Computer Vision and Image Understanding 139 (2015) 29–39

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Histogram of 3D Facets: A depth descriptor for human action and hand gesture recognition☆

Chenyang Zhang, Yingli Tian*

The City College of New York, 160 Convent Avenue, New York, NY, USA

ARTICLE INFO

Article history:
Received 18 February 2014
Accepted 31 May 2015
Available online 6 June 2015

Keywords:
Computer vision
RGBD image processing
Gesture and action recognition
Time sequence representation
Histogram of 3D Facets

ABSTRACT

The recent successful commercialization of depth sensors has made it possible to effectively capture depth images in real time, and thus creates a new modality for many computer vision tasks including hand gesture recognition and activity analysis. Most existing depth descriptors simply encode depth information as intensities while ignoring the richer 3D shape information. In this paper, we propose a novel and effective descriptor, the Histogram of 3D Facets (H3DF), to explicitly encode the 3D shape information from depth maps. A 3D Facet associated with a 3D cloud point characterizes the 3D local support surface. By robust coding and circular pooling 3D Facets from a depth map, the proposed H3DF descriptor can effectively represent both 3D shapes and structures of various depth maps. To address the recognition problems of dynamic actions and gestures, we further extend the proposed H3DF by combining it with an N-gram model and dynamic programming. The proposed descriptor is extensively evaluated on two public 3D static hand gesture datasets, one dynamic hand gesture dataset, and one popular 3D action recognition dataset. The recognition results outperform or are comparable with state-of-the-art performances.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

3D shape representation is a significant component of object categorization and action recognition. Compared to 2D image-based appearance representation, 3D depth-map-based representation is more robust to viewpoint and pose changes and holds great promise for modeling physical-related attributes such as positions, poses, shapes, and scene contexts. Over the last few years, the successful commercialization of depth sensors and corresponding development toolkits have made 3D shape information more accessible for computer vision applications [4,8,9,11]. Compared to traditional RGB cameras, RGBD cameras provide more information about object sizes, shapes, poses, and positions and capture strong boundary clues and spatial layouts, especially in environments with cluttered backgrounds and large illumination changes. The depth sensors have motivated recent research efforts to explore object and human gesture recognition by using 3D information [5,6,10]. However, these methods for 3D depth-map-based hand gesture recognition have only applied the existing 2D feature descriptors to the depth images, such as Gabor filter bank [5] or contour matching [6].

In order to directly and effectively capture and encode 3D shape information from depth maps, we propose a novel characteristic descriptor named Histogram of 3D Facets (H3DF). In 3D depth maps, a 3D cloud point together with its surrounding points is defined as a "3D Facet", which includes the informative local surface pattern surrounding the cloud point. Each facet is modeled by a small plane. Then a spatial centric pooling strategy is applied to organize the collection of facet planes based on their normal orientations to describe the current region of interest (ROI), which forms the final H3DF descriptor. In applications of hand gesture recognition and human activity recognition, a region of interest may be an image patch describing a hand gesture or a body part. To integrate the static depth map descriptor with temporal information in depth video sequences, we propose two approaches: (1) we approximate the depth video sequence as an ordered collection of a number of representative frames. The optimal collection of representative frames is selected by minimizing a sequential loss function defined by using only selected frames to represent the whole video using Dynamic Programming (DP). (2) We capture and represent the local temporal structure patterns via N-gram modeling. The N-gram model can be viewed as a collection of "visual word transitions," which is insensitive to different temporal structures caused by different execution rates.

Compared to existing depth-map descriptors, our proposed H3DF depth-map descriptor has three advantages: (1) it explicitly captures the 3D shape patterns conveyed by depth maps. (2) It applies a compact representation to describe a depth map compared to other 2D

☆ This paper has been recommended for acceptance by Tinne Tuytelaars.

* Corresponding author.
E-mail addresses: czhang10@citymail.cuny.edu (C. Zhang), ytian@ccny.cuny.edu (Y. Tian).

http://dx.doi.org/10.1016/j.cviu.2015.05.010
1077-3142/© 2015 Elsevier Inc. All rights reserved.

feature descriptors, *e.g.* Histogram of Orientated Gradients (HOG) [1]. 3) Compared to existing surface normal-based descriptors such as HONV [20] and HON4D [18], H3DF utilizes a circular grid for spatial pooling to encode more information such as shape and local depth patterns, which implicitly manifests the importance of the center part and makes the descriptor more robust to external contour deformations. Compared with the earlier conference version of this paper [21] which only demonstrated the effectiveness H3DF for static hand gesture recognition, we further extend the proposed H3DF descriptor to handle temporal sequences for recognizing dynamic hand gestures and human activities. By utilizing dynamic programming-based temporal segmentation and N-gram-based representation [22], we generate more robust representations for depth video sequences by combining H3DF with temporal structure information. We evaluate the proposed descriptor on two public datasets of hand gesture recognition: the NTU hand digits dataset [6] and the ASL finger spelling dataset [5]; one dynamic hand gesture data set: the MSR 3D gesture dataset [16], and one popular action recognition dataset: the MSRAction3D [4]. The recognition results on all the tasks demonstrate that our approach outperforms or is comparable to state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 reviews the related work on depth map-based human action and hand gesture recognition. Section 3 describes the procedures of computing the H3DF descriptor and how to apply H3DF to static image-based hand gesture recognition. Section 4 presents the modeling and implementation details of dynamic programming-based temporal segmentation and N-gram-based temporal pattern exploration with the proposed H3DF. Section 5 provides the implementation of H3DF for sparse representation-based hand gesture and human action recognition. Experimental results and discussions are presented in Section 6. Finally, we conclude the remarks of this paper in Section 7.

2. Related work

Due to the explicit 3D structure representation of objects and human body parts from depth maps, many research efforts have been made in depth map-based hand gesture recognition and human activity recognition. This is especially true since the release of low-cost 3D sensors (*e.g.* Microsoft Kinect) and associated software development kits (*e.g.* Microsoft Kinect SDK) and the success in real-time body joints position estimation [9]. Early research has focused on applying existing 2D image representations on 3D depth data, such as bag-of-3D-points by Li et al. [4], which samples representative 3D cloud points from depth maps for action recognition; Histogram-of-3D-gradient-orientations (derived from Histogram-orientation-gradients (HOG) [1]); and extending 2D interest point detectors to depth maps [33]. In [13], projections of 3D depth maps onto three 2D orthogonal planes are stacked as three depth motion maps, and then HOG descriptors were computed from the depth motion maps as the global representations of human actions. This method transfers a sequence of 3D depth maps to a 2D image that is further treated as a gray image without explicitly encoding 3D shape information. Recently, researchers have paid more attention to intrinsic features from depth images. Surface-normal, as a natural and explicit description of a local 3D volume, has been used in depth image descriptors [18,20] and graphics [31], and has demonstrated its potentials in activity recognition [18] and object recognition [20]. Our work builds upon this technique.

Hand gestures serve as a significant component of human computer interaction (HCI) because they convey information that covers multiple function categories in communication [12]. As a first step of hand gesture recognition, hand detection and tracking is either done by skin color or shape-based segmentation, which can be inferred from the given RGB images [2]; or directly resolved by leveraging the depth information [24,34,35]. Based on detection and tracking of hand regions, both dynamic and semantic features are extracted and utilized for gesture recognition [12]. Because of its intrinsic vulnerability to background clutter and illumination variation, RGB-based hand gesture recognition usually requires a clean background, which limits its application. Bergh and Gool [10] successfully used a Time of Flight (ToF) camera combined with RGB camera to recognize four simple hand gestures. In [6], Ren et al. employed a template matching-based approach and recognized hand gestures using a histogram distance metric of finger-Earth mover's distance with near-convex estimation [7]. Pugeault and Bowden [5] employed Gabor filter features at different scales and orientations to recognize characters in American Sign Language (ASL). However, none of these methods makes use of the rich geometry information conveyed by the depth maps.

Our proposed method can be categorized as a crafted feature leveraging surface normals. Extensive experiments demonstrate that our method, with robust coding and two-dimensional circular pooling, can capture the ample 3D surface geometry information and is discriminative in the representation of static or dynamic hand gesture recognition and action recognition. Compared to previous work, our descriptor is as discriminative as the learned features but has the advantage of keeping a very compact (low-dimensional) size.

