

# Representative Body Points on Top-View Depth Sequences for Daily Activity Recognition

Shu-Chun Lin, An-Sheng Liu, Tang-Wei Hsu and Li-Chen Fu, *Fellow, IEEE*

Department of Electrical Engineering

National Taiwan University

Taipei, Taiwan

{r02921013, d00921006, r03921002, lichen} @ntu.edu.tw

**Abstract**— in this paper, a novel feature for activity recognition from vertical top-view depth image sequences is firstly proposed. Most of previous works are focusing mainly on the side-view depth or color image sequences, which unfortunately may encounter occlusion problems. Therefore, top-view camera setting is adopted in our research. Based on the idea of computed tomography (CT) from medical imaging, the depth images are segmented to different layer along the transverse plane. The representative body points which are found from the centroids of the regions on each slice. And those points will be a meaningful descriptor for the activity posture. Dynamic time warping algorithm is also applied to address the different sequence length problem. Finally, a SVM classifier is trained to classify our activities. To verify our performance, a new Top-View 3D Daily Activity Dataset is constructed. In our experiments, a challenging cross-subject test is conducted, and the performance of our representative body points is demonstrated. The result shows that the accuracy can achieve up to 97%, which is promising while being compared with those from the state-of-the-art methods in the literature.

**Keywords**—component; action recognition; top-view; depth; dynamic time warping

## I. INTRODUCTION

In computer vision field, activity recognition has received considerable attention in recent years. It helps us to understand the human activities and can be applied in many fields, e.g., surveillance system, human-computer interaction, and medical health monitoring. The application scenarios considered in this paper are mainly focused for the monitoring purpose. For this application, the goal is to recognize activities by the underlying system non-intrusively. In other words, users' daily activities should not be disturbed and users should not be asked for facing specific orientations while the recognition process proceeds.

In the previous years, activity recognition research mainly focused on recognizing human activities from image sequences captured by the RGB camera. However, it still remains a challenging problem because RGB images suffer from some disadvantages. For example, the foreground is hard to be segmented from the cluttered background, and RGB images are sensitive to the illumination change. Some researches on activity recognition used sensors with higher precision like multi-layer laser range finder or motion capture system.

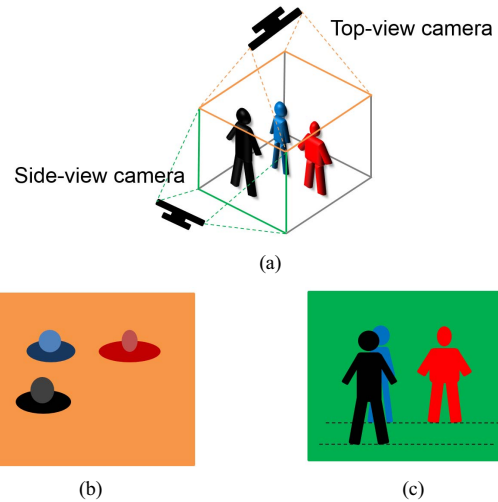


Figure 1. An illustration of (a) top-view and side-view camera settings, (b) top-view depth images, and (c) side-view depth images.

Although these sensors would provide accurate human body images, but the price and the cost of these sensors are too high and too complicated to implement in the real-life scenarios. With the emergence of the low-cost depth cameras such as Kinect and Xtion Pro, there is an increasing interest in recognizing activities with depth cameras these years. Depth image sequences have advantages like providing the pure geometry and shape cues and it's more insensitive to illumination change than RGB sequences. What's more, to address the privacy issue, RGB cameras usually raise this concern because the appearances will be observable clearly in the images from RGB camera while capturing human activities. And this issue will make users uncomfortable under surveillance. On the contrary, depth images shows shapes or silhouettes only which make users feel more relax and also will not affect the normal activities in the surveillance area.

