




Dynamic hand gesture recognition using motion pattern and shape descriptors

Meng Xing¹ · Jing Hu¹ · Zhiyong Feng²  · Yong Su¹ · Weilong Peng¹ · Jinqing Zheng³

Received: 20 October 2017 / Revised: 13 August 2018 / Accepted: 15 August 2018 /

Published online: 10 September 2018

© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

The key problems of dynamic hand gesture recognition are large intra-class (gesture types, without considering hand configuration) spatial-temporal variability and similar inter-class (gesture types, only considering hand configuration) motion pattern. Firstly, for intra-class spatial-temporal variability, the key is to reduce the spatial-temporal variability. Due to the average operation can improve the robustness very well, we propose a motion pattern descriptor, Time-Wise Histograms of Oriented Gradients (TWHOG), which extracts the average spatial-temporal information in the space-time domain from three orthogonal projection views (XY, YT, XT). Secondly, for similar inter-class motion pattern, accurate representation of hand configuration is especially important. Therefore, the difference in detail needs to be fully captured, and the shape descriptor can amplify subtle differences. Specifically, we introduce Depth Motion Maps-based Histograms of Oriented Gradients (DMM-HOG) to capture subtle differences in hand configurations between different types of gestures with similar motion patterns. Finally, we concatenate TWHOG and DMM-HOG to form the final feature vector Time-Shape Histograms of Oriented Gradients (TSHOG) and verify the effectiveness of the connection from quantitative and qualitative perspective. Comparison study with the state-of-the-art approaches are conducted on two challenge depth gesture datasets (MSRGesture3D, SKIG). The experiment result shows that TSHOG can achieve satisfactory performance while keeping a relative simple model with lower complexity as well as higher generality.

Keywords Dynamic hand gesture recognition · Hand configuration · Spatial-temporal variability · Motion pattern descriptor · Shape descriptor

✉ Zhiyong Feng
zyfeng@tju.edu.cn

¹ School of Computer Science and Technology, Tianjin University, Tianjin 300072, China

² School of Computer Software, Tianjin University, Tianjin 300072, China

³ Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

1 Introduction

Hand gestures are widely used as intuitive and convenient ways of communications in our daily life, and hand gesture recognition has been broadly applied in human computer interfaces, robot control, and augmented reality, etc. [28]. Hand gestures are conceptually divided into static gestures and dynamic gestures. Compared to static gestures, dynamic hand gestures usually provide richer communication channels because of motion information is incorporated, and are thus more difficult to recognize.

Approaches for dynamic hand gesture recognition can be broadly categorized into two groups, depending on whether they use RGB information or depth information. Due to the characteristics of RGB data, the results of RGB-based [1, 2, 9, 31, 37] methods always sensitive to clutter, lighting conditions, and skin color. As the imaging technique advances, e.g. the launch of Microsoft Kinect, more recent research methods on dynamic hand gesture recognition have been performed using depth maps that captured by such cameras. Comparing with conventional RGB images, depth maps provide a large body of advantages. For example, (1) the depth maps provide the 3D structure and shape information. (2) Depth maps are insensitive to illumination changes and color changes [42]. (3) The backgrounds of depth maps are fairly clean. Therefore, the Kinect sensor gives a broader scope for dynamic hand gesture recognition.

The most challenging problems of dynamic hand gesture recognition are large intra-class spatial-temporal variability and similar inter-class motion pattern. **Large intra-class spatial-temporal variability:** without considering hand configuration, the gestures with the same movement process can differ in many respects including velocity, the shape of hand, duration, integrality, and distance between hand and depth camera. **Similar inter-class motion pattern:** when only considering hand configuration, many different gestures have very similar motion process. The only difference is that the hands need to maintain different configurations in the movement.

Recently, researchers have presented numerous effective depth-based methods [3, 15, 18, 24, 25, 32, 35, 36, 38, 39, 42] and achieved important progresses in dynamic hand gesture recognition. However, previous depth-based methods capture the superimposed shape information or spatial-temporal information of depth sequences without specifically reinforce the robustness of descriptors to spatial-temporal variability. Therefore they are influenced by many factors (including the shape of hand, velocity, duration, integrality, and the distance between hands and depth camera) and failed to achieve satisfactory results (the recognition rate of a lot of methods is below 98% on MSRGesture3D dataset). Inspired by [40], we not only view a depth sequence as a stack of spatial texture slices (XY) along the T-axis, but also view the depth sequence as a stack of spatial-temporal texture slices (XT/YT) along the X/Y-axis. XT and YT slices provide information about the space-time transitions [40]. Details are shown in Fig. 2. Therefore, capturing the shape information of all the slices in order along every axis (T/Y/X) can obtain the space-time characteristic of depth sequence. Furthermore, averaging shape information can increase the robustness of shape features varying along the axis. This motivates us to design a more discriminatory descriptor by capturing average shape information on three projection views to alleviate intra-class spatial-temporal variability. Specifically, the robustness of descriptor to shape variability (e.g. shape of hand, integrality, and the distance between hand and depth camera) can be enhanced by averaging all features along T-axis (XY slices). In the same way, by averaging all features along X/Y-axis (YT/XT slices), the robustness of descriptor to spatial-temporal variability (e.g. velocity, duration) can be enhanced. In addition, modified Histograms of Oriented Gradients (mHOG) algorithm is used to capture the shape information of each plane in this paper.

Furthermore, there are many gestures whose motion pattern are very similar, and only have slight differences in hand configuration. For example, in MSRGesture3D both “Where” and “Bathroom” are one hand repeatedly swinging left and right with the same amplitude, and the only difference is the hand configuration. Specifically, gestures “Where” maintain fist while index finger stretches, gestures “Bathroom” maintain fist while thumb slightly protruding between index and middle fingers. In this occasion, accurate representation of hand configuration is especially important. The difference in detail needs to be fully captured, and the shape descriptor can amplify subtle differences. Therefore, we introduce Depth Motion Maps-based Histograms of Oriented Gradients (DMM-HOG) to capture subtle differences in hand configurations between different gestures with similar motion patterns. Specifically, (1) “DMM” is used to calculate the global shape information of depth sequence so that the subtle shape differences of every depth map are accumulated. (2) HOG algorithm is used to enhance the shape information of DMMs (DMM_f , DMM_s , DMM_t). (3) Finally, three features of DMMs (DMM_f -HOG, DMM_s -HOG, DMM_t -HOG) are concatenated to construct DMM-HOG.

The final descriptor, TSHOG, is constructed by concatenating TWHOG and DMM-HOG. Through analyzing the performance of our method on two challenge depth gesture datasets (MSRGesture3D [18] and SKIG [21]), we observed that: (1) TWHOG is robust to spatial-temporal variability and can effectively extract motion patterns from dynamic hand gestures. (2) DMM-HOG enhances the inter-class discriminability of the final descriptor by extracting subtle hand configurations differences between different types of gestures with similar motion patterns. (3) Connecting TWHOG and DMM-HOG can effectively distinguish the part of gestures that cannot be classified by the previous method. The result indicates that connecting the two descriptors is effective, and the two descriptors complement each other in some complex situations (spatial-temporal variability and differences in hand configurations are in severe imbalances). We analyzed the effectiveness of the connection from the quantitative and qualitative perspective (Section 3.2).

The outline of the paper is organized as follows. After a brief survey of related work in Section 2, our approach is introduced in Section 3. Section 4 contains experimental results and analysis of our method. Section 5 concludes the paper and looks at the future work.

2 Related work

Recent advances in depth sensing provide a new area for dynamic hand gesture recognition and have attracted a lot of research efforts. Some works have been published to address dynamic hand gesture recognition from depth sequences. We mainly focus on efforts which are closest to our work, including: (1) Depth based methods. (2) Feature fusion approaches.

Depth based methods have been widely utilized for dynamic hand gesture recognition. **Space domain based methods:** In 2012, Kurakin et al. [18] use cell occupancy features and silhouette feature of depth maps to train action graphs. Action graph is utilized to model the temporal dynamics of the key postures of dynamic hand gesture. They get a purely data-driven system which can be used to recognize any other gestures. In [36], the depth sequence is summarized in a motion map, which is the average difference between the depth frames. Consequently, a single (HOG) descriptor [8] is extracted from the motion map. In 2015, Chen et al. [3] combine DMMs-based Local Binary Pattern (LBP) and Kernel-based Extreme Learning Machine (KELM) classifier to recognize the hand gestures. These methods can capture shape information of depth sequences in subtle

meanwhile sensitive to shape of hand, integrality, and the distance between hand and depth camera. In addition, these methods collapse the temporal variations, and thus suffers when the temporal order is of significance. **Space-time domain based methods:** In [32], Random Occupancy Pattern (ROP) features are extracted from depth sequences, and a sparse coding approach is utilized to encode these features. They are less sensitive to occlusion because they only encode information from the regions that are most discriminative for the given gesture. In 2014, Yang and Tian [35] group local hypersurface normals into polynormal, and aggregate low-level polynormals into the Super Normal Vector (SNV). Surface normals are utilized as local features of hand gestures, which show robustness to occlusions. Oreifej and Liu [25] proposed a new descriptor as HON4D for activity recognition which describes the depth sequences using a histogram capturing the distribution of the surface normal orientation in a 4D space of time, depth and spatial coordinates. In 2018, Jiang et al. [15] obtain VME-sequences by controlling the overlap of sub-depth sequences. Then calculate Multi-Temporal DMM-LBP of each VME-sequence in three views to encode the patch descriptors which result to a compact feature representation. By introducing time information, these methods enhanced the performance of gesture recognition at a certain level. However, introduced time information is heavily influenced by factors such as velocity, duration. At the same time, these methods did not alleviate the problems of the space domain based approaches. Hence, they also failed to achieve satisfactory results. **Convolutional Neural Network based:** In [14], a unique multi-layer perception of neural network is built for classification by using back-propagation learning algorithm. In [22], a hand gesture recognition system was introduced, which utilizes depth and intensity channels with 3D convolutional neural networks, and realized automatic detection and classification of dynamic hand gestures in real-world [13]. In 2016, Kim Y et al. [16] investigate the feasibility of recognizing human hand gestures using micro-Doppler signatures measured by Doppler radar with a deep convolutional neural network (DCNN). However, unlike image classification, improvement brought by end-to-end deep ConvNets remains limited compared with traditional hand-crafted features for video-based action recognition [33]. For long-range temporal structure not been considered as a critical factor in deep ConvNet frameworks.

