Histogram of 3D Facets: A Characteristic Descriptor for Hand Gesture Recognition

Chenyang Zhang, Xiaodong Yang, and YingLi Tian
Department of Electrical Engineering
The City College of New York, CUNY
{czhang10, xyang02, ytian}@ccny.cuny.edu

***Abstract*—The availability of 3D sensors has recently made it possible to capture depth maps in real time, which facilitates a variety of visual recognition tasks including hand gesture recognition. However, most existing methods simply treat depth information as intensities of gray images and ignore the strong 3D shape information. In this paper, we propose a novel characteristic descriptor, i.e., Histogram of 3D Facets (H3DF), to explicitly encode the 3D shape information from depth maps. We define a 3D facet as a 3D local support surface associated with each 3D cloud point. By robust coding and pooling 3D facets from a depth map, the proposed H3DF descriptor can effectively represent the 3D shapes and structures of various hand gestures. We evaluate the proposed descriptor on two challenging 3D datasets of hand gesture recognition. The recognition results in the context of both decimal digits and letters in American Sign Language (ASL) demonstrate that our approach significantly outperforms the state-of-the-art methods.**

Keywords - depth map; 3D feature representation; histogram of 3D facets (H3DF); hand gesture recognition

I. INTRODUCTION

Hand gesture recognition, as a significant component of Human Computer Interaction (HCI), has appealed many efforts invested from the research field of computer vision in recent decades for its strong potential in numerous applications, such as game interaction and sign language recognition. However, hand gesture recognition is still a challenging task due to the wide range of poses and considerable intra-class variations, e.g., rotation, scaling, viewpoint change and hand articulations.

As the release of a collection of commodity depth sensors and corresponding development toolkits, research related to 3D depth map has attracted more attentions in recent years [3-7, 9-11, 13, 14, 16, 17]. RGBD cameras have demonstrated their capability to provide more information of object size, shape, and position. Compared to the traditional RGB camera, research on 3D depth map has significant advantages for its availability to discover strong clues in boundaries and 3D spatial layout even in cluttered background and weak illumination. Particularly, those traditional challenging tasks such as object detection and segmentation become much easier with the depth information added in [11]. These new findings have also been motivating recent research to explore hand gesture recognition by using 3D information [6, 7, 13]. However, to the best of our knowledge, most of previous work of 3D depth map based hand gesture recognition only focused on 2D features, e.g., Gabor filters [6] and contour matching [7], which were developed for 2D images.

General Framework: In order to directly and effectively capture and encode strong 3D shape and structure information from depth maps, we propose a novel characteristic descriptor named *Histogram of 3D Facets* (**H3DF**). In 3D depth maps, we define a local support surface of a 3D cloud point as a ***3D facet***, which captures informative local surface properties of each cloud point. In order to capture this information, we first code each facet based on the normal orientation of a local 3D plane and then apply a concentric spatial pooling strategy to aggregate a collection of facets from a region of interest. In the application of hand gesture recognition, a region of interest refers to the image patch that covers a hand gesture.

Main Contributions: Compared to existing 2D image descriptors, our proposed H3DF descriptor based on the depth map has two main advantages: 1) it explicitly captures the 3D informative shape properties conveyed by the depth map; 2) it applies a compact global representation to describe a depth image compared to other 2D global descriptors, e.g., Histogram of Oriented Gradients (HOG) [1]. We evaluate the proposed descriptor on two public datasets of hand gesture recognition: the NTU Hand Digits dataset [7] and the ASL Finger Spelling dataset [6]. The experimental results on both datasets demonstrate that our approach significantly outperforms the state-of-the-art methods.

The remainder of this paper is organized as follows. Section II reviews related work on hand gesture recognition and depth maps. In Section III, we provide the detailed procedures of building the proposed H3DF descriptor. Section IV describes hand gesture recognition using H3DF. A variety of experimental results and discussions are presented in Section V. Finally, we conclude the remarks of this paper in Section VI.

II. RELATED WORK

Hand gesture recognition serves as an important component in HCI due to the conveyed information covers multiple functions, such as conversational gesture, controlling gestures, manipulative gesture, and communicative gestures [15]. As the first step of hand gesture recognition, hand

detection and tracking are usually implemented by skin color or shape based segmentation, which can be inferred from RGB images [2]. Based on the detection and tracking result, either dynamic or static features can be extracted for gesture recognition [15]. However, because of the intrinsic vulnerability against background clutters and illumination variations, hand gesture recognition on 2D RGB images usually requires a clean background, which limits its applications in the real world.

The release of 3D sensors (e.g., Microsoft Kinect) and associated software development kits makes it practical to capture depth maps and human body skeleton joints in real time. This has facilitated a variety of visual recognition tasks, e.g., human activity analysis, object recognition and segmentation, hand gesture recognition, and etc. Li *et al.* [5] sampled a set of representative 3D cloud points from depth maps for human action recognition. But the direct usage of massive cloud points incurred a great amount of data that resulted in expensive computations in clustering training samples of all classes. A compact feature representation of EigenJoints was proposed in [16] based on body joints to recognize human actions. But it is not trivial to extract key joints from depth maps of a hand. Yang *et al.* [17] projected 3D depth maps onto three 2D orthogonal planes that were stacked as Depth Motion Maps (DMM). HOG was then computed from DMM as a global representation of human action. This method transferred 3D depth maps to 2D images which were further treated as gray images without explicitly considering the 3D shape information. Lai *et al.* [3] combined HOG features from both RGB and depth channels to improve object recognition in multiple views. They also simply dealt with 3D depth maps as 2D gray images.

