Action Recognition in Intelligent Environments using Point Cloud Features Extracted from Silhouette Sequences

Radu Bogdan Rusu, Jan Bandouch, Zoltan Csaba Marton, Nico Blodow, Michael Beetz
Intelligent Autonomous Systems, Technische Universität München
{rusu, bandouch, marton, blodow, beetz}@cs.tum.edu

***Abstract*— In this paper we present our work on human action recognition in intelligent environments. We classify actions by looking at a time-sequence of silhouettes extracted from various camera images. By treating time as the third spatial dimension we generate so-called space-time shapes that contain rich information about the actions. We propose a novel approach for recognizing actions, by representing the shapes as 3D point clouds and estimating feature histograms for them. Preliminary results show that our method robustly derives different classes of actions, even in the presence of large variability in the data, coming from different persons at different time intervals.**

I. Introduction

As robots and humans are to share the same workspaces and cooperate with each others, the robots must be capable of recognizing and interpreting the actions and movements of the humans. This requires in many cases the recognition of full body motions and actions. Robots must hand over objects to humans and receive objects from them, and they must approach people differently depending on what they are currently doing. Recognizing an action is not only necessary to understand the current behavior of a human but it is also a pointer to its future intentions (e.g. recognize the situations where he might need help).

In this paper, we propose to apply methods originally developed for the acquisition of 3D environment models from laser scans to improve the performance of vision based action recognition methods. A number of approaches to action recognition propose the interpretation of space-time shapes as depicted in Figure 1. The problem of action recognition then becomes the problem of classifying a given space-time shape with respect to a set of a priori learned classes of actions.

By treating the space-time shapes representing actions as 3D objects, we require techniques for solving the problem in a different domain: that of 3D object recognition. Thus, we propose to apply methods for robustly identifying sets of informative features, which can solve the recognition problem.

The main contribution of this paper is the application of our robust feature histograms[1] to the problem of action recognition. We assume that the feature histogram representation of space-time shapes of the same actions are sufficiently similar and those of different actions sufficiently different that we can classify the actions reliably based on them. We describe how this method can be implemented and applied to the visual data and empirically evaluate the recognition results.

In particular, we propose a system for action recognition from cameras placed in the environment. Given only silhouette data that is relatively easy to obtain from static camera images, we detect actions by comparing features extracted from a time-sequence of silhouettes to exemplars stored in a training database. Treating the silhouette-sequences as 3D point cloud data, we propose the use of robust point-based geometry techniques which have proven to be reliable for solving numerous problems[1]. We show that such features can be successfully applied also to the action recognition domain.

The remainder of this paper is organized as follows. The next section gives a brief overview on related work followed by a description of the acquisition of silhouette data in section III. The creation of the 3D space-time shapes (*action shapes*) is presented in section IV. Section V describes our implementation for computing robust feature histograms for the action shapes. We discuss experimental results in section VI and conclude in section VII.

II. Related Work

Several approaches for human action recognition have been proposed. Some of them aim at extracting human pose data such as articulated joint angles [2], motion trajectories or motion descriptors based on optical flow [3] for each frame. Action classification is then done by inferring the most likely action given the pose data for a sequence of frames. These approaches require highly sophisticated methods to succeed in the feature extraction step, and are prone to noisy input data or cluttered environments.

Other approaches use silhouettes respectively contours as features. These are usually easy to obtain by means of background subtraction when using static cameras (e.g. [4]). While some approaches discard temporal information by trying to identify individual keyframes using the eigenshapes of the silhouettes for example, we believe that (temporal) motion information is important for successfull action recognition. Bobick et al. [5] have introduced *temporal templates* for action recognition, where extracted silhouettes are overlayed on a single image with fading gray values according to the distance in time. Such 2.5D approaches work well for some actions, but are unable to capture all relevant details of the motions as more recent silhouettes partly overwrite older motion information. Another approach

is to consider the temporal domain as a third spatial dimension and to create 3D shapes of actions to make them better distinguishable (Figure 1). This has been proposed independently by Gorelick et al. [6] as *space-time shapes* and by Yilmaz and Shah [7] as *spatiotemporal volumes*. We have adopted this approach in our work. It should be noticed that there is a view dependency when using silhouettes from a single camera, which is not a problem in our scenario. Weinland et al. [8] proposed *motion history volumes* as view independent 4D approach to action recognition, but they require multiple calibrated cameras (at least 5) to capture 3D shapes of humans for each timestep using space carving techniques.