Recently, features fusion methods improve the performance of depth based dynamic hand gesture recognition significantly. In [29], 3D Spherical Histograms of Oriented Normal Vectors (3DS-HONV) and Depth based Histogram of Optical Flow (DHOF) are combined as a descriptor and then a spatial-temporal representation of the presented descriptors is obtained via sparse coding concepts. In [38], Edge Enhanced Depth Motion Map (E2DMM) and Dynamic Temporal Pyramid are combined to capture shape information and temporal structure of the depth sequences. In [10], multiple depth based descriptors are combined for hand gesture recognition. The descriptors included the hand region's edge distance and elevation, the curvature of the hand's contour, and the displacement of the samples in the palm region. Although these methods have achieved good results through feature fusion, few articles analyze the necessity of feature fusion from the point of view of the problem and provide experimental evidence.

In this paper, TWHOG mitigates intra-class spatial-temporal variability by extracting the average spatial-temporal information in the space-time domain, DMM-HOG alleviates similar inter-class motion pattern by extracting subtle hand configuration differences between different types of dynamic gestures. In addition, the complementarity of two descriptors in partial gestures helps us to identify complex situations that cannot be solved by previous methods. We verified the effectiveness of the complementarity of two descriptors in partial gestures from the quantitative and qualitative perspective (Section 3.2).

3 Our approach

The overview of the pipeline proposed in this paper is illustrated in Fig. 1, where dynamic hand gestures are treated as sequences of frames changing over time. In Section 3.1, we discuss the construction stage of Time-Shape Histograms of Oriented Gradients (TSHOG) in detail. The complementarity between the Time-Wise Histograms of Oriented Gradients (TWHOG) and Depth Motion Maps-based Histograms of Oriented Gradients (DMM-HOG) in partial gestures is analyzed in Section 3.2.

3.1 Construction of TSHOG

In this section, we describe our method in detail. Specifically, construction stages of TWHOG and DMM-HOG are discussed in Section. 3.2.1 and Section 3.2.2. Connecting two descriptors is introduced in section 3.2.3.

3.1.1 Construction of TWHOG

Our goal is to compute a descriptor, which is able to mitigate intra-class spatial-temporal variability. There are two conditions for descriptor to meet: (1) the descriptor is able to capture effective space-time information. (2) The descriptor is robust to spatial-temporal variability. Inspired by [40], the variation of shape along three axes (T, Y, X) reflects the spatial-temporal characteristics of depth sequences. Furthermore, averaging shape information can increase the robustness of shape features varying along the axis. Hence, we capture shape information along three axes and averaging all shape features that along the same axis to enhance the robustness of descriptor to spatial-temporal variability.

A sequence of depth maps could be represented by fourth-order tensor, denoted as:

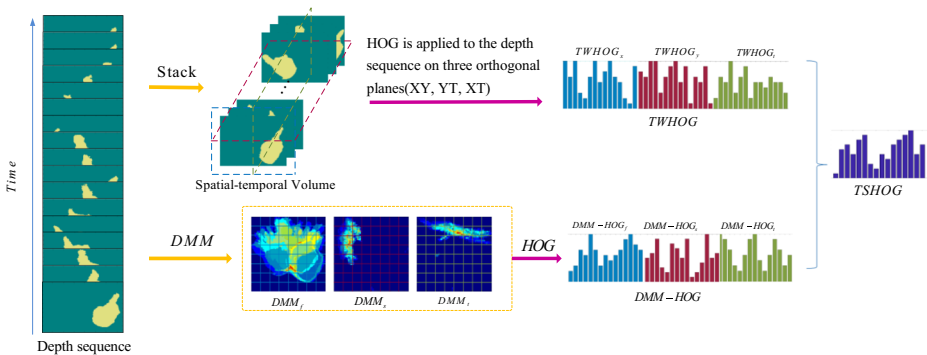


Fig. 1 The overview of the feature extraction. (1) mHOG algorithm is applied on each slice of depth sequence in three projection views (X, Y and T as axes). (2) In each view, mHOG features are averaged and then obtain three components of TWHOG. (3) Connecting three components ($TWHOG_x/TWHOG_y/TWHOG_z$) to establish TWHOG. (4) Modified DMMs are applied to build energy maps of the whole depth sequences in three projection views (X, Y and Z as axes). (5) HOG algorithm is applied on three energy maps and obtain three components of DMM-HOG. (6) Connecting the three components ($DMM-HOG_x/DMM-HOG_y/DMM-HOG_z$) to build DMM-HOG. (7) TWHOG and DMM-HOG of depth sequence are concatenated into a feature vector, named as TSHOG, as the final representation of the depth sequence

$D \in R^{X \times Y \times Z \times T}$ (The X , Y and T represent the length, width and frame number of the depth sequence, respectively. Z represents the maximum of depth value.) Using *Matlab* style notation:

- (1) We refer to the t -th slice of this tensor in *front* view as $D_{:, :, t}$, which is a $X \times Y$ image (matrix) with depth values as the pixel values.
- (2) We refer to the x -th slice of this tensor in *side* view as $D_{x, :, :}$, which is a $Y \times T$ image (matrix) with depth values as the pixel values.
- (3) We refer to the y -th slice of this tensor in *top* view as $D_{:, y, :}$, which is a $X \times T$ image (matrix) with depth values as the pixel values.

Constructing TWHOG descriptor needs three stages: (1) extracting modified Histograms of Oriented Gradients (mHOG) features along three axes. (2) Processing mHOG features. (3) Connecting components. The overview of TWHOG extraction is illustrated in Fig. 2, and details are discussed below.

mHOG algorithm mHOG is constructed by removing the step of constructing block in classical Histograms of Oriented Gradients (HOG) algorithm [8]. This operation is to enhance the robustness of mHOG to subtle shape changes. Specifically: the gradient image of the image patch (using a centered mask $[-1, 0, 1]$) is divided into rectangular cells along the x - and y - directions. A 50% overlap between the cells is used. Within each cell, an orientation histogram is generated by quantizing the angles of each gradient vector into a pre-defined number of bins. These resulting histograms are concatenated to form the final spatial feature vector.

Extracting mHOG feature along three axes mHOG algorithm is applied on XY ($D_{:, :, 1} \cdots D_{:, :, T}$), YT ($D_{1, :, :} \cdots D_{X, :, :}$) and XT ($D_{:, 1, :} \cdots D_{:, Y, :}$) slices of depth sequence. Specifically, along T-axis (XY slices) the gradient image of depth map (using a centered mask $[-1; 0; 1]$) is divided into rectangular cells along the X and Y directions with no overlap. Within each cell, an orientation histogram is generated by quantifying the angles of each gradient vector into a pre-defined number of bins. These histograms divide $L2$ -norm of themselves and concatenated to form the final feature vector $mHOG_{t,i}$ ($mHOG_{t,i}$ represents the feature of $D_{:, :, i}$, $mHOG_{x,i}$ represents the feature of $D_{i, :, :}$, $mHOG_{y,i}$ represents the feature of $D_{:, i, :}$). The same operation is conducted on XT and YT planes and obtains $mHOG_{x,1} \cdots mHOG_{x,Y}$ and $mHOG_{y,1} \cdots mHOG_{y,Y}$ respectively.

Processing mHOG features Calculating the average of mHOG features along three axes. Specifically, along T-axis (XY planes), averaging mHOG features ($HOG_{t,1} \cdots HOG_{t,T}$) to obtain $TWHOG_t$. The same operation is conducted along on Y/X-axis (XT/YT planes) to obtain $TWHOG_x$ and $TWHOG_y$, respectively.

$$TWHOG_v = \frac{1}{V} \sum_{i=1}^V mHOG_{v,i} \tag{3-1}$$

Where i represents the frame index, $mHOG_{v,i}$ is the mHOG feature of i -th frame when v as axis ($v \in \{t; x; y\}$). V ($V \in \{T; X; Y\}$) denote the end frame index.

