3D Human Pose Search using Oriented Cylinders

Selen Pehlivan and Pınar Duygulu
Bilkent University,Department of Computer Engineering
06800, Ankara, Turkey
{pselen,duygulu}@cs.bilkent.edu.tr

Abstract

In this study, we present a representation based on a new 3D search technique for volumetric human poses which is then used to recognize actions in three dimensional video sequences. We generate a set of cylinder like 3D kernels in various sizes and orientations. These kernels are searched over 3D volumes to find high response regions. The distribution of these responses are then used to represent a 3D pose. We use the proposed representation for (i) pose retrieval using Nearest Neighbor (NN) based classification and Support Vector Machine (SVM) based classification methods, and for (ii) action recognition on a set of actions using Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) based classification methods. Evaluations on IXMAS dataset supports the effectiveness of such a robust pose representation.

1. Introduction

Nowadays, multi-camera systems are becoming affordable to be used widely. Being more reliable and robust, they are preferred against single camera systems in many applications, especially for surveillance. Human action recognition and monitoring in multi-camera systems has a special importance, but still remains as a challenge.

The natural and most common way to store human poses is to use 3D volumes obtained by re-construction from multiple views [13, 11, 7, 18]. The challenging part is to find a representation which is efficient and also robust to view point changes, or to different sizes and styles of human bodies in searching for poses or actions. While 3D shape representation is a well studied area [1, 9, 10, 4], human poses are more challenging than any rigid body object due to articulated structure of human bodies. The high number of degree of freedoms on the human body, causing many different potential configurations, requires search algorithms specific to articulated structures of human poses.

In this study, our objective is to introduce a robust representation for finding human body parts in any configuration to recognize 3D poses and further 3D action sequences. We consider body parts as a set of cylinders with various orientations and sizes, and represent poses as distribution of these cylinders.

Before describing the details of the proposed method, we briefly discuss previous work on multi-camera action recognition in Section 2. The pose representation is then introduced in Section 3. Next, in Section 4, we present the methods used for action recognition. After providing experimental results in Section 5, we conclude with the summary and discussions in Section 6.

2. Related Work

The study done by Mikić *et al*. [13] presents a framework to model human poses and track them using Bayesian Network. In the study, human poses are computed using 3D reconstruction and modeled as ellipsoids and cylinders using twist based approach.

Another group of studies propose appearance based methods in 3D. Some of them can be applied for estimating poses in addition to action recognition. Huang and Trivedi [7] present such an approach for gesture analysis over voxel data. They introduce 3D cylindrical shape context and model spatial-temporal information by HMM. In the study, they utilize from more samples in arbitrary views for training the system. Another work is that of Cohen and Li [5] allowing view-invariant pose identification over volumetric data. The reference points laying on a bounding cylinder called reference shape are used to encode a pose. An adaptation of the study is used in [14]. In this case, an optimized cylinder inscribed in volumetric human pose is used as the reference shape and Dynamic Time Warping is applied for recognizing action categories. In both studies, a single cylindrical shape is used as reference to reveal the distribution of sampled points over pose surface. Instead, we define a volumetric pose as a distribution of high-response regions to cylindrical kernels of different sizes in the form of voxel grid.

Two parallel studies based on view-invariant features from 3D data are done by Canton-Ferrer et al. [3] and Wein-

2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops
978-1-4244-4441-0/09/$25.00 ©2009 IEEE

