Multimedia Tools and Applications (2024) 83:43325–43347
https://doi.org/10.1007/s11042-023-17319-0

# A new weighted multi-scale descriptor for hand gesture recognition

**Beiwei Zhang**[1] · **Wen Ding**[1] · **JiaSheng Ye**[1]

Received: 29 November 2022 / Revised: 29 June 2023 / Accepted: 29 September 2023 /
Published online: 12 October 2023


**Abstract**
Image-based hand gesture recognition is a very challenging problem as the hand is a smaller object with complex articulations compared to the entire human body. It occupies a little portion in the image and is more easily affected by segmentation errors, and hence needs delicate description. This paper suggests a new weighted multi-scale feature descriptor (WMD) along the contour of the hand for robust hand gesture recognition using depth images. Firstly, the weight factor is estimated for each contour point by 2D Gaussian smoothing function and Prewitt operator to relate it with its neighbors and highlight its importance. Then the WMD descriptor is constructed via 1D left-side and right-side Gaussian smoothing considering the contour points are more sensitive than those inner points of the hand and depend on each other when used to recognize the gestures. Granularity of the descriptor is characterized by multiple scales with different standard deviations of the Gaussian function. And its invariants to translation, rotation and scaling transformations are proved theoretically and validated experimentally. Finally, extensive experiments on our self-established ten-gesture dataset and two public datasets have been carried out by comparing the proposed algorithm with three distance-based and two CNN-based hand gesture recognition methods. The encouraging results demonstrate that our method outperforms the others and achieves a good combination of accuracy (more than 95%) and computational efficiency (averaging 0.054s per frame).

**Keywords** Gaussian smoothing · Prewitt operator · WMD descriptor · Hand gesture recognition

## 1 Introduction

Human behavior analysis has attracted more and more interests in the field of artificial intelligence and machine learning in recent years. Generally, it can be classified into several categories including hand gesture recognition, human action analysis and

✉ Beiwei Zhang
zhangbeiwei@nufe.edu.cn

[1] School of Information Engineering, Nanjing University of Finance and Economics, Nanjing 210023, China

facial expression analysis. As an active research topic, the hand gesture recognition aims to identify the most perceptually similar hand gesture from its predefined hand gesture dataset. It has found many practical applications, such as augmentation reality, human computer interaction, automatic surveillance and some quietness-required environments [1–3], etc. One typical application is to use the recognized gesture as an efficient way to retrieve further information regarding the hand gesture. Each gesture in the dataset can have additional features such as its skeleton information and 3D shape, which can be associated with the query hand gesture and retrieved in real-time. This process eliminates the requirement of calculating those details from scratch and relaxes large amount of computational resources, especially in case of mobile phones.

Motivated by the widely possible application areas, many efforts have been devoted to the advances of hand gesture recognition. The reported work can be classified into different categories from different perspectives. According to the involved sensors, there are 2D RGB camera-based, wearable device-based and depth-sensor-based algorithms. As the RGB camera provides three basic color components of the video, the algorithms in the first category are typically affected by external environments such as illumination, skin color and cluttered background [4]. To overcome the problem of possible skin-like objects and avoid sensitivity to lighting conditions, Dardas et al. [5] described a module of skin detection and contour comparison algorithm for hand detection. The major limitations of this type of algorithms are the absence of 3D structure information and sensitive to color variations of human clothing or background, which obviously decreases their robustness and accuracy during ROI detection and segmentation. In the second category, the wearable devices such as accelerometers, magnetic trackers and data gloves, are involved in obtaining three-dimensional movement information at the granularity of the fingers for gesture recognition. For example, the dataset of the hand movements were captured by two DG5-VHand data gloves while data labeling was implemented with a camera to synchronize hand motion with their corresponding sign language words in [6]. And a comprehensive review work was provided in [7] where a variety of wearable sensing modalities for activity classification were investigated. In general, the merits in these strategies are low-complexity of data preprocessing and feature extraction while the demerits lie in that they are only suitable for handling some simple gestures. When the gesture becomes a bit complex, its recognition accuracy will be obviously reduced. What's more, the invisibility of the interface for the users is impeded which brings lots of inconvenience and cumbersome as lots of cables may be involved in some cases.

In the last category, an inexpensive depth camera e.g. the Kinect sensor or Intel RealSense, is usually used to collect visual information for the input of the algorithms for human activity recognition. Compared with RGB image, independent objects with depth data could be detected and segmented easily, and their shape structure could be estimated ignoring disturbance of complex background on the platform of Kinect sensor. Instead of wearing data gloves or any other auxiliary equipment, this type of algorithm enables a natural and uninvaded fashion of interaction during working. In this sense, the third category possesses the advantages of the other two. Therefore, more and more researches pay attention to this platform in recent years and the authors can be referred to [8–10] for a comprehensive review work. Based on the above analysis, this paper will employ the Kinect sensor to capture depth data as input for the suggested hand gesture recognition system.

## 2 Related work

Practical applications of Human behavior analysis have to meet some requirements including real-time performance, high recognition accuracy and robustness, etc. In the literature, researchers try to reach acceptable balance among those issues. Depending on different kinds of input data, the reported algorithms can be divided into skeleton-based algorithms and depth-based algorithms. The former uses 3D coordinates of the joints to represent the model of human full body and is suitable for human activity recognition. In Thanh and Chen [11], the discriminative pattern of skeleton data was extracted as local features and the key frames were constructed based on skeleton histogram to classify skeleton sequences in human action recognition. To improve the stability and recognition accuracy, the spatial-temporal descriptions from Kinect skeleton data are employed, e.g. the angular representation [12] and skeletal shape trajectories [13]. It is known that the skeleton information carries highly concise details and is more suitable for human body tracking. For a small object, such as a human hand which occupies a very small portion of the image with complex articulations, it is difficult to detect and segment as pointed out in [14]. In practice, this type of work also suffers from contour distortions since little noise or slight variations in the contour would severely perturb the topology of its skeletal representation. In this sense, the depth-based algorithms with more detailed depth information manifest its advantages in many situations. As human activity recognition depends on whole body parts, the hand gesture recognition is more computationally efficient with only data around the hand need to be handled.

According to the involved classifier, the reported algorithms mainly include distance-based algorithm, probability-based algorithm and CNN-based algorithm. The distance-based algorithm is early employed for human behavior analysis and dynamic time warping is the most used technique [15]. Another approach is to use SVM and multiclass SVM as shown in [16]. The probability-based algorithm is a statistical model and the classifier of HMM with Markov assumption is often used as in [17, 18]. However, it is difficult to define proper hidden and observing states for those gestures as they are formed by a complex intersection of different features or joints. Lastly, CNN-based algorithm essentially is a machine learning technique such as convolutional neural network and Recurrent Neural Network [19–22]. The advantage of the machine learning technique lies in that it is able to extract hierarchical features automatically via convolution and pooling operation to hold more abstract knowledge. This avoids the process of delicate feature engineering. The disadvantage is that the extracted features lack of specific physical meaning, so it is difficult to visualize and analyze their characteristics. During training the deep network, carefully tuning the hyper-parameters is needed as well as the decision of number of hidden layers, the right number of neurons to use in the hidden layers, and strategy for preventing overfitting. In addition, the computational complexity is high which limits its real-life applications. For small human action recognition datasets, the machine learning methods may not provide satisfactory performance.

It is known that the semantic meaning of a hand gesture is delivered by its movement or shape. Different hand gestures are mainly differentiated by relative postures of the hand and fingers as well as their contour shapes. Ren et al. in [23] estimated the contour as time series curve to characterize the Euclidean distances between the hand contour and the palm center, where the key issue is to choose a starting and ending point of the curve. Since then, various physical features for the gesturing hand have been suggested ( [24–27] to site a few), among which He et al. [26] proposed an improved local sparse representation algorithm and

Wang [27] constructed the features for gesture recognition with peak values as well as valley values from the trend of slope difference distribution of the contour points. As showed by Wang, the stability and accuracy for calculating the peak and valley values depend heavily on the quality of the contour since the first and second derivative operation are involved.

The feature engineering is a key step towards human behavior analysis and different kinds of features have been reported including physical features and statistical features. Kim et al. [28] proposed an adaptive local binary pattern from depth images for hand tracking. In [29], the finger-lets, stroke-lets or other characteristics were extracted from its depth information. Calado et al. [30] suggested a geometric model-based approach to gesture recognition which supports the visualization and physical interpretation of the recognition process. As a statistical tool, the 3D histograms of textures from a sequence of depth maps were computed for gesturing hand descriptor in Zhang and Yang et al. [31]. In their work, the depth sequences were first projected onto three orthogonal Cartesian plane to form three projected maps, then the sign-based, magnitude-based and center-based descriptor salient information were extracted respectively. Similarly in Reza et al. [32], the weighted depth motion map was proposed to extract the spatiotemporal information by an accumulated weighted absolute difference of consecutive frames and the histogram of gradient and local binary pattern were exploited for the feature descriptor.