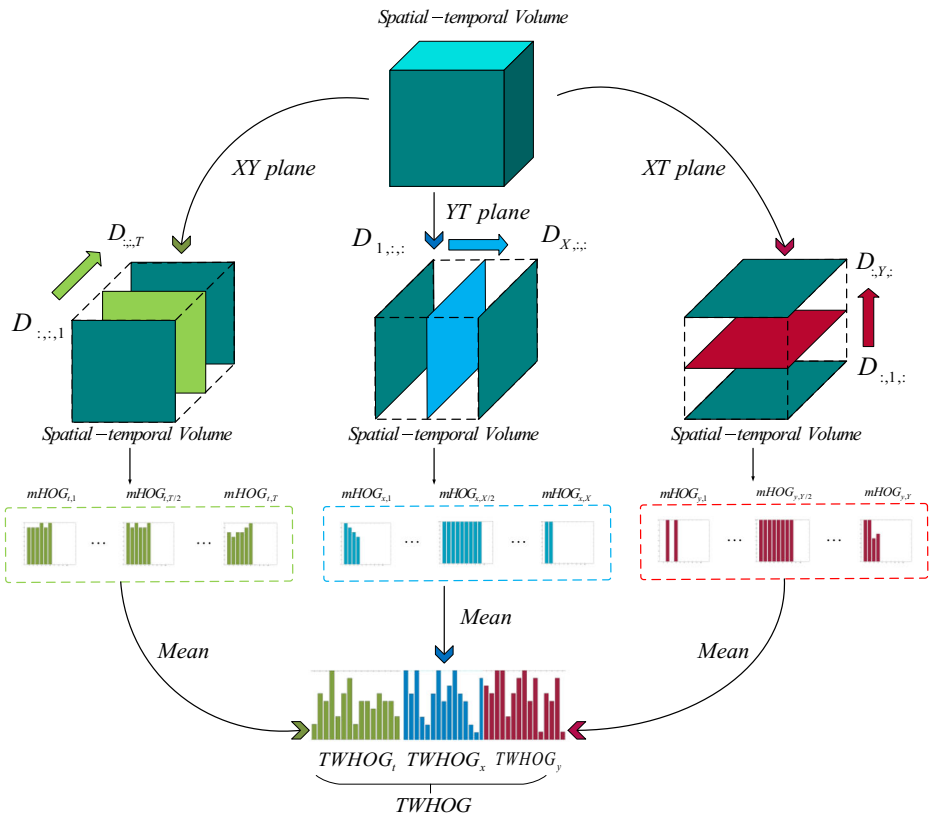


Fig. 2 Extraction of motion pattern information. The depth sequence is not merely viewed as a stack of spatial texture slices (XY) along T-axis, but also a stack of spatial-temporal slices (XT/YT) along X/Y-axis. Along each axis (T, Y, X), mHOG features of each slice are calculated and averaged to obtain the TWHOG component. Then the three components are concatenated to establish TWHOG

Fusing components Piling $TWHOG_t$, $TWHOG_x$ and $TWHOG_y$ into a composite vector to obtain TWHOG. Fusion details are as follows:

$$TWHOG = \left[\begin{bmatrix} 1 \times 1024 \\ \dots \\ TWHOG_t \end{bmatrix} \begin{bmatrix} 1 \times 1024 \\ \dots \\ TWHOG_x \end{bmatrix} \begin{bmatrix} 1 \times 1024 \\ \dots \\ TWHOG_y \end{bmatrix} \right] \quad (3 - 2)$$

The dimension of $TWHOG_t$ / $TWHOG_x$ / $TWHOG_y$ depends on parameters setting.

Note that in the construction of TWHOG, depth values of the depth map are considered as pixel values. And TWHOG obtains motion pattern of gesture by capturing shape information on spatial texture planes and spatial-temporal texture planes. In contrast, in the construction of Histogram of Oriented 4D Normals (HON4D) [25] and Histogram of 3D Facets (H3DF) [39], depth values are used directly to compute the normal vector distribution of hands surface. Therefore, TWHOG is a new descriptor and different from the temporal HOG or spatial HOG. The experimental results on MSRGesture3D dataset show that TWHOG perform better in dynamic hand gesture recognition than HON4D.

3.1.2 Construction of DMM-HOG

There are many different types of gestures in real-world applications that have very similar motion patterns. The only difference is that the hands need to maintain a different configuration in the movement. In order to broaden scope of application of our method. We introduce the DMM-HOG [36] to capture shape information in subtle. In addition, we use the modified the Depth Motion Maps (DMM) algorithm [4] to make the DMM-HOG more sensitive to shape change.

The construction of the DMM-HOG contains three stages: (1) calculating DMMs. (2) Enhancing shape information. (3) Connecting components. Details are discussed below.

Calculating DMMs The concept of DMMs was considered in [4] where the procedure for generating DMMs was modified. In this paper, we adopt the algorithm of DMMs described in [4] due to it is more sensitive to shape change in subtle (By removing the threshold in the step of calculating the absolute value of adjacent frames). Specifically, given a depth video sequence with N frames, each frame in the video is projected onto three orthogonal Cartesian planes (XY, XZ, YZ) to form three 2D projected maps, denoted by map_f , map_s , and map_t . DMMs are then generated as follows:

$$DMM_{\{f,s,t\}} = \sum_{j=1}^{N-1} |map_{\{f,s,t\}}^{j+1} - map_{\{f,s,t\}}^j| \tag{3-3}$$

Where j is the frame index. The procedure of DMMs construction is illustrated in Fig. 3.

Enhancing shape information HOG algorithm [8] is used to enhance the shape information in the DMMs. This stage including three steps: First, use Gaussian filter kernel to remove the noise from each DMMs. Second, according to the parameter setting, each DMM is split into many rectangular cells with no overlap. Third, within each cell, an orientation histogram is generated by quantifying the angles of each gradient vector into a pre-defined number of bins. These resulting histograms divide $L2-norm$ of themselves and concatenated to form the final feature vector. The vectors of DMM_f , DMM_s and DMM_t are labeled as $DMM-HOG_f$, $DMM-HOG_s$ and $DMM-HOG_t$.

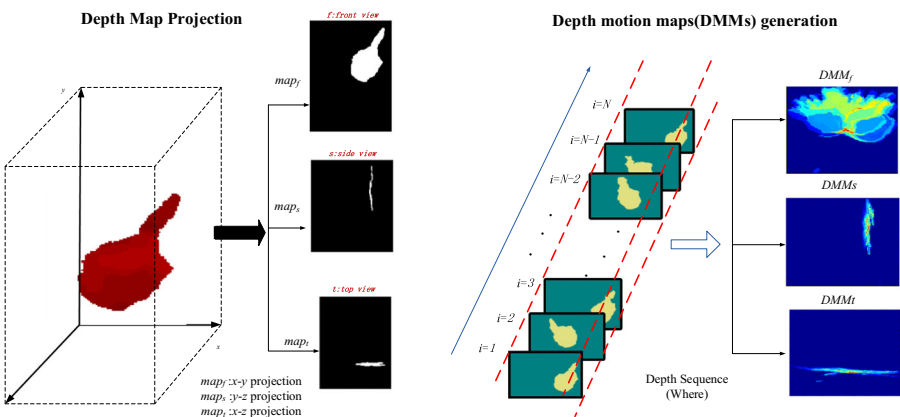


Fig. 3 The procedure of DMMs (DMM_f , DMM_s and DMM_t) construction

Fusing components Piling $DMM-HOG_f$, $DMM-HOG_s$ and $DMM-HOG_t$ into a composite eigenvector to obtain DMM-HOG [36]. Details are as follows:

$$DMM-HOG = \left[\begin{array}{c} 1 \times 1024 \\ \dots\dots\dots \\ DMM-HOG_f \end{array} \right] \left[\begin{array}{c} 1 \times 1024 \\ \dots\dots\dots \\ DMM-HOG_s \end{array} \right] \left[\begin{array}{c} 1 \times 1024 \\ \dots\dots\dots \\ DMM-HOG_t \end{array} \right] \tag{3 - 4}$$

The dimension of $DMM-HOG_f$ / $DMM-HOG_s$ / $DMM-HOG_t$ depends on parameters setting.

3.1.3 Connecting TWHOG and DMM-HOG

The $DMM-HOG_f$, $DMM-HOG_s$ and $DMM-HOG_t$ features are simply stacked into a composite feature vector labeled as DMM-HOG. TWHOG are constructed at the same way, and $TWHOG_b$, $TWHOG_x$ and $TWHOG_y$ were normalized before concatenation. The TWHOG and DMM-HOG are concatenated into a composite feature vector labeled as TSHOG, which is the final discriminatory representation of the depth sequences. Final descriptor are established as follows:

$$TSHOG = \left[\begin{array}{c} 1 \times 3072 \\ \dots\dots\dots \\ TWHOG \end{array} \right] \left[\begin{array}{c} 1 \times 3072 \\ \dots\dots\dots \\ DMM-HOG \end{array} \right] \tag{3 - 5}$$

The dimension of $DMM-HOG/TWHOG$ depends on parameters setting.

Note that such direct fusion of TWHOG and DMM-HOG might not perform well because the data range of these features are different. Hence, we map different features into a comparable range through normalization. Specifically, suppose there are N depth sequences in the database. The two types of features can form an $N \times D$ feature matrix $F = f_{ij}$, where f_{ij} is the j -th feature component in feature vector f_i , and each fused feature vector is of D dimensions. We normalize the entries in each column $f_{.j}$ to the same range $(-1, 1)$ so as to ensure that each individual feature component receives equal weight in determining the similarity between two vectors.