3. Histogram of 3D Facet (H3DF)

The computation procedures of the new 3D feature descriptor, Histogram of 3D Facets (H3DF) is illustrated in Fig. 1. Given a depth image, we first delimit its in-plane rotation by normalizing the dominant orientation of the depth image. Then for each 3D point of the image associated with its neighbor points (a 3D Facet), we compute its normal vector and then encode the normal vector to represent the current 3D Facet. The encoding is processed by projecting the normal vector onto three orthogonal planes (*i.e. xy*, *yz*, *xz*) and quantizing each projection. To generate a compact description of the whole image, we design a concentric spatial pooling to organize all encoded 3D Facets into a compact descriptor vector to capture the spatial layout and local structure of the depth image.

3.1. Gradient-based orientation normalization

One challenge of hand gesture recognition is the large appearance variations when hand rotates. To make H3DF rotation invariant, we conduct gradient-based orientation normalization for an input depth image or patch. For each depth patch as shown in Fig. 2(a), the dominant orientation (denoted as θ) of the hand depth patch is first computed based on its shape and gradients. We then rectify the 3D cloud points set (denoted as P) to obtain orientation-corrected 3D cloud point set P' of its salient orientation with the following equation:

$$P' = PR(\theta)^T, \tag{1}$$

where P and P' are $K \times 3$ matrices as the collection of K 3D points; $R(\theta)^T = R(-\theta)$ represents an in-plane correction rotation matrix.

Let D be the depth image patch before orientation correction, we define a pixel-to-point mapping $I(\cdot)$, as it takes a 2D coordinate as input and outputs a 3D coordinate, where $P = I(D)$, and its inverse mapping $I^{-1}(\cdot)$, *vice versa*, where $D = I^{-1}(P)$. Together with Eq. (1), we have the corrected patch as:

$$D' = I^{-1}(I(D)R(\theta)^T), \tag{2}$$

which provides the orientation correction of a depth image patch of dominant orientation θ. As illustrated in Fig. 2(a), depth images (D) of the same hand gesture may significantly vary due to rotation. Dominant orientations (see Fig. 2(b)) can be detected based on *Gradient Consensus* to rectify the images to more similar corrected images (D') as shown in Fig. 2(c).

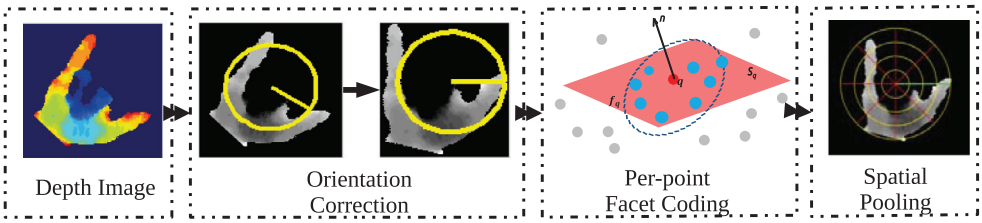

Fig. 1. Pipeline of the proposed Histogram of 3D-Facets (H3DF) modeling for single depth frame. H3DF utilizes surface normals and centric spatial pooling together to encode a depth frame.

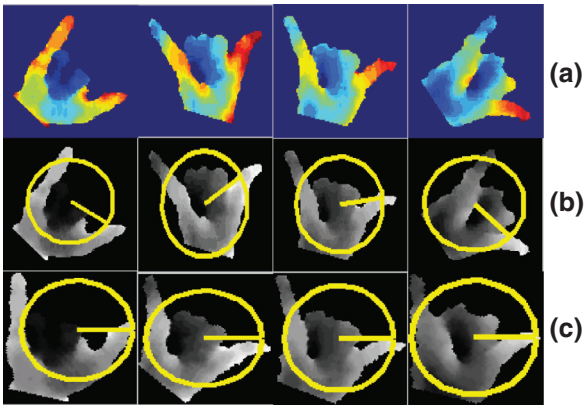

Fig. 2. Examples of gradient-based orientation correction results of hand gestures. (a) Significant appearance variations of the same hand gesture when hand rotates. (b) Estimated dominant orientations are illustrated as yellow orientated circles. (c) Orientation normalized depth patches with removed appearance variations.For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

In order to estimate the dominant orientation θ and achieve in-plane rotation invariance, we compute the dominant depth gradient orientation as the normalization used by most local image descriptors [3]. A dominant orientation corresponds to the largest bin of the histogram of gradient angles, weighted by gradient magnitudes and smoothed by a Gaussian filter. As suggested in [3], each local maximum bin with a value above 80% of the largest bin is retained as well. Thus, each depth image might be associated with multiple orientations which are considered as multiple samples in our training set. As for a testing image with multiple dominant orientations, to avoid decision ambiguity, we choose only the key angle corresponding to the largest gradient angle bin.

3.2. Defining a 3D Facet

To model a 3D object in a depth image, in addition to the outer contour, 3D surface properties and different shape patterns such as bumps and grooves provide rich and discriminative information. In some cases, the outer contour cannot be defined, and features inside the contour convey relative plentiful details. Unlike previous research by applying existing 2D visual descriptors to obtain a compact representation we propose a novel 3D surface feature descriptor which can directly represent the rich information conveyed by 3D object surfaces.

As shown in Fig. 3, we propose 3D Facets to model the shape details of a 3D surface. A 3D Facet associated with a cloud point q is determined by a local support surface defined by its surrounding cloud point set f_p:

$$f_p = \{q' | q, q' \in Q, \|q' - q\|_p \leq \sigma\}, \tag{3}$$

where σ is a threshold to control the size of the support region around the cloud point q, applying a locality constraint that only neighbor points can contribute to f_q. We then fit a plane S_q according to f_q such that the sum of distances between each point in f_q and the fitted plane is minimized. The normal vector $\mathbf{n}$ of a fitted plane S_q is then calculated as the representation of a 3D Facet. The normal fitting can be computed as a least-squares solution to the stack of N equations of the form $\mathbf{n}^T p_i = 1$ where N is the number of cloud points p_i in the 3D Facet f_q. When we set N equal to 4, there is an analytical solution for the normal, which will be discussed in later sections.

Additionally, in Eq. (3), the parameters p together with threshold σ can jointly control the granularity of sampling surrounding points of q. In this work, we utilize two particular forms of them:

- $(p, \sigma) = (1, 1)$: Bi-linear (analytical solution) or 4-neighbor (least-squares solution)
- $(p, \sigma) = (\text{inf}, \text{a})$: a $\times$ a patch. (least-squares solution)

In the first case, the difference between *"Bi-linear"* and *"4-neighbor"* is that the former one excludes the center point (*i.e.*, q) where the latter one does not. In the second case, the Chebyshev (l_∞) distance is used to define the supporting area as a patch in the corresponding 2D depth map. The difference of different selections of (p, σ) will be discussed in Section 3.4.

3.3. 3D Facet coding

A 3D Facet can be represented by using $[n_x, n_y, n_z, d]^T$, where the first three coefficients are the normal vector $\mathbf{n} = [n_x, n_y, n_z]^T$ of the facet plane and the forth attribute d is the Euclidean distance from origin point to the plane. Although all four coefficients are used to fix a plane, in this paper we focus on the orientation rather than the distance of the plane, thus d is not coded and is highly dependent on the distance of an object to a camera. Therefore, a 3D Facet is only coded by its normal vector. The procedure of coding is angular-based using the orientation of each 3D Facet as illustrated in Fig. 3 (b–d).

First, the normal vector (the vector $\mathbf{n}$ colored in red in Fig. 3(b)) is projected into three orthogonal planes, *i.e.*, *xy*, *yz*, and *xy* planes as shown in Fig. 3(c and d). Since the 3D point set is mapped from a 2D depth image, every cloud point corresponds to a pixel in the 2D depth image. Consequently, all the 3D points actually locate in front of the surface they formed (namely, the normal is pointing outward). So we can safely assert that all the normal vectors are pointing outward, in other words, their z-attributes are always non-negative.