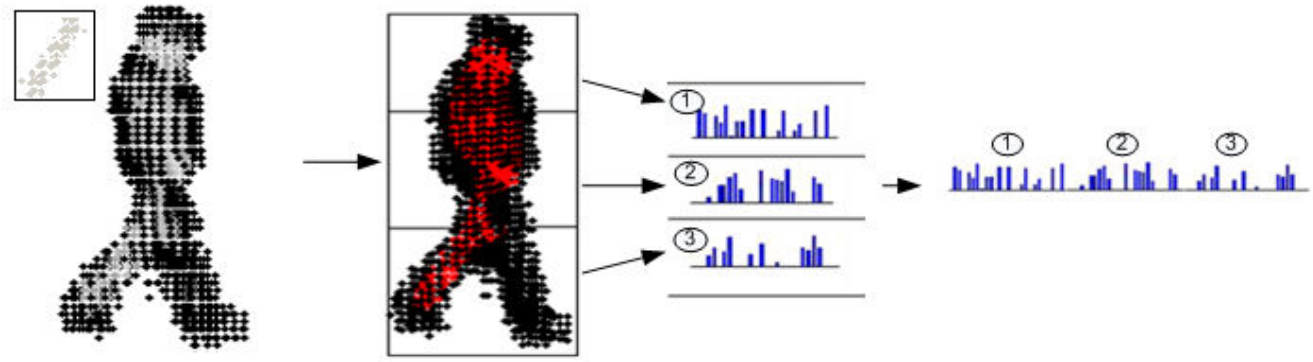

Figure 1. Representing 3D poses as distribution of cylinders: First, a set of 3D kernels in different orientations and sizes are searched over volumetric pose data. Then, regions having a score more than a threshold are selected. The bounding volume is computed and then divided into three parts through the vertical axis. From each sub-volume, high response regions are selected and histograms holding distribution of oriented cylinders are created. The combination of the three histograms constructs the feature vector to be used as a pose representation.

land et al. [17] extending motion templates [2] to three dimensions, called Motion History Volumes (MHV). MHV represents a whole action as a single volumetric data, that is also the temporal memory of the action.

Another strategy is to use volumetric key poses. Weinland et al.[6] present a probabilistic method based on exemplars using HMM. Exemplars are key poses extracted by reconstruction from action sequences. A similar work *Action Net* [12] is a graph based approach that models 3D poses as transitions of 2D views.

3. Representation of 3D Poses

A human pose is characterized by configuration of body limbs having cylinder like shapes in different sizes and orientations. This notion plays an important role while structuring our representation. We model poses as a set of cylinders with various sizes and orientations. Motivated by the success of representing 2D human poses as distribution of rectangles over complicated models [8], we use the distribution of these cylinders as a compact representation for 3D human poses.

In order to find cylinder-like structures over body, we generate a set of 3D kernels and measure the correlation of these kernels with human poses represented in the form of volumetric data. High response regions are likely to correspond to body parts at specific orientations and sizes. Rather than using complex models, we count the frequencies of these responses for different orientations and sizes. A simple localization is provided by partitioning the bounding volume into sub-volumes and then combining the corresponding histograms. Figure 1 illustrates the overall process for representing poses. In the following, the details of the proposed method are described.

3.1. Forming and Applying 3D Kernels

We assume that 3D poses are provided in the form of voxel grid. In order to locate cylinders, we construct 3D kernels in a similar format to search over volumetric data appropriately. To model a cylinder of specific size and orientation, a new kernel is formed as a grid data consisting of voxels that are located inside this cylinder. This results in a cylinder-like voxel grid.

A set of kernels K is constructed from a specified size of cylinder by rotating it around its local axis α° apart (see Figure 2 for an example). While rotating, the symmetry of kernel reduces almost one half of the search space in 3D. Therefore, the number of kernels in K with α° apart can be computed as follows:

$$|K| = 1 + (\frac{90}{\alpha} - 1)(\frac{360}{\alpha}) + (\frac{180}{\alpha}) \qquad (1)$$

where, the first and the third components are the number of different kernels in vertical and horizontal positions of the cylinder respectively. The second component is rotation by $\frac{360}{\alpha}$ times around the vertical axis for each α° apart from the vertical axis.

Limbs in the human body are in various sizes. Although longer cylinders may be more discriminative compared to shorter cylinders to locate a limb, due to noise factors we may loose some limbs if the cylinder is too long. For this reason, we construct a set of oriented kernels for various sizes defined by the length of the cylinders to detect limbs in various lengths. Although by changing the radius of the cylinders limbs in different thicknesses can be detected, empirically we have observed that a thick limb can be represented as a collection of thin cylinders, and therefore we only choose a single radius value for all cylinder types.

After generating a set of kernels, we convolve each one

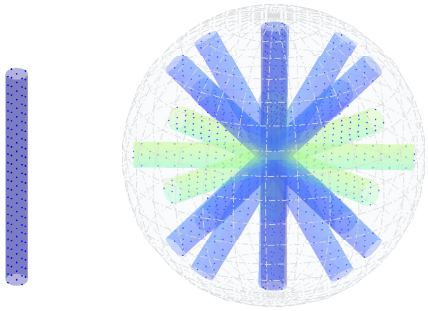

Figure 2. Left is the kernel in the form of voxel grid and right is the set of kernels with various orientations around the local axis 45° apart. The number of kernels for this size is 13.

of the kernels with volumetric pose data and obtain responses for each voxel. High response voxels correspond to body parts having same orientation with the applied kernel. To detect voxels with high response, the responses are scaled in the range of $[0, 1]$ and then, regions having a score more than a threshold are selected. In the experiments, we choose 0.8 as the threshold. This threshold may change depending on whether the search is performed over dense or sparse data.

