# Human Activity Recognition using the 4D Spatiotemporal Shape Context Descriptor

Natasha Kholgade and Andreas Savakis

Department of Computer Engineering
Rochester Institute of Technology, Rochester NY 14623
natasha.kholgade@gmail.com, andreas.savakis@rit.edu

**Abstract.** In this paper, a four-dimensional spatiotemporal shape context descriptor is introduced and used for human activity recognition in video. The spatiotemporal shape context is computed on silhouette points by binning the magnitude and direction of motion at every point with respect to given vertex, in addition to the binning of radial displacement and angular offset associated with the standard 2D shape context. Human activity recognition at each video frame is performed by matching the spatiotemporal shape context to a library of known activities via k-nearest neighbor classification. Activity recognition in a video sequence is based on majority classification of the video frame results. Experiments on the Weizmann set of ten activities indicate that the proposed shape context achieves better recognition of activities than the original 2D shape context, with overall recognition rates of 90% obtained for individual frames and 97.9% for video sequences.

## 1 Introduction

Human activity recognition has important applications in the areas of video surveillance, medical diagnosis, smart spaces, robotic vision and human computer interaction. Several strategies have been employed in activity recognition to develop features such as principal components [1, 2] and skeletonization [4, 5, 6] from the human subjects in video sequences, and to recognize these features by HMMs, neural networks and nearest neighbors. Improvement in recognition can be obtained when spatial and temporal features are combined into one representation. Pioneering work in this field was done in [3] by using motion energy and motion history images. The authors of [16] use spatiotemporal features such as sticks, balls and plates developed as solutions of the Poisson equation on a volume constructed by concatenating frames of a subject's silhouette, obtained by background subtraction, along the temporal axis. The authors of [21] use spatiotemporal features found by linear filters for activity recognition; these have also been used by [18] in a hierarchical model (of the constellation of bags-of-features type) together with spatial features found with the 2D shape context [7]. In [20], the authors store distances between all combinations of frame pairs in a self-similarity matrix and analyze the inherent similarities in action recognition. In [21], a two-stage recognition process is used together with local spatiotemporal discriminant embedding; in the first stage, the silhouette is projected into a space where discrimination is enhanced between classes that are further apart in the spatial domain, and if such discrimination is not obtained, then in the second stage, a short segment of frames centered at the present frame is used to form a temporal subspace for discrimination.

The precursor to the descriptor proposed in this paper is the 2D shape context [7]. The 2D shape context is a shape descriptor that bins points in the contour of an object using a log-polar histogram. For every point along the contour, such a log-polar histogram is maintained that counts how surrounding points fall within various sections. Contour points are binned according to their radial distance and angular offset from the point under consideration. The use of these two criteria for binning makes its histograms two-dimensional resulting in the 2D shape context, but typically, the histograms are vectorized, and histogram vectors for various points are stacked to form a matrix. In [7] and [8], in-depth description of the shape context and its use in applications such as matching numerical digits, objects and trademarks has been provided. In the area of human subject representation, the 2D shape context was used for human body configuration estimation in still images [9], pose estimation in motion sequences [11], and action recognition [10]. Extension of this descriptor to 3D was proposed in [12] where the log-polar histogram is stretched out into the third dimension to form a cylindrical histogram. A spherical-histogram based 3D shape context has been used in [13] for action recognition on a spatiotemporal volume formed (as in [16]) by concatenating silhouettes along the temporal dimension. Points for the 3D shape contexts are usually obtained from the surface of the volumes (3D objects or spatiotemporal); however, in [14], they have been generated by using all the voxels within the 3D object.