Then, we evenly deploy m (for *xz* and *yz* planes) and n (for *xy* plane) bin centers on different planes. Each normal projection votes to two nearest bin centers (indices are colored red in Fig. 3(c and d)). The benefit of this local soft assignment strategy over a hard assignment (in which each normal projection only votes to the nearest bin center) is that the loss of information can be significantly reduced and thus the coded feature vector is much more informative. The weights of each normal vector assigned to the two nearest bin centers are given as:

$$w_i = \frac{\sin\theta_j}{\Sigma_k \sin\theta_k}, i, j, k \in I_2, j \neq i, \tag{4}$$

where θ_i is the angular offset between the normal projection and the bin center indexed i. I_2 is the bin center indices that composed by

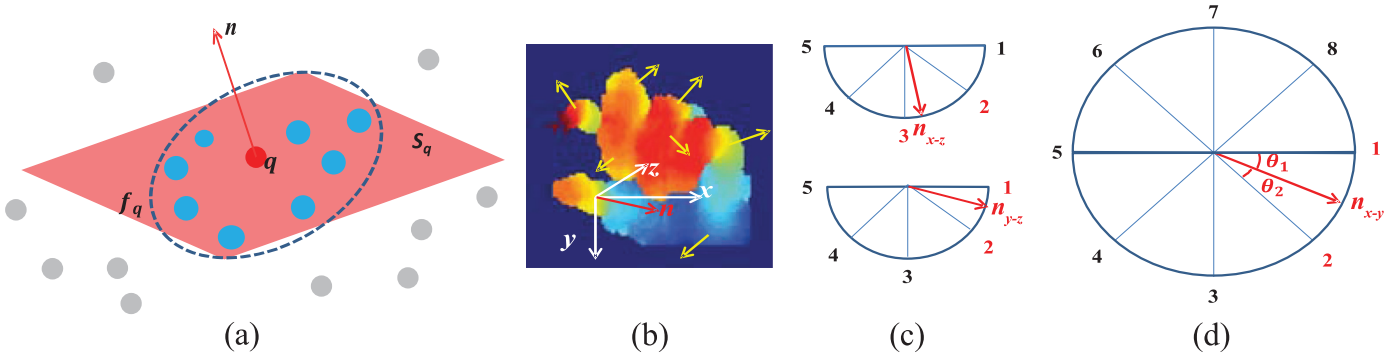

Fig. 3. (a) Computing the 3D Facet S_q of a cloud point q according to its neighbor cloud point set f_q. The pink plane is the fitted plane S_q and the blue region indicates the local constraint. The normal vector n is used as the representation of the 3D Facet. (b) The normal vector n is encoded by projecting onto three orthogonal planes in (c) (xz and yz) and (d)(xy). As n_z is non-negative, the projected normal orientation ranges in xz and yz (c) are both $[0, \pi]$, but $[0, 2\pi]$ in the $x-y$ plane (d). A soft assignment strategy is employed to weight the two nearest orientation bins as shown in (d).For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

two nearest bin centers (c_1, c_2). Therefore the encoded 3D Facet is represented as a vector of length $2m + n$, in which there are up to six non-zero elements.

3.4. 3D Facet pooling to generate H3DF

Once all encoded 3D Facets are computed, we design a concentric spatial pooling scheme to group these 3D Facets from the image patch into a compact H3DF descriptor as shown in Fig. 4. Another perspective of the proposed spatial pooling is to capture the information of facets arrangement coordinated in the center. In this phase, we address the boundary information as in [6].

For a spatial grid centered at (p_x, p_y), the bin index (a, b) of a pixel in the depth image $D(i, j)$ can be determined by the spatial distance $\|i - p_x, j - p_y\|_2$ and the angle $\arctan((j - p_y)/(i - p_x))$, where $a \in [1, A]$ and $b \in [1, B]$ and A, B are the spatial bin dimensions. Therefore, the dimension of the final H3DF descriptor of the image patch is $A \times B \times (2m + n)$.

The proposed pooling strategy is inspired by the invariant property of shape context in modeling rotations and scales of exterior contours [32]. However, our usage of circular bins is beyond modeling exterior contours. Circular bins intrinsically put more weight in modeling interior parts of a depth object and thus it enables H3DF to capture more local depth patterns such as holes and bumps. Besides, bigger outer bins are capable to capture the shape information and robust to subtle shape variants. The usage of circular bins is a key difference between other surface normal-based descriptors [18,20] by discriminating information intrinsically from interior parts and exterior parts of a depth object.

4. Representing temporal sequence using H3DF

Traditionally, Temporal Pyramid (TP) is used to extend an image representation model (*e.g.*, bag of words) to represent a video sequence. However, TP is sensitive to time, speed, and state-composition variances within each video sequence. The phenomenon can be intuitively illustrated in Fig. 6. In particular, if two sequences share very similar contents but are not well aligned, they will be far from each other in the metric space generated by temporal pyramid matching.

To overcome this issue and adapt H3DF to accommodate varied temporal structures, we propose two methods: 1) Dynamic Programming-based (DP) temporal segmentation to dynamically partition a video into cohesive sub-sequences and 2) N-gram bag-of-phrase-based representation [Ngram].

Algorithm 1: DP temporal segmentation, $(c, \hat{S}) = \text{DP\_TS}(\mathbf{V}, K)$.

Input: video sequence $\mathbf{V}$, number of partitions K
Output: optimal partitions $\hat{S}$, cost c

```
if K==1 then
    Ŝ = ∅;
    c = Σ_{v∈V} ‖v − μ(V)‖_2^2;
    return Ŝ, c
end
c = ∞;
for i ∈ {1, ..., |V| − 1} do
    (c_1, S_1) = DP_TS(V(1 : i), K − 1);
    (c_2, S_2) = DP_TS(V(i + 1 : end), 1);
    if c_1 + c_2 < c then
        c = c_1 + c_2;
        Ŝ = [S_1, i, S_j];
    end
end
return c, Ŝ
```

4.1. Dynamic programming-based representation

The pipeline of DP-based representation is illustrated in Fig. 5. Let $\mathbf{V} = \{vec(I_1), vec(I_2), \ldots, vec(I_t)\}$ be a sequential set of t frames with each frame I_i of dimension $M \times N$, *i.e.*, $vec(I_i) \in \mathbb{R}^d, I_i \in \mathbb{R}^{M \times N}$. A K-segmentation S of the video is a partition S of the frames into K non-overlapping contiguous segments, *i.e.*, $S = (s_1, \ldots, s_k), s.t. \bigcap_{i=1}^{k} s_i = \emptyset, \bigcup_{i=1}^{k} s_i = V$. The optimal segment $\hat{S}$ is defined as:

$$\hat{S} = \text{argmin}_S \left(\sum_{s \in S} \sum_{t \in s} \|t - \mu_s\|_2^2 \right), \tag{5}$$

where μ_s is the mean of samples in each segment s.

This optimization problem is well-known to be efficiently solved by dynamic programming [15]. We implement the DP-based temporal segmentation in a recursive manner, as is detailed in Algorithm 1. The algorithm can be practically accelerated by using a cache to store intermediate solutions to sub-problems. The source code is available at our research website [1].

This description is robust to dynamic warping of a video sequence. For example, as shown in Fig. 6, since the initial hand gesture occupies 50% of total frames, the evenly TP-based method generally assigns a large weight to the initial pose. However, because the overall

[1] http://media-lab.engr.ccny.cuny.edu/~zcy/#Code4Fun

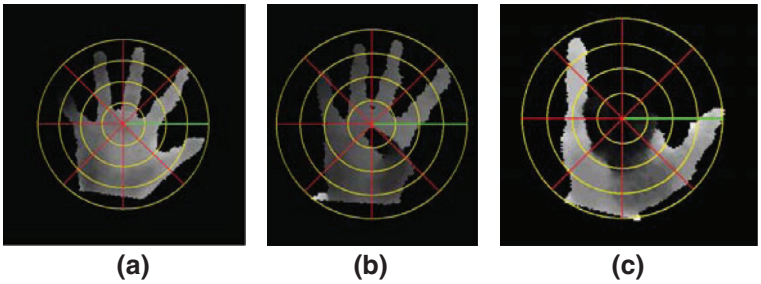

Fig. 4. Illustration of the first phase spatial pooling for creating H3DF descriptors. The region of interest of the depth image or patch is divided into 4×8 bins which are determined by both radial and angular offsets. (a), (b), and (c) are from three different hand gestures. Red line segments illustrate angular bin boundaries, yellow circles illustrate off-center radial distance bin boundaries, and green line segment shows the normalized patch orientation.For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.

