Robust 3D Action Recognition with Random Occupancy Patterns

Jiang Wang[1], Zicheng Liu[2], Jan Chorowski[3], Zhuoyuan Chen[1], and Ying Wu[1]

[1] Northwestern University
[2] Microsoft Research
[3] University of Louisville

Abstract. We study the problem of action recognition from depth sequences captured by depth cameras, where noise and occlusion are common problems because they are captured with a single commodity camera. In order to deal with these issues, we extract semi-local features called random occupancy pattern (ROP) features, which employ a novel sampling scheme that effectively explores an extremely large sampling space. We also utilize a sparse coding approach to robustly encode these features. The proposed approach does not require careful parameter tuning. Its training is very fast due to the use of the high-dimensional integral image, and it is robust to the occlusions. Our technique is evaluated on two datasets captured by commodity depth cameras: an action dataset and a hand gesture dataset. Our classification results are superior to those obtained by the state of the art approaches on both datasets.

1 Introduction

Recently, the advance of the imaging technology has enabled us to capture the depth information in real-time, and various promising applications have been proposed [1–4]. Compared with conventional cameras, the depth camera has several advantages. For example, segmentation in depth images is much easier, and depth images are insensitive to changes in lighting conditions. In this paper, we consider the problem of action recognition from depth sequences.

Although skeleton tracking algorithm proposed in [1] is very robust for depth sequences when little occlusion occurs, it can produce inaccurate results or even fails when serious occlusion occurs. Moreover, the skeleton tracking is unavailable for human hands thus cannot be utilized for hand gesture recognition.

Therefore, we aim at developing an action recognition approach that directly takes the depth sequences as input. Designing an efficient depth sequences representation for action recognition is a challenging task. First of all, depth sequences may be seriously contaminated by occlusions, which makes the global features unstable. On the other hand, the depth maps do not have as much texture as color images do, and they are too noisy to apply local differential operators such as gradients on. These challenges motivate us to seek for features that are semi-local, highly discriminative and robust to occlusion.

In this paper, we treat a three-dimensional action sequence as a 4D shape and propose random occupancy pattern (ROP) features, which are extracted from

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 872–885, 2012.

randomly sampled 4D subvolumes with different sizes and at different locations. Since the ROP features are extracted at a larger scale, it is robust to noise. At the same time, they are less sensitive to occlusion because they only encode information from the regions that are most discriminative for the given action.

An Elastic-Net regularized classification model is developed to further select the most discriminative features, and sparse coding is utilized to robustly encode the features.The feature encoding step further improves the proposed method's robustness to occlusion by modeling the occlusion noise as the sparse coding reconstruction errors. The proposed approach performs well on the depth sequence dataset and is robust to occlusion. The general framework of the proposed method is shown in Fig. 1.

We evaluate our method on two datasets captured by commodity depth cameras. The experimental results validate the effectiveness of the proposed method.

Our main contributions are as follows: First, we propose a computationally efficient method to perform action recognition from depth sequences. Second, a novel weighted sampling scheme is proposed to effectively explore an extremely large dense sampling space. Third, we propose to employ sparse coding to deal with the occlusion in the depth sequences.

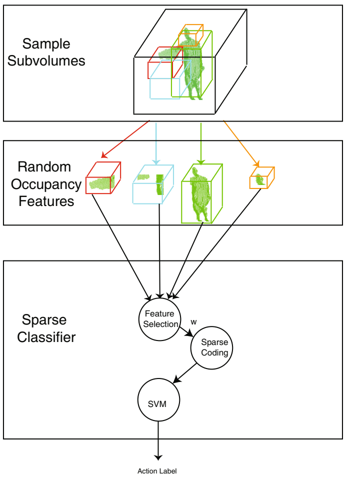


Fig. 1. The framework of the proposed method. The 3D subvolumes are shown for illustration purpose. In the implementation, 4D subvolumes are employed.