Imported from SIFT in computer vision which helps in reliable matching between different views of the same object, the concept of multi-scale features for gesturing hand is employed recently [33–36]. Huang and Yang in [34] suggested a multi-scale descriptor considering the area of major zone, length of major segment and central distance within different sizes of circles along the contour points. In their method, it is important to choose a proper scaling number and a starting point to align all points on the shape contour. The redundancy exists severely between different scales of features as the group of circles overlap with each other. Instead of employing multiple circles, the Euclidean distance between the centroid of the shape and the furthest point on the contour of the shape was used as the radius of the minimum circumscribed circle and then the circle region was partitioned into several bins using concentric circles and equal angle intervals in Lazarou et al. [35]. Another kind of feature descriptor is proposed in Sahana et al. [36] which calculated the number of peaks for each circular signature considering the ROI centroid as the center of those multi-radii circles. Obviously, this type of feature is sensitive to viewpoints of the sensor as the estimated area of the hand region as well as length value vary abruptly for some viewpoints and the computational complexity is considerably high which will weaken the performance of subsequent hand gesture classifier.

In this work, we will aim at a comprehensive, robust and discriminative feature for hand gesture recognition. Here a new weighted multi-scale feature descriptor (WMD) is suggested considering both 2D neighborhood and 1D contour curve within the depth image. With the segmented hand ROI region, the weight factor is estimated for each contour point by 2D Gaussian smoothing function and Prewitt operator to relate it with its neighbors and highlight its importance. Then the feature descriptor is constructed via 1D Gaussian smoothing considering the contour points in the hand should not be independent from each other when used to recognize the gestures. Granularity of the descriptor is characterized by multiple scales with different standard deviations of the Gaussian function. And its invariants to translation, rotation and scaling transformations are proved theoretically and validated experimentally. Compared with those descriptors reported in the literature, the WMD descriptor is a contour-emphasized descriptor. Extensively experiments on our ten-gesture dataset and two public dataset have been carried out comparing the proposed algorithm with three feature-based and two CNN-based hand gesture recognition methods.

The results show that our method outperforms the other methods and provides a good combination of accuracy and computational efficiency for real-time applications.

The remainder paper is structured as follows. The framework for the hand gesture recognition is introduced in Section 3. Section 4 elaborates the WMD descriptor and its invariants. Section 5 presents some experimental results and analysis for both the WMD descriptor and the recognition framework. Finally, this work is concluded briefly in Section 6.

## 3 System framework

The system framework for the hand gesture recognition system on the whole can be separated into three components: image preprocessing, feature extraction and pattern recognition. The Kinect sensor is employed to capture depth images as input for the system, which visualizes depth information from the sensor to the concerned gesturing hand. For each depth image, the ROI of gesturing hand and its contour points are supposed to have been segmented and extracted by the solution suggested in Dominio et al. [37]. Then the weighted multi-scale feature descriptor is constructed with the weights of contour points and ratios of Gaussian smoothing along the contour points. To test the performance of the WMD descriptor, the classification modules with Hausdorff Distance, Dynamic Timing Warping (DTW) distance and SVM model are respectively trained and employed to recognize hand gestures with the descriptor as input data. The framework of the proposed system is briefly shown in Fig. 1.

## 4 Construction of WMD descriptors

### 4.1 Gaussian scale space

The scale-space theory has been established as a well-founded and promising multiresolution technique in image structure analysis for 2D, 3D and time series. The basic idea is to embed the original signal into a one-parameter family of gradually smoothed signals, in which the fine scale details are successively suppressed with increasing scales, just as what the SIFT algorithm does. The multiresolution technique agrees with what human's eyes do when identifying an object from far away to near. Therefore, constructing a multi-scale descriptor for the recognition of the object is an important tool for global and local feature extraction in the field of computer vision.

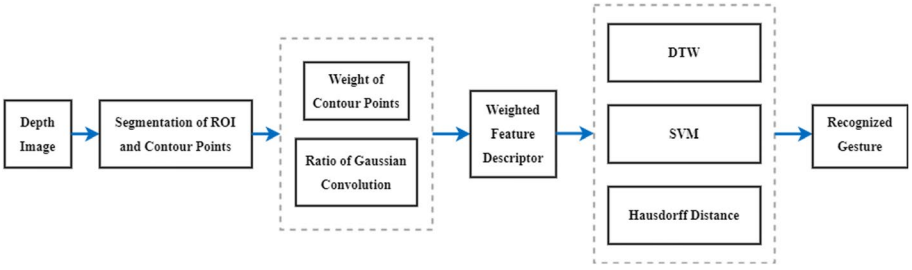



**Fig. 1** Framework of the hand gesture recognition system

In this work, a multi-scale descriptor is constructed for the depth image of the gesturing hand via Gaussian smoothing operation with different standard deviations since it is a scale-invariant kernel function. The original depth image has the most detailed information and its features will be anti-pyramidally encoded less and less with the increase of standard deviations to simulate large-scale global characteristics. In other words, the coarse feature is extracted when large standard deviation is used as the curve of the Gaussian function becomes smooth and fine feature is obtained corresponding to a small deviation. Considering different types of hand gestures exhibit different poses and they may exist some degree of similarity, the range of standard deviation is set as $\sigma \in (0, 0.5]$ and n scales as $\sigma_1 = 2^{-1}, \sigma_2 = 2^{-2}, \cdots, \sigma_n = 2^{-n}$.

### 4.2 Weights of contour points

It is known that the extracted contour points play an important role in hand gesture recognition and their discriminant information can be checked by considering their neighbors. The nearer the neighbors are, the higher influences they have. This phenomenon can be pictured by difference of Gaussian smoothing with the contour point as its center. Next, we will analyze the information of a contour point in two-dimensional space and assign its weight via sigmoid function.

Assuming the sequence of contour points makes up a closed curve, and it can be parameterized in complex domain as

$$s(t) = x(t) + i \cdot y(t) \tag{1}$$

where $i = \sqrt{-1}$ and $t \in [0, 1)$ is its normalized index.

According to the definition of Prewitt operator, two m × m masks denoted by $P_x(m)$ and $P_y(m)$, can be generated along x-axis and y-axis whose elements at r-th row and c-th column be formulated as

$$P_{x,r,c}(m) = \begin{cases} 1 & c < {}^{m}/_{2} \\ 0 & c = {}^{m}/_{2} \\ -1 & c > {}^{m}/_{2} \end{cases} \text{ and } P_{y,r,c}(m) = \begin{cases} 1 & r < {}^{m}/_{2} \\ 0 & r = {}^{m}/_{2} \\ -1 & r > {}^{m}/_{2} \end{cases} \tag{2}$$

Let $G_\alpha(m)$ be a m × m matrix generated by 2D Gaussian function with α standard deviation. Then two kernel templates can be respectively defined as

$$K_x(\mathrm{m}) = G_\alpha(m) * P_x(m) \tag{3}$$

and

$$K_y(\mathrm{m}) = G_\alpha(m) * P_y(m) \tag{4}$$

where $*$ represents the convolution operation.

Given an arbitrary contour point $s(t)$, from the depth image, it is easy to construct the $s(t)$-centered block with size m × m. Let it be $\mathrm{B}(x, y)$. We have

$$\mathrm{d_x} = \mathrm{K_x(m)} * \mathrm{B(x, y)} \tag{5}$$

and

$$d_y = K_y(\mathrm{m}) * \mathrm{B}(x, y) \tag{6}$$

In Eqs. (5) and (6), the Gaussian function acts as a weighted smoothing operation and the Prewitt operator as a derivative operation with the value of m is set as 3 in this work. Therefore, $d_x$ and $d_y$ carry the information of the contour point along x-axis and y-axis. We can define its weight factor via sigmoid function as

$$\mathrm{W}(t) = \frac{1}{1 + e^{-A}} \tag{7}$$

where $A = \sqrt{d_x^2 + d_y^2}$ is the amplitude value. In general, the larger the difference of the contour point and its neighbors, the more information it provides and hence the heavier its weight factor is. Equation (7) agrees with this observation.

