# Group Action Recognition Using Space-Time Interest Points

Qingdi Wei[1], Xiaoqin Zhang[1], Yu Kong[2], Weiming Hu[1], Haibin Ling[3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing 100190, P.R. China
[2]Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, P.R. China
[3]Center for Information Science and Technology, Computer and Information Science Department, Temple University, Philadelphia, PA, USA
{qdwei,xqzhang,wmhu}@nlpr.ia.ac.cn, {kongyu}@bit.edu.cn, {hbling}@temple.edu

**Abstract.** Group action recognition is a challenging task in computer vision due to the large complexity induced by multiple motion patterns. This paper aims at analyzing group actions in video clips containing several activities. We combine the probability summation framework with the space-time (ST) interest points for this task. First, ST interest points are extracted from video clips to form the feature space. Then we use k-means for feature clustering and build a compact representation, which is then used for group action classification. The proposed approach has been applied to classification tasks including four classes: badminton, tennis, basketball, and soccer videos. The experimental results demonstrate the advantages of the proposed approach.

## 1 Introduction

Understanding group activities is an important problem in computer vision with many applications, such as video surveillance and monitoring, object-level video summarization, human-computer interaction, video indexing and browsing, and digital library organization. Despite many research efforts on the human activity analysis, group action classification remains a challenging task and has not been widely studied for the following reasons:

(i) It is difficult to find an effective descriptor for group human action, because there are usually many people performing different actions individually.

(ii) When using local features to describe an individual action, there are too many objects to track. In addition, the environmental background for group action is often highly cluttered.

(iii) Nuisance factors, such as the number of people in the group action, mutual-occlusion and self-occlusion, irregularity of camera parameters, also cause additional difficulties.

In this paper, we combine the probability summation framework with the space-time (ST) [1] feature point for group activity recognition. First, space-time interest point features are extracted to describe the group action, as shown

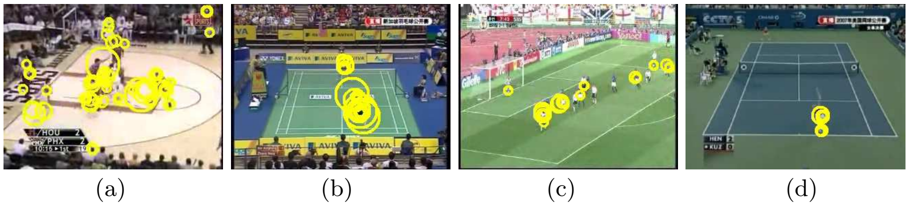

(a) (b) (c) (d)

**Fig. 1.** Local space-time features detected for sports: (a) basketball; (b) badminton; (c) soccer; (d) tennis.

in Fig. 1. Then, the k-means clustering algorithm is employed to cluster action features into the action codebook. Finally, a testing video is classified by computing the probability summation, which ignores the number of people in the group action. our purpose is to distinguish the group actions, just as basketball and soccer. High-precision work, such as attack and defend, will be done in future.

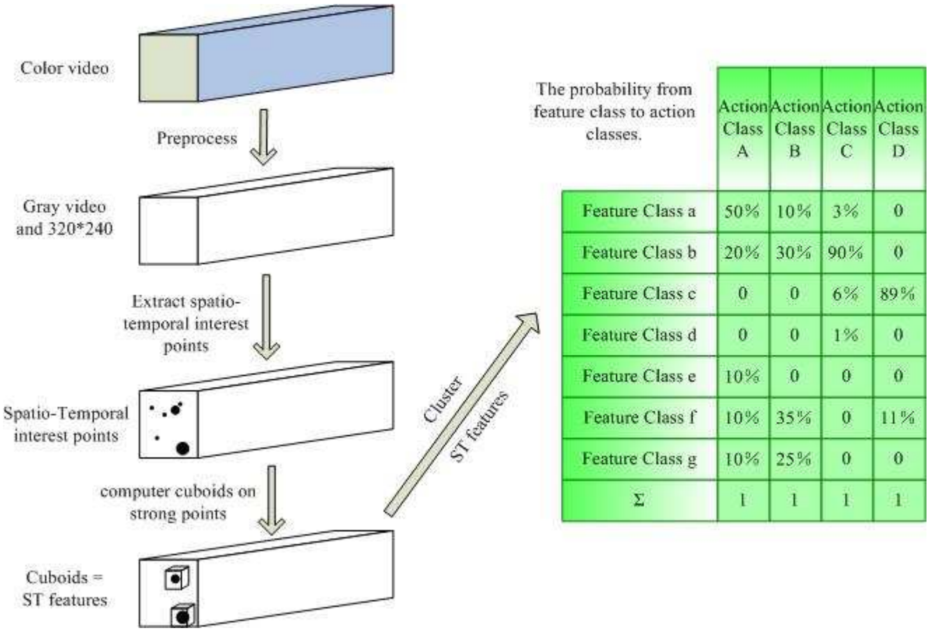
The highlights of our work include: 1) the ST features provide an effective description for the group action in a video clip without tracking and action segmentation; 2) the probabilistic framework provides an effective way for integrating global information of the feature, which makes our method robust to the image noises and invariant to translation, rotation and scaling; and 3) the probabilistic framework is computationally very efficient. In fact, the system has a linear time complexity, given previously extracted features.