2 Related Work

The Haar wavelet-like features have been successfully applied in [5] for face detection. A boosted classifier is learned by applying AdaBoost algorithm [6] on a very large pool of weak classifiers. The weak classifier pool is constructed based on the features extracted from the rectangles at all possible locations and scales. While this approach is successful for 2D images, when the data becomes 3D or 4D, the number of possible rectangles becomes so large that enumerating them or performing AdaBoost algorithm on them becomes computationally prohibitive. [7] utilizes 3D occupancy features and models the dynamics by an exemplar-based hidden Markov model. The proposed ROP feature is much simpler and more computationally efficient than Haar-like features, while achieving similar performances in depth datasets.

Randomization has been applied in [8] and [9] to address this problem. [9] employs a random forest to learn discriminative features that are extracted either from a patch or from a pair of patches for fine-grained image categorization. [8] applies a random forest to mine discriminative features from binary spatial arrangement features for character recognition. Their work demonstrates the effectiveness of randomization in dealing with this problem. We exploit randomization to perform action recognition from depth sequences, in which the data is sparse and the number of the possible subvolumes is much larger.

[10] and [11] also employ randomization in the learning process. These approaches randomly map the data to features with a linear function whose weights and biases are uniformly sampled. Their empirical and theoretical results show that their training is much faster than AdaBoost's, while their classification accuracy is comparable to AdaBoost's. The proposed approach also randomly maps the data to features. Unlike [10] and [11], however, our approach exploits the neighborhood structure of depth sequences, and extracts features from the pixels that are spatially close to each other. Furthermore, we propose a weighted sampling technique that is more effective than uniform sampling.

Recently, a lot of efforts have been made to develop features for action recognition in depth data. [12] represents each depth frame as a bag of 3D points on the human silhouette, and utilizes HMM to model the temporal dynamics. [13] uses relative skeleton position and local occupancy patterns to model the human-object interaction, and developed Fourier Temporal Pyramid to characterize temporal dynamics. [14] also applies spatio-temporal occupancy patterns, but all the cells in the grid have the same size, and the number of cells is empirically set. [15] proposes a dimension-reduced skeleton feature, and [16] developed a histogram of gradient feature over depth motion maps. Instead of carefully developing good features, this paper tries to learn semi-local features automatically from the data, and we show that this learning-based approach achieves good results.

3 Random Occupancy Patterns

The proposed method treats a depth sequence as a 4D volume, and defines the value of a pixel in this volume $I(x, y, z, t)$ to be either 1 or 0, depending on whether there is a point in the 4D volume at this location. The action recognition is performed by using the value of a simple feature called *random occupancy patterns*, which is both efficient to compute and highly discriminative.

This paper employ four dimensional random occupancy patterns to construct the features, whose value is defined to be the soft-thresholded sum of the pixels in a subvolume:

$$o_{xyzt} = \delta(\sum_{q \in \text{bin}_{xyzt}} I_q) \tag{1}$$

where $I_q = 1$ if the point cloud has a point in the location q and $I_q = 0$ otherwise. $\delta(.)$ is a sigmoid normalization function: $\delta(x) = \frac{1}{1+e^{-\beta x}}$. This feature is able to capture the occupancy pattern of a 4D subvolume. Moreover, it can be computed in constant complexity with the high dimensional integral images [17].

As shown in Fig. 1, we extract ROP features from the subvolumes with different sizes and at different locations. However, the number of possible simple features is so large that we are not able to enumerate all of them. In this paper, we propose a weighted random sampling scheme to address this problem, which will be described in Section 4.

4 Weighted Sampling Approach

Since the number of possible positions and sizes of a 4D subvolume is extremely large and the information of these features is highly redundant, it is neither necessary nor computationally efficient to explore all of them. In this section, we propose a random sampling approach to efficiently explore these 4D subvolumes.