Given a 3D shape representation of an action, robust features need to be extracted for good classification results. The 3D object recognition community has developed different methods for computing multi-value features which describe complete models for classification: curvature based histograms [9], spin image signatures [10], or surflet-pair-relation histograms [11]. All of them are based on the local estimation of surface normals and curvatures and describe the relationships between them by binning similar values into a global histogram. A high number of histograms per object is required by [9], but the method can cope with up to 20% occlusions. The 4D geometrical features used in [11] and the spin image signatures in [10] need a single histogram and achieve recognition rates over 90% with synthetic and CAD model datasets, and over 80% with added uniformly distributed noise levels below 1% [11]. All of the above show promissing results, but since they have only been tested against synthetic range images, it's still unclear how they will perform when used on noisier real-world datasets.

We extend the work presented in [11] by computing robust feature histograms for a sequence of 3D silhouettes. Our implementation reduces the theoretical computational complexity of the algorithm by a factor of 2, and is shown to be robust in the presence of noisy data.

III. Action Shapes from Silhouettes

Two-dimensional silhouette data provides a rich source of information regarding the pose of a human. People are capable to correctly guess a human pose simply by looking at the silhouette shapes. However, actions can only be recognized by looking at motion information, as the temporal alignment of poses becomes important. We therefore classify actions by comparing their visible space-time shapes [6] with learned training data.

Space-time shapes are generated from a sequence of 2D silhouette contours by shifting silhouettes in the third spatial dimension according to their position on the timeline. Figure 1 (center) shows a space-time shape where the shape surface was generated using a *marching cubes* algorithm for visualization purposes. For obtaining a good classification accuracy as well as for scale invariance, the sizes of the silhouettes are normalized with the maximum distance between points along the vertical direction (i.e. the shape's height). Before stacking the silhouettes in the third dimension, they are aligned by their centers of gravity. We also remove frames with no significant change in the silhouette appearance from the shape, so that we achieve an invariance regarding to the speed at which an action is carried out, as show in section IV.

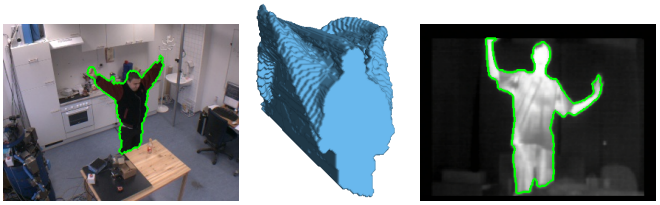

Fig. 1. Silhouette extracted from a camera image using background subtraction (left); Space-time shape generated from a sequence of silhouettes (center); Silhouette extracted from a thermal camera using thresholding (right).

One advantage of space-time shapes is that they are simple and fast to generate. Extraction of silhouettes in the case of static cameras can be done using background subtraction techniques. There, a model of the background is learned in advance for each pixel, so that for each image a binary per-pixel classification is performed to separate the foreground from the background. We use the method proposed by Kim et al. [4], which is basically a fast approximation to a *mixture of gaussians* model of the background. Robust methods that are capable to extract silhouettes from moving cameras have also been proposed recently [12], and could easily be incorporated. After segmenting the foreground, noise can be reduced by applying morphological opening and closing operations. The silhouette contours are extracted after selecting the biggest connected component and filling all holes in the component.