Recently, hand gesture recognition based on depth maps has gained growing interests. Bergh and Van Gool [13] used a Time of Flight (ToF) camera combined with a RGB camera to successfully recognize four hand gestures by simply using small patches of hands. Ren *et al.* employed a template-matching based approach to recognize hand gestures through a histogram distance metric of Finger Earth Mover Distance (FEMD) through a near-convex estimation [7, 8]. However, their method only considered the outer contour of fingers but ignored the palm region that also provides important shape and structure information for complex hand gestures. Pugeault and Bowden [6] employed responses from Gabor filters on different scales and orientations as the feature to recognize the letters in American Sign Language (ASL). However, none of these methods explicitly use the ample 3D information conveyed by the depth maps. In this paper, we propose a discriminative descriptor which aims to explicitly capture the 3D surface information.

III. Computation of Histogram of 3D Facets (H3DF)

In this section, we describe the detailed computational procedures of our proposed 3D descriptor, i.e., Histogram of 3D Facets (H3DF). The pipeline is illustrated in Fig. 1. The input is a depth map that covers the region of interest. We first compute its dominant orientation to normalize the appearance variation induced by the in-plane rotation as demonstrated in Fig. 1(b). A 3D facet is then extracted from every discrete 3D cloud point (each one corresponds to a pixel on the depth image) and coded as shown in Fig. 1(c). A concentric spatial pooling is in the end applied to aggregate all the coded 3D facets into the H3DF descriptor represented by a feature vector in Fig. 1(d-e).

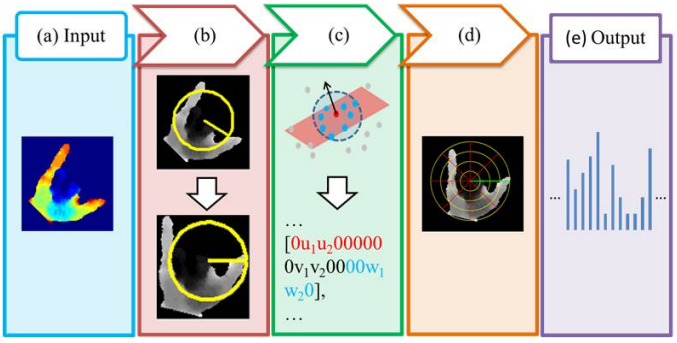

Figure 1. The pipeline of computing Histogram of 3D Facets (H3DF): (a) the input depth map of a region of interest, (b) orientation normalization, (c) facet coding, (d) concentric spatial pooling, and (e) the output of H3DF descriptor.

A. Gradient-based Orientation Normalization

One big challenge for hand gesture recognition is the large intra-class variation incurred by hand rotations. As illustrated in Fig. 2(a), the depth maps of the same gesture can significantly vary due to the in-plane rotation. To make the H3DF descriptor invariant to the rotation change, we perform a gradient-based orientation normalization for the input depth map. For each hand depth map as shown in Fig. 2(a), the dominant orientation θ of the image is computed based on the depth gradient.

In order to estimate the dominant orientation θ and achieve in-plane rotation invariance, we compute the dominant depth gradient orientation as the normalization employed by most local descriptors [4] in 2D images. The dominant orientation corresponds to the largest bin of the histogram of gradient orientations weighted by gradient magnitudes and smoothed by a Gaussian window. As suggested by [4], each local maximum bin with a value above 80% of the largest bin is retained as well. Therefore, each depth map might be associated with multiple orientations which are considered as multiple samples in our training set. As for a testing map with multiple dominant orientations, we only choose the key angle corresponding to the largest gradient angle bin. In such a way we can ensure the training set to cover as many samples as possible and for each testing image there is only one sample to avoid the decision ambiguity.

After computing the dominant orientation θ, we can rectify the 3D cloud points P to obtain the corrected 3D cloud points P' using the following equation:

$$P' = P \times R(-\theta) \quad (1)$$

where P and P' are two $K \times 3$ matrices of K 3D points; $R(-\theta)$ represents the in-plane rotation correction matrix.

Let D be the depth image before orientation correction, we define a pixel-to-point mapping $I(\cdot)$, which takes a 2D coordinates and outputs its 3D coordinates, i.e., $P = I(D)$ and its inverse mapping $D = I^{-1}(P)$. Combined with Eq. (1), we have the corrected depth image as:

$$D' = I^{-1}\big(I(D) \times R(-\theta)\big) \tag{2}$$

where D' is the depth image corrected by the dominant orientation θ.

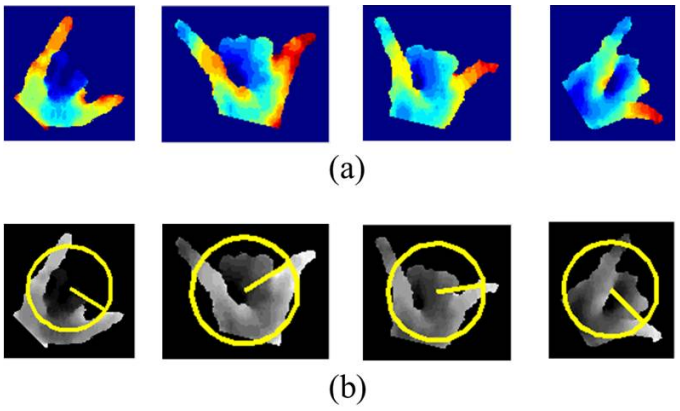

Figure 2. Orientation normalization of depth maps: (a) depth maps of the same hand gesture demonstrate considerable intra-class variations due to rotation; (b) the estimated dominant orientations denoted by the yellow lines.