Many efforts have been devoted to the action recognition using depth image sequences [1-7] and there are activity datasets such as MSR Action 3D, and MSR Daily Activity 3D. These researches are focusing on the side-view camera settings and the application scenarios are usually human machine interaction. Side-view recognizing human activities by depth

This research is supported by the Ministry of Science and Technology, Taiwan, R.O.C. under the Grants No. MOST 103-2627-E-002-001 and MOST 102-2218-E-002-009-MY2

image sequences is still a challenging problem due to large variations in human appearance, posture and body size within the same class. It also suffers from various factors such as discontinuous points and the uncontrollable noisy data in depth images. However, for the application of surveillance systems, side-view camera settings are not suitable because of the following problems. First, people are occluded by other people or other objects easily, as shown in Figure 1. (a), (c). And to recognize human activities with side-view camera settings, users have to face the cameras all the time and this requirement will must influence users' daily activities, which is not intuitive and violates the idea of surveillance systems.

To fit the requirements that users do not have to face the cameras and images are insensitive to the illumination changes, top-view depth camera settings are adopted in our concerned surveillance systems. The comparison between top-view camera setting and the side-view one can be seen in Figure 1. Images captured by top-view cameras have less chance to be occluded by other objects, which is so-called *inter-occlusion*, as defined in [8], and these images usually give a better overview of the overall environment. However, there are some new challenges to solve for top-view depth image sequences. First, the *self-occlusion* (*intra-occlusion*) in depth images is more severe and fewer data points can be obtained. Second, as same as the side-view depth images, top-view depth images also suffer from the discontinuous and noisy points, and such disadvantage makes the local features very unreliable.

In this paper, under this severe self-occlusion situation, a novel method to find the representative points of the body parts seen from the top-view depth image sequences are proposed. The images are segmented into different layers. The representative points of body parts are chosen according to the region on each slice. Moreover, Dynamic Time Warping (DTW) algorithm is applied for aligning the different video length. To verify our method, a new Top-View 3D Daily Activity Dataset is constructed and the experimental results show that our performance is up to 97%, which outperforms those the state-of-the-art approach, called Space-Time Occupancy Patterns (STOP) [4].

The rest of this paper is organized as follows. A brief review of the existing approaches is provided in Section II. Section III presents the feature extraction methods and DTW algorithm. Our new dataset for top-view depth camera settings and experimental results are demonstrated in Section IV. Finally, the conclusion is drawn in Section V.

## II. RELATED WORK

There are numerous researches related to side-view activity recognition using depth image sequences. As far as our goal is concerned, the related technical papers in the literature can be divided into two different classes, the first one is about reviews of applications based on top-view camera settings, and the second one is about investigation of activity recognition using depth image sequences.

As mentioned above, the technical results in the first class suggested that the top-view camera settings can solve occlusion problems. In [8-10], the result of recognize human postures from still images captured by top-view RGB cameras. In [8],

posture descriptors are learned for each posture category and the matching scores returned by the posture descriptor are used to recognize the postures. Noted that [8] showed that recognizing postures from vertical top-view images is harder than from side-view images due to the severe self-occlusion, which is also our main challenge. Both [11] and [12] detect and track humans from the depth camera under top-view settings. The detection scheme is proposed based on the graph-based segmentation and the head hemi-ellipsoid model in [11]. There are also applications in human identification [13]. However, due to the limitation of this setting, there's no other application to activity recognition based on vertical top-view depth camera so far.

The results in the second class are about researches on side-view activity recognition using depth image sequences. It has explored a number of features which range from local interest points [1], skeleton joints [2, 3], occupancy pattern [4], depth motion map [5, 6], to surface normal [7].

Generally speaking, the local interest point methods with depth image sequences, such as DSTIP [1], are not reliable due to the existence of noisy, discontinuous of the depth images. Contrarily, the human body joints of the skeleton are more reliable interest points. Therefore, methods based on human body joints of skeleton have been proposed in [2]. However, the body skeleton isn't available in the vertical top-view setting because most of the body parts cannot be seen anymore. On the other hand, compared with the local interest points and skeleton joints, the methods based on the occupancy pattern are less sensitive to noise. For example, [4] introduced the space-time occupancy pattern to represent the depth image sequences. There, the 4D subvolumes are divided to cells and the points falling into the cells are counted. They also claim that their method can tolerate the intra-class difference more than the other methods. And, the depth motion map methods transformed the 3D cloud points to the 2D space [5], [6], and the so-called depth motion maps are the accumulated differences between every two adjacent depth maps. In [5], the HOG features are extracted from depth motion maps and are taken as the representation of the activity sequences. HON4D [7] quantized the 4D space by polychorons and found the 4D surface normal to represent the depth image sequences.