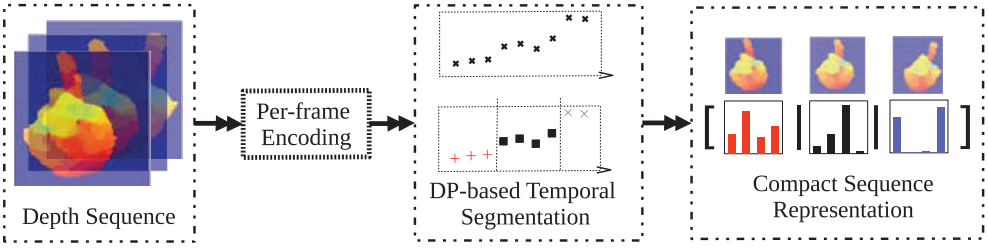

Fig. 5. Pipeline of the proposed DP-based video sequence representation example. DP-based temporal segmentation is used to partition each depth video sequence into a fixed number of segments, while the sum of within-segment intra-variances is minimized. A compact video representation is the concatenation of pooled H3DF codes of all segments.

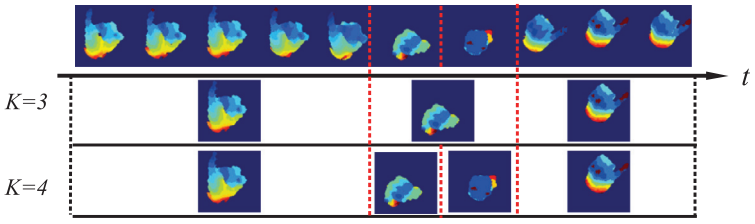

Fig. 6. For a depth video sequence, dynamic programming-based temporal segmentation computes an optimal segmentation in terms of minimum representative error. We illustrate the idea with a dynamic American Sign Language (ASL) gesture for character "*j*" and two segmentations with number of segments, *K*, set to 3 (middle row) and 4 (bottom row) respectively. In particular, the DP-based segmentation is a better representation than the temporal pyramid since it can overcome the uneven gesture distribution, *e.g.* in the example case, initial pose occupies almost 50% of total frames.

representative error is minimized (Eq. 5) in our proposed DP-based temporal segmentation, only the most representative frames are selected, while dynamically tuning the partition boundaries and thus the selected representatives are more informative and generic.

4.2. N-gram bag-of-phrase based representation

In the N-gram bag-of-phrase model, instead of building a global representation of the whole temporal structure of a time sequence T, we attempt to discover local patterns of the time sequence. Using the same notation as in the previous section for DP, the time sequence $T = (t_1, \ldots, t_n)$ is characterized by its local N-grams, *i.e.,* the tuples constructed by every consecutive N signals. For example, if $N = 2$, the bi-grams of the time sequence are $\{(t_1, t_2), (t_2, t_3), \ldots, (t_{n-1}, t_n)\}$.

The N-gram model has been successfully used in speech recognition and natural language processing [22]. In computer vision, the N-gram model is used to generate bag-of-phrases model [23] and is effective in image retrieval because it conveys more temporal information than the traditional bag-of-words model. In our work, as shown in Fig. 7, we propose to use the bag-of-phrases model to represent video sequences, with each N-gram (a visual and phrase) describing a local pattern of the action. In particular, with $N = 2$, let $B = \{b_1, b_2, \ldots, b_t\}$ be the sequence of image (frame) representations, b_i is the bag-of-visual-words representation of frame i, the sequence B is then modeled as a non-sequential set of tuples $\{(b_1, b_2), (b_2, b_3), \ldots, (b_{t-1}, b_t)\}$. Each tuple (b_i, b_{i+1}) is simply represented by their concatenation $[b_i^T, b_{i+1}^T]^T$. To fix the dimensions of representations of video sequences, we compute the codewords of the set of concatenations using Sparse Coding and then use a max-pooling to generate a histogram of codewords for each video.

5. Implementation details

To evaluate the effectiveness of the proposed H3DF, we applied it to the applications of hand gesture recognition and human action recognition. Here, we introduce the implementation details of H3DF.

5.1. Pooling center selection

In both hand gesture representation and human action recognition, how to select the center point (p_x, p_y) for the hand or a body part is an essential step which can greatly affect the recognition. An ideal center point should be relatively stable for similar objects and robust to minor shape changes. One option is to use the centroid of the convex hull of a shape. We prefer to find a center on the object rather than on the background, while centroid cannot be ensured when the shape is neither convex nor near-convex (as shown in Fig. 8). Therefore, we propose to use an interior center instead. The procedure of

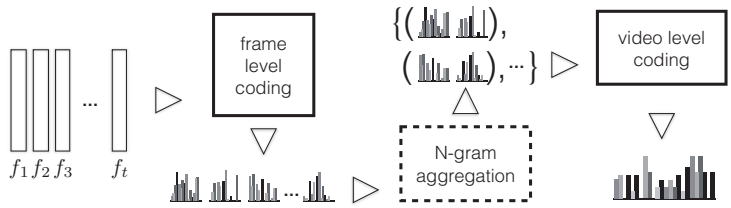

Fig. 7. Illustration of a two-layer bag-of-phrases model for video description. Firstly, a bag-of-words model based on K-means is used to generate a representation vector for each frame in the video (frame level coding). Secondly, a bag-of-phrases model based on Sparse Coding is used to generate a representation vector for the video (video level coding). The final output is a histogram of N-gram codewords.

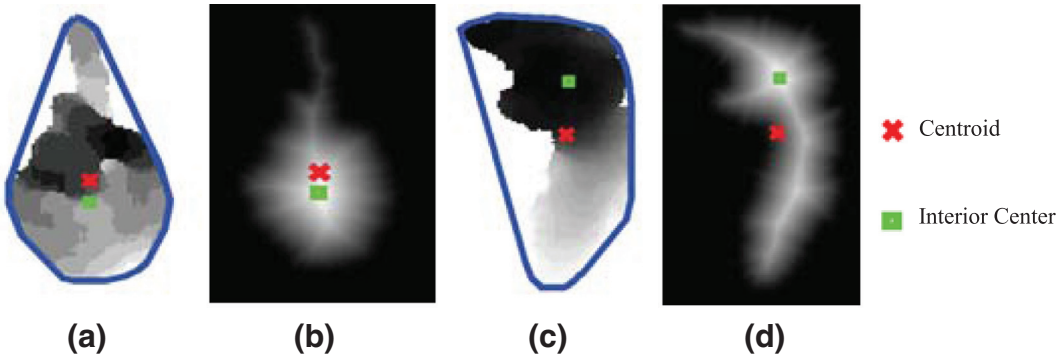

Fig. 8. (a) and (c) are two depth images of hand gestures and associated convex hulls. (b) and (d) are the Euclidean distance transform maps of (a) and (c), respectively. For near-convex shapes such as in (a), the centroid and interior center are similar, but for non-convex shapes such as in (c), the interior center can be ensured to locate within the object and robust to extensions and branches, such as fingers. Brightness of pixels in distance maps (b,d) indicates the Euclidean distance from the nearest boundary pixel to the corresponding pixel locations.

computing the interior center is as: first the depth map is transferred to a binary map(setting foreground pixels as 1 and background pixels as 0), second Euclidean distance transform [17] is applied on the binary map and the "brightest" point is selected as the interior center. The benefit of selecting the interior center rather than the centroid is that the center locates inside the boundary and the major part thus is more robust to minor shape changes such as extensions and branches.

5.2. Normal estimation methods

Here we discuss two estimation methods of the normal vector of a 3D Facet: bilinear estimation (analytical) and least-squares (plane-fitting) estimation. Bilinear normal estimation is suitable for a grid-organized 3D point set (or 2D depth image). Similar to bilinear interpolation, it takes the four neighbors and calculates the two orthogonal line segments that each connects two of them. Given the 3D Facet whose center is at $(i, j, d_{i,j})$, it computes a vector as the normal of this 3D Facet such that this vector is orthogonal to two line segments, one which connects points $(i-1, j, d_{i-1,j})$ and $(i+1, j, d_{i+1,j})$, while the other connects 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 depth image calculation where 3D points are organized as gridded depth pixels. However, when considering 3D point clouds with non-uniform density, this approach will not work.