### 4.3 Invariant of ratio gaussian smooth function

This section talks about Gaussian smoothing operation along the curve of contour points in one-dimensional Gaussian space and shows its invariant to affine transformation.

For an arbitrary contour point $s(t)$, its left-side and right-side neighbors with $w$ as the window size can be represented by

$$S_L = \{s(t + l)| - w \leq l < 0\} \tag{8}$$

and

$$S_R = \{s(t + l)|0 < l \leq w\} \tag{9}$$

Taking $s(t)$ as the starting point, its left and right sequences of vectors can be constructed with those points in $S_L$ and $S_R$ as the ending points, i.e. $s(t + l) - s(t)$. The sequences of vectors would illustrate the shape variation along the contour of the gesturing hand. It is obviously different for different hand gestures or different contour point. Compactly, their intersection angles can be used to depict the shapes of hand gestures in a geometric meaning. We will next construct a robust feature descriptor with these vectors.

Let $g_\sigma(w)$ be one-dimensional w-length vector generated by Gaussian function with standard deviation $\sigma$. We obtain

$$\delta_L(t, \sigma) = \sum\nolimits_{-w \leq l < 0} g_\sigma(l)(s(t + l) - s(t)) \tag{10}$$

$$\delta_R(t, \sigma) = \sum\nolimits_{0 < l \leq w} g_\sigma(l)(s(t + l) - s(t)) \tag{11}$$

where $\delta_L(t, \sigma)$ and $\delta_R(t, \sigma)$ respectively represent the left and right Gaussian-weighted mean vectors at the contour point $s(t)$ under Gaussian function with standard deviation $\sigma$.

With the consecutive value of $t$, the above two equations smooth the left and right in the sliding windows by Gaussian function. This operation agrees with the fact that different neighbors have different effects on the results and suppresses Gaussian noise in the data at the same time. From Eq. (1), they are expressed in the complex space consisting of real part and

imaginary part. The real part of normalized $\delta_L(t,\sigma)$ carries its amplitude and so does $\delta_R(t,\sigma)$. From Eqs. (10) and (11), their ratio is defined as

$$z(t,\sigma) = \frac{\delta_L(t,\sigma)}{\delta_R(t,\sigma)} \tag{12}$$

**Property 1** *The equation in* (12) *is invariant to scaling, translation and rotation transformation.*

***Proof*** Let, $T$, and $R$ respectively denote the scaling, translation and rotation transformation. Without loss of generality, let and. Let be the transformed version of. Next, we will show that the value of Eq. (12) is equal to each other before and after the above transformation.

From (10), we have

$$\begin{aligned}\delta'_L(t,\sigma) &= \sum\nolimits_{-w\le l<0} g_\sigma(l)\big(s'(t+l) - s'(t)\big)\\ &= \sum\nolimits_{-w\le l<0} g_\sigma(l)(\lambda(Rs(t+l)+T) - \lambda(Rs(t)+T))\\ &= \lambda R \sum\nolimits_{-w\le l<0} g_\sigma(l)(s(t+l) - s(t)) = \lambda R\delta_L(t,\sigma)\end{aligned}$$

Similarly,

$$\begin{aligned}\delta'_R(t,\sigma) &= \sum\nolimits_{0<l\le w} g_\sigma(l)(s'(t+l) - s'(t))\\ &= \lambda R\delta_R(t,\sigma)\end{aligned}$$

Then we have

$$z'(t,\sigma) = \frac{\delta'_L(t,\sigma)}{\delta'_R(t,\sigma)} = \frac{\lambda R\delta_L(t,\sigma)}{\lambda R\delta_R(t,\sigma)} = \frac{\delta_L(t,\sigma)}{\delta_R(t,\sigma)} = z(t,\sigma)$$

Therefore, we reach the conclusion that the equation in (12) is invariant to scaling, translation and rotation transformation.

### 4.4 Weighted multi-scale descriptor

In Eq. (12), the amplitude of $z(t,\sigma)$ is equal to the intersection angle between $\delta_L(t,\sigma)$ and $\delta_R(t,\sigma)$, which can be calculated by inverse cosine function of its real part as what follows

$$\theta(t,\sigma) = \text{acos}(\text{real}(z(t,\sigma))) \tag{13}$$

Equation (13) compactly visualizes the geometric relationship between the left and right sequences of vectors at$s(t)$. With consecutive values of $t$, it reveals the evolution of the shapes and curvatures along the contour points corresponding with different scales. Therefore, it can be used as a description for the gesturing hand. To limit the ranges, the cosine value of $\theta(t,\sigma)$ is employed for constructing the descriptor. Considering the weight for each contour point given in Section 4.2, we can define

$$f(t) = W(t) * \left[\text{real}\big(z\big(t,\sigma_j\big)\big)\right]\ j = 1,2,\cdots,n \tag{14}$$

where $f(t)$ collects the weighted feature values in the scale space. Finally, for all the contour points, we have

$$F = \left[f(t)\right] 0 \leq t < 1 \tag{15}$$

Equation (15) gives the weighted multi-scale descriptor (WMD) for a gesturing hand. For a hand depth image with contour points of number N, the dimension of its stacked descriptor is $N * n$. From the **Property 1** given in subsection 4.3, it is invariant to scaling, translation and rotation transformation while encodes all information for the contour points extracted together with their 1D and 2D neighbors.

### 4.5 Algorithm for hand gesture recognition

The major parts in the algorithm of hand gesture recognition consist of the estimation of weights of contour points and then construction of the weighted multi-scale descriptor. To test the performance of the descriptor, three tools of similarity measurements, i.e. dynamic time warping (DTW), Support Vector Machine (SVM) and Hausdorff distance, are involved and compared as the recognition engine. Summarily, the suggested algorithm is shown in algorithm 1 where the functions of GetWeight and GetWMDescriptor respectively represent the procedure of estimating the weights and weighted multi-scale descriptors given in subsection 4.2 and subsection 4.4.

## 5 Experiments and analysis

In this section, we validate the performance of the proposed WMD descriptor and hand gesture recognition system in three aspects: (1) demonstrating the robustness of our descriptor to affine transformations; (2) showing the influences of different scales and different number of contour points on the descriptors; (3) testing the accuracy of hand gesture recognition with the descriptor and compared with state-of-the-art methods under three different datasets.

```
Input: Depth image, ContourPoints, Path of Model Files, number of Classes
Output: Possible recognized gesture
Function HandGestureRecognition(DepthImage, ModelFilePath)
1: model=LoadModel(ModelFilePath)
2: Similarity=[]
3: For i=0 to numClasses
4:    W=GetWeight( Depth image, ContourPoints)
5:    WMD=GetWMDescriptor(W, ContourPoints, Scales)
6:    Similarity[i]=EstimateSimilarity(WMD, Model)
7: EndFor
8: Return max( Similarity)
9: EndFunction
```

**Algorithm 1** Hand gesture recognition procedure

### 5.1 Experimental datasets

**Self-established dataset** Fig. 2 shows the 10 kinds of hand gestures to be recognized in this work, respectively represent the ten digital numbers ranging from zero to nine, from left to right and up to down. Ten students were invited to perform these gestures before the Kinect sensor to collect their depth images. To promote the variety and representative of those samples, the students stood at about three different positions, say 80 cm, 120 and 150 cm considering the effective range of the sensor. Their hands were suggested to be placed in the front of their body for the ease of visualization of the sensor and hand region segmentation. Each kind of hand gestures was repeated 20 times by one person. In this way, the experimental dataset contained a total of 2000 samples for both training and testing. It should be noted that the last row in Fig. 2 gives their segmented hand region.

**NTU dataset** For comparison, we use the challenging public NTU hand gesture dataset where the hand gestures are collected by Kinect sensors. This dataset was collected from 10 subjects and includes 10 gesture classes. Each subject performed the same gesture in 10 different poses, thus the dataset had 10(people) × 10(gestures) × 10(poses) = 1000 samples. This is a very challenging real-life dataset with cluttered backgrounds. Moreover, the samples of the same gesture class had variations in hand orientation, scale, articulation, etc. The 10 kinds of hand gestures with corresponding shape samples are shown in Fig. 3.

**Senz3D dataset** The Creative Senz3D dataset is performed by four different people and each with 11 different gestures repeated 30 times. A group of samples for those gestures defined from G1 to G11 are given in Fig. 4. In total, it contains 1320 gesture samples. For each sample, color, depth, and confidence frames are available with the resolutions of 640×480, 320×240 (short 16 bit) and 320×240 (short 16 bit) respectively.