In our setup we use up to five static cameras plus a thermal infrared camera that can be used for silhouette extraction. Silhouette extraction from thermal cameras is performed by applying a simple thresholding operation (Figure 1 right side). Data from each camera is treated separately, but the extracted features from the space-time shapes can be combined for classification to achieve better view invariance.

In the next chapter we will discuss how to delete unimportant frames in order to neglect the effect of different speeds in the execution of a movement.

IV. Action Shape Minimization

We assume that the silhouette of a human person can be successfully segmented during a motion as shown in Figure 2, and converted to a 2D $\langle x, y \rangle$ point cloud. Our goal is to acquire a sequence of silhouettes for every action, build a 3D point cloud with z as the time axis, and then exploit the data in that point cloud using 3D point-based geometry methods.

Even though the movement from one frame to the other is constrained, the resulting silhouettes do not always overlap by aligning them using their gravity centers, because the body position changes. This results in minor displacements of the 2D point clouds from each other.

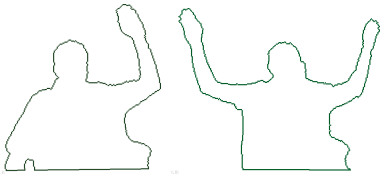

Fig. 2. Segmented silhouettes of human motions

We solve the displacement problem by registering each frame to the next one in the sequence using an ICP-based algorithm[1]. Figure 3 shows the resulting 3D point clouds created by stacking the registered silhouettes representing different actions, time-wise, on top of each other.

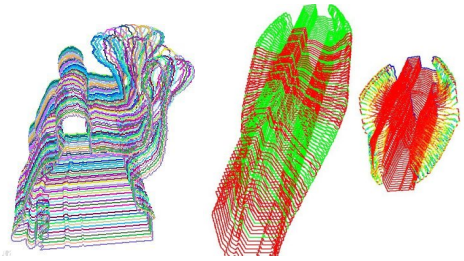

Fig. 3. Left and center: 3D point clouds representing actions of human motions with z as time, right: resulted point cloud after pruning the *duplicate* frames from the dataset in the center of the image.

Because of the stacking process however, double frames can appear, mostly because the subject did not move between two consecutive frames. These double frames are irrelevant with respect to our problem and might create an additional computational burden. Therefore, we proceed to remove them as follows, for each two consecutive frames:

1) for each point p_j, with $j = \overline{1,N}$ in frame f_i we search for the closest corresponding point q_k, with $k = \overline{1,M}$ in frame f_{i+1} and compute an Euclidean distance metric between them;
2) we compute the Frobenious norm in distance space of f_i and f_{i+1} as: $\|f_{i+1} - f_i\|_F = \sqrt{\sum_1^N |q^p i - p_i|^2}$
3) we select f_{i+1} as being different (i.e. *unique*) than f_i if $\|f_{i+1} - f_i\|_F \leq d_{thresh}$

For finding the closest correspondence of a given point p, a fast k-d tree structure in two dimensions was employed. Selecting $d_{thresh} = 0.2$ empirically in our experiments gave good results. The results after pruning the duplicate frames are shown in Figure 3, where relevant frames from the middle sequence are colored in green (and shown again as the resulted sequence in the rightmost part of the image), while the frames which will be pruned in red. The computational time decrease obtained by removing irrelevant frames is highly dependant on the input sequence and can vary greatly. In our experiments, we achieved reduction rates from 24.15% up to 91.47%.

The informative feature extraction from the generated action shapes is the topic of the next chapter.

V. Feature Histograms

In order to efficiently obtain informative features, we propose the computation and usage of a histogram of values[1] which encodes the neighborhood's geometrical properties much better, and provides an overall scale and pose invariant multi-value feature. The feature histogram has already given good values for the problem of 3D objects classification obtained from laser data in indoor environments[1].