B. *Defining 3D Facets*

To model an object in a 3D depth map, the 3D surface properties, such as bumps and grooves provide significant information, especially when the outer contour is not sufficient or discriminative to perform classification. As shown in Fig. 5(a-b), the 3D surfaces of the thumb constitute of an informative region to differentiate the two hand gestures that share similar visual patterns.

Since these surface changes from depth maps can be visualized as intensity variations in gray images, it is a natural way to directly use the existing state-of-the-art 2D image descriptors as feature representations. For example, Yang *et al.* [17] employed a global HOG to describe motion energy distributions from DMM that were generated from projections of 3D depth maps on to three orthogonal 2D planes. However, such methods only transfer the 3D depth patterns to 2D images rather than explicitly modeling the 3D information. In this paper, we propose a novel 3D descriptor based on 3D surface to directly represent the information conveyed by 3D depth maps.

A 3D facet is used to model the property of a 3D surface, as illustrated in Fig. 3. A 3D facet associated with a cloud

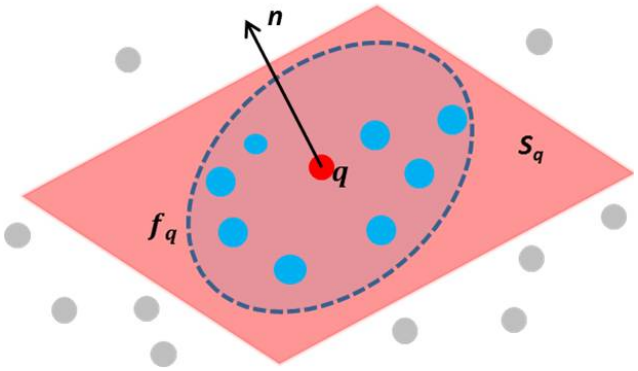

Figure 3. Computing the 3D facet $\boldsymbol{S_q}$ of a cloud point $\boldsymbol{q}$ according to its neighbor cloud point set $\boldsymbol{f_q}$. The pink plane is the fitted plane $\boldsymbol{S_q}$ and blue region indicates the locality constraint. The normal vector $\boldsymbol{n}$ is used as the representation of the 3D facet.

point q is determined by a local support plane defined by its surrounding cloud point set f_q:

$$f_q = \{q' \mid \|q' - q\|_p \le \sigma\} \tag{3}$$

where σ is a locality constraint to control the range of a local support region of the cloud point q. A plane S_q is then fitted according to f_q following one of the two metrics in Eq. (4). We then compute the normal vector n of S_q as the representation of the 3D facet.

Additionally, in Eq. (3) the parameters p together with locality σ jointly control the granularity of sampling surrounding the point of q. In this paper, we utilize two forms of them:

$$(p, \sigma) = \begin{cases} (1, 1) \\ (\infty, \alpha) \end{cases} \tag{4}$$

The first form uses norm $p = 1$ (i.e., Manhattan distance) and locality $\sigma = 1$ to select four adjacent neighbors, which leads to a normal estimation more sensitive to the local variance of a center point q. The second form uses norm $p = \infty$ (i.e., Chebyshev distance) and locality $\sigma = \alpha, \alpha \ge 1$ to define a local supporting region. The different settings of (p, σ) will be discussed in Section V-B. According to our experiments, the first setting with bi-line estimation (see Section V-B) performs better, which shows the H3DF descriptor favors the fine-granularity representation.

C. *Coding 3D Facets*

A 3D facet defined as a 3D plane which can be represented by $[n_x, n_y, n_z, d]^T$, where the first three coefficients are the normal vector $n = [n_x, n_y, n_z]^T$ of a 3D facet and the forth one d is the Euclidean distance from the plane to the origin coordinate. Although all four coefficients are needed to determine a local surface, in this paper we only concentrate on

the orientation rather than the absolute distance of a local surface. Thus we code a 3D facet only using its normal vector n. The procedures of coding each 3D facet are illustrated in Fig. 4.

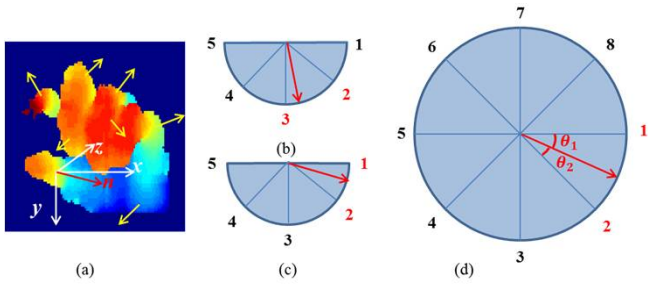

Figure 4. Code a 3D facet denoted by the red colored arrow in (a) by projecting the normal n onto three orthogonal planes in (b-d). As n_z is non-negative, the projected normal orientation ranges in xz (b) and yz (c) planes are $[0,\pi]$, but $[0,2\pi]$ in xy plane (d). The soft assignment is used to weight the two nearest orientation bins (d).

In coding each 3D facet, we first project its normal vector n (the arrow colored in red) in Fig. 4(a) onto three orthogonal planes, i.e., xy, xz, and yz planes. As shown in Eq. (2), a cloud point of 3D depth map can be mapped onto a 2D depth image. So each cloud point corresponds to a pixel in the 2D depth image, which means the pixels on the 2D depth image originate from those 3D points that locate in the front surface, i.e., the normal is pointing outward. Therefore we can safely assert that the n_z attribute of all the normal vectors are non-negative.