3.2. Distribution of Cylinders

We define a pose descriptor as distribution of oriented cylinders. First, we form a set of kernels with different sizes and orientations as mentioned previously. Then, we present a representation storing the frequency of high response regions to each kernel.

Although distribution of cylinders reveals crucial information, localization is needed for a better representation. For this purpose, we fit a bounding volume to each pose and divide the bounding volume into N equal sized sub-volumes through the vertical axis. This process corresponds to dividing the height of an actor's pose into partitions. Empirically, we have found that $N = 3$ gives the best performance. The computed histograms of each sub-volume is then combined to obtain a single feature vector for a given pose.

4. Recognition of Actions

In this section, we present methods for recognizing actions using proposed pose representation. Considering actions as sequence of poses, we evaluate two methods for recognition: a) Dynamic Time Warping(DTW) and b) Hidden Markov Model(HMM).

4.1. Dynamic Time Warping

In our problem, actions are sequence of poses where each pose is modeled as a distribution of cylinders using proposed representation. The concatenation of 1-D feature vectors per pose results in a 2-D representation per action. Action sequences performed by different actors vary in time and speed. The most common way to handle similarities among time series is to use Dynamic Time Warping(DTW) [15]. However, DTW for 2-D series is an NP-complete problem.

Instead, we make use of the approach of Ikizler *et al*. [8] to find similarities between 2-D representation of action sequences. In this approach, DTW is applied to compare 1-D series located at the same bin location of two different action representations. This is to measure the fluctuations through time at the same bin location corresponding to number of high response cylinders of a specific orientation and size. We sum the results of DTW for all bin comparisons and find an optimum match.

Throughout an action, some body regions in the form of voxel grid do not change. Therefore, data series at some bins of the histogram rarely change. We measure the variance at each bin location to measure the amount of change for a cylinder with specific size and orientation through time. Then, we set the 1-D series at that bin location to a zero vector, if variance is below a threshold.

4.2. Hidden Markov Model

Second method that we apply for action recognition is Hidden Markov Models(HMM). In our problem, we have a set of actions consisting of consecutive frames. A frame at time t only depends on a previous frame at time $t - 1$. Thus, HMM is a good technique to model our problem. In our case, we model the problem using discrete HMMs [16].

In this approach, we first quantize all the poses in training actions using k-means clustering algorithm into a set of codewords which we refer to as pose-words. Then, we construct HMM models per action class using 3 states. An unknown action sequence is assigned to a class giving the highest likelihood for its HMM model. Similar to DTW, the variance of each bin is computed to find uniform series. These are set to zero vector.

5. Experimental Results

5.1. Dataset

We test our pose descriptor on publicly available INRIA Xmas Motion Acquisition Sequence (IXMAS) dataset [17]. We choose 5 actions performed by 12 different actors 3 times in different orientations. These actions are walk, wave, punch, kick and pick up.

Multi-view action videos are recorded by 5 cameras and

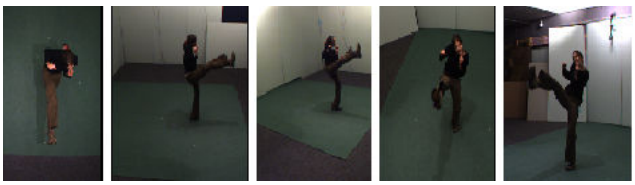

Figure 3. Example views from IXMAS dataset. Multi-camera system has 5 cameras.

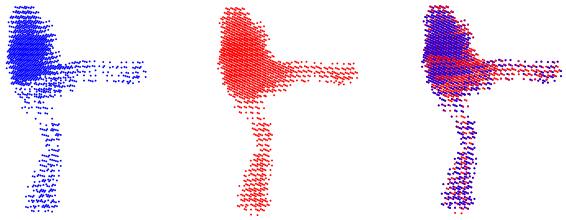

Figure 4. From left to right: the actual kick pose, enhanced kick pose, the difference between them. We fill the missing voxels using close morphology operator with a 3D sphere structuring element. In the third figure, we represent the actual pose and the enhanced version as overlapped to clearly represent the added voxels after closing operation.

videos are used to construct volumes. Volumes from multiple views are extracted using shape from silhouette technique. Results are taken using volumes in the form of [64x64x64] voxel grid.