As stated before, severe self-occlusion and discontinuous points in top-view depth images are our major challenges. The methods based on skeleton and body joints cannot be applied in the vertical top-view case directly. Therefore, we hereby propose a method to find the representative body points from top-view images for the purpose of activity recognition.

## III. ACTIVITY FEATURE EXTRATION

### A. Feature Extraction

For top-view depth image sequences, the self-occlusion takes place from time to time. However, still there are some patterns that can be used to recognize the underlying activities. For different activities, there will be different parts seen by the top-view cameras. Taking the walking activity as example, only head and shoulder can be seen in top-view depth images; on the other hand, the sit-down activity can not only show the head and should, but also reveal the backside and the upper

legs on the top-view depth image. Given this observation, an algorithm to find the representative points of these body parts will be proposed in the following

Each depth image in a sequence is segmented into  $K$  layers as the computer tomography (CT) slides where the depth spacing between two adjacent layers is set to be a fixed value, *distance*, and the number  $K$  is a priori chosen parameter ( $K=3$  in our experiment). After that, the region of each slide can be found based on the classic contour finding algorithm. Every centroid points of each region in the  $k$ -th layer are viewed as the candidates of interest points for the body parts in our algorithm.

Given a sequence of  $L$  depth images ( $S_1, \dots, S_L$ ), each image  $S_t$  is segmented into  $K$  layers from the least depth value, and these  $K$  segmented images are noted as  $f_k(x, y), k \in [1, K]$ . For  $k$ -th layer image  $f_k(x, y)$ , the centroid points at the  $i$ -th region  $C_k^i$  are represented as  $P_k^i(X_k^i, Y_k^i, Z_k^i)$ , and they are calculated as the follow equations:

$$\begin{aligned} X_k^i &= (m_{10})_k^i / (m_{00})_k^i. \\ Y_k^i &= (m_{01})_k^i / (m_{00})_k^i. \\ Z_k^i &= \frac{\sum_{(x,y) \in C_k^i} f_k(x, y)}{|C_k^i|}. \end{aligned}$$

where  $(m_{00})_k^i = \text{area}_k^i$ ,  $(m_{10})_k^i$  and  $(m_{01})_k^i$  are the moments of the region  $C_k^i$  and these can be calculated by the following equations:

$$\text{area}_k^i = (m_{00})_k^i = \sum_{(x,y) \in C_k^i} B_k(x, y).$$

where  $B_k(x, y)$  is the binary images of  $f_k(x, y)$ .

$$\begin{aligned} (m_{10})_k^i &= \sum_{(x,y) \in C_k^i} x B_k(x, y). \\ (m_{01})_k^i &= \sum_{(x,y) \in C_k^i} y B_k(x, y). \end{aligned}$$

Noticed that only the regions  $C_k^i$  with  $\text{area}_k^i > \text{threshold}$  will be considered. This area based criterion is to ensure that the region caused by the noise points will not be considered in our algorithm.

To find the representative points of body parts, the centroids are found subject to the following constraints:

- The head layer is the layer segmented starting from the least depth value  $Z_{\min}$ , which is the lowest point of entire depth value in a bounding box of top-view human detector [11]. And the interval of the head layer

is defined as  $Z_{\min}$  to  $Z_{\min} + \text{distance}$ , where the *distance* depth spacing between two adjacent layers. However, the least depth value sometimes will be inaccurate simply due to the sensor noise. In order to solve this problem, the least depth value  $Z_{\min}$  has to choose as the least depth value with enough cloud points.

- There is only one representative point in the head layer. The physical meaning of this representative point is actually the point of our head center.

$$N_H = 1.$$

Let  $N_k$  denote the number of the centroid points considered in the  $k$ -th layer, and the index of the head layer is noted as  $H$  in our case.