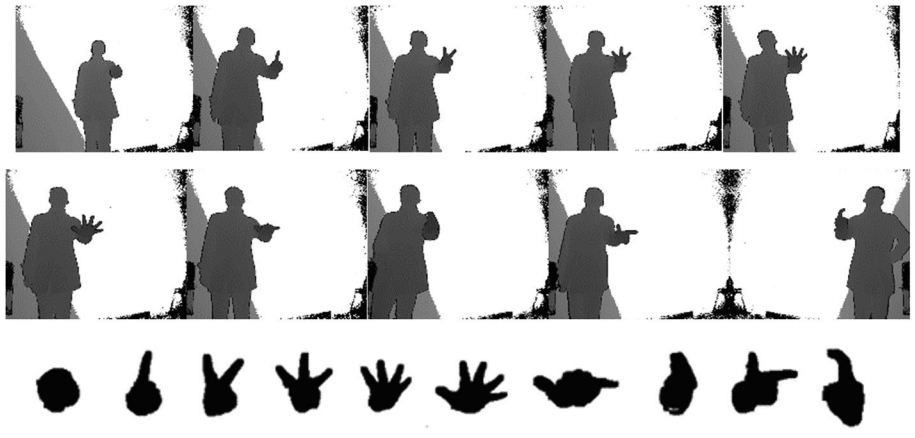


**Fig. 2** One group of depth images for ten types of gestures with extracted Hand Regions given in the last row

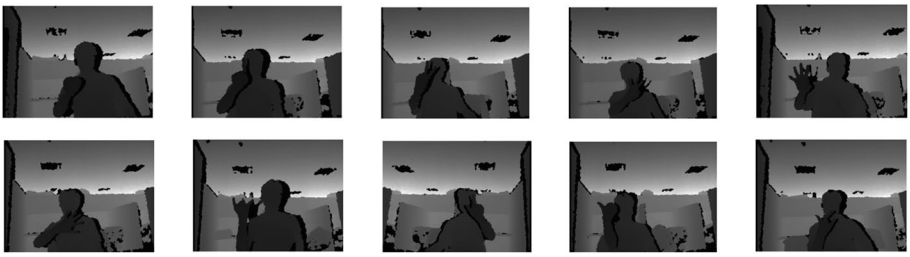


**Fig. 3** Samples from the public NTU hand gesture dataset

**Fig. 4** Samples from the public Senz3D dataset

### 5.2 Alignment of starting point

A stable starting point chosen from the sequence of contour points is crucial for the performance of the WMD descriptor since the classifier needs an aligned version. In general, the farthest point to the centroid of the hand ROI is used as the starting point. But it is sensitive to the rotation transformation of the ROI. There are some researchers employ auxiliary equipment, e.g. wearing a black belt on the wrist, to provide landmark information. In this work, we take a natural and uninvaded way for determining the starting point. Firstly, the major orientation of ROI is estimated as

$$\theta_o = \frac{1}{2} atan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right) \tag{16}$$

where $\mu_{11}$, $\mu_{20}$ and $\mu_{02}$ denote the 2nd order central moments. Then the starting point can be defined as the intersection point of the orientation line with the contour points. Figure 5 showed the estimated orientation denoted by the red arrowed lines. To show their invariance to rotation transformation, the orientations were re-estimated with the rotated hand images. The original images together with 10-degree, 20-degree and 30-degree versions were presented from left to right in Fig. 5. It can be seen that the estimation of orientations is considerably stable and the intersection points are located at the same positions.

### 5.3 Invariances of WMD descriptor

A well-behaved shape descriptor should have the tolerance of rotation, translation, scaling, noise and small degrees of deformation of the shape. It is known that the

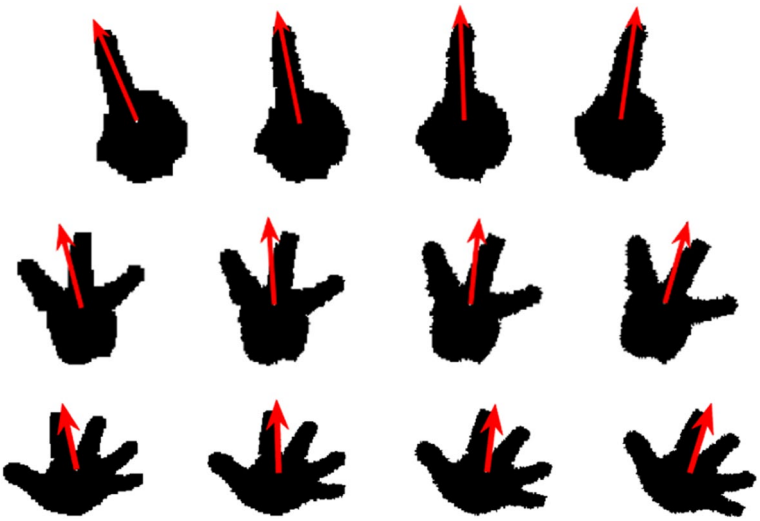


**Fig. 5** Major orientation estimated (denoted by Red arrowed lines) with hand ROI for the original and rotated images

translation transformation has no influence on the pose and relative positions of the gesturing hand as well as those contour points since it is just a pure shift of different regions in the depth images. The corresponding descriptors are obviously invariant to translation transformation. Therefore, this experiment is carried out to validate the invariant of rotation and scaling transformation. For clear demonstration, Fig. 6 gave the weighted multi-scale descriptors in three different standard deviations of Gaussian smoothing under four scaling-transformations (including the original one), where the first column represented the same hand gesture in different transformations and the second column illustrated their WMD descriptors. In these figures, the lines in red, green and blue were from the cases of $n$=-6, $n$=-5 and $n$=-3. Although the gesturing hand was heavily zoomed, strong similarities were observed among the corresponding plots in each column, which verified the robustness and invariants of the suggested descriptor.

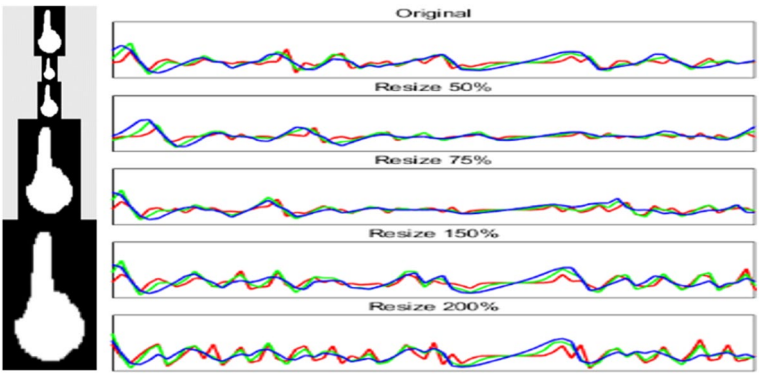



**Fig. 6** WMD descriptors) under scaling transformations (The lines in different colors are from different standard deviation of Gaussian functions)

The lines from the first two scales (in red and green) were close to each other, and the third one ( in blue) was a bit more distorted since it came from a larger standard deviation for Gaussian smoothing operation with less difference of weights in the neighbors. The quantitative evaluation was implemented as well and the KL-divergence values calculated between the transformed version and the original one were summarized in Table 1. From this table, we can find that the values of KL-divergence between the corresponding plots of are all very small numbers which quantitatively verifies that the descriptor is invariant to this transformation.

As to the rotation transformation, four different rotation angles, i.e. $\pm$ 15 and $\pm$ 30 degrees, were respectively applied to the original image as shown in the first column of Fig. 7, where the WMD descriptor was extracted for each image and illustrated in the second column. It can be seen that there exist high correlations among these figures. Similarly, the quantitative evaluation was implemented as well and the KL-divergence values summarized in Table 2. It is observed that the KL-divergence between the corresponding plots falls in a very small range which quantitatively verifies that the descriptors are invariant to these rotation transformations.

