datasets, we extracted descriptors also on flipped versions of videos due to the low cardinality of the datasets.

We make a baseline using a standard Fisher Vectors pipeline using linear kernels. This is equivalent to our framework with a single cluster containing all the features.

We run a set of experiments to show how the trajectory clustering step affects classification accuracy. We report classification results varying the number of clusters in Table 2 and Table 3.

Accuracy of our method is not strongly dependent on the amount of clusters extracted per sequence. In both dataset the best performance is obtained using 10 clusters.

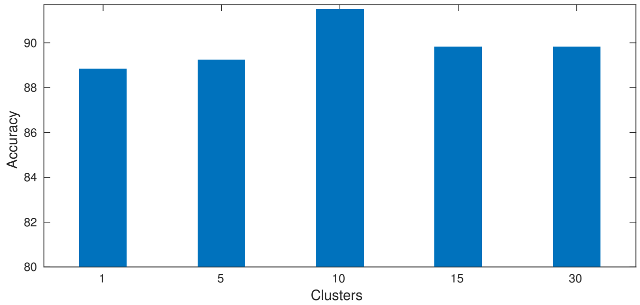


Figure 4: Accuracy values varying number of clusters for the MICC-SOCACT4 dataset.

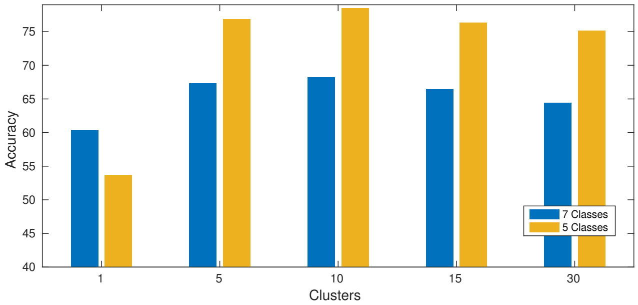


Figure 5: Accuracy values varying number of clusters for the Volleyball Actions 2014 dataset.

Method	Accuracy
Our	**91.0**
Fisher Vector Baseline	87.6
Karaman *et al.* [7]	90.4
Lan *et al.* [11]	83.7
Kovashka *et al.* [10]	87.3
Klaser *et al.* [9]	86.7
Wang *et al.* [24]	85.6
Yeffet *et al.* [27]	79.3
Rodriguez *et al.* [18]	69.2

Table 1: Comparison with the state of the art on the UCF Sports dataset. Results are reported as mean per-class accuracy over the 10 classes.

We first show a comparison with the state-of-the-art on UCF Sports Actions, which highlights the good behaviour of our method respect to other known approaches. We slightly outperform the state-of-the art of [7], without making use of pooling maps to weight high saliency areas of the scenes[7]. UCF Sports actions are performed by individual athletes, so the clustering step is able to put in evidence the salient subset of trajectories without additional external information.

We then report results of our method on the smaller MICC-SOCACT4 dataset. Soccer actions are often defined by collective behaviors. On this dataset our Fisher Vector baseline already improves over [2] by a large margin as is shown in Table 2. Nevertheless our correspondence kernel can boost the accuracy further obtaining 92.50%, especially raising the accuracy on “Goal Kick” and “Placed Kick”; in both these actions there is a single player performing a discriminative motion: kicking the ball from a fixed position, while other players are less involved in the action. For this reason our clustering can isolate this actions and better match the respective spatio-temporal structures.

In Table 3 we report a comparison of our method with previous work and our baselines on the Volleyball Activity Dataset. It can be seen that our baseline based on a single Fisher vector per video perform worse than [23]. Our clustering based baseline improves over the FV baseline by 8% (and by 25% on 5 classes). Some player activities are better recognized in isolation as can be seen in the confusion matrix while other are better recognized exploiting context. From Figure 6 it is clear that collective activities as “Block”,“Defence” and “Attack” are better captured by a global representation (FV), while individual actions like “Setting” and “Service” are better recognized by our correspondence kernel.