In this paper, we propose a 4D spatiotemporal shape context (STSC) descriptor, which captures both spatial and temporal description of the object based on its contour shape and motion over consecutive frames. For a contour point under consideration, in addition to the radial distance and angular offset of surrounding points used for spatial representation, the STSC uses two more criteria, namely the magnitude and direction of the velocity at the surrounding points for development of histograms. The use of velocity magnitude allows distinction between fast and slow moving objects for example, running versus walking, while the use of direction allows distinction between the trajectories of object parts, for example, bending versus jumping. Since there are two additional criteria of magnitude and direction for binning velocity information, the resulting histograms in the STSC are four-dimensional. The four-dimensional histograms for various points are vectorized and concatenated into a matrix for processing.

For activity recognition applications, the advantage of the STSC descriptor over the traditional 2D shape context is that it incorporates local motion information that makes it possible to distinguish between similarly-shaped stances of activities that are separable due to the subject's motion. In comparison with the 3D shape context in [12], it requires fewer frames and can be applied for activity recognition in multiple-action sequences. The spatiotemporal shape context is applicable to other areas besides activity recognition, such as expression recognition, semantic annotation and content-based video retrieval.

This paper is organized as follows. Section 2 outlines the feature extraction process for generating the spatiotemporal shape context, Section 3 presents a comparison of results for activity recognition on the Weizmann dataset [16] using the 2D and spatiotemporal shape contexts, and comparison of our results against those from other works on the same dataset, and Section 4 includes concluding remarks.

## 2 Methodology

This section outlines the process of generating the STSC descriptor and the methodology for using it for activity recognition.

### 2.1 Extraction of points and generation of motion vectors

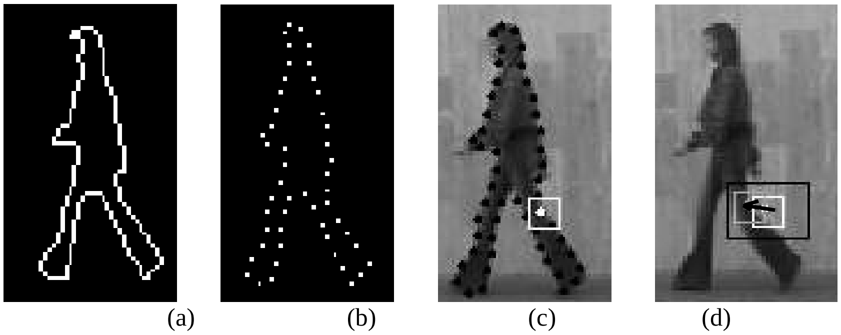

**Fig. 1.** Motion vector extraction for a point along the boundary of a walking subject: (a) contour of the subject, (b) 50 equidistant points from the contour; (c) for a given silhouette point a 7x7 window (white) around the point is selected; (d) the best match 7x7 window (light gray) in the *next* frame within a 13x19 search region (dark gray) is found and the displacement between the centers (black arrow) represents the motion vector

For each video frame, background subtraction is performed by subtracting a background frame obtained by concatenating subject-free halves of the corresponding subject's walking video. After background subtraction, the contour of the subject's silhouette is obtained at each frame using the chain code. Uniformly spaced points on the silhouette contour are selected for generating the STSC. The number of silhouette points *N* is chosen to be 50 in this paper, but it is a parameter that may be varied to optimize performance or efficiency. To estimate the motion vectors, a 7x7 square centered at each contour point is selected in the current video frame and the nearest 7x7 square within the next video frame is found by searching within a 13x19 window. The displacement between the centers of their centers provides a motion estimate for that particular point. Fig. 1 shows the process of getting motion vectors. Examples of the resulting contour points and their motion vectors are provided in Fig. 2.