**Table 1** KL-divergence for the scaling transformations

| Scales | Resize 50% | Resize 75% | Resize 150% | Resize 200% |
|---|---|---|---|---|
| Scale $n$=-6( Red) | 0.0834 | 0.0571 | 0.0926 | 0.1047 |
| Scale $n$=-5(Green) | 0.0902 | 0.1004 | 0.0639 | 0.0836 |
| Scale $n$=-3(Blue) | 0.0963 | 0.1108 | 0.0858 | 0.0772 |

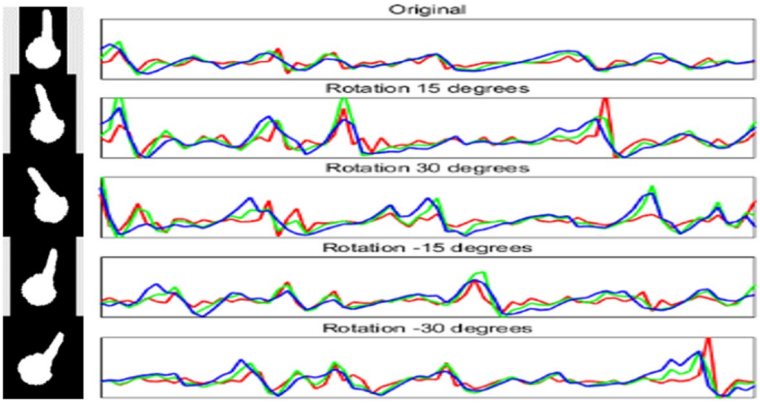



**Fig. 7** WMD descriptors under rotation transformations (The lines in different colors are from different standard deviation of Gaussian functions)

**Table 2** KL-divergence for the rotation transformations

| Scales | Rotation 15$^o$ | Rotation 30$^o$ | Rotation $-$ 15$^o$ | Rotation $-$ 30$^o$ |
|---|---|---|---|---|
| Scale $n$=-6( Red) | 0.1014 | 0.0462 | 0.0876 | 0.0981 |
| Scale $n$=-5(Green) | 0.1242 | 0.0819 | 0.0750 | 0.1004 |
| Scale $n$=-3(Blue) | 0.0754 | 0.0971 | 0.0864 | 0.1358 |

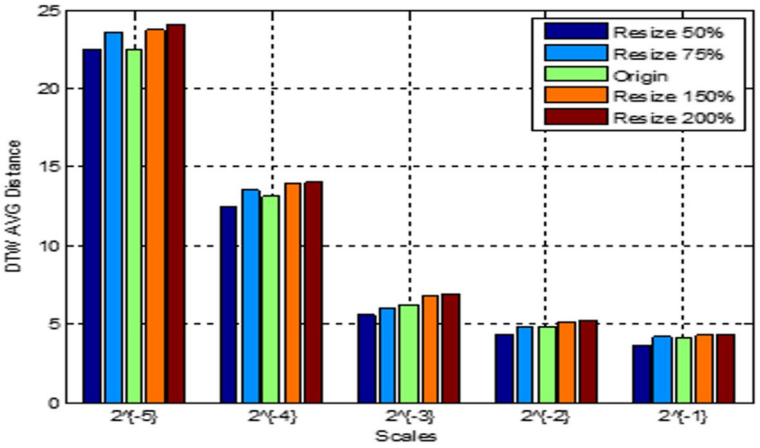



**Fig. 8** DTW avg. distance VS scaling transformation (The values are close to each other for the same scale and vice versa)

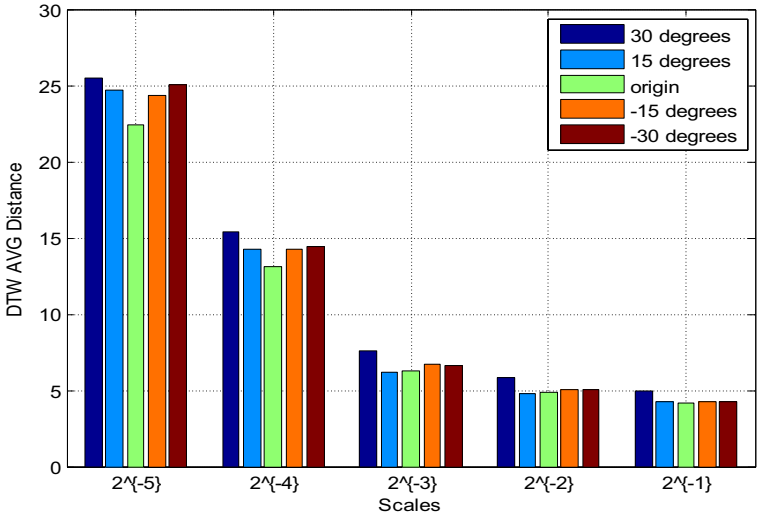



**Fig. 9** DTW avg. distance VS rotation transformation (The values are close to each other for the same scale and vice versa)

For further quantitative validation of the robustness of WMD descriptors, the similarities between the descriptor of transformed gesturing hand and those from the training dataset were estimated by dynamic time warping. The averaged values of the accumulated distances for scaling and rotation transformations were illustrated respectively in Figs. 8 and 9, where five scales in the scale space were taken here and different colors represented different transformations as given in the legend of each figure. It is observed that highly similar results are obtained in each group of transformation and obviously different from each other for different scales. This demonstrates that the suggested WMD

descriptor is also robust and invariant to those transformations in terms of DTW accumulated distance and provides high discriminative capacity.

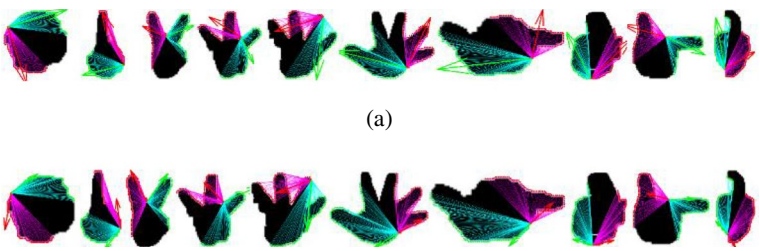
### 5.4 Effects of different scales and sliding windows

Mathematically, the Gaussian function exhibits different shape for different standard deviations. The larger the deviation is, the wider the curve and the lower its peak are. When performing Gaussian smoothing, the distribution of the weights will be approximately even. Therefore, the corresponding Gaussian smoothed image will become more and more blurred and lead to large-scale WMD descriptor finally. This is a simulation of observing an object from far away to capture its global features. To visualize its influence on the descriptors, their amplitudes were estimated following Eq. (13). The results were given in Fig. 10a and b corresponding to the cases of $\sigma = 2^{-1}$ and $\sigma = 2^{-4}$ where the arrowed green bold-line and red bold-line respectively represent the Gaussian-smoothed mean vectors for the left and right half parts given one contour point. On the whole, the amplitudes get smaller and smaller with the increase of standard deviations. This agrees with the phenomenon of human vision that the object seems to be big when near and small when far.

Similarly, Gaussian smoothing with the same standard deviation on different sizes of sliding windows will have different influence on the results of Gaussian-smoothed mean vectors as well as the WMD descriptor subsequently. The descriptor corresponding to smaller size of sliding windows is equivalent to observe it closely and focus on details of gesturing hand. Figure 11a and b presented the results with the sliding windows of 50-point and 100-point sizes under $\sigma = 2^{-1}$. It is observed that the amplitudes with shorter sliding windows are obviously larger than those with longer ones, which means that the curved surface is flattened and coincides with what we expect.

### 5.5 Parameter sensitivity

As showed above, different standard deviations and sizes of sliding windows will have different effects on the Gaussian-smoothed mean vectors and subsequently the WMD descriptors as well as the performance of gesture recognition algorithms. Here, two experiments were carried out to find the optimal balance for those parameters, where half samples in our self-established dataset were randomly selected for training and the remainders for testing. Three different classifiers for the recognition of hand gestures, i.e. SVM, DTW and HSDF,



(a)

(b)

**Fig. 10** Gaussian-smoothed mean vectors in cases of **a** $\sigma = 2^{-1}$ and **b** $\sigma = 2^{-4}$. The amplitudes in (**a**) are less than those in (**b**) corresponding to large-scale descriptor and vice versa