Plane fitting-based normal (least squares) estimation is more general and can be used in the situations where point density is non-uniform. It takes the center of a 3D Facet along with its neighbor points in a certain range, which we define as its local support surface. Then a plane is fitted using them. Despite its ability to generalize, there is a risk of losing detail when the size of the local support surface is enlarged.

5.3. Sparse representation based classification

To further explore the discriminative power of the proposed H3DF descriptor and its compatibility with different classification schemes, we apply two classification methods with H3DF to recognize hand gestures or human actions: linear-SVM and Sparse representation-based Classification (SRC), which is proposed by Wright et al. [14] with good performance in face recognition. A brief review of SRC is provided as follows: given C as the set of class labels, we have $A = [A_{C_1}, A_{C_2}, \ldots, A_{C_c}]$ as the dictionary of training samples. In our approach, A is the matrix of vectored H3DF descriptors, *i.e.*, $A_{C_i \in C} = [vec(x_1^{C_i}), vec(x_2^{C_i}), \ldots, vec(x_n^{C_i})]$, where $x_j^{C_i}$ is the j^{th} H3DF vector of gesture or action class i. For a query descriptor y, the SRC via l_1-minimization is:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \|\alpha\|_1 \quad s.t. \quad \|y - A\alpha\|_2 \leq \lambda. \tag{6}$$

Therefore, the classification rule is:

$$\text{identity}(y) = \underset{C_i}{\operatorname{argmin}}\, r_{C_i}(y), \tag{7}$$

where the class-wise reconstruction residual $r_{C_i}(y)$ is computed as:

$$r_{C_i}(y) = \|y - A\delta_{C_i}(\hat{\alpha})\|_2, \tag{8}$$

where δ_{C_i} is the characteristic function that selects the coefficients associated with that class.

Runtime of H3DF: computing an H3DF descriptor is fast. Without preprocessing, calculation of the H3DF for a 100 by 100 depth patch is about 2 ms with a Matlab implementation on one Intel Xeon Core (2.13 GHz). H3DF is thus feasible for real-time applications.

6. Experimental results

6.1. Static hand gesture recognition

6.1.1. Datasets and experiment set-up

We apply H3DF descriptors for hand gesture recognition from static depth images on two 3D datasets: the NTU hand digits dataset [6] and the ASL finger spelling dataset [5]. Both datasets were captured by a Kinect camera. The NTU hand digits dataset [6] contains a total of 1000 depth images of 10 hand gestures of decimal digits 0–9 from 10 subjects with 10 samples for each gesture. The ASL finger

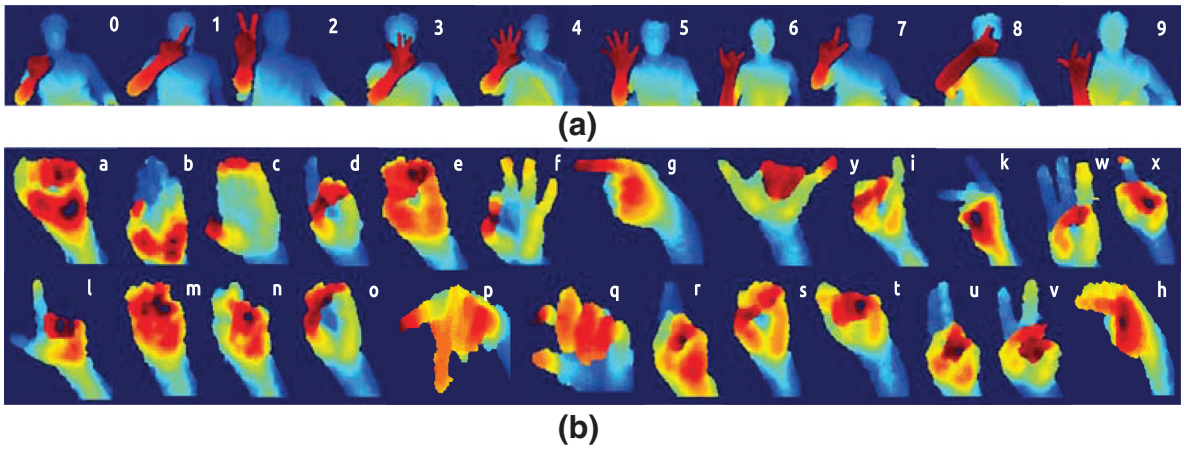

Fig. 9. (a) Sample depth images from the NTU hand digits dataset for digits 0–9 [6]. (b) Sample depth images from the ASL finger spelling dataset for english character from *"a"* to *"z"* (without *"j"* and *"z"*) [5].

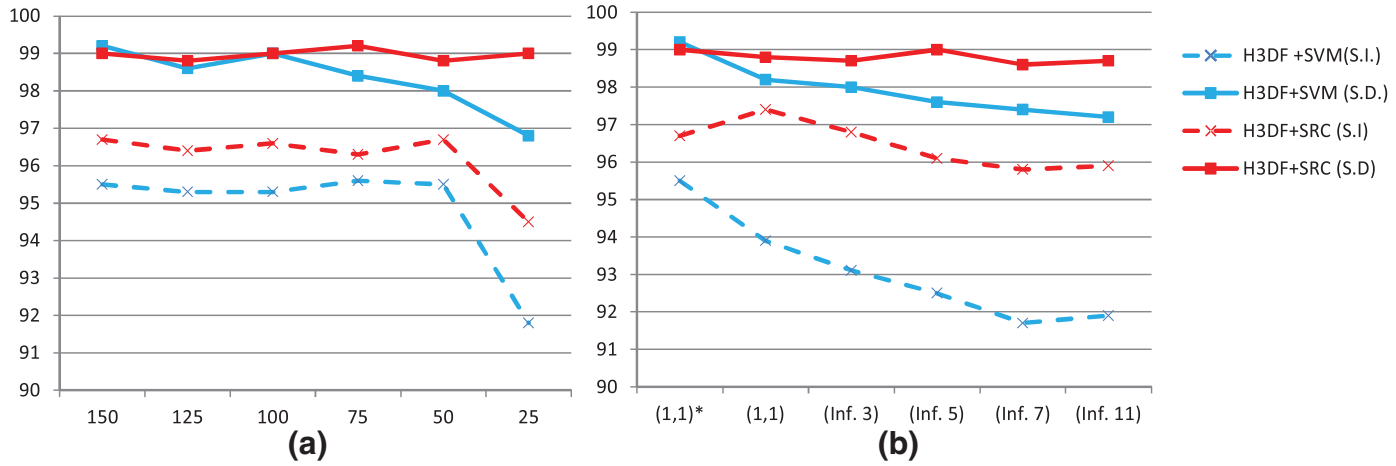

Fig. 10. Accuracies of hand gesture recognition on the NTU hand digits dataset [6] of (a) resolutions of hand-patches, and (b) different methods and parameters of normal estimation. Subject Independent (S.I.) and Subject Dependent (S.D.) accuracies of H3DF with both SVM and SRC [14] classifiers are shown.

spelling dataset [5] captures hand gestures in 24 different categories, each of which represents one English character from *"a"* to *"z"* while *"j"* and *"z"* are excluded since these two characters are performed in ASL using motion. Compared with the NTU hand digits dataset, this dataset is much larger, containing about 60, 000 depth images from 5 subjects. Some images of the datasets are shown in Fig. 9. While the ASL finger spelling dataset only provides segmented hand regions, we obtain the hand regions of the NTU hand digits dataset based on the depth information since the hand is always the most front body part facing to the camera [21].

For static hand gesture recognition, to explore the effect of subjective variance, we conduct two types of experiments. One is a subject-independent test, in which we use a "leave-one-out" strategy, *i.e.*, for a dataset with N subjects, we use $N-1$ subjects for training and the rest one subject for testing. This process is repeated for each subject and the averaged accuracy is reported as the overall accuracy. The other is a subject-dependent test in which all subjects appear in both the training and testing data, but no video appears in both training and testing.