The rest of the paper is organized as follows: In Section 2 we discuss the related work. In Section 3 we deal with the major problem, in which action features and the action recognition method are respectively presented in Section 3.1 and Section 3.2. Experimental results are shown in Section 4, and Section 5 is devoted to conclusion.

## 2 Related Work

Generally, there are two major parts in a human action recognition system: human action representation, recognition strategy.

Laptev [1] propose space-time interest points for a compact representation of video data, and explore the advantage of using space-time interest points to describe human action. The ST feature does not need any segmentation or tracking of the individual performance of the action. With this property, space-time features have recently shown considerable success on action recognition [2], [3], [4], [5], [6]. Niebles, Wang and Fei-Fei [3] extract space-time interest points to represent a video sequence as a collection of space-time words, and use a probabilistic Latent Semantic Analysis (pLSA) model to recognize human actions. Schuldt [2] and Laptev [6] also use the ST feature but they prefer the codebook and the bag-of-words. However, most existing methods for action recognition mentioned above focus on individual actions. Efros, Berg, Mori et al. [7] takes soccer video as their experiment data but they also merely recognize single person's action. Kong et al. [8] use optical flow as action features to recognize group


Color video
Preprocess
Gray video and 320*240
Extract spatio-temporal interest points
Spatio-Temporal interest points
computer cuboids on strong points
Cuboids = ST features
Cluster ST features

| The probability from feature class to action classes. | Action Class A | Action Class B | Action Class C | Action Class D |
|---|---|---|---|---|
| Feature Class a | 50% | 10% | 3% | 0 |
| Feature Class b | 20% | 30% | 90% | 0 |
| Feature Class c | 0 | 0 | 6% | 89% |
| Feature Class d | 0 | 0 | 1% | 0 |
| Feature Class e | 10% | 0 | 0 | 0 |
| Feature Class f | 10% | 35% | 0 | 11% |
| Feature Class g | 10% | 25% | 0 | 0 |
| Σ | 1 | 1 | 1 | 1 |

**Fig. 2.** the training phase

action in soccer videos. This work is limited on soccer videos, and can only handle three categories of group actions. In [9], Ali and Shah track individual targets in density crowd using a scene structure based force model.

Recognition strategy is another important part in an action recognition system. In the literature, there is a large number of work referring to this problem, and many impressive results have been obtained over the past several years, such as Hidden Markov Models (HMMs), Autogressive Moving Average (ARMA) [10], Conditional Random Fields (CRFs) [11], Finite State Machine(FSM) [12], [13] and their variations [14], semi-Markov model [15], 1-Nearest Neighbor with Metric Learning [16], ActionNets [17], LDCRF [18]. Wang and Suter [19] presented the use of FCRF in the vision community, and demonstrated its superiority to both HMM and general CRF. Ning, Xu, Gong and Huang [20] improved standard CRF model and its variations performance on continuous action recognition, by replacing the traditional random fields model with a latent pose estimator. Boiman and Irani [21] proposed a graphical Bayesian model which describes the motion data using hidden variables that correspond to hidden ensembles in a database of spatio-temporal patches. Vitaladevuni, Kellokumpu and Davis [22] presented a Bayesian framework for action recognition through ballistic dynamics.