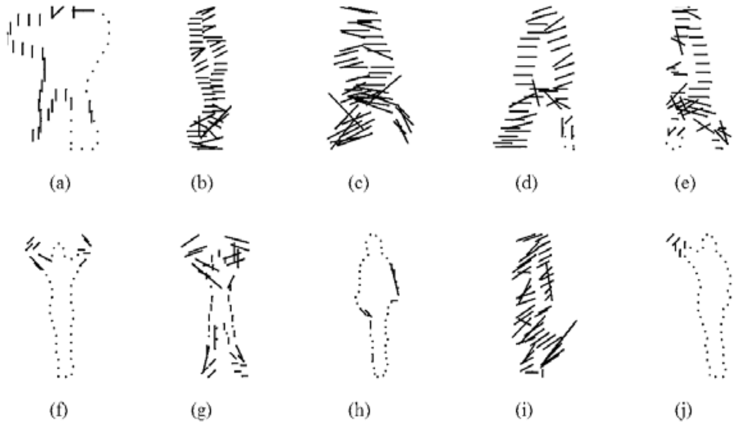

**Fig. 2.** Boundary points, with 50 points per silhouette, and corresponding motion vectors for (a) bend, (b) run, (c) jump forward, (d) side-shuffle, (e) walk, (f) wave with two hands, (g) jumping jacks, (h) jump in place, (i) skip, (j) wave with one hand

## 2.2 Generation of the Spatiotemporal Shape Context

The STSC can be generated by specifying the number of bins required along the radial, angular, motion magnitude and motion direction dimensions denoted by $n_r$, $n_\theta$, $n_{vr}$, and $n_{v\theta}$ respectively, and the bounds for the distance and motion magnitudes, namely $r_{min}$, $r_{max}$, $v_{max}$, and $v_{min}$=0. Given points $P_i$ and $P_j$ on the contour of an object, the individual bin for the distance $r$ between $P_i$ and $P_j$ is determined as:

$$b_r = \arg\min_i \left( \left| r - t_r(i) \right| \right) \text{ subject to } r < t_r(i) \tag{1}$$

where $t_r$ is the $n_r$-element threshold vector given as

$$t_r(i) = \exp_{10}\left( \log(r_{\min}) + \frac{i}{n_r}\left(\log(r_{\max}) - \log(r_{\min})\right)\right) \tag{2}$$

Similarly, the individual bin for the motion change magnitude $v_r$ at point $P_j$, is determined as:

$$b_v = \arg\min_i \left( \left| v_r - t_{vr}(i) \right| \right) \text{ subject to } v_r < t_{vr}(i) \tag{3}$$

where $t_{vr}$ is an $n_{vr}$-element threshold vector given by Equation (4).

$$t_{vr}(i) = \frac{(i-2)\cdot v_{\max}}{n_{vr} - 2} \tag{4}$$

The vectors $t_r$ and $t_{vr}$ contain the thresholds for the various bins within which the $r$- and $v_r$-values of the point $P_j$ get categorized. The selected bins form an upper cap on these values. Bins for angular offset $\theta$ and motion change direction $v_\theta$ are obtained as

$$b_\theta = 1 + \left\lfloor \frac{\theta \cdot n_\theta}{2\pi} \right\rfloor \tag{5}$$

$$b_{v\theta} = 1 + \left\lfloor \frac{v_\theta \cdot n_{v\theta}}{2\pi} \right\rfloor \tag{6}$$

The index for the final bin in the STSC-vector for $P_i$ in which to place $P_j$ is given as:

$$f = (b_r - 1)n_\theta n_v n_{v\theta} + (b_\theta - 1)n_v n_{v\theta} + (b_v - 1)n_{v\theta} + b_{v\theta} \tag{7}$$

Let $H_i$ be the shape context vector for $P_i$; then $H_i$ is incremented by one at position $f$ to reflect the inclusion of $P_j$ in the shape context vector description of $P_i$:

$$H_i(f) := H_i(f) + 1 \tag{8}$$

The process is repeated for all points to get a matrix of size $N \times n_r\, n_\theta\, n_{vr}\, n_{v\theta}$ which forms the STSC. In our implementation, $n_r$ = 5, $n_\theta$ = 12, $n_{vr}$ = 5, $n_{v\theta}$ = 12. These parameters may be varied in order to obtain optimal performance. Fig. 3 gives an image version of the STSC matrix; it tends to be sparse.