We then evenly deploy u (for xz and yz planes) and v (for xy plane) orientation bins on the three projected planes. The projected normals of each 3D facet vote to their two nearest bins as shown in Fig. 4(b-d). The benefit of this soft assignment strategy over a hard one (i.e., only vote to the nearest bin) is to avoid the boundary effect to make the histogram more stable, as well as to reduce the information loss in quantization. The weights of each normal vector assigned to its two nearest bins are given as:

$$w_i = \frac{\sin\theta_{1-i}}{\sin\theta_0 + \sin\theta_1} \qquad (5)$$

where θ_0 and θ_1 are the angular offsets between the projected normal and its two nearest bin centers indexed by c_0 and c_1. As shown in Fig. 4(d), the two nearest bin centers are $c_0 = 1$ and $c_1 = 2$ that are colored in red. The weights of a projected normal vector assigned to the two nearest orientation bins are inversely proportional to the sine of its offset angle. Therefore each coded 3D facet is represented as a feature vector with the length of $2u + v$, where there are six elements are non-zero, as illustrated in Fig. 1(c).

D. Pooling 3D Facets

After coding the 3D facets of a depth map, a concentric spatial pooling scheme is used to group these coded 3D facets from the entire hand gesture region to generate the H3DF descriptor, as illustrated in Fig. 5. The concentric spatial grid configuration is determined by the radius quantization number A and the angular quantization number B (e.g., $A = 4$ and $B = 8$ in Fig. 5). Therefore the overall dimension of the H3DF descriptor is $A \times B \times (2u + v)$.

It would be easy to distinguish Fig. 5(a) and Fig. 5(c) by using the contour matching as in [7]. However, this is not the case for Fig. 5(a) and Fig. 5(b) because their outer contours share a large portion of the same curves. Our concentric spatial pooling can alleviate this problem since we consider the interior region of the folded thumb which makes the corresponding spatial bins more discriminative.

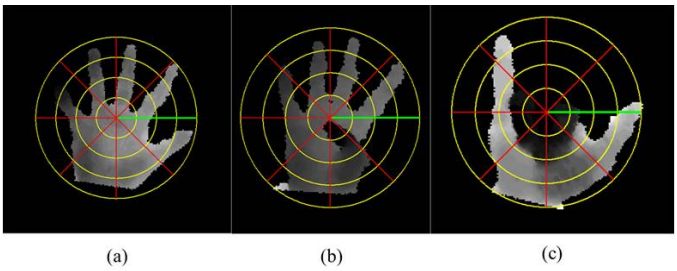

Figure 5. The concentric spatial pooling scheme to group coded 3D facets. The entire hand region is quantized as 4 bins in radius and 8 bins in angular. (a-c) correspond to samples of three different hand gestures.

IV. H3DF-BASED HAND GESTURE RECOGNITION

To evaluate the robustness of our proposed H3DF descriptor, we apply it to solve the hand gesture recognition problem. We first discuss how to segment the hand regions from depth images and then detail the implementations of H3DF in the context of hand gesture recognition.

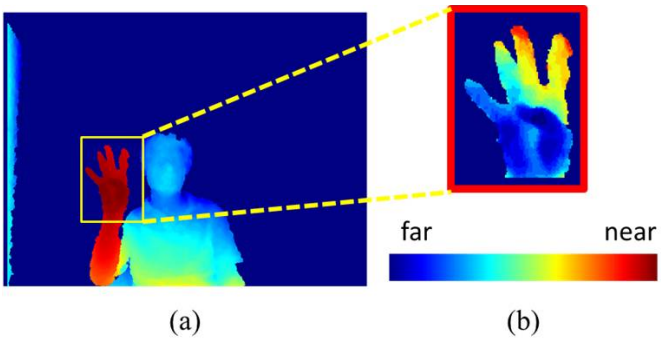

Figure 6. Segmentation of the hand region (b) from a 640×480 depth image in (a).

A. Hand Region Segmentation

The hand region extraction from a depth image can be done in several ways, such as to retrieve a hand joint using a pose estimator [9, 10] or to employ a hand tracker [12] using skin color. In the HCI setting as in [6, 7], a reasonable assumption is to ensure the hand is always the most front body

part facing to the camera. In our work, we inherit this assumption to pre-process 3D depth maps to segment hand regions based on the depth information. As a special case in [7], all the subjects have worn a black hand-wrist band to facilitate hand segmentation. We also utilize this clue in our experiment to obtain accurate hand regions.

As demonstrated in Fig. 6, we first find the nearest point as the one with the shortest camera-object distance from the depth map and record its value as d_{near}. The point cloud is then thresholded according to the range of $[d_{near}, d_{near} + t]$, where t is the distance threshold. In our experiments, we set $t = 100$ mm. The points falling in this range are considered as within a hand region.

B. Hand Gesture Recognition

In this section, we present the details on parameter settings of the proposed H3DF descriptor in recognizing hand gestures. Before computing the descriptor, we normalize the hand region into an image patch with the fixed size of 150×150. The effect of different patch sizes of hand regions is also studied in Section V-C.

We have two metrics in estimating the normal of each 3D facet employ as shown in Eq. (4). The estimated normal n is projected onto 3 orthogonal planes as n_{xy}, n_{xz}, and n_{yz}. We set $u = 5$ orientation bins for n_{xz} and n_{yz}, and $v = 8$ bins for n_{xy}, respectively. Thus each 3D facet is coded as a feature vector with the dimension of $5 + 5 + 8 = 18$.