Before comparison with the state-of-the-art approaches, we start by discussing the influences of 1) different approaches to estimate the normal of a 3D Facet, 2) different resolutions of extracted depth map, and 3) different numbers of grids while pooling encoded 3D Facet to generate the final descriptor. We discuss the issues using the NTU hand digits dataset [6].

6.1.2. Normal estimation and hand patch resolution

Here, we first analyze the influence of different resolutions of extracted depth maps as well as the robustness of proposed descriptors against resolution. We set different resolutions ranging from 150×150 to 25×25 for the normalized hand regions. As shown in Fig. 10(a), results in terms of overall classification accuracy of both leave-one-out subject-independent and subject-dependent tests are above 90%, which demonstrates the robustness of the proposed H3DF descriptor for different resolution of the normalized hand regions. Besides, as the resolution decreases, the performances are relatively stable, except in the case of 25×25 resolution. In all the following experiments of static hand gesture recognition, we use 150×150 as default patch size unless otherwise noted.

As shown in Fig. 10(b), we also study the influence of different choices of normal estimation methods. We compare the bilinear normal estimation method with plane fitting-based method of different patch sizes (As shown in Fig. 10 (b), the analytical ($(1,1)^*$) approach performs best ($(1,1)^*$) indicates the analytical solution for normal computation.) For the plane-fitting approach with different sizes of local support surface, performances of both subject-dependent and subject-independent tests decrease and become stable when the size of the local support surface is greater than 7×7. This observation has demonstrated that for this particular problem, bilinear operator is more suitable and the proposed 3D descriptor favors more details rather than less noise. We use the bilinear estimation approach

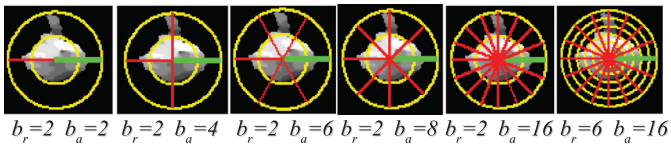

Fig. 11. Illustration of pooling bins layouts with different radial bin (b_r) and angular bin (b_a) settings.

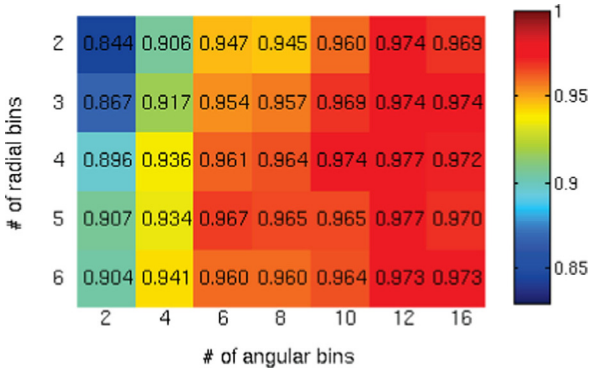

Fig. 12. Recognition accuracies of different pooling granularity settings (y-axis for b_r, x-axis for b_a) on the NTU hand digits dataset [6].

in all following experiments. In addition, we observe that subject-dependent tests perform better than subject-independent tests and are less affected by increases in the local support surface size. Additionally, proposed H3DF combined with sparse representation-based classification (SRC) [14] performs better than linear SVM. Thus, the default classifier is SRC in the rest of this paper, unless otherwise noted.

6.1.3. Discussion of pooling granularity

To explore how the pooling granularity affects the discriminative power of our proposed descriptor, we first conduct different settings of radial bin layouts (number of bins $= b_r$) and angular bin layouts (number of bins $= b_a$). Some of the pooling grids are illustrated in Fig. 11. Since inside each cell of the pooling grid, the encoded facet vectors (dimension $= 18$) are pooled together by taking the average, thus the total dimension of the final H3DF descriptor is proportional to the product of $b_r \times b_a$.

The recognition accuracies are illustrated in Fig. 12. We observe that low pooling granularity (from upper-left corner to bottom-right corner, pooling granularity increases) associates with relative low recognition accuracy. As granularity increases, recognition accuracy tends to increase and gradually reaches a stable value. We set a default of ($b_r = 4, b_a = 10$) unless otherwise noted because this is an appropriate trade-off between feature length and discriminative power based on our experiments.

6.1.4. Comparison with the state-of-the-arts

To compare our proposed H3DF feature descriptor with the benchmark methods as well as traditional 2D HOG descriptor on both datasets, we compare the proposed H3DF with the benchmark methods as well as the traditional 2D Histogram of Gradients (HOG) descriptor on static hand gesture recognition. In our implementation of HOG, we evenly separate the normalized region of interest into 8×8 non-overlapping patches and each patch has eight orientation bins. Thus the dimension of each HOG descriptor is 2046. The average accuracies on the NTU hand digits dataset are shown in Table 1. Our method outperforms the benchmark method and the traditional 2D HOG descriptor for both subject-independent and subject-dependent tests. Compared with [6], our H3DF feature descriptor contains more information, such as folded thumb in palm than only contour

Table 1
Performance comparison of different methods on the NTU hand digits dataset [6].

Approach	Subj. ind. test(%)	Subj. dep. test(%)
Ren et al. [6]	93.9	N/A
HOG [1]	93.1	94.6
H3DF+SVM	94.5	99.2
H3DF+SRC	97.4	99.0

Table 2
Performance comparison of different methods on the ASL finger spelling dataset [5].

Approach	Subj. ind. test(%)	Subj. dep. test(%)
Pugeault and Bowden [5]	49.0	N/A
HOG [1]	65.4	96.0
Keskin et al. [24]	84.3	97.8
H3DF+SVM	73.3	99.0
H3DF+SRC	77.2	99.9
denseH3DF+SVM	83.8	100.0

information. Our method performs 3.5% higher than [6] and 4.3% higher than 2D HOG descriptor in the subject-independent test. As can be predicted, performances in subject-dependent test are much higher than in subject-independent test, where our method achieves 99.2% (H3DF+SVM) and 99.0% (H3DF+SRC) classification accuracy. Recently, classification results on this dataset are saturated (99 and 100% reported in [36]) via combining over three kinds of features which are specifically designed *for hand-shape only*. Since, the proposed H3DF descriptor is a generic descriptor and can be used for multiple purposes such as action recognition and object recognition, we will not directly compare it with the fusion mechanism as proposed in [36].

Compared with the NTU hand digits dataset [6], the ASL finger spelling dataset [5] contains more complicated (24 gesture categories *vs.* 10 gesture categories) and realistic (all gestures are as in American Sign Language (ASL)) hand gestures. The ASL finger spelling dataset is also much larger (over 60,000 images) than the NTU hand digits dataset (1,000 images).

We follow the same experiment setting as previous stated. The average accuracies of both subject-dependent and subject-independent tests are shown in Table 2. Our descriptor achieves 77.2% average accuracy in the subject-independent test, which significantly outperforms [5] with 28.2% higher accuracy, partially because we perform orientation correction before coding. Compared with the traditional 2D HOG descriptor, which is also with orientation correction, our method still achieves 11.8% higher accuracy and demonstrates the effectiveness of the proposed H3DF descriptor in describing 3D depth images than just applying an existing 2D descriptor. The main confusions are cause by gestures with very similar poses or shapes such as *"p"* and *"q"* where hand poses are almost the same and the only difference is the layout of two fingers (see Fig. 9 (b) for hand gestures) or *"m"* and *"n"*, which share quite similar shapes.