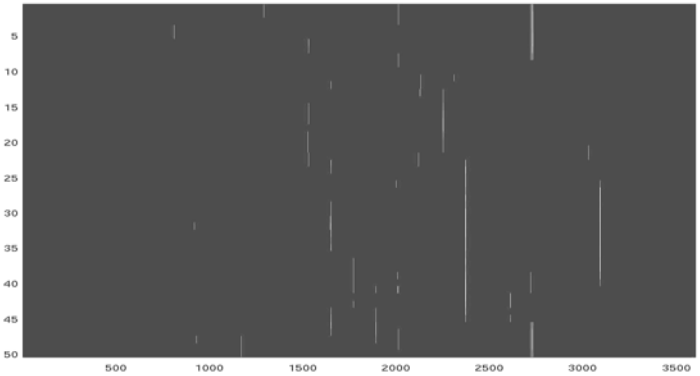

**Fig. 3.** Spatiotemporal shape context (STSC) developed for the frames in Fig. 1: the STSC matrix shown here has 50 rows corresponding to the 50 points in Fig. 1 (b), and 3600 columns corresponding to $n_r\ n_\theta\ \ n_{vr}\ n_{v\theta}$ = 5x12x5x12; each row has a total of 50 points that have been binned according to their radial displacement from the point corresponding to that row, angular offset, magnitude of motion vector and direction of motion vector

### 2.3 Matching and Classification with the Spatiotemporal Shape Context

For a given pair of images or video frames, a match value can be attributed by computing the STSCs of the objects in the two images, and by finding an optimal correspondence between the two STSCs. This correspondence is essential prior to matching. Since rows of the STSC correspond to points from the object contour, row-wise ordered matching is not guaranteed between STSCs of two image pairs, even if they are visually similar in shape and motion. To introduce an ordered match, an optimal permutation must be computed for the rows of one STSC. The authors of [7] compute the correspondence between their 2D shape contexts with the Hungarian algorithm [15], which we adopt in our implementation. The Hungarian algorithm computes correspondence using an $N$x$N$ cost matrix as input which can be obtained by computing pair-wise distances between each of the $N$ rows in one STSC to each of the $N$ rows in the other STSC. The distances are computed using the $X^2$ metric; if $H_i$ is the $i$-th row in the first STSC, and $H_j$ is the $j$-th row in the second STSC then the element of the cost matrix at the $(i,j)$-th location is given using the $X^2$ metric as:

$$C_{ij} = \frac{1}{2}\sum_{k=1}^{N}\frac{\left(H_i(k) - H_j(k)\right)^2}{H_i(k) + H_j(k)} \tag{9}$$

Once the Hungarian algorithm is used to compute the optimal correspondence between rows of the two STSCs, then a match value can be assigned to the pair by computing the sum of $X^2$ distances between optimally corresponded rows of the STSCs. A small match value indicates that two activity stances are similar in their shapes and motion, while a large match value indicates the opposite. This allows $k$-NN classification to be used on the match values in order to identify the activity. For the purpose of classification, a library of action stances and their STSCs can be maintained, and one can compute the STSC of an input frame from itself and the next adjacent frame using the equations listed in section 2.2. Match values can then be computed between the input frame's STSC and the STSCs in the library after optimal correspondence is found using the Hungarian algorithm, and the $k$ nearest match values can be used to classify the input frame's action. Such a classification is done for every frame in the sequence and the activity for the entire video is found by majority voting on the results of individual frames within the video.