In the concentric spatial pooling, we divide the normalized hand region into 32 spatial bins, i.e., 4 radius quantization bins and 8 angular quantization bins. The H3DF descriptor is therefore with the dimension of $32 \times 18 = 576$. The Support Vector Machines (SVM) with linear kernel is used as the classifier in our experiments.

V. Experimental Results

In this section, we evaluate the proposed H3DF descriptor on two public datasets of hand gesture recognition and extensively compare with the state-of-the-art methods on each dataset.

A. Datasets and Experimental Setup

1) Datasets

The samples of the NTU Hand Digits dataset [7] and the ASL Finger Spelling dataset [6] are illustrated in Fig. 7. Both datasets are captured by a Microsoft Kinect camera. The NTU Hand Digits dataset [7] contains 1,000 depth maps of 10 hand gestures (i.e., decimal digits from ***0*** to ***9***) from 10 subjects with 10 samples for each hand gesture. The ASL Finger Spelling dataset [6] captures 60,000 hand gestures from 5 subjects. It includes 24 English letters from ***a*** to ***z***, but with ***j*** and ***z*** discarded as these two letters in ASL are dynamic. Compared to the NTU Hand Digits dataset, the ASL Finger Spelling dataset is more diverse and much larger.

Unlike the NTU Hand Digits dataset release the whole depth maps, the ASL Finger Spelling dataset only provides the hand regions after segmentation, as we can see in Fig. 7. So for the ASL Finger Spelling dataset, we skip the pre-processing step of hand segmentation as described in Section IV-A.

2) Experimental Setup

We conduct two types of experiments: 1) subject-independent test which uses the *leave-one-out* strategy, i.e., for a dataset with N subjects, $N - 1$ subjects are used for training and the rest one for testing. This process is repeated for every subject and the average accuracy is reported; 2) subject-dependent test where all subjects are used in both training and

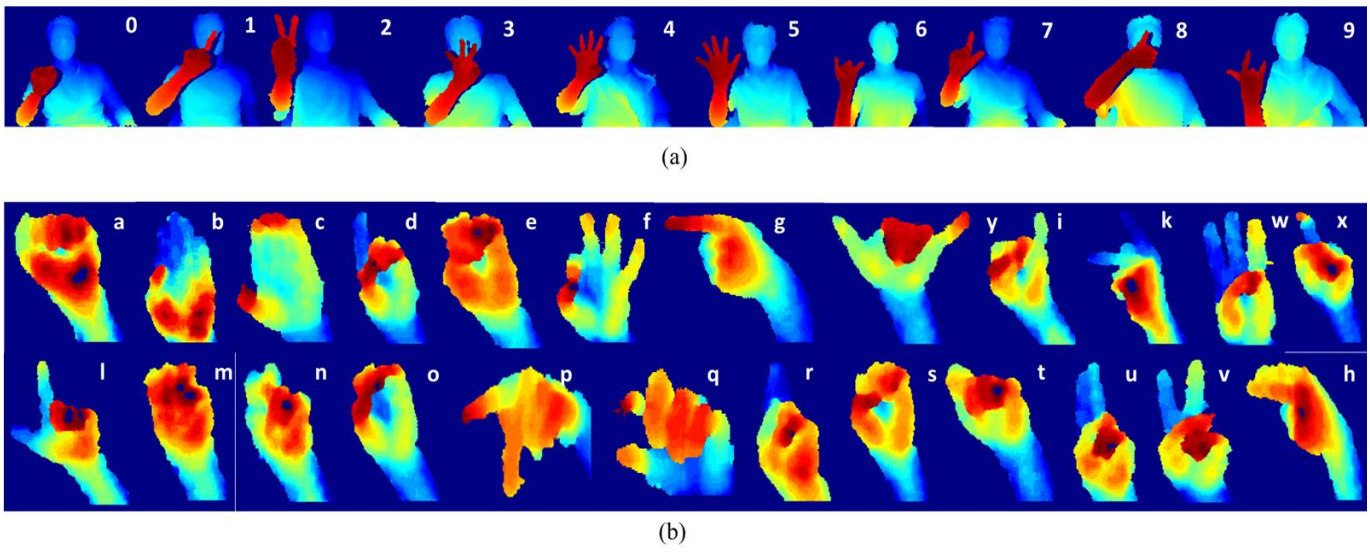

Figure 7. Samples (digits from ***0*** to ***9***) of depth maps from the NTU Hand Digits dataset [7] in (a). Samples (letters from ***a*** to ***z*** without ***j*** and ***z***) of depth maps from the ASL Finger Spelling dataset [6] in (b).

testing, where the whole dataset is evenly split for training and testing.

We start by discussing two important parameters in our method: 1) different approaches to estimate the normal of a 3D facet in Section V-B and 2) different sizes of the normalized hand regions in Section V-C. After determining the two parameters, we compare our proposed H3DF descriptor with the benchmarks and the traditional 2D image based HOG descriptor. In the implementation of HOG, we evenly separate the normalized patches into 8×8 non-overlapping cells and each cell has 8 orientation bins. The feature vectors of four different normalizations, i.e., *L1-norm*, *L2-norm*, *L1-sqrt*, and *L2-Hys* are concatenated as the final HOG representation as in [17]. The HOG descriptor is therefore with the dimension of $8 \times 8 \times 8 \times 4 = 2048$.