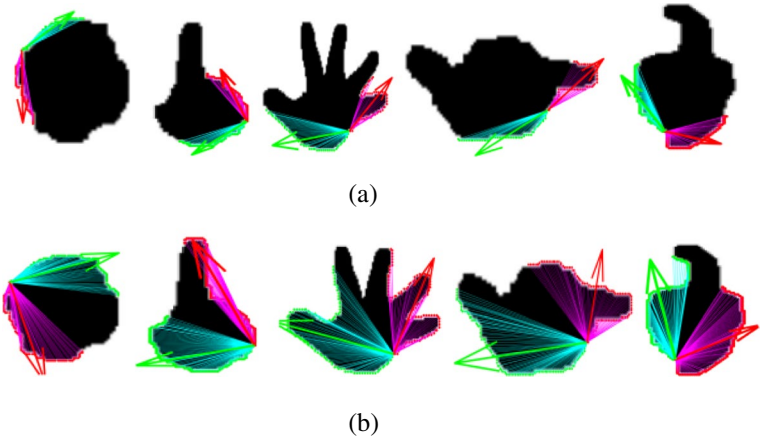
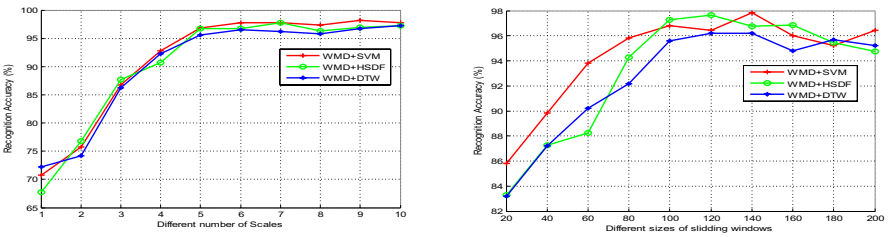


**Fig. 11** Gaussian-smoothed mean vectors VS different sizes of windows: **a** w = 50 **b** w = 100. The amplitudes in (**a**) are greater than those in (**b**) corresponding to small-scale descriptor and vice versa

were implemented with the suggested WMD descriptor. Each experiment was repeated fifty times and total average recognition accuracies were estimated. In the first experiment, the number of scales for one-dimensional Gaussian smoothing operation was ranged from one to ten with fixed size of sliding window and the result was shown in Fig. 12a. It is observed that the averaged accuracy increases rapidly with the increasing number of scales and becomes steady when the number is equal to or great than five. For example, around 70% of accuracy was obtained with the three classifiers when one scale was used, i.e. $n = 1$. The reason is that the WMD descriptor only encodes the coarsest characteristic along the hand contour and does not carry enough discriminative information. The accuracy is increased with more scales as both coarse and fine characteristics will be encoded by the descriptor and its discriminative capacity is enhanced. However, the fine characteristics are sensitive to the noise or disturbance introduced during ROI segmentation. As a result, the recognition accuracy oscillates in the range of 95 and 100%. Therefore, we take five scales as an optimal balance in the following experiments considering the accuracy and computational complexity.

In the second experiment, the size of sliding window was varied from 20 to 200 with fixed scales and the testing result was given in Fig. 12b. We can see that different sizes of sliding windows have different performance on the hand gesture recognition and the




(a) AVG accuracy vs number of scales (b) AVG accuracy vs size of sliding window

**Fig. 12** AVG accuracy VS number of scales and size of sliding window. **a** AVG accuracy vs. number of scales **b** AVG accuracy vs. size of sliding window

**Table 3** Brief comparison of similar algorithms

| Algorithms | Pros | Cons |
|---|---|---|
| Huang et al. [34] | • FMD is derived from area, arc length and central distance<br>• Different scales via various radii of circles<br>• The scale invariant is reached via normalization | • Repeated computation of hand area<br>• Values of FMD may fluctuate abruptly in case of improper size of circles |
| Lazarou et al. [35] | • ARB via number of contour points, accumulative distance and average distance<br>• The scale invariant is reached via normalization<br>• Rotation tolerance via overlapping and accumulative ARB | • Inner points of hand area not considered<br>• Very limited rotation transformation<br>• Centroid of hand area required |
| Sahana et al. [36] | • multi-radii circular signature constructed by counting frequency of occurrence of 1's in each partition<br>• Computational efficiency | • Highly Sensitive to centroid of hand area<br>• Number of partitions required |
| Proposed method | • Gaussian-smoothing angle related WMD descriptor is suggested<br>• The WMD is robust with Gaussian smoothing<br>• Naturally rotation, translation and scaling invariance | • Inner points of hand area not considered |

optimal range of size was between 80 and 120 points. The averaged accuracy was reduced outside this range. The WMD descriptor will concentrate on detailed information and be sensitive to noise and disturbance when size of sliding window is small. On the other hand, the descriptor with large size mainly captures global characteristic and ignore local information, which will bring loss of information to some extends. As a conclusion, the optimal size of sliding window is set as 100 points based on this experiment.

### 5.6 Performance evaluation

This section focuses on detailed performance of the proposed WMD descriptor on hand gesture recognition. For comparison, we implement the proposed algorithm together with three most recent benchmark methods, including Huang et al. [34], Lazarou et al. [35] and Sahana et al. [36] respectively denoted by ALG1, ALG2 and ALG3, as they follow a similar mechanism. To be a fair game, all the experiments are carried out on the same dataset with the same platform. In ALG1, different scales of circle regions centered at each of the contour points were employed to extract the area, major segment and distance information as characteristics of the hand gesture. Basically, it is a multi-resolution analysis along the hand contour. The descriptor corresponding to larger size of circle encodes coarse information and smaller size captures detailed information. In this sense, it is highly similar with the proposed WMD descriptor. The major difference lies in that the WMD descriptor is derived from ratio of Gaussian smoothing operation and invariant to scaling transformation. The main contribution in ALG2 is a new descriptor that is constructed via angular–radial bins within the concentric circles of the hand ROI. The multi-resolution analysis is achieved by using different angular widths and different number of concentric circles. On the other hand, the gesture descriptor in ALG3 is based on circular signal sampling and peak frequency. In summary, theoretical comparison of the above algorithms can be found in Table 3.

The first experiment was carried out on our self-established dataset by comparing the proposed algorithm with the three benchmark distance-based methods where the suggested WMD descriptor were used as input for Hausdorff distance, DTW algorithm and SVM model, respectively denoted by ALG6, ALG7 and ALG8. Besides, two CNN-based methods including the deep architecture proposed in [38] and YOLOv3, respectively denoted by ALG4 and ALG5, were implemented for further comparison in this experiment. The true positive rate for each category given in Table 4 and their confusion matrix in Fig. 13

**Table 4** Comparison of the proposed algorithm with Benchmark methods on self-established dataset (%)

| Different Gestures | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Mean/S.D. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALG1 | 99 | 98 | 95 | 96 | 97 | 96 | 95 | 96 | 96 | 95 | $96.3 \pm 1.34$ |
| ALG2 | 97 | 96 | 97 | 95 | 97 | 96 | 95 | 97 | 95 | 96 | $96.1 \pm 0.88$ |
| ALG3 | 100 | 98 | 96 | 96 | 96 | 97 | 95 | 96 | 95 | 96 | $96.5 \pm 1.51$ |
| ALG4 | 98 | 95 | 96 | 95 | 96 | 96 | 95 | 94 | 93 | 95 | $95.3 \pm 1.34$ |
| ALG5 | 99 | 94 | 93 | 96 | 94 | 96 | 92 | 95 | 95 | 94 | $94.8 \pm 1.93$ |
| ALG6 | 100 | 96 | 97 | 95 | 98 | 97 | 96 | 97 | 97 | 98 | $97.1 \pm 1.37$ |
| ALG7 | 99 | 98 | 98 | 97 | 96 | 96 | 95 | 96 | 96 | 97 | $96.8 \pm 1.22$ |
| **ALG8** | **99** | **96** | **98** | **98** | **96** | **97** | **98** | **97** | **96** | **96** | $\mathbf{97.1 \pm 1.10}$ |