## 3 Results

Recognition was performed on the ten activities of the Weizmann dataset [16]. Two experiments were conducted for the purpose of comparison: one using the 2D shape context, and one using the STSC. For each experiment, a library of the corresponding shape contexts was maintained by hand-picking 10 frames per subject per activity, and computing shape contexts for each frame. The spatiotemporal shape context calculation was done using silhouette points on each frame and motion information between successive frames. Classification was done using the leave one out technique, i.e. for every test subject, library shape contexts corresponding to all activities for that subject were left out. During testing, shape contexts were computed from each frame in the video sequence and were matched to the library using *k*-NN with *k*=1, i.e. the closest matching library shape context was selected. The activity corresponding to the closest match was used to classify the input frame, and for a video sequence, majority vote of the activity classifications for individual frames were used to make the activity decision for the video. Results from the two experiments were used to generate confusion matrices that contain percentages of frames classified as various activities for individual video frames and entire video sequences.

**Table 1.** Results from individual frame recognition using the 2D shape context

| | **B** | **JF** | **R** | **SS** | **W** | **V2** | **JJ** | **JP** | **S** | **V1** |
|---|---|---|---|---|---|---|---|---|---|---|
| **B** | **93.5** | 4.3 | 0 | 0 | 0 | 0 | 0 | 1.1 | 0 | 1.1 |
| **JF** | 5.1 | **79.5** | 1.6 | 0.2 | 3.1 | 0 | 0.2 | 0.2 | 10.1 | 0 |
| **R** | 0 | 2.0 | **72.4** | 1.3 | 8.4 | 0 | 0 | 0.2 | 15.7 | 0 |
| **SS** | 0.5 | 0.2 | 0.2 | **86.2** | 0.7 | 0.2 | 0.9 | 11.1 | 0 | 0 |
| **W** | 0 | 5.0 | 3.6 | 0 | **88.3** | 0 | 0.3 | 0 | 2.8 | 0 |
| **V2** | 0 | 0 | 0 | 0.3 | 0 | **95.9** | 2.8 | 0.8 | 0 | 0.2 |
| **JJ** | 0 | 0 | 0 | 2.5 | 0 | 2.6 | **87.4** | 7.5 | 0 | 0 |
| **JP** | 0.8 | 0 | 0 | 5.3 | 0 | 0 | 2.4 | **91.3** | 0 | 0.2 |
| **S** | 0 | 7.4 | 31.1 | 0 | 6.5 | 0 | 0.2 | 0 | **54.8** | 0 |
| **V1** | 1.9 | 0 | 0 | 0 | 0 | 0 | 0 | 1.5 | 0 | **96.6** |

Table 1 provides results for classifying individual video frames and entire video sequences respectively using the 2D shape context. The nomenclature is as follows: B=bend, JF=jump forward, R=run, SS=side-shuffle, W=walk, V2=wave with two hands, JJ=jumping jacks, JP=jump in place, S=skip and V1=wave with one hand. Individual frame classification rates for most activities are above 80%; however, jumping, running and skipping show lower classification. There is room for improvement in the recognition of several actions, since bending and jumping confuse with each other due to the hunched-back posture of the subject. Several stances of profile-based actions such as walking, running and jumping overlap with one another. Similarly, shapes of frontally-faced actions such as side-shuffling, jumping jacks, waving with two hands and jumping in place overlap. Some frames of waving with one hand confuse with bending due to incorrect silhouette extraction that causes the elbow to connect with the head. Correct video classification is attained for all actions except skipping which is recognized at a rate of 70%. An overall frame recognition rate of 86% and video recognition rate of 96.8% is obtained.

Table 2 provides results of individual frame recognition and video recognition when the STSC is used during classification. We observe that the mismatches due to the 2D shape context between bending and jumping, and between walking and running (on account of similarities in shape) are reduced when STSC is used. Mismatches of several jumping and running frames with the skipping action are reduced as well.

Mismatches of jumping in place with two-handed waving and jumping jacks, and of side-shuffling with jumping in place are reduced considerably. Some misclassifications are introduced for jumping jacks frames with jumping in place, jumping in place frames with one-handed waving, and side-shuffling with walking, primarily due to the added similarity of motion vectors. The classification rate for the bending activity is much higher, and the classification rate for waving with one hand now attains 100% rate. All videos are classified correctly except skipping which is recognized at a rate of 80%. Overall recognition rates of 90% for frames and 97.9% for videos are obtained. Thus, we observe that there is an overall increase in classification over the original shape context.