Method	Accuracy
Our Fusion	**92.5**
Our clustering	91.5
Fisher Vectors Baseline	88.8
String Kernel+SVM[2]	73.0
NN+NWD [2]	54.0

Table 2: Mean per class accuracy of our method compared with [2] on the MICC-SOCACT4 dataset.

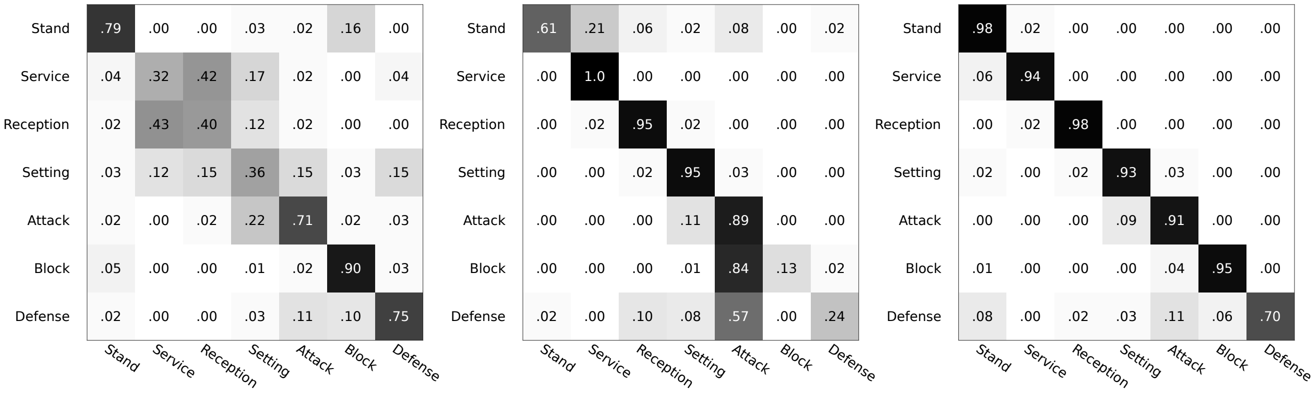
Method	Acc. 7 Cl.	Acc. 5 Cl.
Our Fusion	**91.2**	**94.1**
Our clustering	68.2	78.5
Fisher Vector Baseline	60.3	53.7
Waltner [23] *et al.*	77.5	90.2

Table 3: Mean per class accuracy results on the VolleyBall Activity dataset compared with [23].

The fusion approach implemented by Eq. 8 is able to obtain accurate results in both setups outperforming the state of the art by more than 14% (and by 4% on 5 classes).

The classification task noticeably benefits from the clustering step, especially in the volleyball setup. However, it appears clearly looking at the confusion matrices in Figure 6 that results are complementary. “Stand”,“Block” and “Defence” need some additional contextual information to be recognized, while clustering focuses on local information located in cluster areas, which is better for the other classes. So we tested a combination of both approaches by means of a kernel fusion. Notice how the kernel fusion allows to distinguish between Block and Attack actions, which are almost totally confused by the clustering method.

On soccer videos classification accuracy took just a small advantage from kernel fusion, compared to clustered Fisher encoding by itself. We can hypothesize that soccer scenes do not benefit from global contextual information because of their structure and dynamic characteristics, with very high camera motion that is not fully compensated by improved trajectories features, while volley sequences need to be analysed both globally and in specific areas to locate the distinctive elements, such as the players disposition.