4.1 Dense Sampling Space

Recall that a ROP feature is extracted from a 4D subvolume. It is computationally prohibitive to extract the features from all possible subvolumes, thus we would like to sample a subset of discriminative subvolumes and extract the features from them. This subsection characterizes the *dense sampling space*, from which we randomly sample subvolumes.

Denote the size of a depth sequence volume to be $W_x \times W_y \times W_z \times W_t$. A subvolume can be characterized by two points $[x_0, y_0, z_0, t_0]$ and $[x_1, y_1, z_1, t_1]$, and is denoted as $[x_0, y_0, z_0, t_0] \sim [x_1, y_1, z_1, t_1]$. A normal subvolume has the property that $x_0 \leq x_1, y_0 \leq y_1, z_0 \leq z_1$, and $t_0 \leq t_1$, and the subvolume is the set of points

$$\begin{aligned} \{[x, y, z, t] : x_0 \leq x \leq x_1, y_0 \leq y \leq y_1, \\ z_0 \leq z \leq z_1, t_0 \leq t \leq t_1\} \end{aligned} \tag{2}$$

Our sampling space consists of all the subvolumes $[x_0, y_0, z_0, t_0] \sim [x_1, y_1, z_1, t_1]$ where $x_0, x_1 \in \{1, 2, \cdots, W_x\}$, $y_0, y_1 \in \{1, 2, \cdots, W_y\}$, $z_0, z_1 \in \{1, 2, \cdots, W_z\}$, $t_0, t_1 \in \{1, 2, \cdots, W_t\}$. If we take $(W_x, W_y, W_z, W_t) = (80, 80, 80, 80)$, the size of the dense sampling space is $80^8 = 1.67 \times 10^{15}$. This dense sampling space is so large that exhaustively exploring it is computationally prohibitive. However, since the subvolumes highly overlap with each other, they contain redundant information, and it is possible to employ randomization to deal with this problem.

4.2 Weighted Sampling

One way to sample from the dense sampling space is to perform uniform sampling. Nevertheless, since in the depth sequences many subvolumes do not contain useful information for classification, uniform sampling is highly inefficient. In this section, to efficiently sample from the dense sampling space, we propose a weighted sampling approach based on the rejection sampling, which samples the discriminative subvolumes with high probability.

To characterize how discriminative a subvolume is, we employ the scatter matrix class separability measure [18]. The scatter matrices include *Within-class scatter matrix* ($\boldsymbol{S}_W$), *Between-class scatter matrix* ($\boldsymbol{S}_B$), and *Total scatter matrix* ($\boldsymbol{S}_T$). They are defined as $\boldsymbol{S}_W = \sum_{i=1}^{c} \sum_{j=1}^{n_i} (\boldsymbol{h}_{i,j} - \boldsymbol{m}_i)(\boldsymbol{h}_{i,j} - \boldsymbol{m}_i)^T$, $\boldsymbol{S}_B = \sum_{i=1}^{c} n_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$, $\boldsymbol{S}_T = \boldsymbol{S}_W + \boldsymbol{S}_B$, where c is the number of the classes, n_i denotes the number of training data in the i-th class, and $\boldsymbol{h}_{i,j}$ denote the features extracted from the j-th training data in the i-th class. $\boldsymbol{m}_i$ denotes the mean vectors of the features $\boldsymbol{h}_{i,j}$ in the i-th class and $\boldsymbol{m}$ the mean vector of the features extracted from all the training data. A large separability measure means that these classes have small within-class scatter and large between-class scatter, and the class separability measure J can be defined as

$$J = \frac{\mathrm{tr}(\boldsymbol{S}_{\mathrm{W}})}{\mathrm{tr}\boldsymbol{S}_{\mathrm{B}}} \tag{3}$$

Denote V as the 4D volume of a depth sequence. For each pixel $p \in V$, we define a neighborhood subvolume centered at p, and extract the 8 Haar feature values from this neighborhood subvolume. These 8 feature values form an 8-dimensional vector which is used as the feature vector $\boldsymbol{h}$ to evaluate the class separability score J_p at pixel p.