Compared with the previous work, we target on group human actions that are more general and complex. As shown in Section 4, we work on four different group actions.

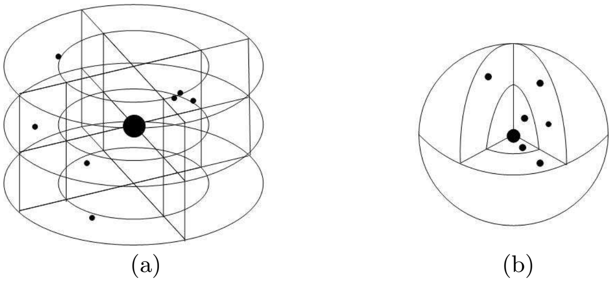


**Fig. 3.** (a) cylindrical histogram (b) spheriform histogram

## 3 Recognition method

We first define symbols used in the work. Denote $V$ as the input video, and $ST = \{st_i\}_{i=1}^{m}$ is denoted as the action feature, where each $st_i$ denotes the cuboids of a space-time interest point in video $V$. The recognition task is usually formed as a classification problem. Specifically, our purpose is to find a classifier $f : f(ST) = c$ that classifies a given sequence to a predefined group action $c \in C = \{1, \cdots, n\}$ , where $C$ is a set of group actions that we are interested in.

Our algorithm consists of two stages which are the training phase and the testing phase. As shown in Fig. 2, the training phase contains the following steps:

1, Preprocess color video clips into gray ones and reduce the resolution to 320 × 240.

2, Extract space-time interest points from each video clip, and retain the cuboids on the interest points with significant feature strengths.

3, Cluster ST features to the feature classes, forming a sample database.

4, The priori probability of each feature class is estimated from the frequency of action classes.

The testing phase contains the similar steps:

1, Preprocess color video clips into gray ones and reduce the resolution to 320 × 240.

2, Extract space-time interest points from each video clip, and computer cuboids on interest points which have significant relative feature strengths.

3, Classify ST features to get their feature class label. Here, each feature class label means a probability to action classes.

4, Sum the probability, and take the maximum, then group action recognition is completed.

### 3.1 Action Feature

The space-time interest point is used for a compact representation of video data and robust to occlusions, background clutter, significant scale changes, and high action irregularities. They have been successfully used as action feature in action recognition tasks [2], [3], [4]. In our work, we detect space-time interest points using a periodic detector. Then a cuboid [4] is extracted at each interest point, which contains the space-timely windowed pixel values. The Euclidean distance metric is adopted to evaluate the similarity of two cuboids.

A cuboid $st_i$ of a space-time interest point is a description of information contained in the neighbor pixels, such as location, scale, histogram of gradient (HOG) and the like. In our system, we choose to concatenate all pixels in the cuboid for the description. Then the video is represented by a discrete set $ST = \{st_i\}_{i=1}^{m}$ , containing $m$ cuboids sampled from the video. PCA is applied to the descriptor for dimension reduction. We also experiment with 3D-shape context, which is similar as [23]. Huang and Trivedi [23] use multilayered cylindrical histogram to describe the human body voxels,while we employ spheriform histogram in our experiment. The $st_i$ is a $8 \times 8 \times 8$ matrix in the process of computing 3D-shape context on the interest points.

### 3.2 Recognition

Using the k-means algorithm, we cluster a large number of cuboids extracted from the training data into $k$ clusters in the training section. We call these clusters as feature classes. Since a feature class rarely comprises of cuboids features belonging to the same action class, it is reasonable that the transformation probability $T_i$ can be estimated by the frequencies of each action class appearing in the corresponding feature class. The transformation probability $T_i$ measures the probability of each action occurs when feature $i$ appears.

$$T_i : feature\ class\ i \begin{cases} action\ class\ 1\ p_{i1} \\ \quad\vdots \qquad\qquad \vdots \\ action\ class\ n\ p_{in} \end{cases} \tag{1}$$

where

$$\sum_{i=1}^{k} p_{ij} = 1 \tag{2}$$

The Eq. (2) can be replaced by $\sum\limits_{j=1}^{n} p_{ij} = 1$ , but it is sensitive to the number of each action class samples. For instance, if one action class seldom appears, it will never appear in the experiment results. In this sense, the Eq. (2) works better. Moreover, the number of feature classes $k$ is important to the result, which we will discuss in section 4.3.