3.2 Complementarity analysis between TWHOG and DMM-HOG in partial gestures

We consider the dynamic gesture recognition as two stages: (1) we use TWHOG, which is robust to space-time variability, to capture the motion patterns of gestures. (2) Shape descriptor DMM-HOG is used to distinguish different types of gestures with similar motion patterns. In a problem perspective, TWHOG and DMM-HOG can complement each other in some complex situations (spatial-temporal variability and differences in hand configurations are in severe imbalances). We combine specific examples to prove this conclusion through qualitative analysis and quantitative analysis respectively.

Qualitative analysis (1) In real-world applications, many different types of dynamic hand gestures have similar movement processes, for example the “Where” and “Bathroom” in Fig. 4 (a). On this occasion, TWHOG unable to identify different types of gestures effectively. Nevertheless, DMM-HOG can capture shape difference in subtle and assists in maximising inter-class distance. (2) The same gesture finished by different “Subjects” with different amplitude and different integrity, hence there is a significant difference in shape information, for example the gestures in Fig. 4 (b). In this case, TWHOG, which based on shape change,

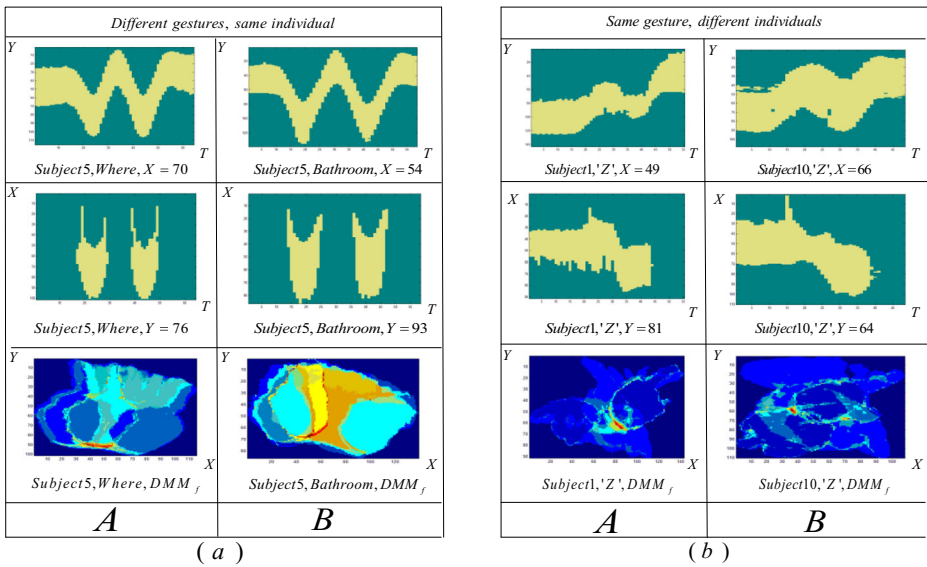


Fig. 4 Depth sequence is a stack of spatial-temporal texture slices (XT/YT), along Y/X-axis. In this figure, spatial-temporal slices and DMM_f of four gestures are given. "Subject5, Where, X = 70" means one Spatial-temporal slice, which along X-axis, of "Where" gesture finished by "Subject5"

can mitigate the influence of "Subject", and conduce to minimise intra-class distance. Therefore, TWHOG and DMM-HOG complement with each other in part of gestures.

Quantitative analysis To further prove TWHOG and DMM-HOG complement with each other in part of gestures. Mutual Information (MI) is introduced to quantitatively validate this conclusion. The Mutual Information (MI) can be a good measure of the redundancy between the two feature representations. The Kraskov I [17] estimator of MI, is a measure that examines the similarity between the neighborhoods of each datum, as defined on each representation via the k-NN rule. That way, the MI estimation can be performed in several scales, revealing interesting properties of the underlying features. Two features are denoted as X and Y, and MI is calculated as follows:

$$MI = \psi(k) - \frac{1}{k} - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N) \tag{3 - 6}$$

Here $\psi(x)$ is the digamma function:

$$\psi(x) = \Gamma(x)^{-1} \frac{d\Gamma(x)}{dx} \tag{3 - 7}$$

It satisfies the recursion $\psi(x + 1) = \psi(x) + 1$ and $\psi(1) = -C$, where $C = 0.5772156\dots$ is the Euler-Mascheroni constant. In X/Y features, collection of the number of neighbor points for each element is n_x/n_y .

Using a subset of MSRGesture3D dataset that consisting only of "Finish" gestures, we compare the Mutual Information between TWHOG and DMM-HOG. The obtained MI for various values of k is illustrated in Fig. 5 (blue line), and the data consists of 20 sequences performed by 5 individuals.

We can easily observe that MI is relatively high at "k = 1", which indicates that both descriptors have adequate gestures-discriminative power. In the area that MI is near-zero, meaning that the examined descriptors carry very different information. As a baseline, we

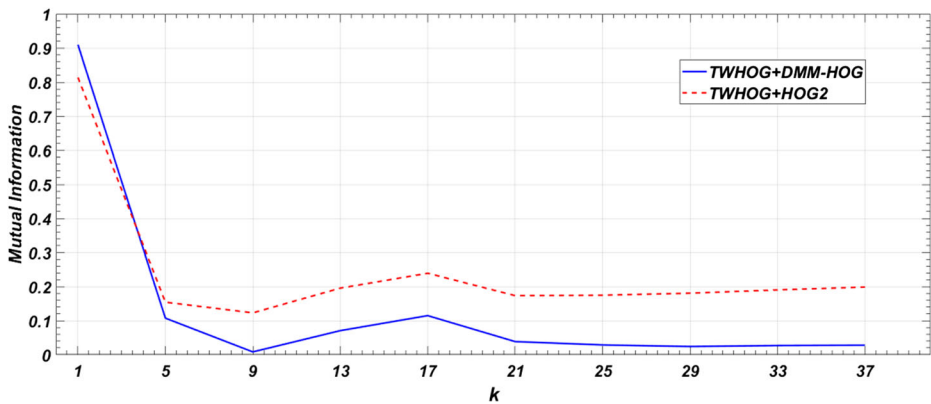


Fig. 5 Mutual Information using *Kraskov I* estimator (our method with baseline)

compare the MI between TWHOG and HOG² (a spatial-temporal descriptor that captures spatial-temporal information by using the HOG algorithm twice in succession on the depth sequence) [24]. We can easily observe that the MI (red dotted line) is generally higher, reflecting there is weaker complementary between TWHOG and HOG² in gestures “*Finish*”.

4 Experimental

In this section, we evaluate our method on publicly available MSRGesture3D dataset [18] and compare with other methods that were conducted with the same experimental setup. Then, we evaluate our approach on a more challenging dataset, SKIG [21], to verify the generalization of our approach. Experimental results show that our method outperforms the state-of-the-art methods on the two datasets.

LIBLINEAR [12] is employed as the linear SVM classifier. We choose the “support vector classification by Crammer and Singer” solver for SVM. The parameter “C” was set to 10. Note that the same parameter setting of SVM as in [41] is used. In addition, to demonstrate the significant advantage of TSHOG, we conduct 15 sets of experiments using different SVM parameters. All results outperform the state-of-the-art methods. The details are given in Table 5.

4.1 MSRGesture3D dataset and setup

MSRGesture3D [18] dataset was captured by a Kinect device. There are 12 dynamic American Sign Language (ASL) gestures (“Z”, “J”, “Where”, “Store”, “Pig”, “Past”, “Hungary”, “Green”, “Finish”, “Blue”, “Bathroom”, “Milk”) and 10 people. Each person performs each gesture 2–3 times. There are 336 files in total, each corresponding to a depth sequence. The hand portion (above the wrist) has been segmented. There are four problems in this dataset impact the performance of dynamic hand gesture recognition seriously. (1) Frame loss (several sequences even only 16 frames, but average number of frames is about 50). (2) Hand positions are unaligned (in several sequences, hands active at the corner). (3) Noise interference. (4) Segmentation is not clean. A region of the interest algorithm is used to reduce the impact of these problems. Specifically, (1) in the construction of DMM_F, a bounding box is set to extract non-zero area (region of interest). (2) According to the

boundary box size obtained in (1), the depth sequence is resize to realize gesture alignment. In order to have a fair evaluation with the other methods, we use the leave-one-subject-out cross-validation scheme proposed by [36], 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. The performance is calculated as the overall accuracy which is the ratio of the correctly recognized gestures over the total number of test sequences. We shows some example frames of MSRGesture3D dataset in Fig. 6.

Parameter setting To improve the computational efficiency, Principle Component analysis (PCA) [34] is utilized to reduce the dimension of the feature vector. The PCA transformed matrix is calculated using the training feature set and then applied to the test feature set. The principal components that account for 100% of the total variation of the training features were considered. The parameters reported in [41] were used here for the sizes of DMMs. The sizes for DMM_f , DMM_s and DMM_t are set to 118×133 , 118×29 and 29×133 , respectively. Experimental parameters are the *cells* and *bins* of mHOG/HOG algorithm which used in the construction of TWHOG and DMM-HOG.