A subvolume should be discriminative if all the pixels in this subvolume are discriminative, and vice versa. Therefore, we utilize the average of the separability scores of all the pixels in the region R as the separability score of R.

The probability that a subvolume R is sampled should be proportional to its separability score J_R, that is,

$$P_{R\ \mathrm{sampled}} \propto J_R = \frac{1}{N_R} \sum_{p \in R} J_p \tag{4}$$

where N_R is the number of pixels in the subvolume R.

We can uniformly draw a subvolume, and accept the subvolume with probability

$$P_{R\ \mathrm{accept}} = \frac{W_x W_y W_z W_t}{\sum_{p \in V} J_p} J_R \tag{5}$$

Note that $P_{R\ \mathrm{uniform}} P_{R\ \mathrm{accept}} = P_{R\ \mathrm{sampled}}$. Therefore, with the rejection sampling scheme, the probability that R is selected is equal to the desired probability as specified in equation (4). The derivation of the acceptance rate can be found in the supplemental material.

Because we are able to compute the average separability score in a subvolume very efficiently using the high dimensional integral image, this sampling scheme is computationally efficient. The outline of the algorithm is shown in Alg. 1.

```
Let $N_S$ denote the number of subvolumes to sample, $J_p$ the separability score at pixel $p$.
Compute $P = \frac{W_x W_y W_z W_t}{\sum_{p \in V} J_p}$
repeat
    Uniformly draw a subvolume $R$.
    Compute the average separability score in this subvolume $J_R$ with integral image.
    Compute the acceptance rate $P_{\text{accept}} = P J_R$.
    Uniformly draw a number $N$ from $[0, 1]$.
    if $N \leq P_{accept}$ then
        Retain the subvolume $R$.
    end
    else
        Discard the subvolume $R$.
    end
until the number of subvolumes sampled $\geq N_S$ ;
```

Algorithm 1. Weighted Sampling Algorithm

5 Learning Classification Functions

Given the training data pairs $(\boldsymbol{x}^i, t^i), i = 1, \cdots, n$, where $\boldsymbol{x}^i \in \mathcal{R}^L$ denotes a training data, and $t^i \in \mathcal{T}$ is the corresponding label, the aim of the classification problem is to learn a prediction function $g : \mathcal{R}^L \rightarrow \mathcal{T}$. Without loss of generality, we assume the classification problem is a binary classification problem, i.e., $\mathcal{T} = \{0, 1\}$. If the problem is a multiclass problem, it can be converted into a binary classification problem with the one-vs-others approach.

An Elastic-Net regularization is employed to select a sparse subset of features that are the most discriminative for the classification. Choosing a sparse subset of features has several advantages. First, the speed of the final classifier is faster if the number of the selected features is smaller. Second, learning a sparse classification function is less prone to over-fitting if only limited amount of training data is available [19].

For each training data sample $\boldsymbol{x}^i$, N_f ROP features are extracted: $h^i_j, j = 1, \cdots, N_f$, and the response is predicted by a linear function

$$y^i = \sum_{j=1}^{N_f} w_j h^i_j \tag{6}$$

Denote $\boldsymbol{w}$ as the vector containing all $w_j, j = 1, \cdots, N_f$. The objective of the learning is to find $\boldsymbol{w}$ that minimizes the following objective function:

$$E = \sum_{i=1}^{n} (t^i - y^i)^2 + \lambda_1 \|\boldsymbol{w}\|_1 + \lambda_2 \|\boldsymbol{w}\|_2^2 \tag{7}$$

where λ_1 is a regularization parameter to control the sparsity of $\boldsymbol{w}$, and λ_2 is used to ensure that the margin of the classifier is large. It has been shown that if the number of features N_f is much larger than that of the training data n, which is the case of this paper, Elastic-Net regularization works particularly well [19]. SPAMS toolbox [20] is employed to numerically solve this optimization problem.