The input data consists of 3D $\langle x, y, z \rangle$ point coordinates. For a given radius r, the algorithm will first estimate the surface normals (see Figure 4) at each point p by performing Principal Component Analysis (PCA) on the data points contained in the sphere with radius r and p as its center. The eigenvector corresponding to the smallest eigenvalue approximates the surface normal at point p. Once the normals are obtained and consistently re-oriented[1], the histogram for the 3D action shape will be computed using the four geometric features as proposed in [11].

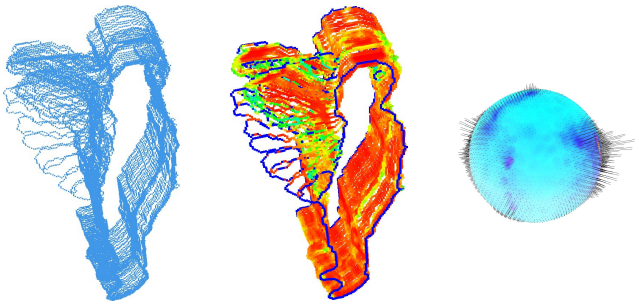

Fig. 4. Action shape (left), estimated surface curvatures (center), and distribution of estimated surface normals on the Extended Gaussian Image.

For every pair of points p_i and p_j ($i \neq j$, $j < i$) in the shape and their estimated normals n_i and n_j, we select a source p_s and target p_t point, the source being the one having the smaller angle between the associated normal and the line connecting the points:

$$\begin{gathered} \text{if } \langle n_i, p_j - p_i \rangle \leq \langle n_j, p_i - p_j \rangle \\ \text{then } p_s = p_i, p_t = p_j \\ \text{else } p_s = p_j, p_t = p_i \end{gathered}$$

and then define the Darboux frame with the origin in the source point as (see Figure5):

$$u = n_s, \; v = (p_t - p_s) \times u, \; w = u \times v.$$

We then proceed by computing four feature values for each pair of points, and categorize the resulted values into a global histogram, where each bin of the histogram at index idx contains the percentage of the source points in the shape which have their features in the interval defined by idx:

$$\left.\begin{aligned} f_1 &= v \cdot n_t \\ f_2 &= ||p_t - p_s|| \\ f_3 &= u \cdot (p_t - p_s)/f_2 \\ f_4 &= atan(w \cdot n_t, u \cdot n_t) \end{aligned}\right\} \Rightarrow idx = \sum_{i=1}^{i \leq 4} cat(f_i) \cdot div^{i-1}$$

[1]see [13] for a general algorithm for consistent normal orientation propagation

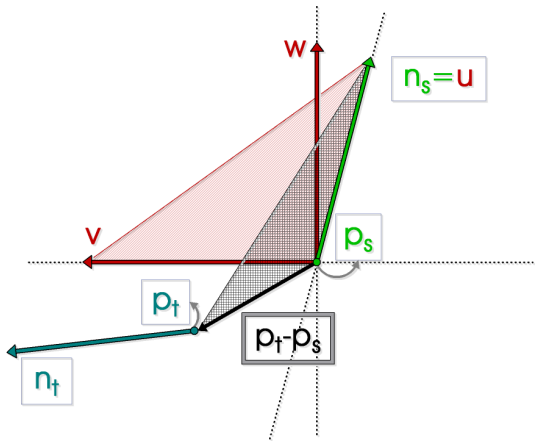

Fig. 5. The computed Darboux frame (vectors u, v and w) placed at the source point.