For TWHOG, we conduct 27 sets of experiments and the optimal parameters are *cells* = (8×8) ; *bins* = 16. Detailed results are shown in Table 1. For DMM-HOG, we conduct 27 sets of experiments and the optimal parameters are *cells* = (7×7) ; *bins* = 18. Detailed results are shown in Table 2. To find the optimal parameters of the final descriptor, we fixed desirable parameters of the TWHOG or DMMHOG and optimize the parameters of the other one. As showed in Tables 3 and 4, two sets of parameters can achieve 100% recognition rate. Due to the complexity of TWHOG and DMM-HOG features both depend on the parameters of HOG. Therefore, to obtain same recognition rate (100%) at MSRGesture3D dataset with lower features dimension, we select *cells* = (8×8) ; *bins* = 16 for TWHOG and *cells* = (9×7) ; *bins* = 16 for DMM-HOG as desirable parameter.

Comparison of methods We compare our method with other nineteen competitive methods that were conducted on MSRGesture3D dataset with the same experimental setup. The recognition outcome of our method as well as nineteen existing methods is shown in the Table 6. It can be noted that our method performs best (100%) in terms of overall classification accuracy. Furthermore, our recognition result indicates that the fusion features, TSHOG, obtain higher discriminatory power compared with TWHOG or DMM-HOG feature.

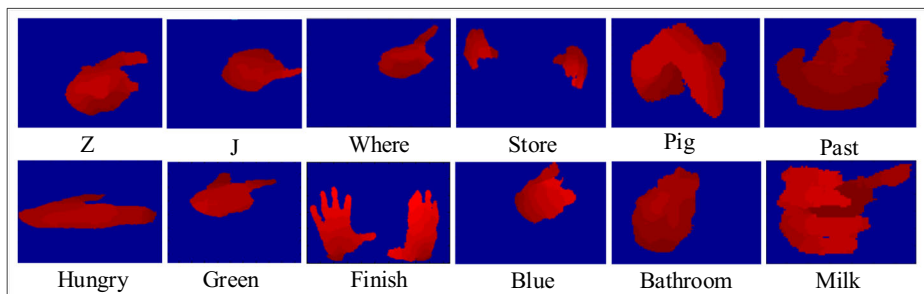


Fig. 6 Some example frames of MSRGesture3D dataset

Table 1 Recognition accuracy (%) of TWHOG with different parameters, on MSRGesture3D dataset

cells	7 × 7	7 × 8	7 × 9	8 × 7	8 × 8	8 × 9	9 × 7	9 × 8	9 × 9	mean
bins = 14	93.1	97.6	95.2	95.2	97.9	95.5	95.5	97.3	94.6	96.41
bins = 16	96.1	97.6	97.3	96.7	98.8	96.7	96.7	97.9	96.7	
bins = 18	96.1	97.9	95.2	94.6	97.9	96.1	96.4	97.0	95.5	

Table 2 Recognition accuracy (%) of DMM-HOG with different parameters, on MSRGesture3D dataset

cells	7 × 7	7 × 8	7 × 9	8 × 7	8 × 8	8 × 9	9 × 7	9 × 8	9 × 9	mean
bins = 14	94.6	95.2	94.6	94.3	93.7	94.0	94.6	94.0	95.2	94.68
bins = 16	94.0	93.4	94.0	94.3	93.4	94.0	94.3	93.1	94.9	
bins = 18	97.3	95.8	97.0	95.5	94.9	94.3	97.0	94.9	94.3	

DLEH² (a spatial-temporal descriptor, which is a fusion of multiple descriptors) [41] employ DLE to capture 3D structure and shape information of hands in detail, and employ HOG² (a spatial-temporal descriptor) to capture Spatial-temporal information. Then, fusing Spatial-temporal information and shape information for dynamic hand gesture recognition. DLEH² is more excellent than the previous method because both incomplete Spatial-temporal and shape information is considered in descriptor construction stage. Comparing with DLEH², TSHOG has three principle advantages. (1) HOG² uses the second order gradient to capture the spatial-temporal information and lead to information loss, yet TWHOG capture spatial-temporal information by the first order gradient can alleviate this problem. (2) In some complex situations (spatial-temporal variability and differences in hand configurations are in severe imbalances) the complementarity of our descriptors is obviously stronger than the DLEH² method. Detailed are shown in Fig. 7. (3) DMM-HOG can capture more effective shape information with lower vector dimension (DLE feature dimension above 10,000, while DMM-HOG feature dimension is about 3000).

Connection performance Classification results in detail is shown in Figs. 8 and 9.

- (1) Using TSHOG, all gestures, subjects are classified with 100% classification accuracy.
- (2) TSHOG perform better than TWHOG or DMM-HOG, which demonstrates that the information carried by the TWHOG and DMM-HOG is complementary in partial gestures (spatial-temporal variability and differences in hand configurations are in severe imbalances).
- (3) For TWHOG, eight out of twelve gestures in the MSRGesture3D dataset are classified with 100% classification accuracy, and “Where” gesture has the lowest recognition rate.

Table 3 Recognition accuracy (%) of TSHOG with different parameters of TWHOG (cells: 7 × 7, bins: 18 for DMM-HOG), on MSRGesture3D dataset

cells	7 × 7	7 × 8	7 × 9	8 × 7	8 × 8	8 × 9	9 × 7	9 × 8	9 × 9	mean
bins = 14	98.8	98.8	98.8	99.7	99.1	99.1	98.2	98.5	98.5	98.97
bins = 16	98.8	98.8	98.8	99.7	99.7	99.1	99.4	99.4	98.5	
bins = 18	98.8	98.8	99.1	99.1	99.1	99.4	98.2	98.8	98.8	

Table 4 Recognition accuracy (%) of TSHOG with different parameters for DMM-HOG (*cells: 8 × 8, bins: 16* for TWHOG), on MSRGesture3D dataset

cells	7 × 7	7 × 8	7 × 9	8 × 7	8 × 8	8 × 9	9 × 7	9 × 8	9 × 9	mean
bins = 14	99.4	99.7	99.1	99.7	99.7	99.7	99.1	99.7	99.1	99.49
bins = 16	99.4	99.7	99.1	99.4	100	99.7	100	99.1	99.1	
bins = 18	99.7	99.7	99.4	99.7	99.4	99.4	99.1	99.4	99.7	

Table 5 Recognition accuracy (%) of TSHOG (Optimal parameter setting) with different SVM parameters

svm_type	C = 1	C = 5	C = 10	C = 15	C = 20
C_SVC	99.7	99.7	99.7	99.7	99.7
Epsilon_SVR	99.4	99.7	99.7	99.7	99.7
Nu_SVR	100	100	100	100	100

This result indicates: i) TWHOG is effective in capturing motion patterns of different types of gestures. ii) It is necessary to introduce DMM-HOG to maximise the inter-class distance of partial gestures.

4.2 SKIG dataset and setup

SKIG dataset contains 2160 gesture sequences collected from six subjects (1080 RGB sequence and 1080 depth sequence, we only use the depth sequence part). All of these sequences are captured

Table 6 Comparison of the proposed method and other methods on MSRGesture3D dataset

Method	Accuracy(%)
Kurakin et al. [18]	87.7
HON4D (Oreifej and Liu [25])	87.29
Random Occupancy Patterns (Wang et al. [32])	88.5
DMM-HOG (Yang et al. [36])	89.20
Edge Enhanced DMM (Zhang and Tian [38])	90.5
HON4D + Ddisc (Oreife and Liu [25])	92.45
HOG2 (Ohn-bar and Trivedi [24])	92.64
Tran et al. [29]	93.31
Rahmani et al. [26]	93.61
DMM-LBP-FF (Chen et al. [3])	93.4
DMM-LBP-DF (Chen et al. [3])	94.6
DMM + KECA (Madany et al. [11])	94.44
Super Normal Vector (Yang et al. [35])	94.74
STPCM (Liang et al. [19])	97.15
HAGR-D (Santos et al. [27])	97.49
Depth Context (Liu et al. [20])	98.21
Chen et al. [5]	98.50
Multi-Temporal DMM-LBP (Jiang et al. [15])	98.80
DLEH2 (Zheng et al. [41])	99.10
DMM-HOG	97.3
TWHOG	98.8
TSHOG (TWHOG + DMM-HOG)	100

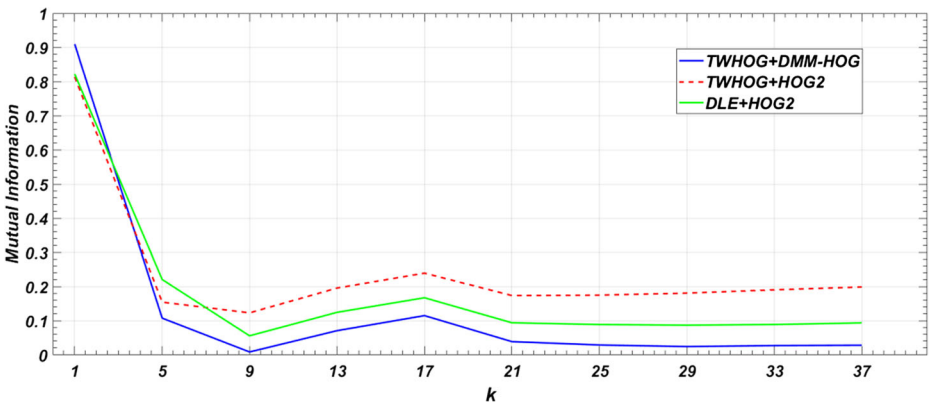


Fig. 7 Mutual Information using *Kraskov I* estimator. (our method with DLEH²)

synchronously with Kinect sensors (including RGB cameras and deep cameras). The data set collects a total of 10 gestures: “round (clockwise)”, “triangular (counterclockwise)”, “up down”, “right-left”, “wave”, “Z”, “cross”, “come here”, “turn-around”, and “pat”. In the collection process, all ten categories use three configurations: fist, index and flat. Using three different backgrounds (*i.e.* wood, white plain paper and paper with characters) and two lighting conditions (*i.e.* light and differential) sequence. We divided all sequences into three subsets: “subject1+subject2”, “subject3+ subject4”, and “subject5 + subject6”. The 3-fold cross-validation scheme as [21] is used to evaluate our method. The performance is calculated as the overall accuracy which is the ratio of the correctly recognized gestures over the total number of test sequences. Fig. 10 shows some example frames of this dataset.