The selected feature $\boldsymbol{f}$ is obtained by discarding the features x_j with corresponding w_j less than a given threshold and multiplying the rest of x_j by weight w_j.

All the training data is utilized as the dictionary $\boldsymbol{A} = [\boldsymbol{f}^1, \boldsymbol{f}^2, \cdots, \boldsymbol{f}^n]$. For a test data with feature $\boldsymbol{f}$, we can solve the following sparse coding problem:

$$\min \frac{1}{2} \|\boldsymbol{f} - \boldsymbol{A}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \tag{8}$$

where $\boldsymbol{\alpha}$ is called the reconstruction coefficients, and λ is a regularization parameter. This model assumes the feature vector $\boldsymbol{f}$ can be represented as the linear combination of the features of the training data plus a noise vector ϵ.

$$\boldsymbol{f} = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \boldsymbol{f}^i + \epsilon \tag{9}$$

The reconstruction coefficients $\boldsymbol{\alpha}$ are employed to represent a depth sequence, and an SVM classifier is trained for action classification.

6 Experimental Results

In this section, we evaluate our algorithm on three datasets captured by commodity depth cameras: the MSR-Action3D datasets [12], Gesture3D dataset. The experimental results show that our algorithm outperforms the existing methods on these datasets, and is not sensitive to occlusion error. The β of the sigmoid function for ROP feature is set to be 10 in all the experiments.

6.1 MSR-Action3D

MSR-Action3D dataset [12] is an action dataset of depth sequences captured by a depth camera. This dataset contains twenty actions: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pick up & throw*. Each action was performed by ten subjects for three times. The frame rate is 15 frames per second and resolution 640 × 480. Altogether, the dataset has 23797 frames of depth maps for 402 action samples. Some examples of the depth map sequences are shown in Fig. 2.

Those actions were chosen to cover various movement of arms, legs, torso and their combinations, and the subjects were advised to use their right arm or leg if an action is performed by a single arm or leg. Although the background of this dataset is clean, this dataset is challenging because many of the actions in the dataset are highly similar to each other. In this experiment, 50000 subvolumes are sampled unless otherwise stated.

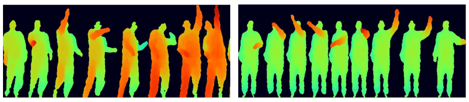


Fig. 2. Sample frames of the MSR-Action3D dataset

In order to ensure the consistency of the scale, Each depth sequence is resized to the same size $80 \times 80 \times 80 \times 10$. The separability scores of the MSR-Action3D are shown in Fig. 3. The regions with high separability scores are human's arms and legs, which is consistent with the characteristics of the dataset, and the beginning and the ending parts of an action is less discriminative than the middle part of an action. Moreover, in Fig. 3(d), it can be observed that the center of the human does not have high separability score, because the center of the human usually does not have many movements and does not contain useful information for classification.

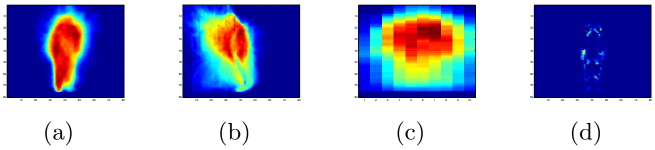


Fig. 3. The projection of the separability scores for MSR-Action3D, where red color means high intensity. (a) the projection of the separability scores to x-y plane. (b) the projection of the separability scores to y-z plane. (c) the projection of the separability scores to y-t plane. (d) the cross section of the separability scores on x-y plane at $z = 40, t = 1$.