In the testing section, we first classify each cuboid in the test video to get a set of feature classes. Then the transformation is applied to the set of feature class: $T(feature\ class)$. Each feature class from test video corresponds to a $1 \times n$ vector (Fig. **??**). So we get an $k \times n$ matrix, action class probability table, after the transformation. In the recognition process, the probabilities of action classes are summed up by Eq. (3), and then the $\max(ap_j)$ is taken as the action class in this video. We call this method as the *probability summation*.

$$\textit{action class probability} : ap_j = \sum_{i=1}^{m} (p_{1j}, \ldots, p_{ij}) \tag{3}$$

In the experiment we also use another method template matching to make a contrast. In the *template matching*, we first study the action class templates in the training section. Each template, which is a histogram of feature class, corresponds to an action class. When a test video comes, we compute the histogram of feature class, and use the Euclidean distance to evaluate the similarity between the test video and the template.

## 4 Experiments

### 4.1 Group Human Action Video Database

For the experiments, we build a video database containing four types of group human actions (tennis, badminton, soccer, and basketball, see Fig. 4) from several real sport videos collected from Internet. The reason why we choose these four kinds of sports is as follows. The soccer and the basketball both involve a number of people with a great deal of movements. The tennis and the badminton have two or four players in the game and share a similarity of movement frequency. Furthermore, the other three, except the basketball, all take green as their main color so that we cannot distinguish them simply through the aspect of color. As is known to all, the movement information and the color information change dramatically in one sport video due to the long shot and close-up. Therefore, to classify them correctly is not a simple task even in the ideal case, let alone the original videos are recorded under complex backgrounds with a moving camera. These video data are 25 fps frame rate, with various resolutions, such as 1280 × 720, 480 × 372. When creating the experimental video database, we downsample all videos to the resolution of 320 × 240 and have a length of one hundred frames on average. Our database contains 1193 videos with their ground label. To the best of our knowledge, this is the first video database with group human actions.

### 4.2 Experimental Method

We divide the video database into two parts, the training part (599 videos), and the testing part (594 videos). Space-time interest points are extracted on all videos(Fig. 4). There are a large number of ST points detected in most videos, for each video, we only choose 50 ST points with the largest feature responses for efficiency. In fact, we find in our experiments that 50 points contain sufficient information for group action recognition. Meanwhile, those videos which fail to contain 50 interest points can also be recognized correctly. The codes for detecting Space-time interest points and extracting cuboids are provided by

**Table 1.** Confusion matrix of template matching

| | badminton | tennis | soccer | basketball |
|---|---|---|---|---|
| badminton | .87 | .08 | .05 | .00 |
| tennis | .09 | .90 | .01 | .00 |
| soccer | .05 | .05 | .81 | .09 |
| baskerball | .14 | .04 | .00 | .82 |

**Table 2.** Confusion matrix of probability summation

| | badminton | tennis | soccer | basketball |
|---|---|---|---|---|
| badminton | .92 | .02 | .06 | .00 |
| tennis | .06 | .94 | .00 | .00 |
| soccer | .05 | .01 | .88 | .06 |
| baskerball | .15 | .06 | .24 | .75 |

piotr_Activity Recognition Toolbox[1]. The experiment is performed on matlab, and is running on a CY2.4 GHz personal computer with 4G memory. It takes 40 seconds to process a 4 seconds video. The most time is spent on the PCA, around 36 seconds.

Furthermore, we obtain 29k cuboids for training with the k-mean algorithm. The number of clusters is set to $k = 200$. So the transformation probability $T$ is a $200 \times 4$ matrix. We normalize each cluster in $T$ by Eq. (2) as the numbers of samples of every group action are uneven.

### 4.3 Results

In the recognition process, we evaluate two approaches by comparing their recognition rates: one is the template matching; the other is the probability summation. Table 1 shows the confusion matrix of template matching, while Table 2 is for probability summation. We can see that the probability summation performs better than the template matching. The reason is that each sport has various motions, for example, serving and spiking. In probability summation, we can correctly classify a sport even when serving or spiking is performed respectively. In comparison, template matching requires that both serving and spiking should be included in the video. The average recognition rate using probability summation is 88.48%. The rate is very promising, especially considering the limit training data and the peculiarities of the action.