5.2. Data Enhancement

Volumetric poses have defects that significantly reduce the recognition rate. Before extracting pose histograms, some techniques are used to enhance volumetric data to eliminate reconstruction defects. In our experiments, we perform morphological closing on volumetric data using sphere structural element with radius 2. The closing operator can close up internal holes corresponding to missing voxels. Data enhancement process is shown in Figure 4.

5.3. Pose Representation

We define a set of kernels with different sizes and orientations. During our experiments, we form the set of kernels with sizes [1x5],[1x10] and [1x20] where $[nxm]$ means a cylinder with radius n and length m. Then, we rotate each kernel to obtain a set of oriented cylinders. These kernels are constructed with 30° apart. The number of kernels per size is 31.

After forming kernels, each one of them is searched over pose volume resulting in scores as kernel responses (see Figure 5) . Then, we divide voxel grid into 3 sub-volumes. The number of high response voxels to these kernels at each sub-volume are stored in a histogram of oriented cylinders. After forming histograms, we concatenate them to be used for pose inference. Please note that, each histogram is normalized prior to concatenation into a single pose vector.

5.4. Pose Retrieval

There are many configurations of body parts that reveal different body poses. Among various kinds of poses, some are more discriminative known as key poses. In the following, we measure the performance of the proposed pose representation to classify key poses. We use two methods to evaluate the descriptor performance.

The simplest method that we use is the nearest neighbor (NN) based classification. We use the Euclidean distance to find the best matched pose. The stand alone and complete performances of various sized kernels are shown in Table 1.

Poses	Accuracy all	Accuracy [1x5]	Accuracy [1x10]	Accuracy [1x20]
walk	96.97	93.94	96.97	96.97
wave	90.91	90.91	93.94	81.82
punch	63.64	48.48	69.70	72.73
kick	87.88	84.85	84.85	75.76
pick	96.97	90.91	96.97	93.94

Table 1. NN-based Classification results: 3 kernel sets with sizes [1x5], [1x10] and [1x20] are constructed with 30° apart. [1x5] means that a voxel grid inside a cylinder with radius 1 and length 5. The fist column lists the names of the selected key poses. The third column gives the performances when all kernels are used. The other columns gives the stand alone performances of each kernel with different sizes.

Nearest neighbor based classifications requires a search over the whole dataset. Another method is multi-class Support Vector Machine (SVM) classification. SVM classifiers are formed using RBF kernel and trained using 3 actors. We evaluate the performances for 5 action classes. The results are presented in Table 2.

Poses	Accuracy all	Accuracy [1x5]	Accuracy [1x10]	Accuracy [1x20]
walk	100	87.50	91.67	95.83
wave	91.67	54.17	95.83	79.17
punch	66.67	66.67	66.67	45.83
kick	100	100	100	95.83
pick	95.83	95.83	95.83	95.83

Table 2. SVM-based Classification results: 3 kernel sets with sizes [1x5], [1x10] and [1x20] are constructed with 30° apart.

The results show that NN-based method is better than SVM based method. While taking more computation time, since all the examples in the data set are searched it is more likely to find a closer example with the NN-based method.