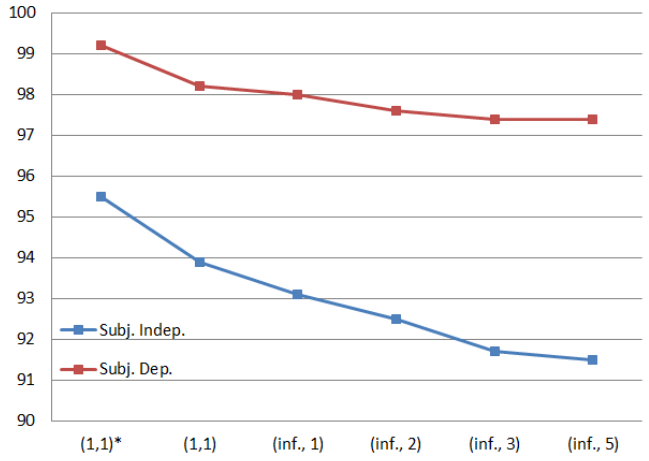

Figure 8. Accuracies (%) of hand gesture recognition on the NTU Hand Digits dataset using different normal computation methods and parameters. X-axis shows different combinations of (p, σ) from Eq. (4). The result denoted by * are obtained by the bi-line estimation and others are from the plane-fitting method.

B. *Normal Estimation of 3D Facets*

In this section, we evaluate different methods and parameters on computing the normal vector of a 3D facet. As in Eq. (4), we have two parameters p and σ to control the granularity of sampling neighboring points. In the first case, we use norm $p = 1$ and locality $\sigma = 1$ to sample the four adjacent neighbors. In the second case, we set norm $p = \infty$ and locality $\sigma = \alpha, \alpha \geq 1$ to sample a set of local support points. A plane-fitting method based on minimizing sum of distances of sampled points can be used to estimate the surface normal. Specifically, in the first sampling case, if the center pixel is removed, we can use the bi-line normal estimation. This method is suitable for a grid-organized 3D point set (e.g., a depth image). It takes the four adjacent neighbors of a pixel and computes the two diagonal 3D lines. Given a 3D facet whose center is $(i, j, d_{i,j})$, the bi-line estimation obtains the normal of this 3D facet as a vector which is orthogonal to the two diagonal 3D lines, i.e., one line connecting the points $(i-1, j, d_{i-1,j})$ and $(i+1, j, d_{i+1,j})$ and the other connecting the points $(i, j-1, d_{i,j-1})$ and $(i, j+1, d_{i,j+1})$. This approach is simple to implement and suitable for the depth image where the 3D points are organized as depth pixels. However, in the case of a 3D point cloud with non-uniform density, this approach will be not capable.

Compared to the bi-line estimation, the plane-fitting method is more general and can be used in the cases where point cloud density is non-uniform. However, it also takes the risk to lose details when the locality is not appropriately set. We conduct an experiment on the NTU Hand Digits dataset to compare the bi-line normal estimation method with the plane-fitting method under different values of p and σ. As can be observed from Fig. 8, the bi-line approach consistently performs better than the plane-fitting method. This observation suggests that the bi-line normal estimation is more suitable for the grid-organized 3D points. Based on this conclusion, we employ the bi-line estimation approach in the following experiments. As for the plane-fitting method with different sampling parameters, the accuracies of both subject-dependent and subject-independent tests are decreased as the increment of localities. The other observation from Fig. 8 is that subject-dependent tests significantly outperforms subject-independent tests and is more stable to the changes of locality.

C. *Resolution of Hand Regions*

In this section, we evaluate the impact of hand region resolution (i.e., normalized patch size) to the recognition results on the NTU Hand Digits dataset. We test different resolutions ranging from 150×150 to 25×25 as shown in Fig. 9. As we can see in this figure, the overall classification accuracies of both subject-independent and subject-dependent tests are over 90% under a variety of resolutions. This observation demonstrates the robustness of the proposed H3DF descriptor to different resolutions of normalized hand regions. In the following sections, we set the default normalized hand region size as 150×150.

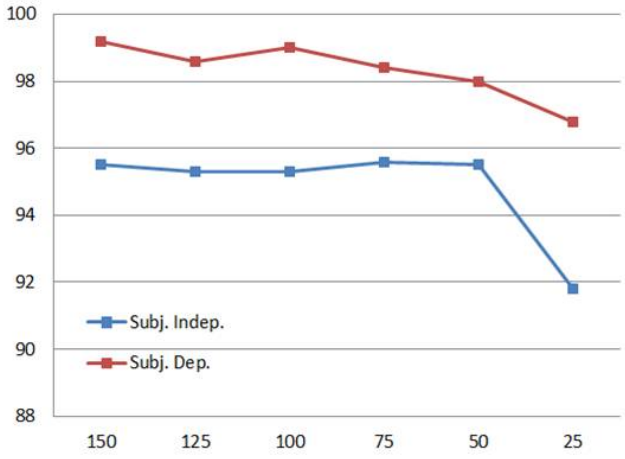

Figure 9. Accuracies (%) of hand gesture trecognition on the NTU Hand Digits dataset under different resolutions of normalized hand regions, from 150×150 to 25×25.