	Stand	Service	Reception	Setting	Attack	Block	Defense
Stand	.79	.00	.00	.03	.02	.16	.00
Service	.04	.32	.42	.17	.02	.00	.04
Reception	.02	.43	.40	.12	.02	.00	.00
Setting	.03	.12	.15	.36	.15	.03	.15
Attack	.02	.00	.02	.22	.71	.02	.03
Block	.05	.00	.00	.01	.02	.90	.03
Defense	.02	.00	.00	.03	.11	.10	.75

(a) Baseline

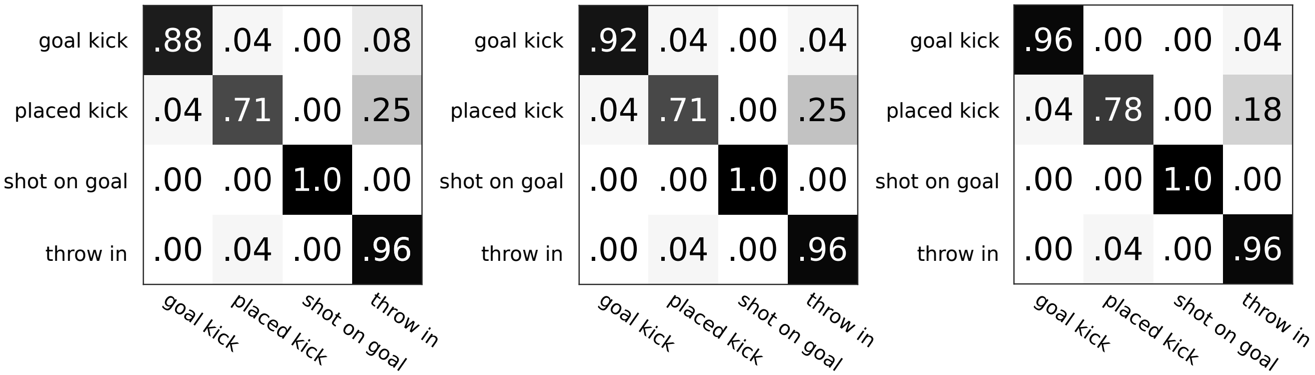
	Stand	Service	Reception	Setting	Attack	Block	Defense
Stand	.61	.21	.06	.02	.08	.00	.02
Service	.00	1.0	.00	.00	.00	.00	.00
Reception	.00	.02	.95	.02	.00	.00	.00
Setting	.00	.00	.02	.95	.03	.00	.00
Attack	.00	.00	.00	.11	.89	.00	.00
Block	.00	.00	.00	.01	.84	.13	.02
Defense	.02	.00	.10	.08	.57	.00	.24

(b) 10 clusters

	Stand	Service	Reception	Setting	Attack	Block	Defense
Stand	.98	.02	.00	.00	.00	.00	.00
Service	.06	.94	.00	.00	.00	.00	.00
Reception	.00	.02	.98	.00	.00	.00	.00
Setting	.02	.00	.02	.93	.03	.00	.00
Attack	.00	.00	.00	.09	.91	.00	.00
Block	.01	.00	.00	.00	.04	.95	.00
Defense	.08	.00	.02	.03	.11	.06	.70

(c) Fusion

Figure 6: Confusion matrices for volleyball.



	goal kick	placed kick	shot on goal	throw in
goal kick	.88	.04	.00	.08
placed kick	.04	.71	.00	.25
shot on goal	.00	.00	1.0	.00
throw in	.00	.04	.00	.96

(a) Baseline

	goal kick	placed kick	shot on goal	throw in
goal kick	.92	.04	.00	.04
placed kick	.04	.71	.00	.25
shot on goal	.00	.00	1.0	.00
throw in	.00	.04	.00	.96

(b) 10 Clusters

	goal kick	placed kick	shot on goal	throw in
goal kick	.96	.00	.00	.04
placed kick	.04	.78	.00	.18
shot on goal	.00	.00	1.0	.00
throw in	.00	.04	.00	.96

(c) Fusion

Figure 7: Confusion matrices for soccer. Our method improves on “Goal Kick” and “Placed Kick” actions.

6. Conclusion

We have proposed a novel method for team sport activities recognition based on local trajectory grouping and matching. Our approach allows to automatically understand what activities are performed in video. Thanks to our cluster set kernel we can compute partial video correspondences effectively without exhaustively matching all local features.