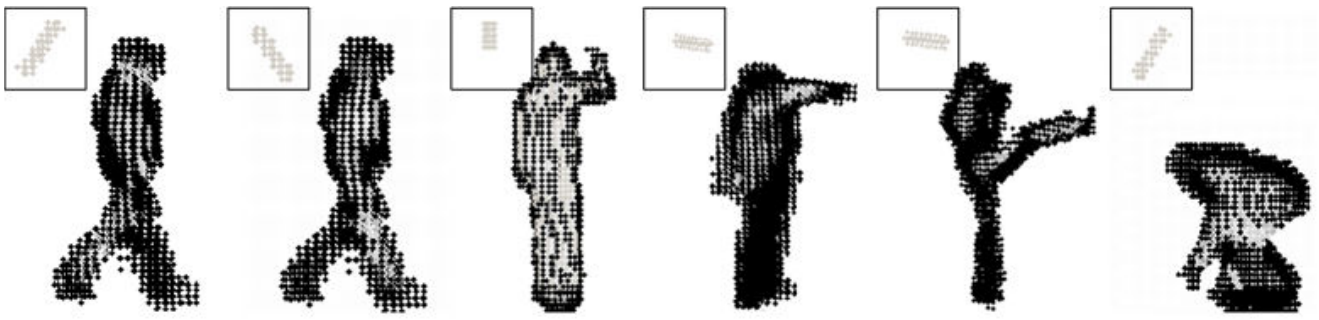

Figure 5. Kernel search results for five classes of 3D key poses: From left to right, poses are walk, wave, punch, kick and pick up. Figure represents poses from an arbitrary view. Gray level voxels are the high response regions for the corresponding kernel that is also drawn on the top left corner of each pose. Lower response regions below 0.8 are not counted. For walking pose, we represent search responses for two kernels in order to show different responses on the same pose. Through the experiments, we construct kernels in various sizes. For example, as shown in the third example a smaller size kernel is successful in finding upper arm of a wave pose. Note that, while using various length kernels, the radius size is fixed to 1 giving 3 voxels width.

It is likely that by increasing the number of training examples SVM-based method will be comparable to NN-based method, but still with a very few examples the results are acceptable.

5.5. Action Recognition

The proposed pose descriptor is also evaluated for action recognition. We select the same set of action classes and test the same kernel configurations used for pose retrieval. We evaluate two methods to classify actions. First, we perform action matching by DTW. DTW gives highest performance for the combination of [1x5] and [1x10] kernels. The confusion matrix can be found in Figure 6.

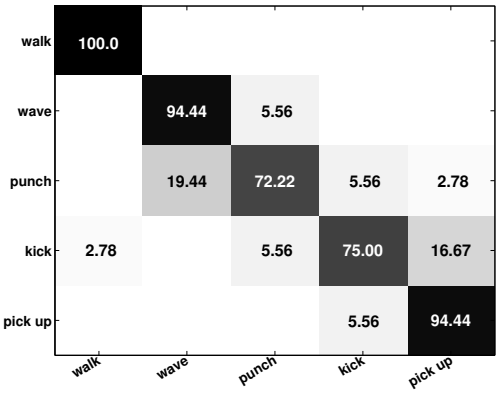

Figure 6. DTW-based Classification results: experiments are done over 5 classes that are performed by 12 actors in 3 different viewpoint. This is the confusion matrix when [1x5] and [1x10] kernels are used.

The second method used for action recognition is HMM. Actions performed by 3 actors in 3 different orientations are used for training and the remaining dataset is used for testing. We quantize training actions using k-means clustering algorithm into 80 pose-words and construct HMM models per action class using 3 states. The recognition performances over 5 classes are shown in Table 3.

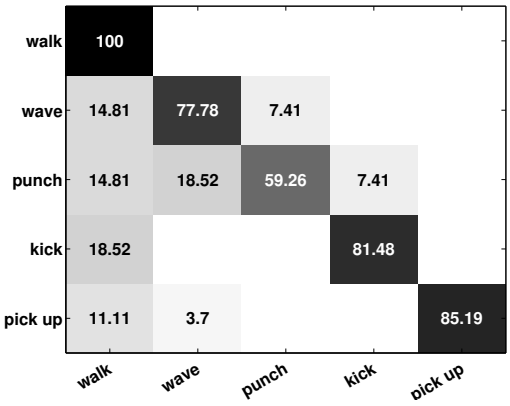

Figure 7. HMM-based Classification results:experiments are done over 5 classes that are performed by 12 actors in 3 different viewpoint. 3 actors are used to train HMM models and remaining used for testing. This is the recognition performance when all kernel types are used.

In our experiments, DTW gives highest recognition rate for 5 classes of actions using kernels with sizes [1x5] and [1x10]. HMM gives the highest accuracy using all kernels. On the other hand, DTW-based action classification requires one-to-one comparison in order to find most similar action. However, HMM only requires a trained model with a set of action samples. Therefore, it has a lower running time.