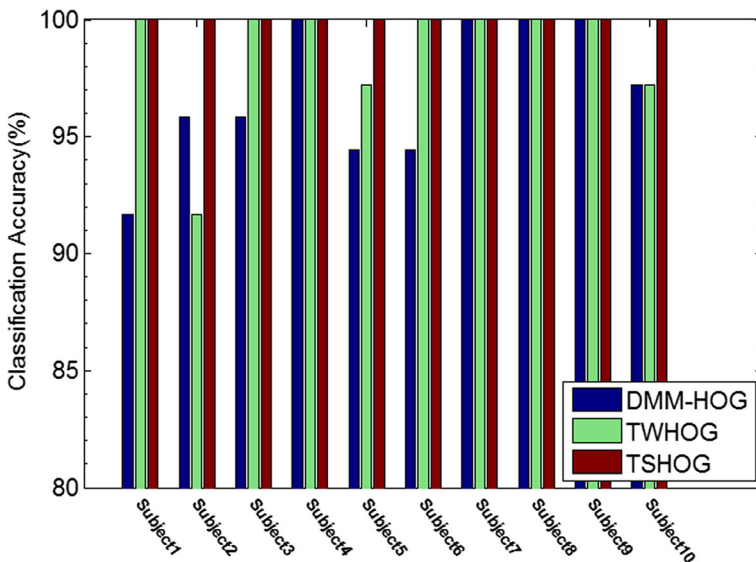


Fig. 8 Classification performance (recognition rates per gesture class) on MSRGesture3D dataset when use DMMHOG, TWHOG or TSHOG alone

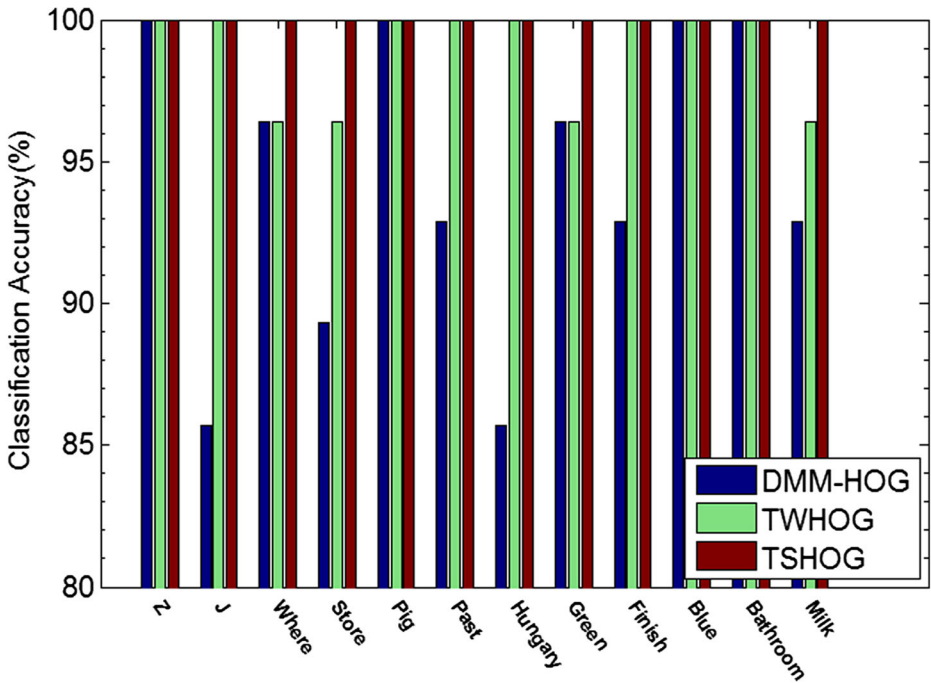


Fig. 9 Detailed classification results, by separately performing our 3 kinds of descriptors on MSRGesture3D dataset, for each fold in the leave-one-out cross-validation

Parameter setting As noted in [41], in our experiments SKIG dataset has the same resolution 240×320 with MSRAAction3D dataset. Hence the sizes for DMM_j , DMM_s and DMM_t are set to 102×54 , 102×75 and 75×54 , respectively. Experimental parameters are the *cells* and *bins* in the mHOG/HOG algorithm, which is used to construct the TWHOG and DMM-HOG descriptors in this paper.

For TWHOG, we conduct 27 sets of experiments and there are four sets optimal parameter: 1) $cells = (7 \times 7); bins = 16.2$ 2) $cells = (7 \times 7); bins = 18.3$ 3) $cells = (8 \times 7); bins = 16.4$ 4) $cells = (9 \times 7); bins = 18$. Detailed results are shown in Table 7. For DMM-HOG, we conduct 27 sets of experiments and get the optimal parameters: $cells = (8 \times 8); bins = 16$. Detailed results are shown in Table 8. To find the optimal parameters of TSHOG, we fixed desirable parameters of the TWHOG or DMM-HOG and optimize the parameters of the other one. As showed in Tables 9

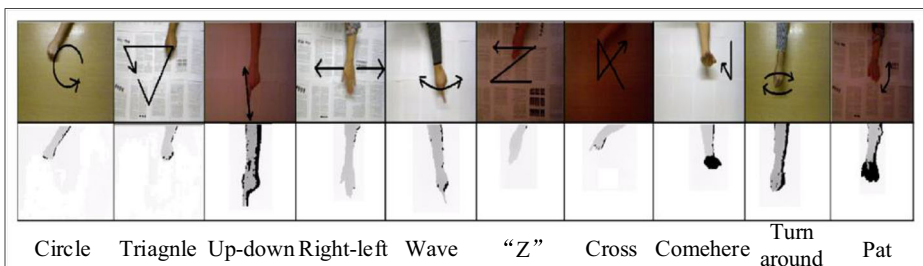


Fig. 10 Some example frames of SKIG dataset

Table 7 Recognition accuracy (%) of TWHOG with different parameters, on SKIG dataset

cells	7×7	7×8	7×9	8×7	8×8	8×9	9×7	9×8	9×9	mean
bins = 14	97.5	96.9	97.4	97.4	97.4	96.9	97.4	97.4	96.0	97.21
bins = 16	97.6	96.8	97.3	97.6	96.8	97.4	97.0	97.5	96.9	
bins = 18	97.6	97.1	97.4	97.3	97.1	97.3	97.6	97.3	96.8	

Table 8 Recognition accuracy (%) of DMM-HOG with different parameters, on SKIG dataset

cells	7×7	7×8	7×9	8×7	8×8	8×9	9×7	9×8	9×9	mean
bins = 14	87.3	88.4	88.1	89.2	90.6	89.1	87.1	86.9	86.6	88.90
bins = 16	88.7	89.7	90.2	90.8	91.6	89.4	88.4	88.8	87.7	
bins = 18	89.1	90.7	89.1	89.7	90.1	90.3	88.0	87.1	87.7	

Table 9 Recognition accuracy (%) of TSHOG with different parameters of TWHOG (*cells*: 8×8 , *bins*: 16 for DMM-HOG), on SKIG dataset

cells	7×7	7×8	7×9	8×7	8×8	8×9	9×7	9×8	9×9	mean
bins = 14	97.8	97.4	97.4	97.4	97.7	97.9	97.7	97.4	97.9	97.64
bins = 16	97.6	97.3	97.7	97.6	97.6	97.6	98.0	97.6	97.5	
bins = 18	98.1	97.5	97.4	97.9	97.2	98.0	98.1	97.8	97.3	

and 10, *cells* = (7×7); *bins* = 18 for TWHOG and *cells* = (7×7); *bins* = 16 for DMM-HOG are desirable parameters of TSHOG and obtain 98.7% recognition rate on SKIG dataset.