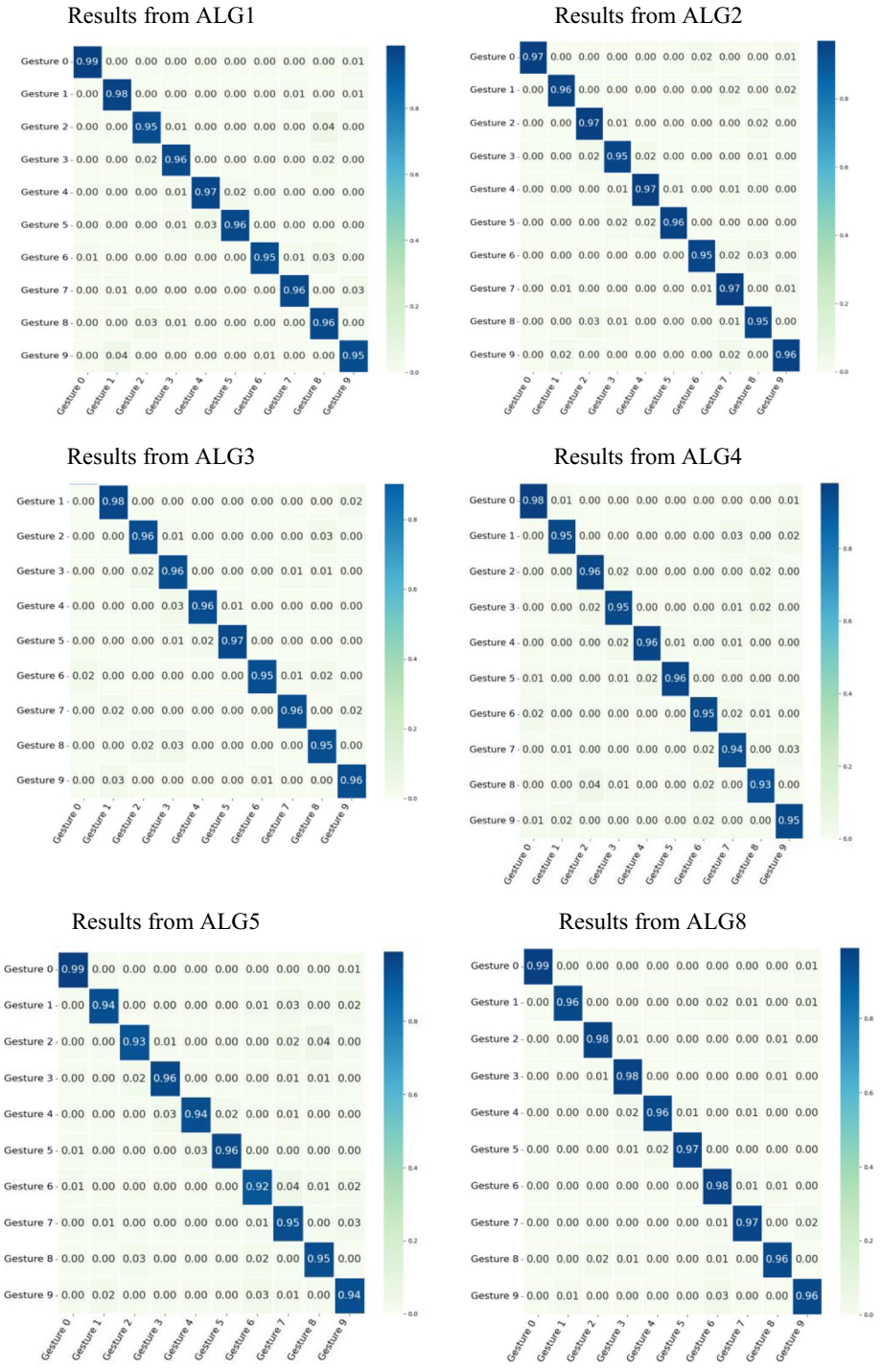



**Fig. 13** Confusion matrices of recognition results for the six algorithms where the horizontal axis and vertical axis respectively represent predicted gestures and true gestures

(a) Gesture 6 (b) Gesture 8 (c) Gesture 7 (d) Gesture 9

**Fig. 14** The misrecognized samples in the experiments. **a** Gesture 6 **b** Gesture 8 **c** Gesture 7 **d** Gesture 9

were used as evaluation metrics to present an overall and detailed performance of those algorithms. It is observed from this table that the proposed algorithm generally gives the highest accuracy of 97.1% followed by ALG3 with 96.5% mean accuracy, ALG1 with 96.3% mean accuracy and ALG2 with 96.1% mean accuracy. The standard deviations from WMD + DTW and WMD + SVM are 1.22 and 1.1 respectively, lower than those of ALG1 and ALG3 but higher than that of ALG2. This is because all the pixels of the hand ROI are involved during feature engineering in ALG2. The inner points are obviously less sensitive to hand shapes and contribute less than those contour points. The two CNN-based methods output a bit poor accuracy because it is prone to over-fitting in case of a small dataset. As can be seen from the confusion matrices in Fig. 13, the hand gestures for six, seven, eight and nine are prone to be confused, among which the true positive rate for the six-gesture is the lowest since its little finger in this gesture is easily overlapped in some viewpoints. This may lead to confusion with seven-gesture or eight-gesture. Figure 14 shows four misrecognized samples in this experiment. The dominant parts in both Fig. 14a and b exhibit nearly circular shapes which would lead to misrecognition if the relative positions of the two fingers are similar to each other. Similarly, the dominant areas in Fig. 14c and d are close to each other which require more discriminative descriptor or optimal parameters in the classifier. On the whole, the proposed WMD + SVM overcomes the shortcomings of the Benchmark methods and outputs the best performance.

In another experiment, these algorithms were implemented on the public NTU hand gesture dataset, in which ten-fold cross validation mechanism was used. Table 5 gave the average recognition accuracy of the ten hand gestures for each algorithm. From this table, we find that there are some gestures provide higher accuracy e.g. Gesture 1, Gesture 5 and Gesture 6, in contrast to Gesture 2 and Gesture 9 as they exhibit similar poses. The proposed algorithm achieves the best performance in terms of average accuracy, of which the WMD + SVM gives as high as $96.6 \pm 1.3\%$. The average accuracy recorded from ALG1, ALG2 and ALG3 are respectively $95.8 \pm 1.8\%$, $95.2 \pm 1.4\%$ and $95.9 \pm 1.1\%$. This further verifies that the suggested WMD descriptor can be combined with different classifiers and present satisfactory performance for various applications.

In the final experiment, we implemented these algorithms on Senz3D dataset again with ten-fold cross validation. Here, six different scales were used to extract more information for the WMD descriptor in the proposed model and the parameters for ALG3 were set as 12 signatures with 12 partitions. The obtained average accuracy was shown in Table 6. It is observed that all the algorithms provide satisfactory results with recognition accuracy above 93% and some gestures including Gesture 1, Gesture 5 and Gesture 8 give a very high accuracy. However, there exist some degree of similarity among Gesture 7, Gesture 10 and Gesture 11 which decreases their recognition rate. As a whole, the WMD + SVM performs best with average accuracy of $96.2 \pm 1.9\%$ followed by ALG3 with $95.6 \pm 1.9\%$.

**Table 5** Comparison of the proposed algorithm with Benchmark methods on NTU dataset (%)

| Different Gestures | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Mean/S.D. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ALG1 | 100 | 96 | 94 | 94 | 96 | 96 | 94 | 96 | 96 | 96 | $95.8 \pm 1.75$ |
| ALG2 | 96 | 94 | 94 | 96 | 96 | 98 | 94 | 94 | 94 | 96 | $95.2 \pm 1.40$ |
| ALG3 | 98 | 95 | 95 | 96 | 97 | 97 | 95 | 96 | 95 | 95 | $95.9 \pm 1.10$ |
| ALG6 | 98 | 96 | 96 | 96 | 96 | 96 | 96 | 94 | 96 | 98 | $96.2 \pm 1.14$ |
| ALG7 | 98 | 94 | 94 | 98 | 96 | 96 | 94 | 94 | 96 | 96 | $95.6 \pm 1.58$ |
| **ALG8** | **99** | **96** | **96** | **96** | **98** | **98** | **96** | **95** | **96** | **96** | $\mathbf{96.6 \pm 1.26}$ |

**Table 6** Comparison of the proposed algorithm with Benchmark methods on Senz3D dataset (%)

| Different Gestures | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Mean/S.D. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ALG1 | 98 | 96 | 94 | 96 | 99 | 95 | 93 | 96 | 93 | 94 | 93 | $95.2 \pm 2.04$ |
| ALG2 | 97 | 95 | 93 | 96 | 98 | 96 | 94 | 95 | 94 | 93 | 94 | $95.0 \pm 1.61$ |
| ALG3 | 98 | 94 | 94 | 97 | 99 | 97 | 94 | 96 | 95 | 93 | 95 | $95.6 \pm 1.91$ |
| ALG6 | 98 | 96 | 95 | 96 | 99 | 96 | 93 | 97 | 93 | 95 | 93 | $95.5 \pm 2.01$ |
| ALG7 | 97 | 95 | 93 | 97 | 98 | 96 | 94 | 96 | 94 | 94 | 94 | $95.3 \pm 1.61$ |
| **ALG8** | **99** | **96** | **94** | **97** | **99** | **97** | **94** | **98** | **95** | **95** | **94** | $\mathbf{96.2} \pm 1.94$ |

## 6 Conclusion

We have talked about a new weighted multi-scale descriptor for hand gesture recognition algorithm based on the Kinect sensor, taking the recognition of ten digital gestures from zero to nine as an example. Firstly, the weight factor is estimated for each contour point by 2D Gaussian smoothing function and Prewitt operator to relate it with its neighbors and highlight its importance. Then the feature descriptor is constructed via 1D Gaussian smoothing considering the contour points in the hand should not be independent from each other when used to recognize the gestures. With a larger deviation, the peak of the Gaussian function will be lower and the distribution of the weights will be approximately even, corresponding to large-scale WMD descriptor for coarse information. The fine information of the gesturing hand will be encoded by Gaussian smoothing with a smaller deviation. The invariances to translation, rotation and scaling transformations of the descriptor are proved theoretically and validated experimentally from different aspects. Extensively experiments on our ten-gesture dataset, NTU dataset and Senz3D dataset have been carried out comparing the proposed algorithm with three distance-based and two CNN-based hand gesture recognition methods. The results show that the proposed algorithm outperforms those algorithms with better robustness and higher recognition accuracy.