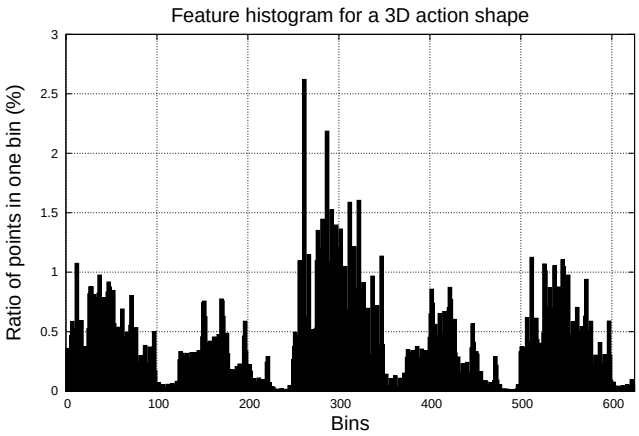

Fig. 6. Example of a resulted histogram for a 3D action shape.

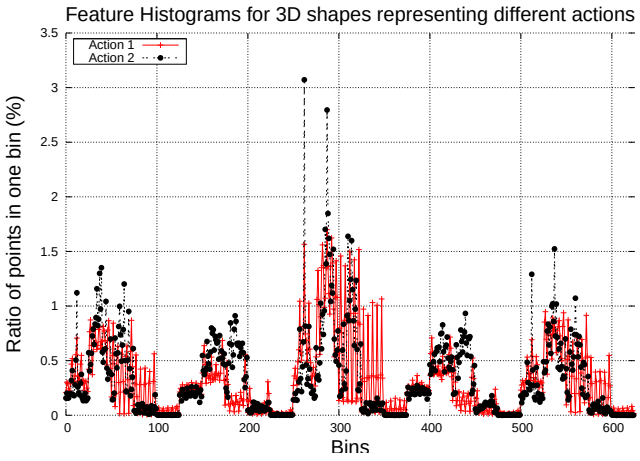

Fig. 7. Two different action shapes and their estimated feature histograms.

where div is the number of subdivisions of the features' value range and $cat(f)$ returns the number of the category in which the feature f falls. This number is defined as the smallest number for which:

$$cat(f_i) \cdot \frac{max(f_i) - min(f_i)}{div} \leq f_i$$

and thus, a number in the $[0, div]$ interval. The number of histogram bins that are formed using these four geometric features is div^4.

To find out the minimum and maximum values of each feature, we have to consider that they are a measure of the angles between the points' normals and the distance vector between them. Because f_1 and f_3 are dot products between normalized vectors, they are in fact the cosine of the angles between the 3D vectors, thus their value is between ± 1, and 0 if they are perpendicular. Similarly, f_4 is the arctangent of the angle that n_t forms with w if projected on the plane defined by $u = n_t$ and w, so its value is between $\pm \pi/2$, and 0 if they are parallel. f_2 is the length of the segment between the two points, which is always positive and for which a maximal value can be set as the diameter of the shape's bounding box.

Because the number of bins is increasing exponentially with the number div of feature categories, we have to select a high enough number for capturing detail, but low enough to reduce the algorithm's computational complexity. Our experiments confirmed the suggestion in [11] to set $div = 5$, as classification of action shapes was performed successfully using the resulted 625 bin histograms.

Figure 6 presents the resulted feature histogram for a given 3D action shape.

The effects of different movements in the feature histogram space for two action shapes is shown in Figure 7.

VI. Discussions and Experimental Results

To evaluate the performance and robustness of our proposed approach, we have performed several experiments in a distributed-sensing kitchen environment[14]. To allow more variation in the data, our subjects were asked to perform several actions in the kitchen, without giving them explicit instructions on how to do them. The list of actions included: (i) opening and closing a cupboard; (ii) opening and closing a drawer; (iii) opening and closing a vertical-door top cabinet; (iv) picking up one object from a table and moving it to another; (v) opening and closing the oven door; and (vi) unscrewing a bottle and drinking from it. Each action had to be performed several times by a subject, while all the other human personnel was asked to step outside the kitchen and they were not allowed to observe the actions of their colleague. After recording several experiments and looking at the data, we noticed a high degree of variability between the way the movements were performed, from subject to subject, but also between the same actions of a subject.