We compare our algorithm with the state-of-the-art method [12] on this dataset, which extracts the contour information from the depth maps, with half of the subjects as training data and the rest of the subjects as test data. Table 1 shows the recognition accuracy. The recognition accuracy is computed by running the experiments 10 times and taking the average of each experiment's accuracy. Our method outperforms this method by a large margin. Notice that our classification configuration uses half of the subjects as the training data and the rest of them as test data, which is difficult because of the larger variations across the same actions performed by different subjects. Our method is also compared with the STIP features. [21], which is a state-of-the-art local feature designed for action recognition from videos. The local spatio-temporal features do not work well for depth data because there is little texture in depth maps. Another method we compare with is the convolutional network. We have implemented a 4-dimensional convolutional network by extending the three-dimensional convolutional network of [22]. Finally we compare with a Support Vector Machine classifier on the raw features consisting of the pixels on all the locations. Although the Support Vector Machine performs surprisingly well on our dataset, the training time of the SVM is very long because the dimension of the features is very high. In contrast, the proposed method is simple to implement and is computationally efficient in both training and testing. Moreover, it outperforms all the other methods including SVM. In addition, we can see that the proposed ROP feature performs comparably with Haar features.

Table 1. Recognition Accuracy Comparison for MSR-Action3D dataset

Method	Accuracy
STIP features [21]	0.423
Action Graph on Bag of 3D Points [12]	0.747
High Dimensional Convolutional Network	0.725
STOP feature [14]	0.848
Eigenjoints [15]	0.823
Support Vector Machine on Raw Data	0.79
Proposed Method (Without sparse coding)	**0.8592**
Proposed Method (Haar Feature)	**0.8650**
Proposed Method (Sparse Coding)	**0.8620**

In order to test the sensitivity of the proposed method to occlusions, we divide each depth sequences into $2 \times 2 \times 1 \times 2$ subvolumes, i.e., we partition each depth sequences into two parts in y, x and t dimensions. Each volume only covers half of the frames of the depth sequences. Occlusion is simulated by ignoring the points that fall into the specified occluded subvolume, illustrated in Fig. 4. We run the simulation with one subvolume occluded, the performance is shown in Table. 2. It can seen that employing sparse coding can greatly improve the robustness.

The confusion matrix is shown in Fig. 5. The proposed method performs very well on most of the actions. Some actions, such as “catch” and “throw”, are too similar to each other for the proposed method to capture the difference.

We also compare the classification accuracy of the proposed sampling scheme to that of the uniform sampling scheme in Fig. 6(a). It can be observed that the weighted sampling scheme is more effective than the uniform sampling scheme. Moreover, the proposed scheme does not suffer from overfitting even when the number of the sampled subvolumes is very large. [23] gives an intuitive proof of the generalization ability of classifier of the randomly generated features.

The depth sequences are downsampled into different resolutions, and we explore the relationship between the resolution of the data and the classification

Table 2. Robustness to occlusion comparison

Occlusion	Accuracy without using sparse coding	Accuracy using Sparse Coding
1	83.047	86.165
2	84.18	86.5
3	78.76	80.09
4	82.12	85.49
5	84.48	87.51
6	82.46	87.51
7	80.10	83.80
8	85.83	86.83

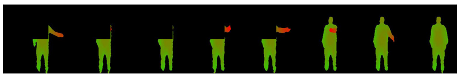


Fig. 4. An occluded depth sequence.

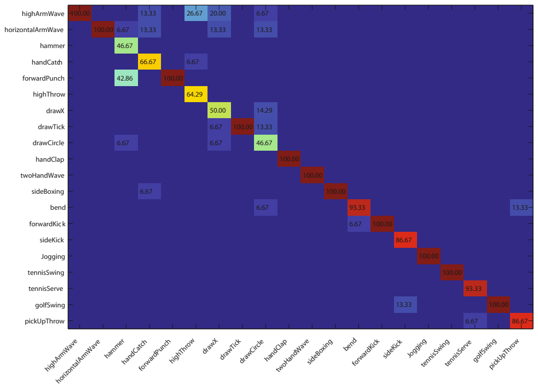
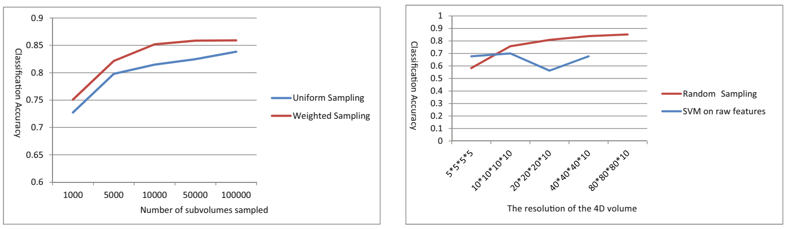