Although hand gesture recognition has witnessed significant advances, it still remains a challenging problem including environmental noise, user's variability and identification of boundary between different gestures. So our future work is twofold. One is to explore more representative features in both spatial and temporal spaces and integrate them with the proposed WMD descriptor for further improving the performance of the algorithm. The other is to develop some interesting HCI applications and deploy it on our mobile robot to understand human's intention and perform some routine housework.

**Acknowledgements** The authors would like to thank anonymous reviewers for their valuable comments and suggestions.

**Data availability** Data will be made available upon request.

## Declarations

**Conflict of interest** There is no conflict of interest.

## References

1. Liu AA, Nie WZ et al (2015) Coupled hidden conditional random fields for RGB-D human action recognition. Sig Process 112:74–82. https://doi.org/10.1016/j.sigpro.2014.08.038
2. Chevtchenko SF, Vale RF et al (2018) A convolutional neural network with feature fusion for real-time hand posture recognition. Appl Soft Comput 73:748–766. https://doi.org/10.1016/j.asoc.2018.09.010
3. Memo A, Zanuttigh P (2018) Head-mounted gesture controlled interface for human-computer interaction. Multimed Tools Appl 77:27–53
4. Liu Y, Jiang J et al (2021) Hand pose estimation from RGB images based on deep learning: a survey. IEEE 7th International Conference on Virtual Reality (ICVR)
5. Dardas NH, Georganas ND (2011) Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. IEEE Trans Instrum Meas 60(11):3592–3607
6. Tubaiz N, Shanableh T et al (2015) Glove-based continuous arabic sign language recognition in user-dependent mode. IEEE Trans Hum-Mach Syst 45:526–533. https://doi.org/10.1109/THMS.2015.2406692
7. Cornacchia M, Ozcan K et al (2017) A survey on activity detection and classification using wearable sensors. IEEE Sensors J 17:386–403. https://doi.org/10.1109/JSEN.2016.2628346
8. Lei W, Du QH, Koniusz P (2019) A comparative review of recent kinect-based action recognition algorithms. IEEE Trans Image Process 29:15–28
9. Song L, Yu G et al (2021) Human pose estimation and its application to action recognition: a survey. J Vis Commun Image Represent 76:103055. https://doi.org/10.1016/j.jvcir.2021.103055
10. Mohamed N, Mustafa M et al (2021) A review of the hand gesture recognition system: current progress and future directions. IEEE Access 9:19
11. Thanh TT, Fan C et al (2012) Extraction of discriminative patterns from skeleton sequences for human action recognition. In: 2012 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future
12. Zhu HM, Pun CM (2013) Human action recognition with skeletal information from depth camera. In: IEEE International Conference on Information & Automation, 26–28. https://doi.org/10.1109/ICIA31444.2013
13. Amor B, Su J, Srivastava A (2015) Action recognition using rate-invariant analysis of skeletal shape trajectories. IEEE Trans Pattern Anal Mach Intell 38:1–13. https://doi.org/10.1109/TPAMI.2015.2439257
14. Liu X, Shi H et al (2020) 3D skeletal gesture recognition via hidden states exploration. IEEE Trans Image Process 29:1–1
15. Kowdiki M, Khaparde A (2021) Automatic hand gesture recognition using hybrid meta-heuristic-based feature selection and classification with dynamic time warping. Comput Sci Rev 39. https://doi.org/10.1016/j.cosrev.2020.100320
16. Wang C, Liu Z, Chan SC (2015) Superpixel-based hand gesture recognition with kinect depth camera. IEEE Trans Multimed 17(1):29–39. https://doi.org/10.1109/TMM.2014.2374357
17. Chen H, Liu X et al (2018) Temporal hierarchical dictionary with HMM for fast gesture recognition. 24th International Conference on Pattern Recognition (ICPR), Beijing, China, pp 3378–3383
18. Raheja JL, Minhas M et al (2015) Robust gesture recognition using Kinect: a comparison between DTW and HMM. Optik, Int J Light Electron Opt 126:1098–1104. https://doi.org/10.1016/j.ijleo.2015.02.043
19. Escobedo EJ, Chavez GC (2020) Multimodal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes. J Vis Commun Image Represent 71. https://doi.org/10.1016/j.jvcir.2020.102772
20. Shin S, Kim WY (2020) Skeleton-based dynamic hand gesture recognition using a part-based GRU-RNN for gesture-based interface. IEEE Access 8:50236–50243. https://doi.org/10.1109/ACCESS.2020.2980128

21. Lai K, Yanushkevich SN (2018) CNN + RNN Depth and skeleton based dynamic hand gesture recognition. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp 3451–3456
22. Guo F, He Z et al (2021) Normalized edge convolutional networks for skeleton-based hand gesture recognition. Pattern Recogn 118. https://doi.org/10.1016/j.patcog.2021.108044
23. Ren Z, Yuan J et al (2013) Robust part-based hand gesture recognition using Kinect Sensor. IEEE Trans Multimed 15:1110–1120. https://doi.org/10.1109/TMM.2013.2246148
24. Wong WK, Juwono FH et al (2021) Multi-features capacitive hand gesture recognition sensor: a machine learning approach. IEEE Sensors J 21:8441–8450. https://doi.org/10.1109/JSEN.2021.3049273
25. Lee DL, You WS (2018) Recognition of complex static hand gestures by using the wristband-based contour features. IET Image Proc 12:80–87
26. He Y, Li G et al (2019) Gesture recognition based on an improved local sparse representation classification algorithm. Clust Comput 22:10935–10946
27. Wang Z (2021) Gesture recognition by model matching of slope difference distribution features. Measurement 181:109590. https://doi.org/10.1016/j.measurement.2021.109590
28. Kim J, Yu S et al (2017) An adaptive local binary pattern for 3D hand tracking. Pattern Recognit 61:139–152. https://doi.org/10.1016/j.patcog.2016.07.039
29. Tang J, Hong C et al (2018) Structured dynamic time warping for continuous hand trajectory gesture recognition. Pattern Recognit 80:21–31
30. Calado A, Roselli P et al (2022) A geometric model based approach to hand gesture recognition. IEEE Trans Syst Man Cybern: Syst 52. https://doi.org/10.1109/TSMC.2021.3138589
31. Zhang B, Yang Y et al (2017) Action recognition using 3D histograms of texture and a multi-class boosting classifier. IEEE Trans Image Process 26:4648–4660. https://doi.org/10.1109/TIP.2017.2718189
32. Reza A, Maryam AA et al (2019) Dynamic 3D hand gesture recognition by learning weighted depth motion maps. IEEE Trans Circ Syst Video Technol 29:1729–1740. https://doi.org/10.1109/TCSVT.2018.2855416
33. Sun Y, Weng Y et al (2020) Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. IET Image Process 14:3662–3668
34. Huang Y, Yang J (2021) A multi-scale descriptor for real time RGB-D hand gesture recognition. Pattern Recognit Lett 144:97–104. https://doi.org/10.1016/j.patrec.2020.11.011
35. Lazarou M, Li B, Stathaki T (2021) A novel shape matching descriptor for real-time static hand gesture recognition. Comput Vis Image Underst 210:103241. https://doi.org/10.1016/j.cviu.2021.103241
36. Sahana T, Basu S et al (2022) MRCS: multi-radii circular signature based feature descriptor for hand gesture recognition. Multimed Tools Appl 81(6):8539–8560
37. Dominio F, Donadeo M, Zanuttigh P (2014) Combining multiple depth-based descriptors for hand gesture recognition. Pattern Recognit Lett 50:101–111
38. Deng M (2020) Robust human gesture recognition by leveraging multi-scale feature fusion. Signal Process Image Commun 83. https://doi.org/10.1016/j.image.2019.115768