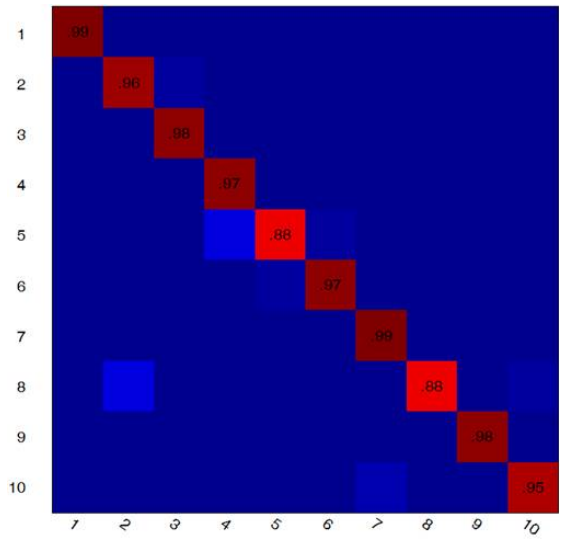

Figure 10. Confusion matrix of our method on the NTU Hand Digits dataset under the subject-independent test.

D. Comparisons to the State-of-the-arts

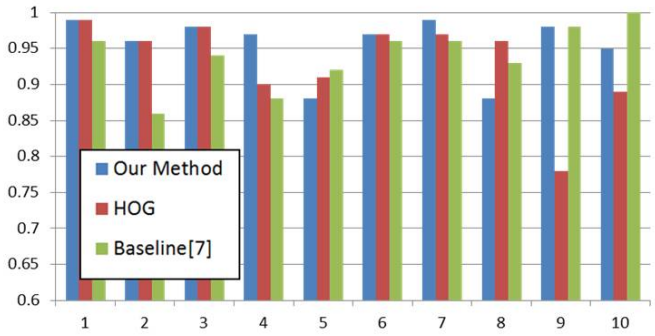

Figure 11. Comparisons of our proposed method with the baseline method (i.e., contour-matching) in [7] and HOG.

In this section, we first compare our method with the benchmark method in [7] and the traditional 2D HOG descriptor on the NTU Hand Digits dataset. The class-wise classification accuracies of subject-independent test are shown in Fig. 11 and the overall accuracies of different methods are shown in Table I. As we can see from these comparisons, our method considerably outperforms the benchmark method in [7] and the traditional 2D HOG descriptor under both subject-independent and subject-dependent tests. Compared to the benchmark contour-matching method in [7], our H3DF descriptor explicitly captures the 3D surface properties such as folded thumb in palm rather than the only outer contour information. In general our method performs 1.6% higher than contour-matching and 2.4% higher than HOG in subject-independent test. In subject-dependent test, our method achieves 99.2% classification accuracy. The confusion matrix of our method in subject-independent test is shown in Fig. 10. As we can see, except for class 5 and 8 (i.e., digits ***4*** and ***7***), the accuracies of other classes are all over 90%, which demonstrates the effectiveness of our proposed descriptor.

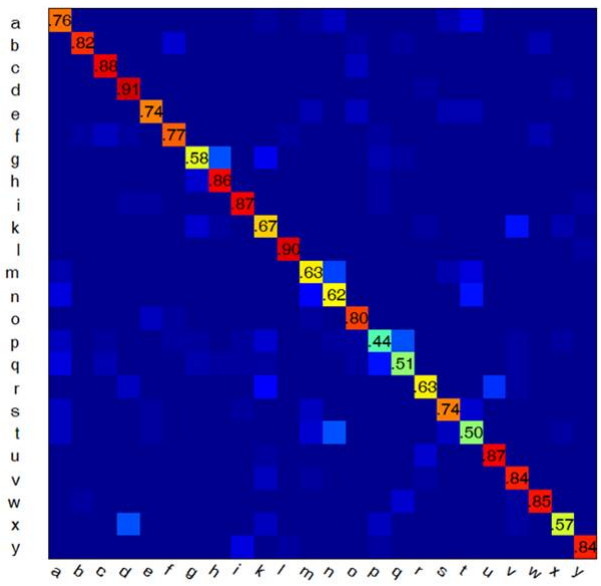

Figure 12. Confusion matrix of our method on the ASL Finger Spelling dataset under the subject-independent test.

TABLE I. COMPARISONS ON THE NTU HAND DIGIT DATASET

Method	Test	Contour-Matching [7]	HOG	Our Method
Mean Accuracy	Subj. Indep.	93.9%	93.1%	**95.5%**
	Subj. Dep.	N/A	96.4%	**99.2%**

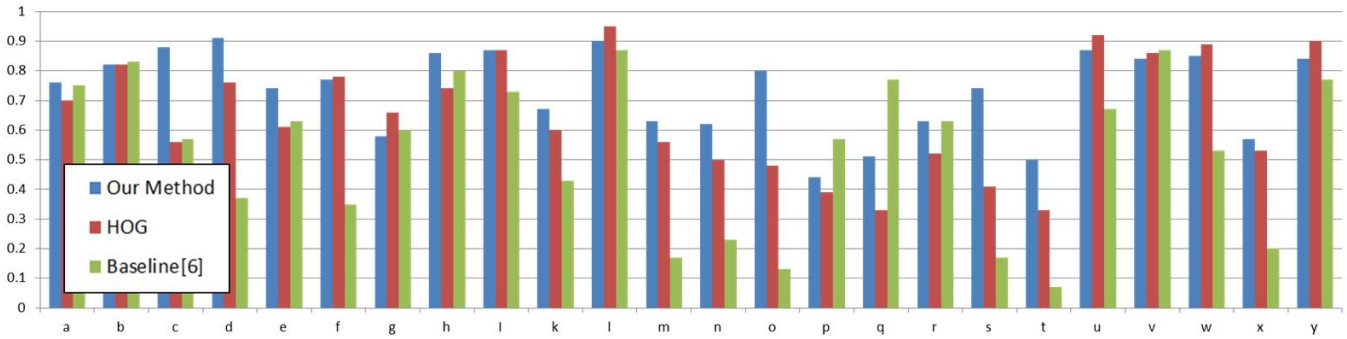

Figure 13. Comparison of our proposed method with the baseline method in [6] and HOG.