Figure 8 depicts the processed histograms for one action shape representing the opening of a cupboard. Three persons were asked to repeat the same action 3 times, at different time intervals, and a fourth person once. Notice that the histograms are matched almost perfectly, even though the subjects participating in the experiment performed the action differently.

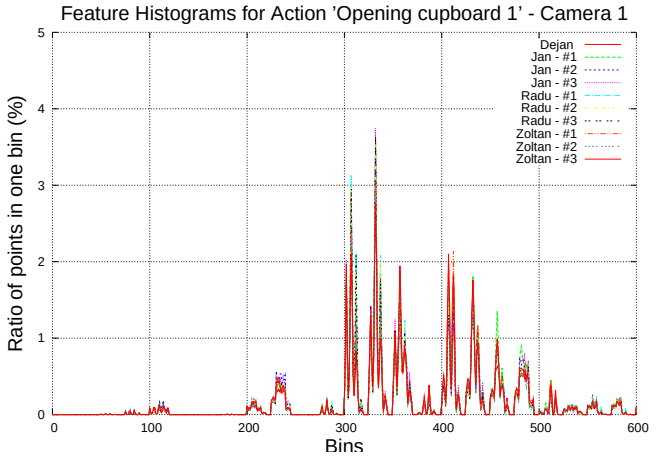

Fig. 8. Feature histograms for the action shape representing the opening of a cupboard. The experiment involved 3 subjects, out of which 3 were asked to repeat the action 3 times.

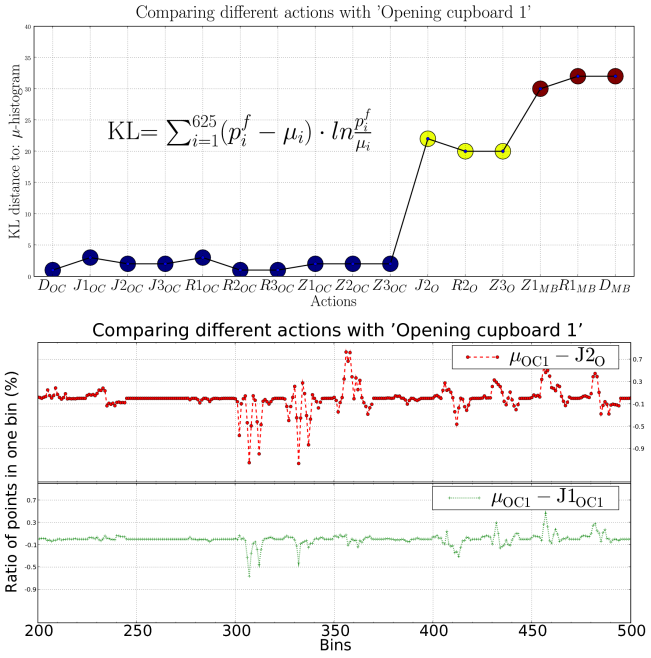

Fig. 9. Comparing different actions using the Kullback-Leibler divergence (top). Differences between the mean μ-histogram of an action and a) the histogram of another action (middle); b) a histogram of the same action performed by another subject (bottom).

To evaluate the histograms of different actions, we computed several distance metrics and compared the resulted values. As indicated by [9], the metrics which gave the best (and similar) results were the Chi-Square (χ^2) divergence and the Kullback-Leibler (KL) divergence. The curve presented in the upper part of Figure 9 shows the values obtained by computing the KL divergence between a set of histograms representing different action-shapes and the mean μ-histogram for one action (opening cupboard). Each of the values shown on the x-axis of the plot is encoded as Nt_{act}, where N is the first initial of the person performing the action, t represents the trial number (i.e. the number of the experiment), and act is an acronym describing the action name. The 3 evaluated actions are: a) OC for opening the cupboard; b) O for opening the oven door; and c) MB for moving a bottle from a table to the another table. The first 10 values on the x-axis are the KL distance values for the histograms in Figure 8, from which the mean μ-histogram was computed. The following three values are obtained by computing the divergence between three histograms representing the action of opening the oven, and the last three values for the action of moving the bottle. Note how the distances appear to be in the same *clusters*, in the sense that their values are very close to each other, thus demonstrating the robustness of our method to categorize actions efficiently.