Fig. 5. The confusion matrix for the proposed method on MSR-Action3D dataset. It is recommended to view the figure on the screen

accuracy. The relationship found is shown in Fig. 6(b). Our observation is that the performance of the SVM classifier may drop when we increase the resolution of the data, but for our random sampling scheme, increasing the data resolution always increases the classification accuracy.

6.2 Gesture3D Dataset

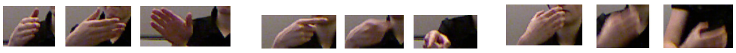
The Gesture3D dataset [24] is a hand gesture dataset of depth sequences captured by a depth camera. This dataset contains a subset of gestures defined by American Sign Language (ASL). There are 12 gestures in the dataset: *bathroom*, *blue*, *finish*, *green*, *hungry*, *milk*, *past*, *pig*, *store*, *where*, *j*, *z*. Some example frames



(a) (b)

Fig. 6. The comparison between different sampling methods, and the relationship between the resolution of data and the classification accuracy for SVM and the proposed sampling method.

of the gestures are shown in Fig. 7. Notice that although this dataset contains both the color and depth frames, only depth frames are used in the experiments. Further description of the gestures can be found in [25]. All of the gestures used in this experiment are dynamic gestures, where both the shape and the movement of the hands are important for the semantics of the gesture. There are ten subjects, each performing each gesture two or three times. In total, the dataset contains 336 depth sequences. The self occlsion is more common in the gesture dataset.

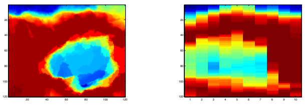


(a) blue gesture (b) green gesture (c) hungry gesture

Fig. 7. The sample frames of the Gesture3D dataset, (a) the blue gesture. (b) green gesture. (c) hungry gesture.

In this experiment, all gesture depth sequences are subsampled to size 120 × 120 × 3 × 10. The leave-one-subject-out cross-validation is employed to evaluate the proposed method. The recognition accuracy is shown in Table 3. The proposed method performs significantly better than the SVM on raw features and the high dimensional convolutional network. Our performance is also slight better than the action graph model which uses carefully designed shape features [24].

The separability score map for Gesture3D dataset is shown in Fig. 8. We observe that the score map of the gestures is very different from that of the actions shown in Fig. 3, because the movement pattern of the gestures and the actions is very different. In gesture depth sequences, the semantics of the gesture are mainly determined by the large movement of the hand, while the human actions are mainly characterized by the small movements of the limbs.



(a) (b)

Fig. 8. The projection of the scores for Gesture3D, (a) the projection of the scores to x-y plane. (b) the projection of the scores to y-t plane.

The confusion matrix is shown in Fig. 9. The proposed method performs quite well for most of the gestures. It can be observed from the confusion matrix that large confusion exists between the gesture "where" and "green". Both gestures involve the movement of one finger, and only the directions of the movement are slightly different.