### 4.4 Discussion

We also use 3D-shape context to do the experiment whose result is shown in Table 3. The first line, 3D-SC-Only distance, means the feature of the 3D-shape context is only binned in distance domain. We use PCA to compress 3D-shape

[1] http://vision.ucsd.edu/∼pdollar/research/cuboids_doc/index.html

**Table 3.** Recognition rate with various parameters

| k | 3D-SC<br>(only distance) | 3D-SC<br>PCA50 | 3D-SC<br>PCA90 | 3D-SC<br>PCA134 | Original-cuboids<br>PCA111 |
|---|---|---|---|---|---|
| 200 | 53.97 | 64.48 | 64.19 | 67.10 | 85.57 |
| 400 | 56.06 | 67.39 | 71.76 | 71.18 | 86.15 |
| 600 | 58.15 | 73.65 | 72.78 | 72.63 | 90.76 |
| 800 | 60.03 | 74.24 | 74.67 | 73.51 | 90.33 |
| 1000 | 59.88 | 77.44 | 75.98 | 78.17 | 91.49 |
| 1200 | 63.06 | 79.62 | 81.37 | 77.87 | 92.50 |
| 2000 | 65.66 | 82.10 | 81.51 | 83.70 | 93.51 |
| 4000 | 67.82 | 85.74 | 85.88 | 85.74 | 95.53 |
| 6000 | 68.47 | 87.05 | 87.05 | 87.34 | 95.38 |
| 10000 | 69.12 | 88.94 | 88.94 | 88.29 | 96.10 |
| 20000 | 70.27 | 90.25 | 90.83 | 90.83 | 95.67 |

context which is a 512-dimension vector. The lines 2-4 respectively correspond to the 50-dim, 90-dim, and 134-dim. The 134-dim contains more-than-90% information. The last line is the original cuboids feature which works well especially when the $k$ is small.

In our method, $k$, the number of clusters is an important factor of affecting the experimental results. As $k$ increases, recognition rate also rises accordingly with the reducing acceleration. As mentioned in section 4.2, we obtain 29k features for training. When the $k = 20000$, it means one cluster contains few samples which is possibly one or two. Together with *probability summation*, our method is just like a primary boost algorithm, in which each feature works as a weak classifier.

## 5 Conclusion

In this paper, our contribution focuses on two points. First, we build a group human action database with thousands of videos and it keeps becoming larger. Second, we try our attempt to analyze group action, and achieve satisfying experimental results.

Our method has the potential to analyze group human action in details, because the ST feature can specifically describe action information.

At present, our database only has four group actions. We will extend our method to more categories of human group action, and enrich our database as well. As future work, we would like to consider the group/global features to enhance our work.

## Acknowledgment

This work is supported by NSFC (Grant No. 60825204, 60672040,) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