This approach proves effective in recognizing activities where individual player actions are important. Our representation is complementary to a global encoding of local features. The fusion of these two representation yields state-of-the-art results in recognition of volleyball and soccer activities.

Our feature grouping allows to localize mid-level spatio-temporal features that are semantically sensible. We further show a method to understand what mid-level features are relevant for a certain action.

Acknowledgments. Lorenzo Seidenari and Francesco Turchini are supported by “THE SOCIAL MUSEUM AND SMART TOURISM”, MIUR project no. CTN01_00034_23154_SMST.

References

[1] I. Atmosukarto, B. Ghanem, S. Ahuja, K. Muthuswamy, and N. Ahuja. Automatic recognition of offensive team formation in american football plays. In *Proc. of Computer Vision in Sports, Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2013. 1, 2

[2] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Video event classification using string kernels. *Multimedia Tools and Applications*, 48(1):69–87, 2010. 1, 2, 6

[3] A. Bialkowski, P. Lucey, P. Carr, S. Denman, I. Matthews, and S. Sridharan. Recognising team activities from noisy data. In *Proc. of Computer Vision in Sports, Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2013. 1, 2

[4] R. Gade and T. B. Moeslund. Sports type classification using signature heatmaps. In *Proc. of Computer Vision in Sports, Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2013. 1

[5] A. Gaidon, Z. Harchaoui, and C. Schmid. Activity representation with motion hierarchies. *International Journal of Computer Vision*, pages 1–20, 2013. 1

[6] C.-C. Hsu, H.-T. Chen, C.-L. Chou, C.-P. Ho, and S.-Y. Lee. Trajectory based jump pattern recognition in broadcast volleyball videos. In *Proc. of International Conference on Multimedia*, MM '14, pages 1117–1120, New York, NY, USA, 2014. ACM. 1

[7] S. Karaman, L. Seidenari, S. Ma, A. Del Bimbo, and S. Sclaroff. Adaptive structured pooling for action recognition. In *Proc. of the British Machine Vision Conference (BMVC)*. BMVA Press, 2014. 1, 6

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proc. of Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014. 1

[9] A. Kläser. *Learning human actions in video*. PhD thesis, Université de Grenoble, jul 2010. 6

[10] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proc. of Computer Vision Pattern Recognition (CVPR)*. IEEE, 2010. 6

[11] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2011. 6

[12] J. Liu, P. Carr, R. T. Collins, and Y. Liu. Tracking sports players with context-conditioned motion models (oral). In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2013. 1

[13] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2013. 1

[14] J. C. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Proc. of European Conference on Computer Vision (ECCV)*. Springer, 2010. 1

[15] D. Oneata, J. Verbeek, and C. Schmid. Action and Event Recognition with Fisher Vectors on a Compact Feature Set. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2013. 1

[16] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proc of. European Conference on Computer Vision (ECCV)*, 2010. 3

[17] J. Revaud, M. Douze, C. Schmid, and H. Jégou. Event retrieval in large video collections with circulant temporal encoding. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013. 1

[18] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *Proc of. Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008. 6

[19] M. J. Roshtkhari and M. D. Levine. Online dominant and anomalous behavior detection in videos. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013. 1

[20] M. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2011. 1

[21] K. Soomro and A. R. Zamir. Action recognition in realistic sports videos. In *Computer Vision in Sports*, pages 181–208. Springer, 2014. 1

[22] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2003. 4

[23] G. Waltner, T. Mauthner, and H. Bischof. Indoor activity detection and recognition for automated sport games analysis. In *Proc. Workshop of the Austrian Association for Pattern Recognition (AAPR/OAGM)*, 2014. 2, 4, 6

[24] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, pages 60–79, 2013. 1, 2, 6

[25] H. Wang and C. Schmid. Action recognition with improved trajectories. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 3551–3558. IEEE, 2013. 1

[26] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012. 1

[27] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *Proc. of International Conference on Computer Vision (ICCV)*. IEEE, 2009. 6