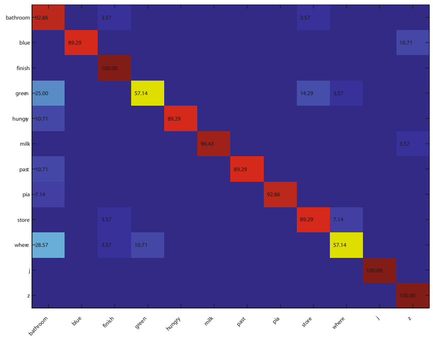


Fig. 9. The confusion matrix of the proposed method on Gesture3D dataset

Table 3. Recognition Accuracy Comparison for Gesture3D dataset

Method	Accuracy
SVM on Raw Features	0.6277
High Dimensional Convolutional Network [22]	0.69
Action Graph on Occupancy Features [24]	0.805
Action Graph on Silhouette Features [24]	0.877
Proposed Method (Without sparse coding)	**0.868**
Proposed Method (Sparse coding)	**0.885**

7 Conclusion

This paper presented a novel random occupancy pattern features for 3D action recognition, and proposed a weighted random sampling scheme to efficiently explore an extremely large dense sampling space. A sparse coding approach is employed to further improve the robustness of the proposed method. Experiments on different types of datasets, including an action recognition dataset and a gesture recognition dataset, demonstrated the effectiveness and robustness of the proposed approach as well as its broad applicability.

Acknowledgments. This work was supported in part by National Science Foundation grant IIS-0347877, IIS-0916607, US Army Research Laboratory and the US Army Research Office under grant ARO W911NF-08-1-0504, and DARPA Award FA 8650-11-1-7149. This work was conducted when the first author was at Microsoft Research Redmond.

References

1. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
2. Hadfield, S., Bowden, R.: Kinecting the dots: Particle Based Scene Flow From Depth Sensors. In: ICCV (2011)
3. Baak, A., Meinard, M., Bharaj, G., Seidel, H.P., Theobalt, C., Informatik, M.P.I.: A Data-Driven Approach for Real-Time Full Body Pose Reconstruction from a Depth Camera. In: ICCV (2011)
4. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient Regression of General-Activity Human Poses from Depth Images. In: ICCV (2011)
5. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. International Journal of Computer Vision 57, 137–154 (2004)
6. Freund, Y., Schapire, R.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. In: Computational Learning Theory, vol. 55, pp. 23–37. Springer (1995)
7. Weinland, D., Boyer, E., Ronfard, R.: Action Recognition from Arbitrary Views using 3D Exemplars. In: ICCV, pp. 1–7 (2007)
8. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Computation 9, 1545–1588 (1997)
9. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR (2011)
10. Rahimi, A., Recht, B.: Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In: NIPS, vol. 885. Citeseer (2008)
11. Huang, G.B., Wang, D.H., Lan, Y.: Extreme learning machines: a survey. International Journal of Machine Learning and Cybernetics 2, 107–122 (2011)
12. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Human Communicative Behavior Analysis Workshop (in conjunction with CVPR) (2010)
13. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In: CVPR (2012)

14. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.M.: STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences. In: 17th Iberoamerican Congress on Pattern Recognition, Buenos Aires (2012)
15. Yang, X., Tian, Y.: EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor. In: CVPR 2012 HAU3D Workshop (2012)
16. Yang, X., Zhang, C., Tian, Y.: Recognizing Actions Using Depth Motion Maps-based Histograms of Oriented Gradients. In: ACM Multimedia (2012)
17. Tapia, E.: A note on the computation of high-dimensional integral images. Pattern Recognition Letters 32, 197–201 (2011)
18. Wang, L., Chan, K.L.: Learning Kernel Parameters bu using Class Separability Measure. In: NIPS (2002)
19. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society 67, 301–320 (2005)
20. Julien Mairal (SPArse Modeling Software),
    `http://www.di.ens.fr/willow/SPAMS/`
21. Laptev, I.: On Space-Time Interest Points. IJCV 64, 107–123 (2005)
22. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. In: ICML. Citeseer (2010)
23. Bartlett, P.: The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. IEEE Transactions on Information Theory 44, 525–536 (1998)
24. Kurakin, A., Zhang, Z., Liu, Z.: A real-time system for dynamic hand gesture recognition with a depth sensor. In: EUSIPCO (2012)
25. (Basic America Sign Language),
    `http://www.lifeprint.com/asl101/pages-layout/concepts.htm`