The second part of Figure 9 presents 2 histograms obtained by: subtracting a histogram for a different action (opening the oven door) from the μ-histogram of the action opening the cupboard (top plot), and subtracting a histogram for the same action but performed by a different person from the same μ-histogram (bottom plot).

The acquisition process of an action shape is clearly dependent on the position of the camera in the environment with respect to the action performed. It is therefore obvious that certain gestures or movements will not be captured in the most relevant way by all cameras, thus their resulted histograms might contain a high degree of ambiguity between classes of actions. Since our environment is instrumented with several cameras, we performed one experiment to see the differences in the resulted histograms for a gesture using two different cameras, one located behind the subject at a $\approx 55^\circ$ angle, and one in front at $\approx 80^\circ$ angle.

Figure 10 show the results of our test, which indicate that the differences between the actions of two subjects (Z1 and J2) captured using camera *cam1* are much smaller than the ones capture by *cam0*. This demonstrates our above analysis, and suggest that using a single camera in such an environment is inadequate for capturing all the aspects of a given action accurately. We plan to extend our processing pipeline to deal with data coming from all cameras in our future work.

To better illustrate the invariance of our method to sampling density and scale, three sequences of similar silhouettes representing 10 frames, 50 frames and 100 frames respectively were taken, and their histograms compared (see Figures 11 and 12). The results validate our method of removing irrelevant frames presented in Section IV, as the feature histograms of the different sequence frames are very similar – almost coinciding. Pruning the frames at an earlier step however, decreases the computation time a lot.

For a comprehensive classification of the recorded 3D action shapes, we have gathered several data sets from multiple subjects, extracted their histograms and used them to train a SVM (Support Vector Machines) classifier. Preliminary results yielded very good result, above 96% in most of the cases.

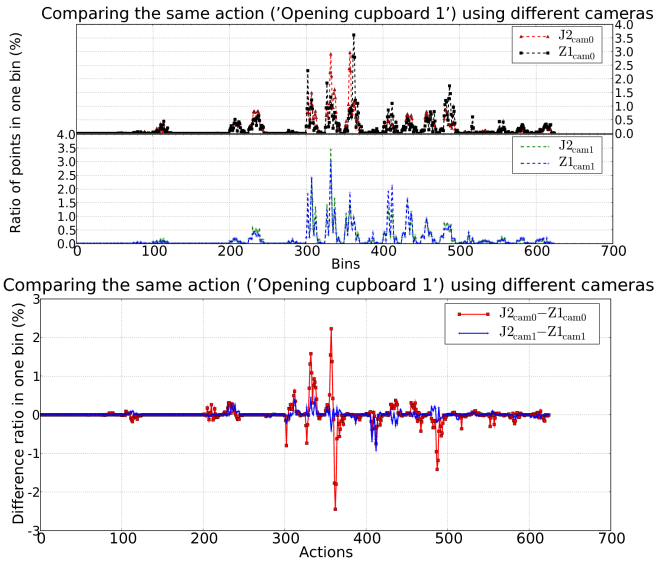

Fig. 10. Comparing the same action using different cameras for two subjects. Notice the differences in the histogram plots for the data captured using one camera (*cam0*) against the ones from another camera (*cam1*).

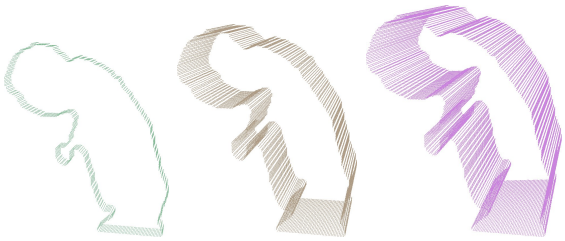

Fig. 11. Silhouettes for still poses of different lengths.