Poses	DTW all	DTW [1x5][1x10]	HMM all	HMM [1x5][1x10]
walk	97.22	100	100	100
wave	94.44	94.44	77.78	62.96
punch	69.44	72.22	59.26	70.37
kick	66.67	75	81.48	62.96
pick	91.67	94.44	85.19	74.07

Table 3. HMM and DTW Classification results respectively: experiments are done over 5 classes that are performed by 12 actors in 3 different viewpoint. The first one is the recognition rates when all kernel types are used. The second one is the recognition rates when [1x5] and [1x10] kernels are used.

6. Summary and Discussion

In this study, we propose a new pose representation using distribution of oriented cylinders. The representation is for 3D poses re-constructed from multiple camera views. Human body consists of cylinder like body parts. Similarly, we model 3D poses as a set of cylinders. Kernels are constructed in various sizes and orientations to find limbs in any configuration. The distributions of high responses are used as our representation.

The proposed descriptor is based on searching cylinders to find body limbs that can change their orientations in different body configurations. Therefore, it is suitable to be used for highly salient actions with observable changes in body configuration. During experiments, we select 5 classes that have discriminative key poses. On the other hand, the data used for experiments is sparse and has defects. It results in lower responses for some body limbs during search. The recognition results will be better over denser and higher resolution data.

Another important point is about scaling volumetric poses. Actions can be performed by different actors. Therefore, volumetric poses vary in terms of size. In this study, we do not scale the volumetric poses and use them as they are. We observe that cylinders with small radius sizes can form larger cylinders. We argue that they include the high response regions returned by larger cylinders. So, the distribution of a small size cylinder will be higher for all kernel types over a bigger pose. So histogram normalization solves this problem without needing a volume scaling.

The same is true for cylinder lengths. However using a shorter cylinder can cause the lost of true poses. We observe that, the cylinders with different lengths return more distinguishing results than different radius sizes. As a result we preserve the radius of cylinders as small as possible and form kernels with various lengths.

An action can be performed by an actor in any orientation. This shifts the histogram bins holding responses of kernels constructed through the vertical axis rotation. In this study, we do not apply any strategy to provide pose alignment. We still obtain high pose retrieval results both by NN-based classification and SVM-based classification. This shows that the pose samples with different orientations are enough to handle viewpoint variations. In the future, PCA-based alignment can also be applied to volumetric data prior to histogram extraction. Moreover, some of body pixels such as torso pixels can be excluded from the computation of the distribution as they result in high responses for more than a cylinder kernel.

References

[1] M. Ankerst, G. Kastenmueller, H. P. Kriegel, and T. Seidl. 3d shape histograms for similarity search and classification in spatial databases. *LNCS*, pages 207–228, 1999. 1

[2] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001. 2

[3] C. Canton-Ferrer, J. R. Casas, and M. Pardas. Human model and motion based 3d action recognition in multiple view scenarios. In *In European Signal Processing Conference*, 2006. 1

[4] D. Y. Chen, X. P. Tian, Y. T. Shen, and M. Ouhyoung. On visual similarity based 3d model retrieval. In *Computer Graphics Forum*, volume 22, pages 223–232, 2003. 1

[5] I. Cohen and H. Li. Inference of human postures by classification of 3d human body shape. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003. 1

[6] D. D. Weinland, E. E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, 2007. 2

[7] K. S. Huang and M. M. Trivedi. 3d shape context based gesture analysis integrated with tracking using omni video array. In *IEEE Workshop on Vision for Human-Computer Interaction (V4HCI), in conjunction with CVPR*, 2005. 1

[8] N. Ikizler and P. Duygulu. Histogram of oriented rectangles: A new pose descriptor for human action recognition. *Image and Vision Computing*, 2009. 2, 3

[9] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *PAMI*, pages 433–449, 1999. 1

[10] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Proceedings of the 2003 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 156–164, 2003. 1

[11] R. Kehl and L. Gool. Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding*, 104:190–209, 2006. 1

[12] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 2

[13] I. Mikić, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision*, 53(3):199–223, 2003. 1

[14] M. Pierobon, M. Marcon, A. Sarti, and S. Tubaro. Clustering of human actions using invariant body shape descriptor and dynamic time warping. In *IEEE Conference on Advanced Video and Signal Based Surveillance*, 2005. 1

[15] L. Rabiner and B. H. Juang. *Fundamentals of speech recognition*. 1993. 3

[16] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 3

[17] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2-3):249–257, 2006. 2, 3

[18] P. Yan, S. M. Khan, and M. Shah. Learning 4d action feature models for arbitrary view action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1