- Only the layers that are 1~2 layers lower than the head layer will be considered normally, which are denoted as  $(H+1)$ -th layer and  $(H+2)$ -th layer, respectively. For each of these two layers, only two representative points are considered, namely,

$$N_{H+1} = N_{H+2} = 2.$$

The reason for this design is that some body parts don't always appear in the depth images while the activity is performed. For example, for the walking activity, the area of a human leg is small and sometimes parts of a leg disappear due to self-occlusion. Therefore, the number of the body points to be considered is set as a certain value so that the noisy parts can be discarded.

Since  $N_H = 1$  and  $N_{H+1} = N_{H+2} = 2$ , some centroids that are not useful should be discarded. For the  $k$ -th layer, the centroids  $C_k^i$  are sorted from the largest to the smallest by the area of that region  $\text{area}_k^i$ . Then, the first  $N_k$  centroids are chosen as the representative points in the  $k$ -th layer. Noticed that if the number of centroids found in the  $k$ -th layer is less than  $N_k$ , zeros will be filled to preserve the dimension of the feature vectors. In particular, these  $N_H + N_{H+1} + N_{H+2}$  points are then considered as the feature vector  $F_t$  for the depth images  $S_t$ .

The examples of the representative points found in different postures are shown in Figure 2. For different poses, the centroid positions are quite different. Notice that although the depth value is hard to show in 2D images, the actually depth value  $Z_k^i$  in the representative points  $P_k^i(X_k^i, Y_k^i, Z_k^i)$  is very informative for top-view activity recognition.

The feature vectors  $F_t$  ( $t=1:L$ ) extracted from each depth image  $S_t$  are concatenated to a feature vector for this activity sequence and the SVM classifier is applied to recognize the activities.

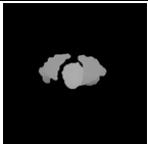
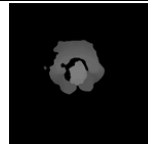
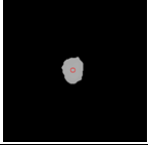
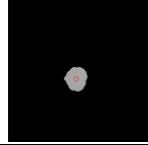
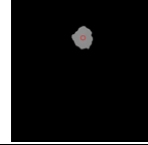
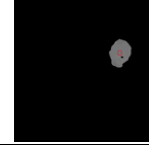
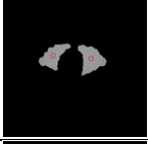
|                | (a) stand                                                                         | (b) sit                                                                           | (c) bending                                                                       | (d) squating                                                                       | (e) mahjong                                                                         | (f) fall down                                                                       |
|----------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| Original image |  |  |  |  |  |  |
| Head Layer     |  |  |  |  |  |  |
| Head Layer+1   |  |  |  |  |  |  |
| Head Layer+2   |  |  |  |  |  |  |

Figure 2. The examples of the representative points, which are the red points, found in different layers. (a) ~ (f) are the different postures consisting of each activity sequence. The number  $K$  is equal to 3 in this experiment.

### B. Dynamic Time Warping

The activity sequence lengths vary from sequence to sequence due to the different styles or different velocities performed by the different subjects. Therefore, the representations of the features are different. However, for the same activity, the feature vectors should exist consistency even there are different subjects or different styles. To align the sequence lengths, the Dynamic Time Warping (DTW) algorithm provides a good solution. DTW [14] is a dynamic programming technique for template matching, which has originally been applied to the field of speech processing. The goal of DTW is to align two sequences of feature vectors by warping the time axis iteratively until an optimal matching is found. DTW is popular because it can efficiently calculate the similarity between two templates and the effects of shifting and distortion can be minimized. With the matching result, the length of the feature vector can be warped into a constant and the SVM classifier can be utilized much easier.

For example, given two temporal sequences  $A$  and  $B$  with lengths  $N$  and  $M$ , respectively, the DTW algorithm calculates the similarity between these two sequences  $A = (A_1, \dots, A_N)$  and  $B = (B_1, \dots, B_M)$ . The goal is to find the matching warping path  $W = \{(p_1, q_1), (p_2, q_2), \dots, (p_K, q_K)\}$  such that the distance is minimized.

$$D(A, B) = \arg \min_W \left( \sum_{k=1}^K d(a(p_k), b(q_k)) \right)$$

where  $D(i, j)$  is the DTW distance between  $A(1:i)$  and  $B(1:j)$ , with the mapping path from  $(1, 1)$  to  $(i, j)$ . The Euclidean distance is used for the measure  $d(\circ)$  here. In order

to get the optimal warping path  $W$ , there are some additional constraints:

- Boundary condition: the start and end indices of the sequence should be aligned to each other, that is, the start and end point define the first and last pair of the warping path element.

$$(p_1, q_1) = (1, 1). \\ (p_K, q_K) = (N, M)$$

- Monotonic condition: For any given node  $(i, j)$  in the path, the possible fan-in nodes are restricted to  $(i-1, j), (i, j-1), (i-1, j-1)$  as illustrated in Figure 3. This constraint can guarantee that the path is non-decreasing, both the  $i$  and the  $j$  indices either stay the same or increase, they never decrease.
- Continuity condition: The increment of the element is limited to one step at a time.

To find the warping path and the final distance, the recursion is applied with the initial condition  $D(1, 1) = d(a(1), b(1))$  and the final distance  $D(n, m)$  and optimal warp path  $W$  can be obtained.

$$D(i, j) = D(i, j) + \min \{D(i-1, j), D(i-1, j-1), D(i, j-1)\}.$$

Figure 3. The illustration of monotonic condition in dynamic time warping.

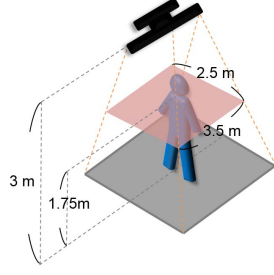


Figure 4. An illustration of the sensor setup. Xtion Pro camera is mounted on the ceiling with 3 meters height and for a person around 1.75 meters high, the sensing range is 3.5 m  $\times$  2.5m

#### IV. EXPERIMENTAL RESULTS

In the following sections, the detail of our new 3D dataset for top-view daily activity recognition will be described. Then the evaluation results of our approach are presented and compared with those from the state-of-the-art work, the Space-Time Occupancy Pattern (STOP)[4]. Experimental results show that our method outperforms the state-of-the-art method in our new dataset.

##### A. Top-View 3D Daily Activity Dataset

To test our method, the new dataset is constructed because most researches focus on the side-view depth images and thus there is no dataset captured by top-view depth cameras. To record our top-view dataset, each sequence is obtained from the Xtion Pro camera mounted on the ceiling with 3 meters height. For a person with height around 1.75 meters, the sensing area is 3.5 meters  $\times$  2.5 meters. The illustration of our system setup can be found in Figure 4. In this dataset, the background had been removed by the background subtraction scheme [15] and the extracted foreground, which contains the subjects, will be transformed into world coordinate then. Nevertheless, the dataset is still challenging as many discontinuity points appear.

This new dataset includes 6 daily activities, namely, *walk*, *stand up*, *sit down*, *squat*, *play mahjong*, and *fall down on the floor*. These 6 activities occur very often for a surveillance system, especially in the elderly health care system. Each activity is performed by 9 different subjects: 7 males and 2 females. The height of these subjects ranges from 1.6 m to 1.9 m, which makes the feature vectors quite different in the same class and thus the dataset is very challenging. Each subject performed every activity three times. Hence, there are totally 162 activity sequences in our Top-View Daily 3D Activity Dataset. Exemplar frames of these six activities in this dataset can be found in Figure 5.

##### B. Action Recognition Result

Before recognizing activities by the representative body points proposed in this paper, the problem of different sequence length is solved by the DTW. The histogram is used for the