Comparison of methods The comparison of our method with several other competitive methods that with the same experimental setup, on SKIG dataset is shown in Table 11. TSHOG obtains the state-of-the-art accuracy of 98.7%, which outperforms all previous methods as showed on this table. In spite of only 0.9% better than the third-best performance method Multi-stream Recurrent Neural Network (MRNN) [23], our method is simpler than MRNN, which is based on Neural Network. Furthermore, MRNN is trained on multiple modalities, including depth, color and optical flow but TSHOG only used depth modality. In addition, compared with DMM-HOG or TWHOG alone as features, 7.1 and 1.1% improvements are achieved by TSHOG on SKIG dataset respectively.

Fusion performance The confusion matrix is shown in Fig. 11. It can be observed from the

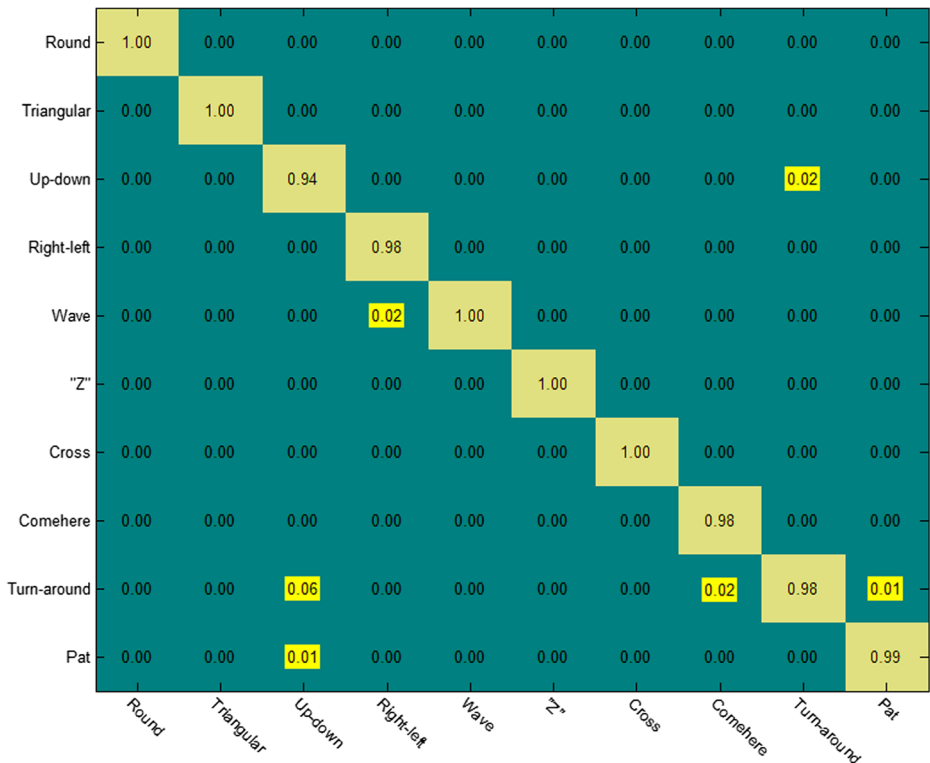
Table 10 Recognition accuracy (%) of TSHOG with different parameters for DMM-HOG (*cells*: 7×7 , *bins*: 18 for TWHOG), on SKIG dataset

cells	7×7	7×8	7×9	8×7	8×8	8×9	9×7	9×8	9×9	mean
bins = 14	98.3	98.1	98.1	98.1	98.1	98.1	97.8	97.9	97.9	97.96
bins = 16	98.7	98.2	97.9	98.5	98.0	97.9	97.8	97.2	97.4	
bins = 18	98.1	97.4	98.1	98.6	98.1	98.1	97.7	97.2	97.4	

Table 11 Comparison of proposed method and other methods on SKIG dataset

Method	Accuracy (%)
RGGP +RGB-D (Liu et al. [21])	88.7
Choi et al. [6]	91.9
4DCOV (Cirujeda et al. [7])	93.8
Depth Context (Liu et al. [20])	95.37
Tung et al. [30]	96.7
MRNN (depth only) (Nishida et al. [23])	95.9
MRNN (Nishida et al. [23])	97.8
DLE (Zheng et al. [41])	94.35
HOG ² (Zheng et al. [41])	94.72
DLEH ² (Zheng et al. [41])	98.43
R3DCNN (Gupta P M X Y S et al. [13])	98.6
DMM-HOG	91.6
TWHOG	97.6
TSHOG	98.7

confusion matrix that all the gestures are classified with above 94% classification accuracy. Five gestures (“Round”, “Triangular”, “Wave”, “Z”, “Cross”) are classified with 100% recognition accuracy. Classification performance in detail is shown in Figs. 12 and 13. As

**Fig. 11** The confusion matrix of proposed method on SKIG dataset

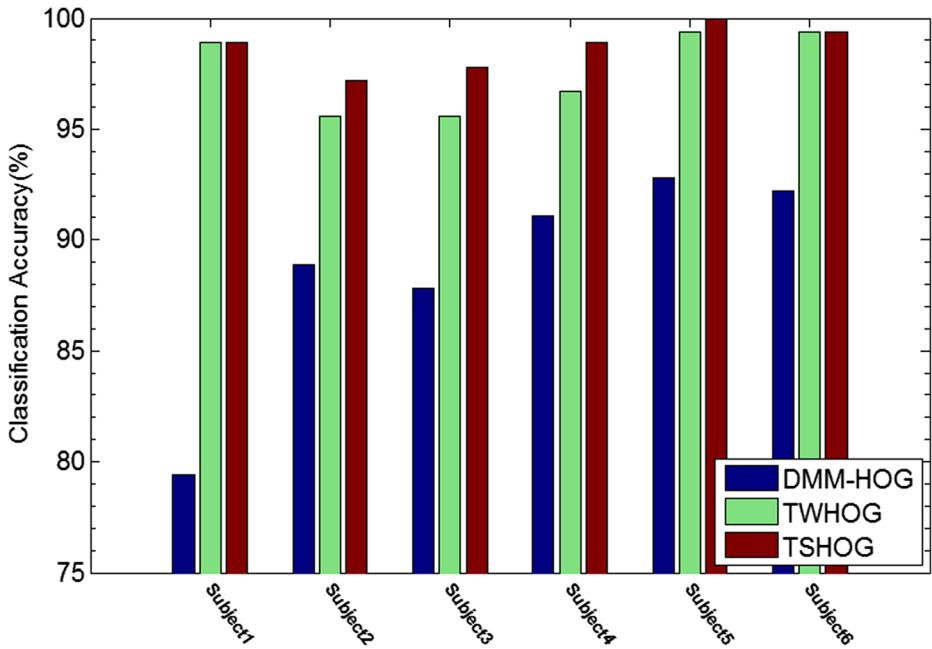


Fig. 12 Detailed classification results, by separately performing our 3 kinds of descriptors on SKIG dataset, for the 3-fold cross-validation

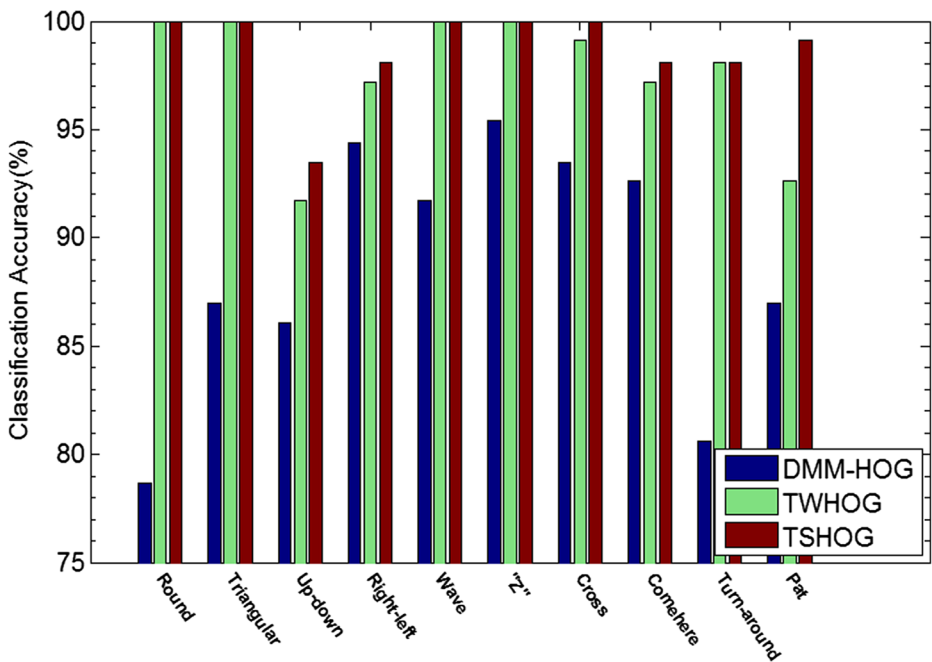


Fig. 13 Classification performance (recognition rates per gesture class) on SKIG dataset when use DMMHOG, TWHOG or TSHOG alone

mentioned above, each type of gesture in the SKIG dataset uses three different configurations (fist, index, flat), hence capturing shape information causes an unnecessary increase in the intra-class distance. This also explains that why fusing TWHOG and DMM-HOG cannot achieve significant improve on SKIG dataset. Furthermore, it can be seen that our method is robust against pose, illumination and background. Experimental results on SKIG dataset demonstrate that our method has good generalization ability.