**Table 2.** Results from individual frame recognition using the STSC

| | **B** | **JF** | **R** | **SS** | **w** | **V2** | **JJ** | **JP** | **S** | **V1** |
|---|---|---|---|---|---|---|---|---|---|---|
| **B** | **98.4** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1.6 |
| **JF** | 0 | **86.6** | 0.9 | 2 | 0.7 | 0 | 0 | 0 | 9.8 | 0 |
| **R** | 0 | 2.5 | **86.1** | 0.2 | 1.3 | 0 | 0 | 0 | 9.9 | 0 |
| **SS** | 0 | 1.9 | 0.5 | **93.3** | 4.1 | 0 | 0.2 | 0 | 0 | 0 |
| **W** | 0.3 | 0.6 | 0.4 | 2.1 | **95.6** | 0 | 0 | 0 | 1 | 0 |
| **V2** | 0 | 0 | 0 | 0 | 0 | **97.1** | 0 | 0 | 0 | 2.9 |
| **JJ** | 0.7 | 0 | 0 | 0 | 0 | 1.4 | **83.1** | 13.7 | 0 | 1.1 |
| **JP** | 0.2 | 0 | 0 | 0 | 0 | 1.1 | 0.2 | **93.8** | 0 | 4.7 |
| **S** | 0 | 20.4 | 17.4 | 1.0 | 3.4 | 0 | 0 | 0 | **57.8** | 0 |
| **V1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **100** |

**Table 3.** Individual frame and video recognition rates of the 2D and STSC compared against other works (- indicates that result was not provided ,* recognition was done on space-time cubes with 8 frames per cube as opposed to individual frames).

| | **Frame recognition rate** | **Video recognition rate** |
|---|---|---|
| **STSC (this paper)** | 90% | 97.9% |
| **2D Shape Context** | 86% | 96.8% |
| **Grunmann *et al* [13]** | - | 94.4% |
| **Gorelick *et al* [16]*** | - | 97.8% |
| **Jhuang *et al* (skip excluded, using C3 features) [17]** | - | 98.8% |
| **Niebles *et al* (skip excluded) [18]** | 55% | 72.8% |
| **Junejo *et al* [19]** | - | 95.3% |
| **Jia *et al* (using LSTDE) [21]** | 90.9% | - |

Table 3 shows the overall frame and video recognition rates for the work in this paper compared against other works done using the Weizmann dataset. It must be noted that for [16], recognition has been done on space-time cubes consisting of 8 frames per cube, and hence cannot necessarily be categorized as frame or video classification. However, our video recognition rates are similar to the space-time cube classification rates. Our frame-based activity recognition with the STSC is on par with that in [21], and is better than [13] and [19]. In [17] and [18], recognition has been done by excluding the skipping action from the dataset: we observe that our frame and video recognition rates even with skipping are higher than those from [18], and since we obtain 100% video classification on all actions except skipping, then if skipping were excluded from the library, we obtain 100% overall video recognition rate which is higher than the rate in [17].

## 4 Conclusions

In conclusion, this paper proposes a new 4D spatiotemporal shape context which is a descriptor for objects in video sequences that encodes information about the shape of the object and the change in its motion from one frame to the next. Using the spatiotemporal shape context for activity recognition demonstrates that it performs at least as well as other leading methods and is better than the shape context for most activities, since it allows the separation of similarly shaped stances using the differences in their motion over frames. The spatiotemporal shape context can be used to represent shape and motion in other areas of object or category recognition, such as content based video retrieval, expression or face recognition in video, and distinction between individual actions in multiple-action sequences. Future work includes analyzing the performance of the spatiotemporal shape context for activity recognition over short segments of video for the purpose of recognizing changes in activities using a sliding window.