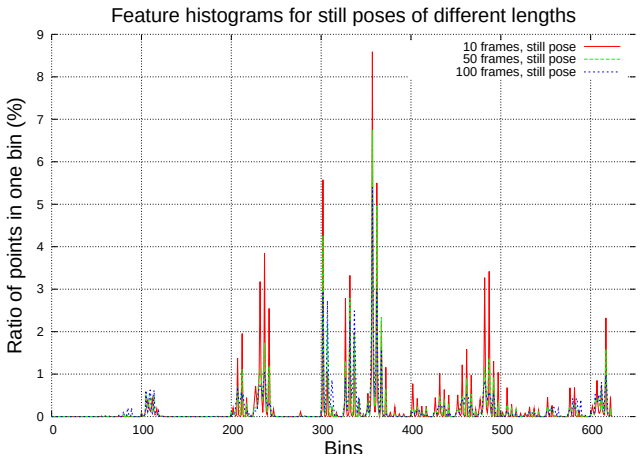

Fig. 12. Robustness of histograms towards different lengths of still poses compared to histograms of a movement. Notice that the histograms overlap almost perfectly.

VII. Summary and Future Work

In this article, we have presented a novel way of reasoning about action recognition, by transforming the problem into a different space, that of 3D point cloud based geometry. The input data is acquired using standard video cameras and 3D action shapes represented by points are generated from it. By minimizing the number of relevant frames and computing informative feature histogram descriptors, classes of actions can be robustly identified and categorized. The proposed approach can deal with large variations in the data, such as actions performed by different persons. We have presented an in-depth discussion and experimental results which look promissing for applications such as ours.

Future work will include fusing the data from multiple cameras at the action shape level, as well as testing the method on larger datasets using more variate actions.

Acknowledgements

This work is supported by the CoTeSys (Cognition for Technical Systems) cluster of excellence.

References

[1] R. B. Rusu, N. Blodow, Z. Marton, A. Soos, and M. Beetz, "Towards 3D Object Maps for Autonomous Household Robots," in *Proceedings of the 20th IEEE International Conference on Intelligent Robots and Systems (IROS), San Diego, CA, USA, Oct 29 - 2 Nov.*, 2007.

[2] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision (IJCV)*, vol. 61, no. 2, pp. 185–205, 2005.

[3] A. Efros, A. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *9th IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2003, pp. 726–733.

[4] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. S. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[5] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[6] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[7] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 984–989.

[8] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249–257, November/December 2006.

[9] G. Hetzel, B. Leibe, P. Levi, and B. Schiele, "3D Object Recognition from Range Images using Local Feature Histograms," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'01)*, vol. 2, 2001, pp. 394–399.

[10] X. Li and I. Guskov, "3D object recognition from range images using pyramid matching," in *ICCV07*, 2007, pp. 1–6.

[11] E. Wahl, U. Hillenbrand, and G. Hirzinger, "Surflet-Pair-Relation Histograms: A Statistical 3D-Shape Representation for Rapid Classification," in *3DIM03*, 2003, pp. 474–481.

[12] G. Zhang, J. Jia, W. Xiong, T. Wong, P. Heng, and H. Bao, "Moving object extraction with a hand-held camera," in *ICCV*, 2007.

[13] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, and W. Stuetzle, "Surface reconstruction from unorganized points," in *SIGGRAPH '92: Proceedings of the 19th annual conference on Computer graphics and interactive techniques*, 1992, pp. 71–78.

[14] M. Beetz, J. Bandouch, A. Kirsch, A. Maldonado, A. Müller, and R. B. Rusu, "The assistive kitchen — a demonstration scenario for cognitive technical systems," in *Proceedings of the 4th COE Workshop on Human Adaptive Mechatronics (HAM)*, 2007.