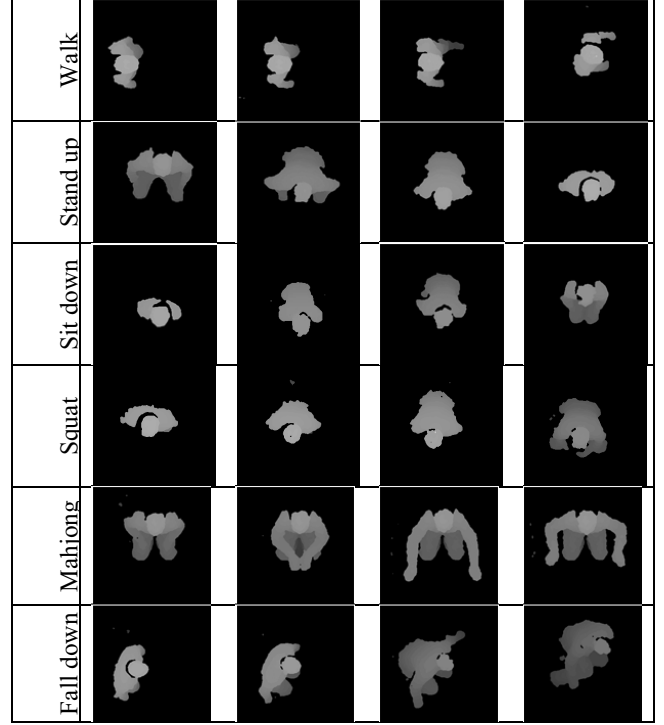


Figure 5. Example frames of six activities from the Top-View 3D Daily Activity Dataset. Each row shows four example frames in each activity. Note that, in the “sit down” and “stand up” activities, they have similar motion and shape; however, they occur in different spatiotemporal order.

DTW algorithm and the warped activity sequences can be obtained. Depth values in top-view images range from 500 mm to 3700 mm and each image is segmented to 3 layers from the least depth value, the head layer, and the spacing depth between two adjacent layers is 220 mm in our experiments.

Our experiment is the challenging cross-subject test, which means that the subjects in the testing dataset cannot be seen in the training dataset. Cross-subject test can evaluate that whether a method can tolerate the intra-class difference or not, that is, different subjects perform the same activity with different styles. Also, this test is more similar as the actually activity recognition problem in the real world because we cannot get everyone’s data to pre-train. In our experiments, 5 subjects, 15 activity sequences were randomly chosen for training and the rest 4 subjects, 12 activity sequences were used as testing. And the entire cross-subject test were done 3 times to get the average result.

To make comparison with other existing methods, STOP approach [4] is implemented. Because the method claimed all of the feature from depth image will be presented into 4d volume and the entire space occupancy projection shouldn’t be effect by the variance of viewpoint, which is suitable for our dataset. The parameters of 4D volume are adjusted for the best performance and partitioned into 10 segments along  $x$ ,  $y$ , 16 segments in  $z$  and 4 segments in  $t$  axis. Note that the depth value  $z$  in side-view sequences like MSR Action 3D dataset are around 500 mm, but the depth value  $z$  in our top-view sequences is much larger, which range from 500 ~ 3700 mm.

TABLE I. THE EXPERIMENTAL RESULT

|                            | Accuracy |
|----------------------------|----------|
| Representative Body Points | 97.7     |
| STOP [4]                   | 92.6     |

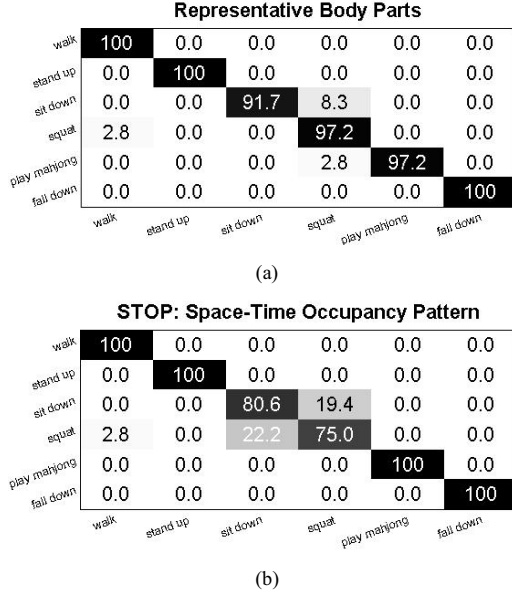


Figure 6. The Confusion matrix of our experimental result. (a) Representative Body Parts approach and (b) STOP [4] approach.

This is the reason why we partition  $z$  axis in 16 segments rather than 10 segments in  $z$  axis as in [4].

The quantification of results is shown in the following section. The confusion matrices of our method and the STOP method are also demonstrated in Figure 6. The label of  $x$  axis is the true class and the label of the  $y$  axis is the predicted result. And the average accuracy is shown in TABLE I.