## Acknowledgements

This research is supported in part by the Eastman Kodak Company and the Center for Electronic Imaging Systems (CEIS), a NYSTAR-designated Center for Advanced Technology in New York State.

## References

1. Masoud, O., Papanikolopoulos, N.: Recognizing Human Activities.In: IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 157-162, Miami, Florida (2003)
2. Niu, F., Mottaleb, M.: View-invariant Human Activity Recognition Using Shape and Motion Features. In: IEEE Sixth International Symposium on Multimedia Software Engineering, Miami, Florida (2004).
3. Davis, J., Bobick, A.: The Representation and Recognition of Human Movement Using Temporal Templates. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 928-934, San Juan, Puerto Rico (1997)

4. Guo, Y., Xu, G., Tsuji, S.: Understanding Human Motion Patterns. In: International Conference on Pattern Recognition, pp. 325-329, Jerusalem, Israel (1994)
5. Fujiyoshi, H., Lipton, A.: Real-Time Human Motion Analysis by Image Skeletonization. In: IEEE Workshop on Applications of Computer Vision, vol. 15, Princeton, New Jersey (1998)
6. Chen, D.Y., Liao, H.Y.M., Shih, S.W.: Continuous Human Action Segmentation and Recognition Using a Spatiotemporal Probabilistic Framework. In: Eighth IEEE International Symposium on Multimedia, pp. 275-282, San Diego, California (2006).
7. Belongie, S., Malik, J., Puzicha, J.: Shape context: A new descriptor for shape matching and object recognition. In: NIPS (2000)
8. Belongie, B., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Contexts. Technical Report, UCB//CSD00 – 1128, Berkeley (2001).
9. Mori, G., Malik, J.: Estimating Human Body Configurations Using Shape Context. In: European Conference on Computer Vision, Copenhagen, Denmark (2002)
10. Kholgade, N., Savakis, A.: Human activity recognition in video using two methods for matching shape contexts of silhouettes. In: SPIE Defense and Security Symposium, Orlando, Florida (2008)
11. Qiu, X., Wang, Z., Xia, S., Li, J.: Estimating Articulated Human Pose from Video Using Shape Context. In: IEEE International Symposium on Signal Processing and Information Technology, Athens, Greece (2005).
12. Kortgen, M., Park, G., Novotni, M., Klein, R.: 3D Shape Matching with 3D Shape Contexts. In: Seventh Central European Seminar on Computer Graphics, Budmerice, Slovakia (2003)
13. Grundmann, M., Meier, F., Essa, I.: 3D Shape Context and Distance Transform for Action Recognition. In: International Conference on Pattern Recognition, Tampa, Florida (2008)
14. Huang, K.S., Trivedi, M.: 3D Shape Context Based Gesture Analysis Integrated with Tracking using Omni Video Array. In: IEEE Workshop on Vision for Human-Computer Interaction in conjunction with IEEE CVPR Conference, San Diego (2005)
15. Kuhn, H.W.: The Hungarian Method for an assignment problem. Naval Research Logistics Quarterly, vol. 2, pp. 83-97 (1995)
16. Gorelick, L., Blank , M., Shechtman, E., Irani, M.: Actions as Space-Time Shapes. PAMI vol. 29, no. 12, pp. 2247-2253 (2007).
17. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A biologically inspired system for action recognition. In: IEEE International Conference on Computer Vision (2007)
18. Niebles, J.C., Fei-Fei, L.: A hierarchical model of shape and appearance for human action classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
19. Junejo, I., Dexter, E., Laptev, I., Perez, P.: “Cross-view action recognition from temporal self-similarities. In: European Conference on Computer Vision (2008)
20. 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, pp. 65-72 (2005)
21. Jia, K., Yeung, D.: Human Action Recognition using Local Spatio-Termporal Discriminant Embedding. In: IEEE Computer Society Conference on Pattern Recognition (2008)