To further explore the capability of the proposed H3DF as a local pattern descriptor, we combine the H3DF with dense sampling as used in DenseSIFT [30] with an evenly dense sampling grid at multiple scales (denseH3DF). In our experiment, we sample keypoints every 4×4 pixels at scales {8, 12, 16}. In each sampling keypoint, we compute the H3DF with radial bin number as 2 and angular bin number as 8. The local descriptor is then encoded using a soft vector quantization with a codebook of 1024 codewords computed from training set. For spatial pooling, we use a 4×4 spatial grid which partition the sampled points into 16 sets. Within each set, the sampled points (codes) are pooled using max pooling. Thus the resulting dimension of the feature vector is $4 \times 4 \times 1024 = 262,144$. We test denseH3DF using a linear SVM and the performance achieves 83.8% in subject

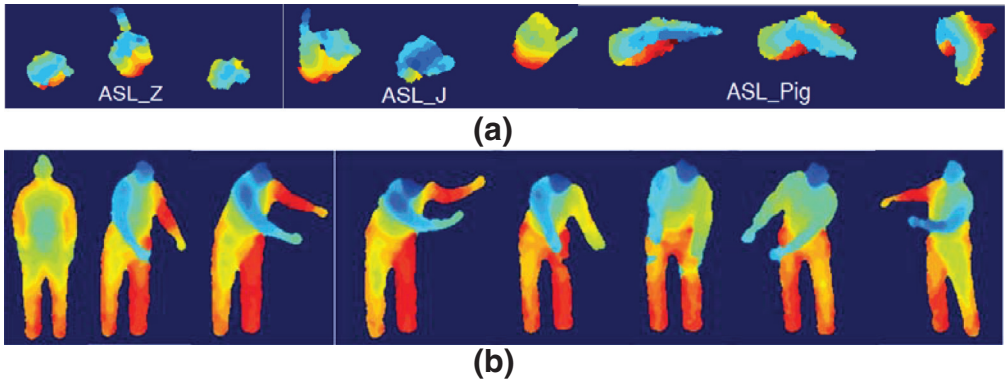

Fig. 13. (a) Sample frames from the MSR 3D gesture dataset for dynamic ASL hand gesture recognition. (b) Sample frames of action "*Golf Swing*" from the MSRAction3D dataset [4].

independent test (Table 2), which is very close to the current best result obtained by Keskin et al. [24] (84.3%). However, [24] is specially designed only for hand poses, not a generic descriptor as H3DF is.

6.2. Dynamic hand gesture recognition and human action recognition

6.2.1. Datasets

To validate our H3DF descriptor together with DP-based temporal segmentation for dynamic hand gesture recognition from video sequences, we employ the MSR 3D gesture dataset. This dataset contains 12 dynamic American Sign Language (ASL) gestures performed by 10 subjects. There is a total of 336 video sequences captured by a Kinect camera. The gesture categories cover ASL gesture signs such as *"Where", "Store", "Pig", etc.* The hand region has been segmented. This dataset was collected by Wang et al. [16] and state-of-the-art performance has been demonstrated by Oreifej and Liu [18]. We normalize each image along its height to 50 pixels for efficiency, while keeping the width/height ratio unchanged. We follow the same setting as in [18], which leaves one subject out for testing and trains on the rest and 10 repeats are processed to generate an averaged accuracy as the reported accuracy.

To further investigate how well our proposed descriptor can cope with more complex spatial-temporal feature descriptions, we also evaluate the H3DF for human action recognition on the MSRAction3D dataset [4], and compare its performance with existing state-of-the-art methods. The MSRAction3D dataset includes 20 action categories such as *"high arm wave", "hand catch", etc.*, which are performed by 10 subjects facing the camera. Each subject performed each action two or three times. The actions in this dataset capture a variety of motions related to arms, legs, torso, and their combinations. Several samples from mentioned datasets are shown in Fig. 13.

6.2.2. Discussion of pooling granularity

Before comparing proposed H3DF descriptor with others on these two datasets, we first conduct experiments to investigate both spatial and temporal pooling granularity on MSR 3D gesture dataset. The experiment settings for spatial pooling granularity are the same as in Section 6.1.3 and the temporal segments number (K) is set to 5 for consistency. The results are shown in Fig. 14, we can observe similar patterns as in Fig. 12, which again validate our default settings for r_a and r_b. A second issue is how temporal pooling granularity affects the recognition accuracy of dynamic gesture recognition. We compare the proposed dynamic programming-based temporal segmentation with traditional evenly partitioning of different numbers of temporal segments (K), the accuracies are shown in Fig. 15. We observe that as K increases, more complementary information is modeled which results in higher accuracies. Five is a good selection for K, because normally *"neutral", "on-set", "peak", "off-set"* and *"neutral"* are the general states of a sequence of action. Dynamic partitioning is consistently a better strategy than even partitioning (except for $K = 4$) because dynamic partitioning is more robust to variance in temporal sequences due to its invariance to action execution speed.

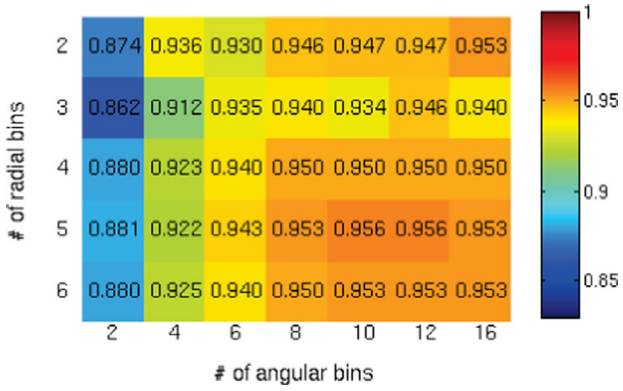

Fig. 14. Recognition accuracies of different pooling granularity settings (y-axis for b_r, x-axis for b_a) on the MSR 3D gesture dataset.

Table 3
Performance comparison of different methods on the MSR 3D gesture dataset [16].

Approach	Avg. recognition rate(%)
H3GO [19]	85.23
ROP [16]	88.50
DMM [13]	89.20
HON4D [18]	92.45
DP-H3DF	95.00

6.2.3. Comparison with the state-of-the-arts

Dynamic gesture recognition: We further evaluate the proposed H3DF descriptor together with DP-based temporal partitioning in the application of dynamic hand gesture recognition on the MSR 3D gesture dataset [16]. We compare our proposed descriptor with several state-of-the-art algorithms for dynamic hand gesture representation such as the Histogram of 3D Gradient Orientations (H3GO) [19] and Histogram of 4D normals (HON4D) [18] which combines surface normals and Fourier transforms to represent spatial-temporal 4D volumes. As shown in Table 3, our framework (DP−H3DF+SRC) outperforms all previous methods (the best recognition rate of our method on the MSR 3D gesture dataset is 95.6% with a different pooling grid setting, but to be consistent, we report the performance with default grid setting here).

Human action recognition: We also evaluate the proposed H3DF descriptor in the application of human action recognition from depth sequences on the MSRAction3D [4] and compare it with existing

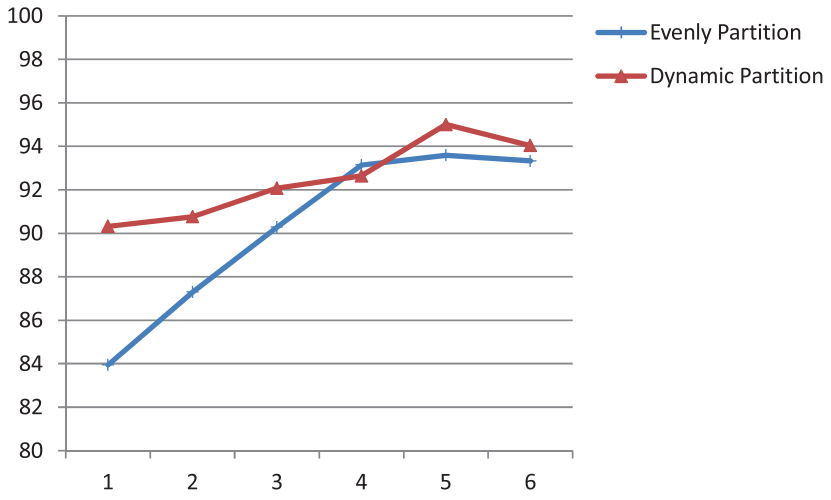

Fig. 15. Recognition accuracies of different temporal strategies and numbers of segments (*i.e.*, *x*-axis shows *K*).

Table 4
Performance comparison of different methods on human action recognition of the MSRAction3D Dataset [4].

Approach	Avg. recognition rate(%)
Bag of 3D points [4]	74.70
HOJ3D [25]	79.00
STOP [26]	84.80
ROP [16]	86.50
Actionlet [27]	88.20
DMM [13]	88.73
HON4D [18]	88.89
DSTIP [28]	89.30
Proposed method	89.45
Pose set [29]	90.00

state-of-the-art methods. Since the actions are not constraint to hand gestures in this dataset, instead of extracting hand patches, we compute H3DF around each skeleton joint and use a codebook with 3000 codewords to encode each H3DF. Then each frame is represented by the max-pooled histogram of all H3DF to generate a bag-of-word representation of that frame. Next we use Bi-gram representations as discussed in Section 4.2 to obtain a set of Bi-grams for each video sequence. Finally, sparse coding is employed to generate a sparse histogram for each video using a dictionary with 1024 basis bi-grams trained.