Noticed that our performance is up to 97%, which is better than the STOP method and our method works very well on most activities. The confusion matrix can tell the incorrectly classified result. On the *sit-down* and *squat* activities especially, the STOP method cannot classify them correctly. On the contrary, there are only one or two confused sequences at each class when the representative body points method is applied. The STOP features for these two activities are similar due to the similar distribution in 4D space. As Figure 5. shown, for those two activities, there are only slightly different in the ending posture of entire sequence. Therefore the feature proposed from STOP method could be ambiguous for these two activities. In contrast, our representative body points are more accurate for these two activities. Therefore, the classification result is better in our case both in confusion matrix and in the average accuracy.

## V. CONCLUSION

In this paper, a new feature, defined as representative body point, for activity recognition is proposed. The set of representative body points are extracted from the segmented layer images along the transverse plane for the top-view depth image sequences. As for aligning the different sequence lengths, dynamic time warping algorithm is also applied. In the evaluation, we constructed a new Top-View 3D Daily Activity Dataset including 6 different activities and each is performed by 9 different subjects three times. In our experiment, the challenging cross-subject test is performed and the experiment

shows that our performance is up to 97% and outperforms the state-of-the-art method STOP [4]. Moreover, there is no severely confused sequence like STOP. This shows that our feature can alleviate the phenomenon of ambiguous activities. Overall, the representative body points in top-view depth image setting not only can handle the problem of occlusion, self-occlusion, and ambiguous sequence, but also can deliver a promising feature for activity recognition and its application.

## REFERENCES

- [1] X. Lu and J. K. Aggarwal, "Spatio-temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on* 2013, pp. 2834-2841.
- [2] W. Jiang, L. Zicheng, W. Ying, and Y. Junsong, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on* 2012, pp. 1290-1297.
- [3] X. Lu, C. Chia-Chih, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference on*, 2012, pp. 20-27.
- [4] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, and M. M. Campos, "STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. vol. 7441, ed: Springer Berlin Heidelberg, 2012, pp. 252-259.
- [5] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," presented at the Proceedings of the 20th ACM international conference on Multimedia, Nara, Japan, 2012.
- [6] C. Chen, R. Jafari, and N. Kehtarnavaz, "Action Recognition from Depth Sequences Using Depth Motion Maps-Based Local Binary Patterns," in *Applications of Computer Vision (WACV), IEEE Winter Conference on* 2015, pp. 1092-1099.
- [7] O. Oreifej and L. Zicheng, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2013, pp. 716-723.
- [8] H. Ninghang, G. Englebiene, and B. Krose, "Posture recognition with a top-view camera," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, 2013, pp. 2152-2157.
- [9] S. Weerachai and M. Mizukawa, "Human behavior recognition via top-view vision for intelligent space," in *Control Automation and Systems (ICCAS), 2010 International Conference on*, 2010, pp. 1687-1690.
- [10] Q. Lin, C. Zhou, S. Wang, and X. Xu, "Human Behavior Understanding via Top-View Vision," *AASRI Procedia*, vol. 3, pp. 184-190, 2012.
- [11] T. Ting-En, L. An-Sheng, H. Po-Hao, H. Cheng-Ming, and F. Li-Chen, "Real-time people detection and tracking for indoor surveillance using multiple top-view depth cameras," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014, pp. 4077-4082.
- [12] M. Rauter, "Reliable Human Detection and Tracking in Top-View Depth Images," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, 2013, pp. 529-534.
- [13] D. Kouno, K. Shimada, and T. Endo, "Person Identification Using Top-View Image with Depth Information," in *Software Engineering, Artificial Intelligence, Networking and Parallel & Distributed Computing (SNPD), 13th ACIS International Conference on* 2012, pp. 140-145.
- [14] Q. K. L. Chinh Huu Pham, Thanh Ha Le, "Human Action Recognition Using Dynamic Time Warping and Voting Algorithm," *VNU Journal of Science*, vol. 30, pp. 22-30, 2014.
- [15] A. B. Klaus Greff, Stephan Krauß, Didier Stricker, Esteban Clua, "A Comparison between Background Subtraction Algorithms using a Consumer Depth Camera," *International Conference on Computer Vision Theory and Applications*, 2012.