TABLE II . COMPARISON ON THE ASL FINGER SPELLING DATASET

Method	Test	Pugeault and Bowden [6]	HOG	Our Method
Mean Accuracy	Subj. Indep.	49.0%	65.4%	**73.3%**
	Subj. Dep.	N/A	96.0%	**98.9%**

We follow the same experiment setting as above on the ASL Finger Spelling dataset. The confusion matrix and class-wise accuracies are shown in Fig. 12 and Fig. 13, both of which are based on subject-independent test. The comparisons with the state-of-the-art methods on both subject-dependent and subject-independent tests are shown in Table II. As shown in Fig. 13, our method is more stable and discriminative than the benchmark, e.g., the lowest accuracy of the method from Pugeault and Bowden [6] is 7% for the letter ***t*** and the lowest accuracy of our method is 44% for the letter ***p***, which indicates that our proposed H3DF descriptor based on 3D facets is more representative. Our method reaches 73.3% average accuracy under subject-independent test, which has significantly outperformed the method of Pugeault and Bowden by 24.3%. This is probably because of the informative 3D surface properties and the orientation correction before coding facet used in our descriptor. Compared to the traditional 2D HOG descriptor, which is also with orientation correction, our method still achieves 7.9% higher accuracy. This suggests that the superiority of explicitly using 3D information in describing 3D depth images over applying the traditional 2D image descriptor. On the other hand, as shown in the confusion matrix in Fig. 12, some letters are still relative hard to classify, such as ***m*** and ***n*** as well as ***p*** and ***q***, whose hand gestures share quite similar patterns (see Fig. 7 for hand gesture samples).

VI. CONCLUSION

In this paper, we have proposed a novel discriminative 3D descriptor that is able to explicitly capture and model ample and discriminative surface information from 3D depth maps. We have applied our proposed descriptor to solve the hand gesture recognition problem. We observe the orientation normalization, robust coding and concentric spatial pooling are critical to handle the large intra-class variances incurred by rotation, scaling, and view point change. We have evaluated the effectiveness of our proposed descriptor on two public hand gesture recognition datasets. The experimental results demonstrate that our proposed approach significantly outperforms the state-of-the-arts. Our future work includes extending this descriptor to the temporal domain to handle dynamic hand gesture recognition from depth videos.

ACKNOLEDGEMENT

This work was supported in part by NSF IIS-0957016 and Microsoft Research.

REFERENCES

[1] N. Dalal, and B. Triggs, "Histogram of Orientated Gradients for Human Detection", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 886-893, (2005)

[2] R. Francois and G. Medioni. "Adaptive color background model-ing for real-time segmentation of video streams". In Int. Conference on Imaging Science, Systems, and Technology, Las Vegas, NA, 1999.

[3] K. Lai, D. Bo, X. Ren, and D. Fox, "A Large-Scale Hierachical Multi-View RGB-D Object Dataset", International Conference on Robotics and Automation (ICRA), 2011.

[4] I. Leptev, "On Space-Time Interest Points", International Journal of Computer Vision, Vol. 64, No. 2, pp. 107-123, Springer, (2005)

[5] W. Li, Z. Zhang and Z. Liu, "Action recognition based on a bag of 3D points", IEEE Workshop on CVPR for Human Communicative Behavior Analysis.

[6] N. Pugeault, and R. Bowden, "Spelling it out: Real-time ASL fingerspelling recognition," IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 1114–1119, (2011).

[7] Z. Ren, J. Yuan and Z. Zhang, "Robust gesture recognition based on finger-earth mover's distance with a commodity depth camera," the 19th ACM International Conference on Multimedia (ACM'MM), pp. 1093-1096, (2011).

[8] Z. Ren, J. Yuan, C. Li and W. Liu, "Minimum near-convex decomposition for robust shape representation," IEEE International Conference on Computer Vision (ICCV), pp. 303-310, (2011)

[9] L. A. Schwarz, A. Mkhitaryan, D. Mateus and N. Navab, "Estimating human 3D pose from Time-of-Flight images on geodesic distance and optical flow", IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 700-706, (2011)

[10] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, "Real-time pose recognition in parts from single depth images", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, page 7, (2011)

[11] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images", European Conference on Computer Vision(ECCV), 2012

[12] H. Trinh, Q. Fan, P. Gabbur and S. Pankanti, "Hand tracking by binary quadratic programming and its application to retail acivity recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1902-1909, (2012)

[13] M. Van den Bergh and L. Van Gool, "Combining RGB and ToF cameras for real-time 3D hand gesture interaction", IEEE Workshop on Applications of Computer Vision (WACV), pp. 66-72,(2011)

[14] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Minning actionlet ensemble for action recognition with depth cameras", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290-1297, (2012)

[15] Y. Wu and T. Huang, "Vision-based gesture recognition: A review", Gesture-based Communication in Human Computer Interaction, pp. 103-115, Springer (1999)

[16] X. Yang and Y. Tian, "EigenJoints-based Action Recognition Using Naïve-Bayes-Nearest-Neighbor", IEEE CVPR Workshop on Human Activity Understanding from 3D Data, pp. 14-19, (2012).

[17] X. Yang, C. Zhang, and Y. Tian, "Recognizing Actions Using Depth Motion Maps-based Histograms of Oriented Gradients", ACM Multimedia, (2012) .