5 Conclusion and future works

This paper has presented an effective feature descriptor, TSHOG, which can capture both the motion pattern and shape information. Specifically, TWHOG mitigates intra-class spatial-temporal variability by extracting the average spatial-temporal information in the space-time domain, DMM-HOG alleviates similar inter-class motion pattern by extracting subtle hand configuration differences between different types of dynamic gestures. We experimentally confirm the efficacy of our proposed approach using two different datasets (100% for MSRGesture3D, 98.7% for SKIG). And experimental results indicate that TSHOG is robust to spatial-temporal variability, illumination and background. The future work mainly focuses on two aspects: (1) enhance the discriminative power of TWHOG descriptor for the gestures with the similar motion. (2) Applying this proposed method to more extensive tasks, such as person ReID.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Ahmed W, Chanda K, Mitra S (2016) Vision based hand gesture recognition using dynamic time warping for Indian sign language [C]/Information Science (ICIS), International Conference on. IEEE 120–125
2. Baraldi L, Paci F, Serra G, et al. (2014) Gesture recognition in ego-centric videos using dense trajectories and hand segmentation [C]/IEEE COMPUTER SOCIETY CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION WORKSHOPS. IEEE
3. Chen C, Jafari R, Kehtarnavaz N (2015) Action recognition from depth sequences using depth motion maps-based local binary patterns [C]/Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on. IEEE 1092–1099
4. Chen C, Liu K, Kehtarnavaz N (2016) Real-time human action recognition based on depth motion maps [J]. *J Real-Time Image Proc* 12(1):155–163
5. Chen C, Zhang B, Hou Z et al (2017) Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features [J]. *Multimed Tools Appl* 76(3):4651–4669.30
6. Choi H, Park H (2014) A hierarchical structure for gesture recognition using RGB-D sensor [C]/Proceedings of the second international conference on Human-agent interaction. ACM 265–268
7. Cirujeda P, Binefa X 4DCov: a nested covariance descriptor of spatio-temporal features for gesture recognition in depth sequences [C]/3D vision (3DV), 2014 2nd international conference on. IEEE 2014(1):657–664
8. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection [C]/computer vision and pattern recognition, 2005. CVPR 2005. IEEE computer Society conference on. IEEE (1):886–893
9. Derpanis KG, Sizintsev M, Cannons KJ et al (2013) Action spotting and recognition based on a spatio-temporal orientation analysis [J]. *IEEE Trans Pattern Anal Mach Intell* 35(3):527–540
10. Dominio F, Donadeo M, Zanuttigh P (2014) Combining multiple depth-based descriptors for hand gesture recognition [J]. *Pattern Recogn Lett* 50:101–111

11. El Madany N E D, He Y, Guan L (2015) Human action recognition using temporal hierarchical pyramid of depth motion map and keca [C]//Multimedia Signal Processing (MMSp), 2015 IEEE 17th International Workshop on. IEEE 1–6
12. Fan RE, Chang KW, Hsieh CJ et al (2008) LIBLINEAR: a library for large linear classification [J]. *J Mach Learn Res* 9:1871–1874
13. Gupta PMXYS, Kautz KKSTJ (2016) Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks [C]. *CVPR*
14. Hasan H, Abdul-Kareem S (2014) RETRACTED ARTICLE: static hand gesture recognition using neural networks [J]. *Artif Intell Rev* 41(2):147–181
15. Jiang M, Jin K, Kong J (2018) Action Recognition Using Multi-Temporal DMMs Based on Adaptive Vague Division [C]//Proceedings of the 2018 International Conference on Image and Graphics Processing. ACM 8–13
16. Kim Y, Toomajian B (2016) Hand gesture recognition using micro-Doppler signatures with convolutional neural network [J]. *IEEE Access* 4:7125–7130
17. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information [J]. *Phys Rev E* 69(6):066138
18. Kurakin A, Zhang Z, Liu Z (2012) A real time system for dynamic hand gesture recognition with a depth sensor [C]//Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European. IEEE 1975–1979
19. Liang B, Zheng L (2015) Spatio-temporal pyramid cuboid matching for action recognition using depth maps [C]//Image Processing (ICIP), 2015 IEEE International Conference on. IEEE 2070–2074
20. Liu M, Liu H (2016) Depth context: a new descriptor for human activity recognition by using sole depth sequences [J]. *Neurocomputing* 175:747–758
21. Liu L, Shao L (2013) Learning Discriminative Representations from RGB-D Video Data [C]//IJCAI. 1: 3
22. Molchanov P, Gupta S, Kim K, et al. (2015) Hand gesture recognition with 3D convolutional neural networks [C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops 1–7
23. Nishida N, Nakayama H (2015) Multimodal gesture recognition using multi-stream recurrent neural network [C]//Pacific-rim symposium on image and video technology. Springer, Cham, pp 682–694
24. Ohn-Bar E, Trivedi M M (2013) Joint angles similarities and HOG2 for action recognition [C]//Computer vision and pattern recognition workshops (CVPRW), 2013 IEEE conference on. IEEE 465–470
25. Oreifej O, Liu Z (2013) Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences [C]//Computer vision and pattern recognition (CVPR), 2013 IEEE conference on. IEEE 716–723
26. Rahmani H, Mahmood A, Huynh DQ, Mian A (2014) Real time action recognition using histograms of depth gradients and random decision forests. In: *Applications of Computer Vision (WACV)*, 2014 IEEE Winter Conference on. IEEE 626–633
27. Santos DG, Fernandes BJT, Bezerra BLD (2015) HAGR-D: a novel approach for gesture recognition with depth maps [J]. *Sensors* 15(11):28646–28664
28. Shen X, Hua G, Williams L et al (2012) Dynamic hand gesture recognition: an exemplar-based approach from motion divergence fields [J]. *Image Vis Comput* 30(3):227–235
29. Tran Q D, Ly N Q (2013) Sparse spatio-temporal representation of joint shape-motion cues for human action recognition in depth sequences [C]//Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on. IEEE 253–258
30. Tung P T, Ngoc L Q (2014) Elliptical density shape model for hand gesture recognition [C]//Proceedings of the Fifth Symposium on Information and Communication Technology. ACM 186–191
31. Wang X, Xia M, Cai H, et al. (2012) Hidden-markov-models-based dynamic hand gesture recognition [J]. *Math Problems Eng*
32. Wang J, Liu Z, Chorowski J et al (2012) Robust 3d action recognition with random occupancy patterns [M]//computer vision–ECCV 2012. Springer, Berlin, pp 872–885
33. Wang L, Xiong Y, Wang Z, et al (2017) Temporal Segment Networks for Action Recognition in Videos [J]. *arXiv preprint arXiv:1705.02953*
34. Wold S, Esbensen K, Geladi P (1987) Principal component analysis [J]. *Chemom Intell Lab Syst* 2(1–3):37–52
35. Yang X, Tian Y L (2014) Super normal vector for activity recognition using depth sequences [C]//Proceedings of the IEEE conference on computer vision and pattern recognition 804–811
36. Yang X, Zhang C, Tian Y L (2012) Recognizing actions using depth motion maps-based histograms of oriented gradients [C]//Proceedings of the 20th ACM international conference on Multimedia. ACM 1057–1060
37. Yuan J, Liu Z, Wu Y (2011) Discriminative video pattern search for efficient action detection [J]. *IEEE Trans Pattern Anal Mach Intell* 33(9):1728–1743
38. Zhang C, Tian Y (2013) Edge enhanced depth motion map for dynamic hand gesture recognition [C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on. IEEE 500–505
39. Zhang C, Yang X, Tian Y L (2013) Histogram of 3D facets: A characteristic descriptor for hand gesture recognition [C]//Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE 1–8

40. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions [J]. *IEEE Trans Pattern Anal Mach Intell* 29(6):915–928
41. Zheng J, Feng Z, Xu C et al (2017) Fusing shape and spatio-temporal features for depth-based dynamic hand gesture recognition [J]. *Multimed Tools Appl* 76(20):20525–20544
42. Zhu Y, Chen W, Guo G (2015) Fusing multiple features for depth-based action recognition [J]. *ACM Trans Intel Syst Technol (TIST)* 6(2):18



Meng Xing received his bachelor degree from Hohai University. Now he is a master student in Tianjin University. Her research interests include Action Recognition, Machine Learning, and Human Computer Interaction.



Jin Hu received her Ph.D. at School of Computer Science and Technology, Tianjin University. She is currently a lecturer in Tianjin University. Her research interests lie in Pattern Recognitin, Services Computing, and Knowledge Management.



Zhiyong Feng received his Ph.D. at Tianjin University. He is currently a professor in Tianjin University. His research interests lie in Artificial Intelligence, Knowledge Engineering, and Services Computing.



Yong Su received his master degree from Xi'an University of Technology, China in 2014. Now he is a Ph.D. candidate in Tianjin University, Tianjin, China. His research interests include Pose estimation, Machine Learning, and 3D reconstruction.



Weilong Peng received his master degree from Tianjin University. Now he is a Ph.D. candidate in Tianjin University. His research interests include Face Recognition, Machine Learning, and 3D face reconstruction.



Jinqing Zheng received her master degree from Tianjin University. Now she is an Assistant Engineer in Institute of Automation Chinese Academy of Sciences. Her research interests include Machine Learning, image processing.