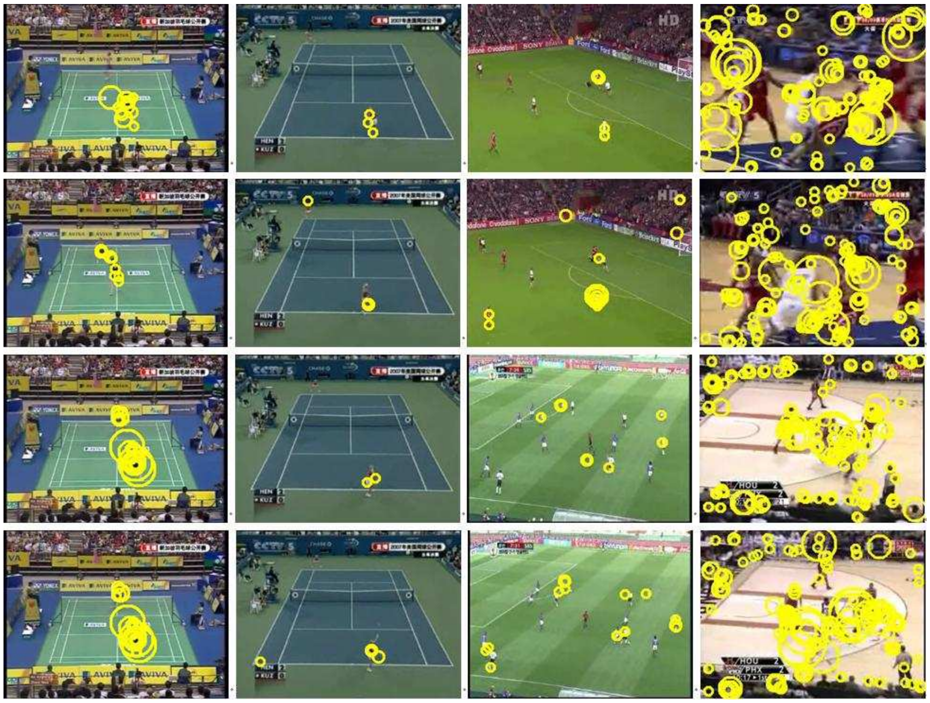


**Fig. 4.** Space-time interest point in the sport video

## References

1. Laptev, I.: On space-time interest points. International Journal of Computer Vision **64** (2005) 107–123
2. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition. Volume 3., Washington, DC, USA, IEEE Computer Society (2004) 32–36 Vol.3
3. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. International Journal of Computer Vision **79** (2008) 299–318
4. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. (2005) 65–72
5. Gilbert, A., Illingworth, J., Bowden, R.: Scale invariant action recognition using compound features mined from dense spatio-temporal corners. In: Proceedings of the 10th European Conference on Computer Vision, Berlin, Heidelberg, Springer-Verlag (2008) 222–233
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8

7. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 726
8. Kong, Y., Zhang, X., Wei, Q., Hu, W., Jia, Y.: Group action recognition in soccer videos. In: 19th International Conference on Pattern Recognition. (2008) 1–4
9. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Proceedings of the 10th European Conference on Computer Vision, Berlin, Heidelberg, Springer-Verlag (2008) 1–14
10. A, V., Roy-Chowdhury, A., Chellappa, R.: Matching shape sequences in video with applications in human movement analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **27** (2005) 1896–1909
11. Sminchisescu, C., Kanaujia, A., Metaxas, D.: Conditional models for contextual human motion recognition. 10th IEEE International Conference on Computer Vision **104** (2006) 210–220
12. Natarajan, P., Nevatia, R.: View and scale invariant action recognition using multiview shape-flow models. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8
13. Zhao, T., Nevatia, R.: 3d tracking of human locomotion: A tracking as recognition approach. In: Proceedings of the 16th International Conference on Pattern Recognition, Washington, DC, USA, IEEE Computer Society (2002) 10546
14. Sminchisescu, C., Kanaujia, A., Li, Z., Metaxas, D.: Conditional models for contextual human motion recognition. In: 10th IEEE International Conference on Computer Vision. Volume 2. (2005) 1808–1815 Vol. 2
15. Shi, Q., Wang, L., Cheng, L., Smola, A.: Discriminative human action segmentation and recognition using semi-markov model. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8
16. Tran, D., Sorokin, A.: Human activity recognition with metric learning. In: Proceedings of the 10th European Conference on Computer Vision, Berlin, Heidelberg, Springer-Verlag (2008) 548–561
17. Lv, F., Nevatia, R.: Single view human action recognition using key pose matching and viterbi path searching. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8
18. Morency, L.P., Quattoni, A., Darrell, T.: Latent-dynamic discriminative models for continuous gesture recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8
19. Wang, L., Suter, D.: Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model. In: IEEE Conference on Computer Vision and Pattern Recognition. (2007) 1–8
20. Ning, H., Xu, W., Gong, Y., Huang, T.: Latent pose estimator for continuous action recognition. In: Proceedings of the 10th European Conference on Computer Vision, Berlin, Heidelberg, Springer-Verlag (2008) 419–433
21. Boiman, O., Irani, M.: Detecting irregularities in images and in video. In: 10th IEEE International Conference on Computer Vision. Volume 1. (2005) 462–469 Vol. 1
22. Vitaladevuni, S., Kellokumpu, V., Davis, L.: Action recognition using ballistic dynamics. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8
23. Huang, K.S., Trivedi, M.M.: 3d shape context based gesture analysis integrated with tracking using omni video array. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, Washington, DC, USA, IEEE Computer Society (2005) 80