Our proposed method has decent performance compared with state-of-the-art methods (Table 4) and achieves comparable performance with the best results [29]. As described in previous sections, our main goal is to propose a generic depth descriptor to handle both static and dynamic gesture and human action recognition. In both dynamic hand gesture and human action recognition, our method achieves better performance than other counterparts.

Discussion: The proposed two approaches of extending H3DF to cope with dynamic representation of a video sequence have different objectives. DP-based partitioning aims to solve the temporal alignment problem caused by different execution rates. The N-gram-based method, on the other hand, is designed to model local transition patterns. For example, to model a sequence of "raising hand", the DP method seeks to end up with 2 (or 3) gestures that can sufficiently summarize the action, *i.e.*, "lowered hand", ("raising hand") and "raised hand"; while N-gram method pursues to capture the motion during "raising hand". In other words, the DP-based method generates *"a sequential collection of gestures"* while the N-gram-based method generates *"a bag of motions"*. The two perspectives are both useful to capture temporal structures and transition patterns. But in practice, we found that the proposed DP method works better with dynamic hand gesture recognition while the proposed N-gram method works better with human action recognition. The reason for this may be the intrinsic difference between hand gestures and action recognition. Hand gestures information is conveyed mainly by the shape of hand while motion is complementary information and human actions are highly performed by drastic motions of body parts. In addition, the l_2 metric used in our DP algorithm is prone to sparse noise but large in magnitude, which is more common in human action recognition.

7. Conclusions

In this paper, we have proposed a novel discriminative 3D descriptor (H3DF) which can effectively capture and model the rich surface shape information of the depth maps. Applying orientation normalization, robust coding and concentric spatial pooling, our H3DF descriptor is robust to translation, view angle and scaling changes. Local H3DF is also able to evolve into denseH3DF for modeling more local patterns. To tackle the task of dynamic hand gesture and human action recognition from depth video sequences, two temporal extension approaches are developed: dynamic programming-based temporal partition and N-gram-based method. The two approaches are applied to build augmented descriptors with robust representative description. We have extensively evaluated the effectiveness of the proposed H3DF descriptor on four public datasets including static hand gesture recognition from single depth image, dynamic hand gesture and human action recognition from depth sequences. The experimental results demonstrate that our proposed approach outperforms or achieves comparable accuracy to the state-of-the-art for action and hand gesture recognition.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments and insightful suggestions that improved the quality of this manuscript. This work was supported in part by NSF grant nos. EFRI-1137172 and IIS-1400802. We also thank Dr. Aries Arditi for his carefully proofreading.

References

[1] N. Dalal, B. Triggs, Histogram of orientated gradients for human detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2005, pp. 886–893.
[2] R. Francois, G. Medioni, Adaptive color background modeling for real-time segmentation of video streams, in: International Conference on Imaging Science, Systems, and Technology, Las Vegas, NA, 1999.
[3] I. Laptev, On space-time interest points, Int. J. Comput. Vis. 64 (2) (2005) 107–123.
[4] W. Li, Z. Zhang, Z. Liu, Action recognition based on a bag of 3D points, in: IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010.
[5] N. Pugeault, R. Bowden, Spelling it out: real-time asl fingerspelling recognition, in: IEEE International Conference on Computer Vision Workshops, 2011, pp. 1114–1119.
[6] Z. Ren, J. Yuan, Z. Zhang, Robust gesture recognition based on finger-Earth mover's distance with a commodity depth camera, in: The 19th ACM international conference on Multimedia (ACM'MM), 2011, pp. 1093–1096.
[7] Z. Ren, J. Yuan, C. Li, W. Liu, Minimum near-convex decomposition for robust shape representation, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 303–310.
[8] L. Schwarz, A. Mkhitaryan, D. Mateus, N. Navab, Estimating human 3D pose from time-of-flight images on geodesic distance and optical flow, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2011, pp. 700–706.
[9] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time pose recognition in parts from single depth images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2011, p. 7.
[10] M.V.d. Bergh, L.V. Gool, Combining RGB and TOF cameras for real-time 3d hand gesture interaction, in: IEEE Workshop on Applications of Computer Vision (WACV), 2011, pp. 66–72.
[11] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 1290–1297.
[12] Y. Wu, T. Huang, Vision-based gesture recognition: a review, Gesture-based commun. in hum. comput. interact. (1999) 103–115.
[13] X. Yang, C. Zhang, Y. Tian, Recognizing actions using depth motion maps-based histograms of oriented gradients, in: ACM International Conference on Multimedia, 2012.
[14] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 31 (2) (2009).
[15] R. Bellman, On the approximation of curves by line segments using dynamic programming, Commun. ACM 4 (6) (1961) 284.
[16] J. Wang, Z. Liu, J. Chorowski, Z. Chen, Y. Wu, Robust 3D action recognition with random occupancy patterns, in: ECCV, 2012.
[17] C. Maurer, R. Qi, V. Raghavan, A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2) (2003) 265–270.
[18] O. Oreifej, Z. Liu, Hon4d: histogram of oriented 4D normals for activity recognition from depth sequences, in: CVPR, 2013.
[19] A. Klaser, M. Marszalek, C. Schmid, A spatio-temporal descriptor based on 3D-gradients, in: British Machine Vision Conference, BMVC'08, Leeds, United Kingdom, 2008.
[20] S. Tang, X. Wang, T. Han, J. Keller, M. Skubic, S. Lao, Z. He, Histogram of oriented normal vectors for object recognition with a depth sensor, in: ACCV, 2012.
[21] C. Zhang, X. Yang, Y. Tian, Histogram of 3D facet: a characteristic descriptor for hand gesture recognition, in: IEEE Conference on Automatic Face and Gesture Recognition (FG), 2013.
[22] P. Brown, P. Desouza, R. Mercer, V. Pietra, J. Lai, Class-based N-gram models of natural language, in: Computational Linguistics, MIT Press, 1992.
[23] G. Pedrosa, A. Traina, From bag-of-visual-words to bag-of-visual-phrases using n-grams, in: IEEE Conference on Graphics, Patterns and Images (SIBGRAPI), 2013.
[24] C. Keskin, F. Kıraç, Y.E. Kara, L. Akarun, Hand pose estimation and hand shape classification using multi-layered randomized decision forests, in: ECCV, 2012.
[25] L. Xia, C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: CVPR Workshop, 2012.
[26] A. Vieira, E. Nascimento, G. Oliveira, Z. Liu, M. Campos, Stop: space-time occupancy patterns for 3d action recognition from depth map sequences, in: CIARP, 2012.
[27] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, in: CVPR, 2012.
[28] L. Xia, C. Chen, J. Aggarwal, View invariant human action recognition using histograms of 3D joints, in: CVPR Workshop, 2012.
[29] C. Wang, Y. Wang, A. Yuille, An approach to pose based action recognition, in: CVPR, 2013.
[30] A. Vedaldi, B. Fulkerson, Vlfeat: an open and portable library of computer vision algorithms, in: Proceedings of the International Conference on Multimedia, ACM, 2010.
[31] H. Pfister, M. Zwicker, J. van Baar, M. Gross, Surfels: surface elements as rendering primitives, in: SIGGRAPH, 2000.
[32] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (24): 509–521.
[33] S. Hadfield, R. Bowden, Hollywood 3d: recognizing actions in 3d natural scenes, in: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[34] C. Qian, X. Sun, Y. Wei, X. Tang, J. Sun, Realtime and robust hand tracking from depth, in: CVPR, 2014.
[35] H. Liang, J. Yuan, D. Thalmann, Parsing the hand in depth images, in: IEEE Trans. on Multimedia (T-MM), 2014.
[36] F. Domino, M. Donadeo, P. Zanuttigh, Combining multiple depth based descriptors for hand gesture recognition, Pattern Recognit. Lett. 50 (2014) 101–111.